Contributions des approches moléculaires in silico à la compréhension des liens structure/fonction des ARN

Fabrice Leclerc

Nancy-Université
Université Henri Poincaré

Département de formation doctorale en biologie **École doctorale BIOSE**
UFR STB

Contributions des Approches Moléculaires *in silico* à la Compréhension des Liens Structure/Fonction des ARN

MÉMOIRE

présenté et soutenu publiquement le 16 décembre 2009 à 14h30,
Amphithéâtre C du LORIA (UMR 7503)

pour l'obtention de l'

Habilitation de l'Université Henri Poincaré – Nancy I

(Spécialité Biochimie)

par

Fabrice Leclerc

Composition du jury

Président :	Le président	
Rapporteurs :	Pr. Gilbert Deléage	de l'Université de Lyon 1
	Dr. Alain Krol	de l'Université de Strasbourg
	Pr. Alain Denise	de l'Université Paris 11
Examinateurs :	Dr. Roland Stote	de l'Université de Strasbourg
	Dr. Christiane Branlant	de l'Université Henri Poincaré de Nancy 1
Invité :	Dr. Bernard Maigret	de l'Université Henri Poincaré de Nancy 1

ARN, RNP, maturation-structure-fonction, Enzymologie Moléculaire et Structurale — UMR 7214

Mis en page avec la classe thloria.

Remerciements

Je remercie les personnes qui m'ont permis d'exprimer mes aptitudes et de développer mes compétences au cours de mon parcours de chercheur, ainsi que les personnes du domaine ou d'autres disciplines avec qui j'ai apprécié d'interagir et desquelles j'ai aussi beaucoup appris.
Je remercie tout particulièrement : Qiang Cui (Dept. de Chimie, University of Madison, Madison), Michael Schaefer (Novartis, Bâle), Andy Ellington (ICMB, University of Texas, Austin), Jamie Williamson (Scripps, La Jolla), Alex MacKerell (Faculté de Pharmacie, University of Maryland, Baltimore), François Major (Dept. d'Informatique et Recherche Opérationnelle, Université de Montréal), Daniel Gautheret (IGM, Université d'Orsay, Paris 11).

Mes remerciements vont aussi de façon plus générale à mes anciens collègues de l'Université de Montréal (1993-1997), de mon ancien laboratoire du "drylab" : François Major, Daniel Gautheret, Serguey Steinberg, et du "wetlab" : Gerardo Ferbeyre, Pascal Chartrand, Benoît Cousineau et tout le département dans son ensemble (plus spécifiquement : Michel Bouvier, Jurgen Sygusch, John Gunn, Kalle Gehring, Stephen Michnick, etc).

Mes remerciements vont aussi de façon plus générale à mes anciens collègues de l'Université de Harvard (1997-1999) et/ou de l'Université de Strasbourg (1999-2000) : Qiang Cui, Paul Lyne, Darrin York, Yaoqi Zhou, Xabier Lopez, Annick Dejeagere, Eric Evensen, Collin Stultz, Michael Schaefer, Emanuele Paci et tout le département de Chimie Biologique de Harvard et tous les anciens membres du Laboratoire de Chimie Biophysique de Strasbourg.

Mes remerciements vont naturellement à mes collègues actuels du laboratoire AREMS (ex-MAEM), ceux avec qui j'interagis le plus : Xavier Manival, Bruno Charpentier, Nathalie Marmier-Gourrier, Audrey Vautrin, Magali Blaud, Arnaud Gruez, Christophe Charron, Isabelle Behm-Ansmant, et l'ensemble des membres du laboratoire.

Je tiens à rendre honneur à différentes personnalités de la communauté des théoriciens en France et notamment : Annick Dejaegere, Tom Simonson et à l'étranger : Jayashree Srinivasan, Dave Ritchie, Alex MacKerell, Bernie Brooks, Charlie Brooks, Bill Smith

Je tiens aussi à rendre honneur à différentes personnalités de la communauté ARN et notamment : Andy Ellington, Jamie Williamson, Joseph D. Puglisi, Tom R. Cech, Olke C. Uhlenbeck, Vickie DeRose.
Je tiens aussi à rendre honneur à différentes personnalités de la communauté des informaticiens : François Major, Alexander Bockmayr, Hélène Touzet.

Enfin, je veux aussi rendre hommage à des personnalités du secteur industriel : Richard Griffey (Isis Pharmaceuticals Inc., Carlsbad), William Brown (BiochemPharma Inc., Montréal), Dipesh Risal (Accelrys, Inc.).

Je dédie ce mémoire à mon Directeur de thèse Robert Cedergren, dit "Bob", qui m'a beaucoup appris à la fois scientifiquement et humainement ;
à mon ancien Directeur de Recherches Martin Karplus qui m'a beaucoup apporté en terme de rigueur scientifique et m'a ouvert de nouveaux horizons ;
à ma Directrice de Laboratoire Christiane Branlant qui m'a permis un "retour aux sources" ;
à mon amie Noëlle Carbonell, Professeur d'informatique, avec qui j'ai toujours eu des échanges scientifiques très stimulants ;
à ma mère qui a toujours une grande curiosité pour ce que je fais même si elle reconnaît elle-même de ne pas y comprendre grandchose ;
à mon épouse qui m'a toujours stimulé scientifiquement par sa vision extérieure de biologiste.

Table des matières

Introduction générale

Les domaines de la modélisation moléculaire et de la bioinformatique, au sens large, recouvrent en fait beaucoup d'approches qui sont développées et appliquées à des systèmes biologiques très différents et pour répondre à des questions biologiques distinctes. En tant que jeune chercheur dans le domaine depuis mon travail de doctorant et de post-doctorant, j'ai eu l'occasion d'utiliser et de développer un certain nombre d'approches et de méthodes afin de répondre à des questions biologiques très diverses mais dont la trame de fond était toujours reliée à une problématique d'étude des liens structure/fonction des ARN. Dans le "RNA World", plusieurs grandes thématiques de recherche existent car elles correspondent à la diversité et versatilité de structures et de fonctions des ARN. Mon intention n'est pas d'en faire une revue exhaustive mais de me focaliser les principaux thèmes concernant mon expérience de recherche et d'encadrement. Ces thèmes peuvent être regroupés sous trois grands axes qui couvrent des problématiques phares dans le domaine ARN dans lesquelles les approches *in silico* ont contribué de façon plus ou moins importante à une meilleure compréhension des liens structure/fonction des ARN.

Le premier axe de recherche concerne l'étude de la réactivité des ARN ; la découverte de l'existence d'ARN catalytiques a bouleversé les dogmes et les conceptions en biologie moléculaire et bien au-delà. Les méthodes faisant appel à la mécanique quantique peuvent contribuer à mieux comprendre la catalyse par les ARN, la catalyse au niveau atomique étant difficile à appréhender et la structure 3D des ARN difficile à déterminer, expérimentalement. Le second grand axe est l'exploitation et l'interprétation de données biologiques par des méthodes de bio-analyse et bio-informatique. Ces méthodes existent depuis un certain nombre d'années (recherche de similarité de séquence, méthodes d'alignement, etc) mais elles étaient davantage utilisées dans une optique d'analyse des données biologiques que d'extraction de connaissances ("data mining" ou fouille de données, etc). L'essor de la génomique s'est accompagné d'un développement parallèle de méthodes de traitement massif de l'information biologique (notamment en RNomics) qui ont joué un rôle dans la découverte récente de l'importance des fonctions biologiques de nombreux petits ARN. Ces découvertes bouleversent à nouveau nos conceptions sur l'expression des gènes et leur régulation. Le troisième axe de recherche est la compréhension des bases moléculaires des interactions ARN/protéines et ARN/ligands en général qui ont des rôles clés dans beaucoup de processus biologiques. L'altération ou la modulation de ces interactions peuvent parfois faire la différence entre un fonctionnement "sain" et un fonctionnement pathologique associé à une maladie. Malgré les avancées réalisées en biologie structurale, les ARN et complexes ARN/protéine restent des objets difficiles à étudier du point de vue expérimentale. Les méthodes de modélisation moléculaire et de bioinformatique structurale offrent alors la possibilité de construire des modèles 3D à partir desquels il est possible de poser des questions et d'avancer des hypothèses. La possibilité de prédire la structure 3D d'ARN et de complexes ARN/protéine est rendue possible par l'amélioration des méthodes de modélisation. D'autre part, elles ouvrent des perspectives dans la conception de ligands à partir d'une structure 3D utilisée comme cible. Les ARN deviennent alors aussi des cibles thérapeutiques potentielles.

L'étude de la réactivité des ARN et les méthodes MQ

Les ARN sont des molécules biologiques qui possèdent une réactivité chimique intrinsèque. A la différence de l'ADN, l'ARN présente un groupe 2'-OH pouvant se comporter comme un nucléophile et qui lui confère en grande partie sa réactivité. D'autres facteurs entrent ensuite en jeu (pH, conformation de l'ARN, présence de cofacteurs : métaux ou nucléobase, etc) qui vont conférer à l'ARN une simple réactivité non-enzymatique ou enzymatique lorsqu'il s'agit d'ARN catalytiques qui ont évolué par rapport à une fonction biologique donnée : les ribozymes. On distingue deux grandes familles de ribozymes qui catalysent chacune un type de réaction de clivage d'un lien phosphodiester où les produits formés comportent soit des extrémités 2'-3'-phosphate cyclique et 5'-OH dans le cas des ribozymes auto-clivables soit des extrémité 5'-phosphate et 3'-OH pour les ribozymes auto-épissables (Fig. 1).

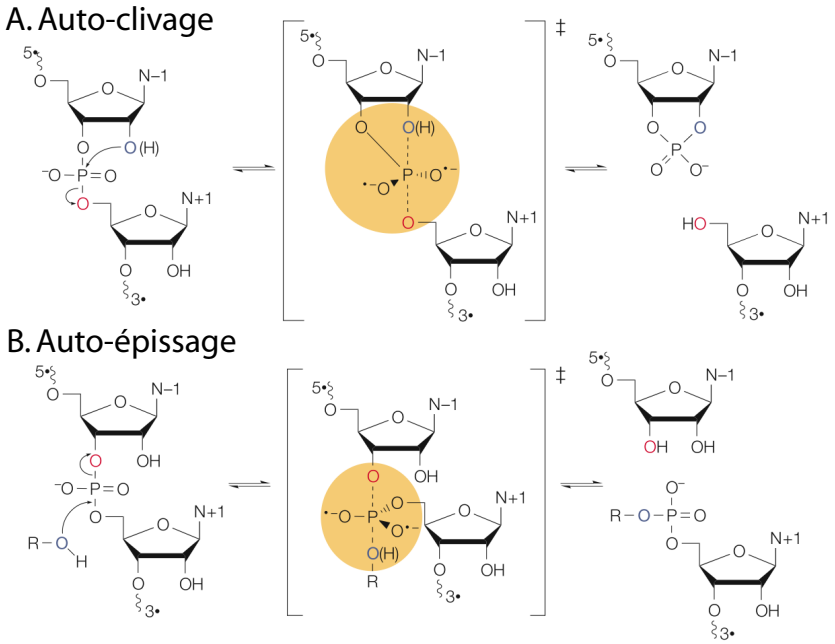


Figure 1. Réactions d'auto-clivage et d'auto-épissage. A. Les ribozymes auto-clivables catalysent une réaction où le nucléophile attaquant est le groupe 2'-OH et le groupe partant le groupe 5'-OR. B. Les ribozymes auto-épissables (ou la RNase P) catalysent une réaction qui diffère par les nucléophile attaquant qui sont un groupe 3'-OH d'une guanosine exogène et 2'-OH d'une adénosine intronique (première étape de la réaction dans les introns des groupes I et II). Chacune des deux réactions procède par une inversion de la configuration stéréochimique des atomes d'oxygène non liant du groupe phosphate subissant l'attaque nucléophile. Le mécanisme réactionnel obéit à une attaque nucléophilique de type S_N2 avec formation d'un état de transition ou intermédiaire trigonale bipyramidale correspondant à une groupe phosphate pentavalent.

L'attribution du prix Nobel de chimie en 1989 à Sidney Altman [1] et Thomas Cech [2] récompensés pour leur découverte respective des propriétés catalytiques de la RNase P (dont le composant enzymatique est l'ARN et non la partie protéique) [3] et de précurseurs d'ARN ribosomiques auto-épissables [4] marque le début d'un intérêt sans cesse grandissant pour les ARN en général et les ribozymes en particulier. Les ARN ne sont plus alors considérés comme de simples intermédiaires dans la transmission de l'information génétique contenue dans l'ADN mais comme des molécules clés à l'origine de la vie servant à la fois de porteur d'une information génétique et de biocatalyseur ("RNA World"). La découverte de nombreux autres ribozymes et plus particulièrement d'une famille de ribozymes auto-clivables baptisés ribozymes à tête de marteau ("hammerhead") [5] va beaucoup faire progresser les connaissances dans le domaine. Du

fait de leur petite taille (moins de 50 nucléotides) comparée à celle de la RNase P ou de ribozymes auto-épissables (près de 400 nucléotides) et de la nature modulaire de leur repliement (ARN à jonction triple), les ribozymes à tête de marteau vont rapidement devenir un prototype de la catalyse ARN. La transformation des ribozymes à tête de marteau naturels, issus de viroïdes, en "nucléases" artificielles en fera un objet d'étude privilégié en enzymologie des ARN mais aussi un outil très utilisé en biologie moléculaire (Fig. 2).

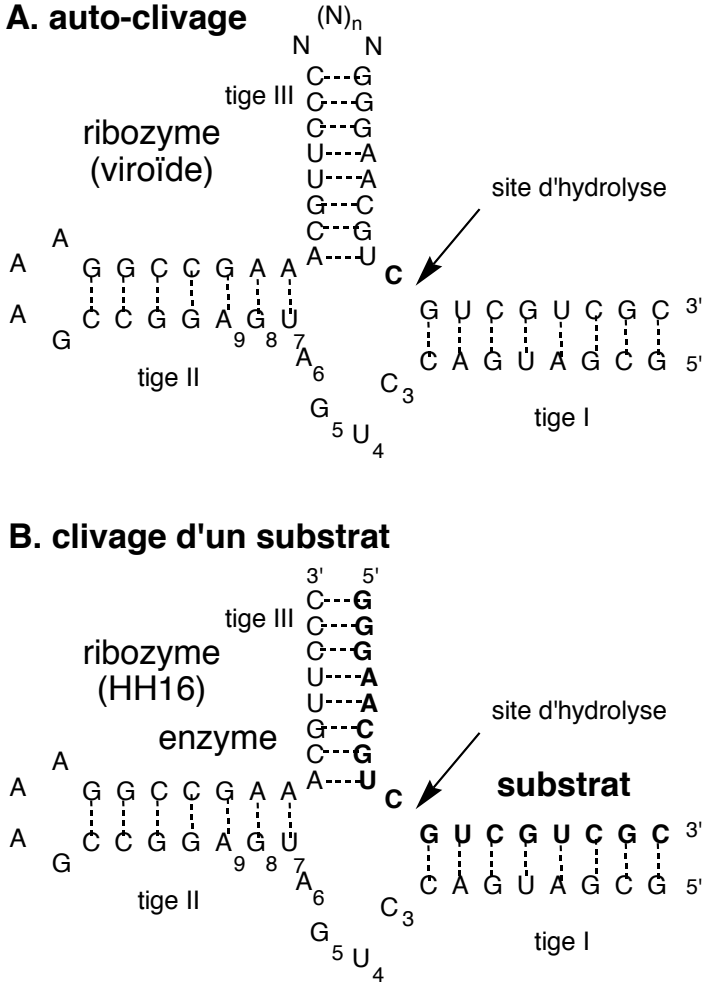


Figure 2. Structures 2D d'un ribozyme à tête de marteau. A. Ribozyme auto-clivable présent dans les viroïdes. B. Ribozyme artificiel (HH16) enzyme/substrat.

Les premiers calculs théoriques sur de petits modèles atomiques destinés à modéliser la réaction catalytique des ribozymes à tête de marteau sont publiés en 1994 [6]. Un modèle minimaliste mimant un groupe ribose phosphate composé de 9 atomes lourds est alors utilisé : il s'agit d'un oxyphosphorane (une groupe phosphate cyclique incluant les carbones C2' et C3' du ribose). Les calculs mettaient en évidence la formation d'états de transition correspondant à une groupe phosphate pentavalent ; ils suggéraient aussi la position de cations Mg^{2+} comme cofacteurs pour optimiser la réaction. L'utilisation de modèles extrêmement simples et limités en taille est caractéristique des approches quantiques (en particulier pour les méthodes de calculs *ab initio*) en raison des coûts élevés en temps de calcul. On cherche donc toujours à simplifier le système au maximum quitte à sacrifier le réalisme du modèle par rapport à la réalité chimique et bio-

chimique. Le système pourra ensuite être compliqué en fonction des besoins. L'avantage d'une telle approche est de pouvoir appréhender un système chimique et biochimique à différents degrés de complexité et d'évaluer les contributions de différents composants ou facteurs dans la catalyse. Les calculs récents effectués sur un modèle à 12 atomes lourds correspondant à un groupe ribose-phosphate non tronqué ont permis d'évaluer les barrières d'énergie dans les réactions non-enzymatiques de transestérification et d'hydrolyses des ARN [7] (Fig. 3). Les chemins réactionnels des 3 mécanismes catalytiques possibles à pH neutre, à pH basique (mécanisme di-anionique, Fig. 3A) ou à pH acide (mécanisme mono-anionique, Fig. 3B) ont été identifiés et comparés dans le contexte d'une réaction non catalysée. L'utilisation d'un modèle plus complet incluant la présence de cations métalliques fournit la possibilité d'évaluer l'influence et le rôle de cations métalliques comme cofacteurs dans une réaction catalysée.

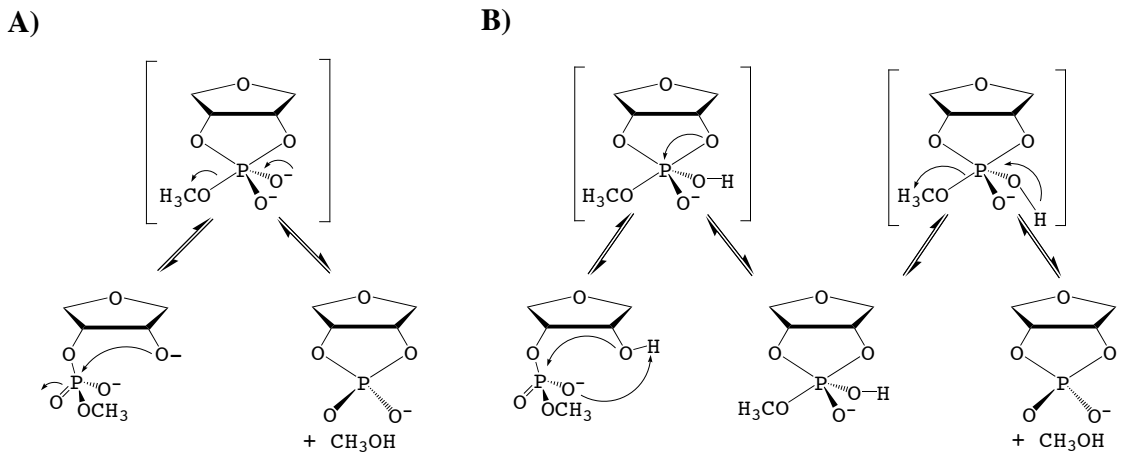


Figure 3. Mécanismes réactionnels pour la transestérification des ARN. A. Mécanisme di-anionique (3 points stationnaires et un état de transition) en conditions basiques. B. Mécanisme mono-anionique (5 points stationnaires et 2 états de transition) en conditions acides.

Plusieurs études structurales amorcées en 1992-1993 par radiocristallographie [8, 9] et spectroscopie [10] vont fournir les premières données structurales sur les ribozymes à tête de marteau. La première structure 3D expérimentale d'un ribozyme biologiquement actif sera déterminée en 1996 [11] et situe le contexte toujours actuel des recherches menées sur ces ribozymes dans l'objectif de comprendre les liens structure/fonction. Une question crucuale est : comment la structure globale du ribozyme et celle de la poche catalytique sont façonnées et comment elles influent sur le mécanisme catalytique ? Ces données susciteront beaucoup de controverses qui ne sont pas totalement résolues à ce jour en raison d'apparentes contradictions entre les premières données structurales d'une part et les données biochimiques d'autre part [12, 13].

La découverte d'autres ribozymes auto-clivables et l'étude de leur mécanisme réactionnel a révélé différentes stratégies catalytiques possibles [5, 15]. Ces stratégies diffèrent par le degré d'optimisation de chacun des 4 processus clés impliqués dans la catalyse de la réaction de transestérification qui conduit au clivage site-spécifique d'un lien phosphodiester (Fig. 4) [16]. Ils correspondent à l'activation du nucléophile (γ) et à l'attaque nucléophilique (α), à la stabilisation de la charge sur les oxygènes non-liant du groupe phosphate (β) suite à l'attaque nucléophile et la formation du groupe phosphate pentavalent et à la stabilisation de la charge sur l'oxygène en 5' du groupe partant (δ). Chacun des 4 processus peut être optimisé à des degrés divers d'un ribozyme à l'autre, ce qui conduit à des performances catalytiques variables ou quelquefois équivalentes même si la catalyse n'est pas optimisée de la même façon dans les 4 processus. D'après

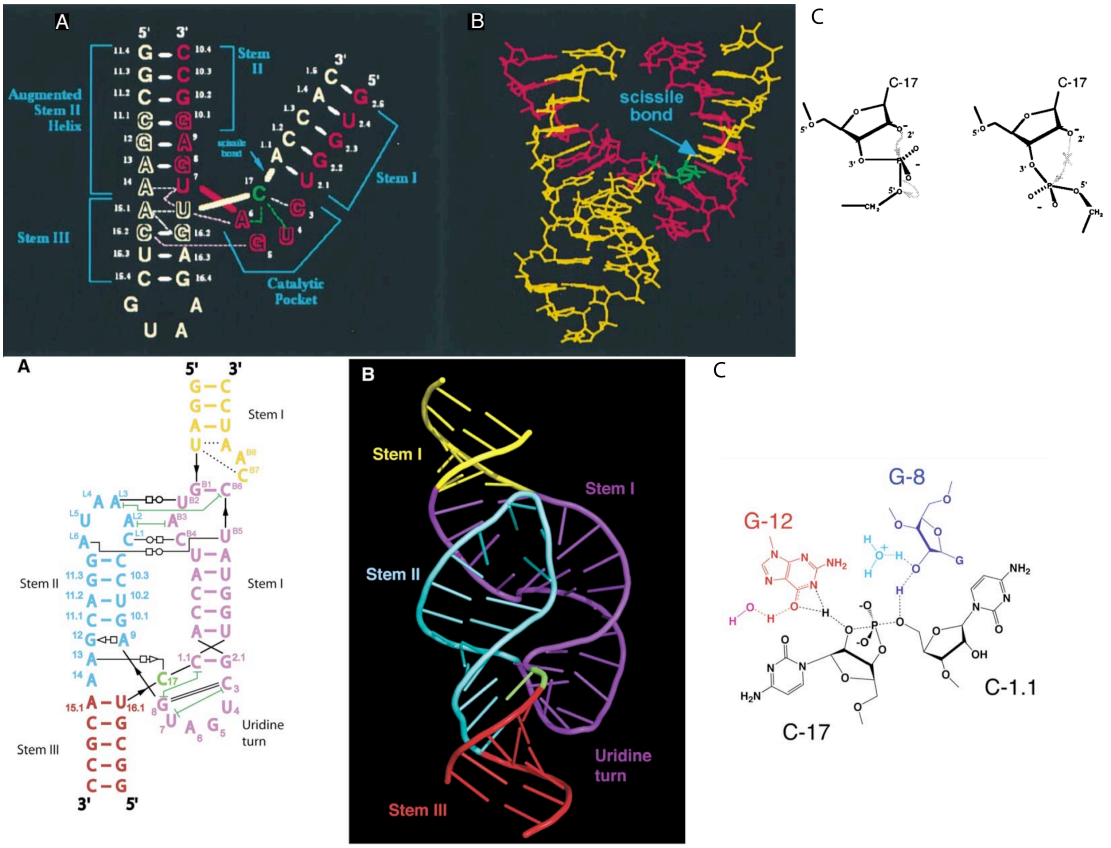


Figure 4. Modèles structure/fonction de ribozymes à tête de marteau. Haut : modèle de catalyse issu des données structurales de 1996 [11]. Bas : modèle de catalyse issu des données structurales de 2006 [14]. A. Structures 2D des ribozymes (d'après Scott *et al.*, 1996 [11]). B. Structures 3D des ribozymes (d'après Scott *et al.*, 1996 [11]). C. Mécanisme réactionnel proposé (d'après Martick *et al.*, 2008 [14]).

les caractéristiques chimiques des groupes chimiques impliqués dans chacun des 4 processus, il est possible de prévoir les limites théoriques de performances catalytiques des ribozymes (Table 1). Ces limites expliquent le fait que la vitesse de la réaction catalysée par les ARN (de l'ordre de la min^{-1}) soit nettement inférieure à celle d'enzymes protéiques comme la RNase A (de l'ordre de la μsec^{-1}) même si elle est largement plus rapide que la réaction non-catalysée (de l'ordre de 10^{-7}min^{-1}). Toutefois, l'existence d'effets synergiques liés à la géométrie du site actif, et les découvertes récentes de l'implication de nucléobases dans la catalyse par les ARN (pouvant jouer notamment le rôle d'acide/base générale [17]) suggère que ces limites peuvent sans doute être repoussées [16, 18]. La compréhension fine des ces limites ouvre des perspectives dans la conception de ribozymes artificiels ayant des performances catalytiques proches de celles des enzymes protéiques.

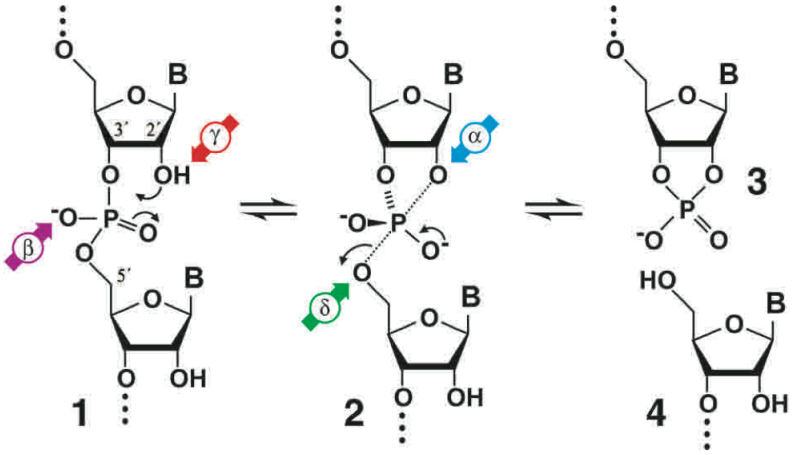


Figure 5. Mécanisme catalytique d'auto-clivage. Les 4 stratégies catalytiques qui peuvent influencer la réaction sont : α l'attaque nucléophilique "in-line" (bleu) ; β la neutralisation de la charge négative sur les oxygènes non-liant du groupe phosphate (violet) ; γ la déprotonation du groupe 2'-OH utilisé comme nucléophile (rouge) ; δ la neutralisation de la charge négative sur l'oxygène 5' du groupe partant (vert). D'après Emilsson *et al.*, 2003 [16].

Table 1 : Combinaisons possibles de stratégies catalytiques d'auto-clivage

Combinaison de stratégies catalytiques				Taux d'accélération maximum (pH neutre)	Vitesse limite de réaction (min^{-1})
absence de catalyse				-	10^{-8}
α				10^{2}	10^{-6}
	β			10^{5}	10^{-3}
		γ		10^{6}	10^{-2}
			δ*	$\geq 10^{6}$	$\geq 10^{-2}$
α	β			10^{7}	10^{-1}
α		γ		10^{8}	10^{0}
α			δ*	$\geq 10^{8}$	$\geq 10^{0}$
	β	γ		10^{11}	10^{3}
	β		δ*	$\geq 10^{11}$	$\geq 10^{3}$
		γ	δ*	$\geq 10^{12}$	$\geq 10^{4}$
α	β	γ		10^{13}	10^{5}
α		γ	δ*	$\geq 10^{14}$	$\geq 10^{6}$
α	β		δ	$\geq 10^{13}$	10^{5}
	β	γ	δ	$\geq 10^{17}$	10^{9}
α	β	γ	δ	$\geq 10^{19}$	10^{11}

la première ligne, en absence de catalyse, correspond à la constante de réaction utilisée comme référence en conditions standards (23°C, 250 mM K^{+}). Les vitesses limites sont les constantes de réaction correspondant à l'optimisation maximale de chacun des 4 processus (Fig. 4). Les astérisques indiquent la vitesse limite de catalyse pour le processus δ catalysé sans le processus β catalysé qui sont prédits comme n'étant pas permis chimiquement. D'après Emilsson *et al.*, 2003 [16].

Fonctions et évolution des ARN : bio-analyse et bio-informatique

L'annotation de génomes, par la recherche de gènes homologues, repose en grande partie sur la recherche de séquences similaires à l'aide de programmes tels que BLAST ou FASTA. Avec le développement des techniques de séquençage et l'accès aux séquences complètes de génomes, ce type de programmes a été largement utilisé pour la recherche de gènes homologues et l'annotation génomique. Les gènes d'ARN, qui peuvent être définis en premier lieu comme des gènes donnant naissance à des transcrits non traduits (ARN non-codant), avaient peu attiré l'attention jusqu'au début des années 2000 [19] lorsque le phénomène d'ARN interférence déjà mis en évidence chez le nématode est détecté également chez les mammifères [20]. D'autre part, les découvertes récentes indiquant que le génome humain est transcrit en très grande partie (à plus de 80%) [21, 22] et qu'il existe de nombreux éléments conservés du point de vue phylogénétique entre génomes parfois très distants (allant des insectes à l'homme par exemple [23]) ont de fortes implications quant à la présence potentielle de nombreux ARNnc encore inconnus.

Avant les années 2000, l'exploitation informatique des données brutes de séquence de génomes complets avait donc surtout été limitée à la recherche de gènes de protéine souvent assimilée aussi à l'identification de phases ouvertes de lecture ou ORF ("Open Reading Frame"). L'annotation systématique des gènes d'ARNnc était limitée aux ARNr dont la séquence primaire est suffisamment conservée pour être détectée par une simple recherche de similarité de séquences de type BLAST, ou à certains autres ARNnc tels que les ARNt qui possèdent une structure secondaire canonique plutôt bien définie et conservée. La recherche d'ARN fonctionnels, ou RNomics, demande beaucoup de ressources lorsqu'elle est abordée par une approche expérimentale de criblage étendu de tous les petits ARN contenus dans une cellule et représente un réel défi technique [24, 25]. Les méthodes bio-informatiques offrent la possibilité de cribler *a priori* des gènes d'ARNnc potentiels qui peuvent ensuite être validés expérimentalement. On pourra distinguer deux cas de figure selon que les ARNnc recherchés sont déjà connus et assez bien caractérisés du point de vue structural ou bien que l'on recherche de nouveaux ARNnc inconnus ou sur lesquels on ne dispose que de très peu d'informations.

Les ARN fonctionnels homologues peuvent avoir une séquence plus ou moins dégénérée d'un organisme à l'autre par rapport à celle de leur ancêtre commun, mais leur structure secondaire est souvent bien définie pour un ARNnc donné. Certains ARNnc présentent des motifs caractéristiques comme une jonction triple dans le cas d'un motif de ribozyme à tête de marteau (Fig. 2). Une approche visant à identifier des gènes potentiels d'ARNnc peut donc consister à rechercher, dans les génomes, des motifs ARN spécifiques. Plusieurs programmes ont été développés pour rechercher des motifs d'ARN bien définis. Certains ont été conçus pour rechercher des motifs structuraux et/ou fonctionnels : des ARNt (tRNAScan-SE) [26], des snoRNA (snoscan [27]), ou d'autres ARN non-codant (ncrnascan [28]), etc. D'autres programmes sont plus généralistes et permettent la recherche de motifs ARN en utilisant : un descripteur qui décrit l'enchaînement des éléments de structure primaire et secondaire (ou même tertiaire) de l'ARN (Rnamot [29], RNAMotif [30]), ou un profil de séquences (ERPIN [31, 32]) qui correspond à un alignement de séquences d'ARN présentant le motif recherché et incluant une information de structure secondaire. D'autres approches encore sont issues de la linguistique ("stochastic context-free grammars" [28]). Ce type d'approche (Rnamot) a été appliquée par exemple à la recherche de motifs ARN dans les génomes et constitue un des premiers exemples de "ribonomics" [33] ou "RNomics" qui est le terme actuel consacré pour faire référence à la recherche d'ARN fonctionnels. Le motif "hammerhead" fait parti des motifs d'ARN fonctionnels qui ont fait l'objet de recherches spécifiques [34] (Fig. 6).

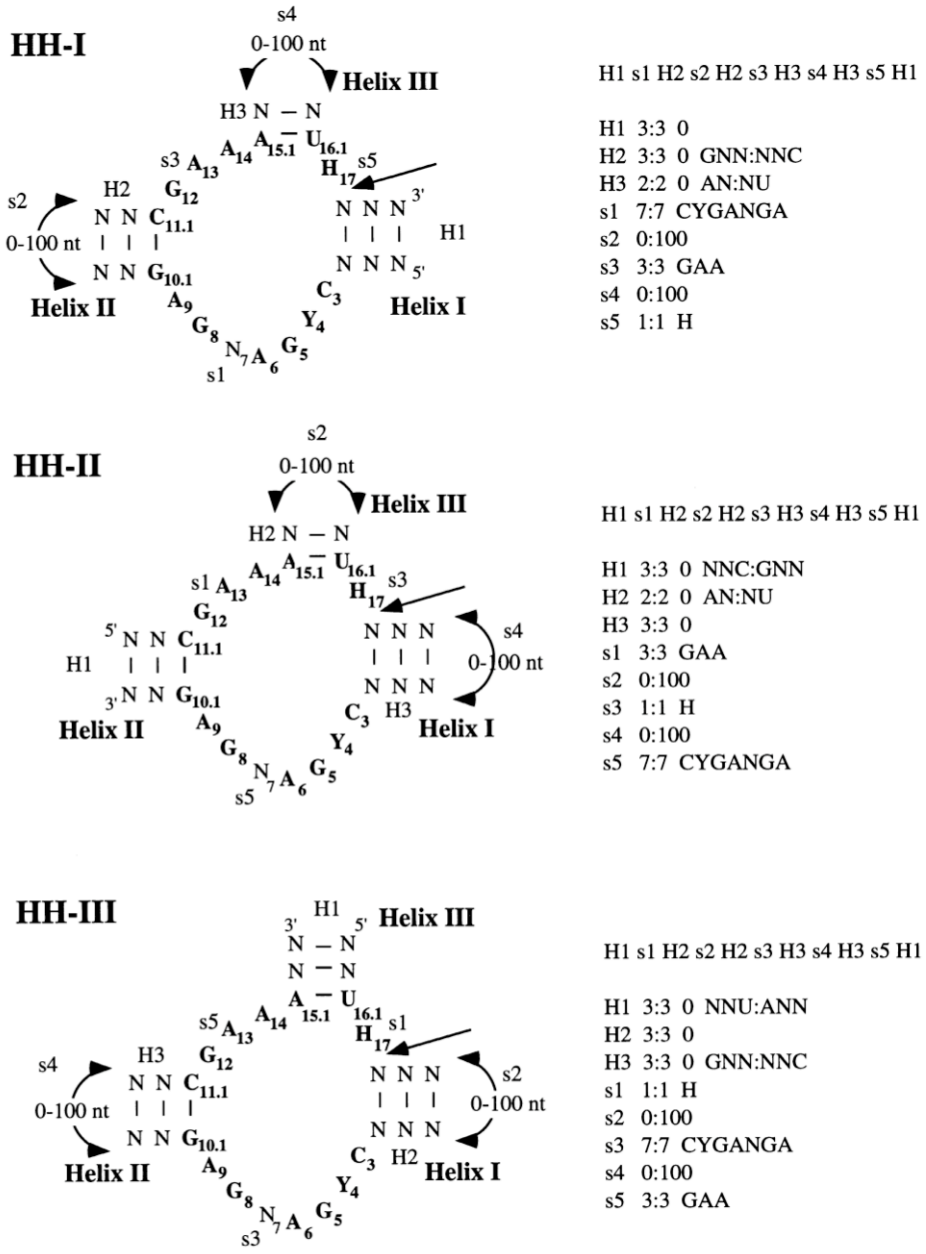


Figure 6. Structures 2D et descripteurs des classes de ribozymes à tête de marteau. Les 3 descripteurs HH-I, HH-II, et HH-III diffèrent par la position de l'hélice en 5'. Chaque descripteur est décrit par ses éléments simple-brin (s) et double-brin (H). Les éléments sont donnés dans l'ordre de la séquence (5' vers 3'), leur longueur (minimale : maximale) et la spécificité de séquence, ainsi que le nombre de mésappariements (pour les éléments H seulement) sont indiqués. Par exemple : HH-I contient les éléments suivants : H1 s1 H2 s2 s3 H3 s4 H3 s5 H1 où H1 est un élément double-brin de longueur fixe sans mésappariement ni séquence spécifique ; H2 est aussi un élément double-brin sans mésappariement mais avec une première paire G-C. H3 est un élément double-brin avec 2 paires, la première étant une paire A-U ; S1 est un élément simple-brin de 7 résidus avec une séquence spécifique ; s2 est un élément simple-brin qui peut varier en longueur entre 0 et 100 résidus, etc. Les résidus en gras indiquent les résidus importants (pour le repliement et l'activité des ribozymes). Le site de clivage est indiqué par une flèche à la position H17. Les noms des résidus sont indiqués en respectant la nomenclature IUPAC, (H) : A, C, ou U ; (N) : A, C, G, or U ; (Y) : C or U (d'après Bourdeau *et al.*, 1999 [33]).

Dans l'état actuel des connaissances, il apparaît que les gènes d'ARNnc sont majoritairement présents dans les régions situées entre les ORF identifiées et annotées dans les banques de séquences (ou alternativement dans certains introns). La présence de signaux d'expression comme un promoteur transcriptionnel en amont du gène, ou la forte structuration des ARNnc codés dans certaines régions du génome peuvent aussi être utilisés comme critères pour rechercher des gènes d'ARNnc. Toutefois ce type de critères intrinsèques à la séquence des ARNnc n'est pas nécessairement suffisant pour l'identification de ces gènes : les ARNnc ne possèdent pas tous un fort degré de structuration interne ; c'est vérifié pour un certain nombre d'organismes, et en particulier chez les archaea. Chez ces derniers, les informations sur les signaux de transcription sont quasiment inexistantes, ce qui limite encore les possibilités de détection de gènes d'ARNnc.

Bases moléculaires de la reconnaissance ARN/ligand : Modélisation 3D

Plusieurs classifications ont été proposées pour répertorier les différents motifs ARN structuraux et fonctionnels rencontrés dans les structures 3D connues [35, 36]. Le nombre total de motifs ARN au sein des organismes vivants, comme celui des repliements (ou "fold") chez les protéines, est probablement limité. Il a été montré que des motifs de liaison à des protéines, à des acides aminés, à des antibiotiques, ainsi que des motifs correspondant à des ribozymes sont potentiellement présents dans de nombreuses séquences naturelles [33]. Avec la détermination de la structure 3D des petites et grosses sous-unités du ribosome [37, 38], plusieurs motifs ARN ont pu être identifiés ou révélés. La présence de motifs communs à des ARN fonctionnels différents montre que le même motif structural peut effectivement se retrouver dans différents contextes fonctionnels pour ses propriétés de liaison à un ligand, en particulier un protéine ou une famille de protéines. L'exemple du motif K-turn est assez démonstratif : il est présent dans des ARN aussi variés que : les ARNr [39], les snRNA (U4) [40], les snoRNA à boîtes B/C et C/D [41], les sRNA à boîtes H/ACA d'archea [42] ou les ARNm (codant la protéine L30) [43].

La détermination de la structure 3D des ARN et de leur(s) complexe(s) avec des ligands est un pas important pour comprendre les bases moléculaires de la fonction des ARN. Pourtant, la structure en tant que telle, si elle fournit les bases à partir desquelles la fonction peut être appréhendée et comprise au niveau moléculaire et atomique, ne suffit pas. Elle doit permettre avant tout d'expliquer la fonction biologique ou d'avancer des hypothèses sur le rôle de tel ou tel résidu par exemple, dont la mutation a un impact sur la fonction, ou l'influence de tel ou tel autre facteur (exemple des ribozymes à tête de marteau ; voir ci-dessus : "L'étude de la réactivité des ARN et les méthodes MQ"). Avant la détermination des structures 3D des sous-unités du ribosome à partir des années 2000 [37, 38], peu de données structurales sur les ARN sont disponibles (Fig. 7). La détermination de la structure 3D des sous-unités du ribosome a été possible grâce aux progrès réalisés dans les méthodes de biologie structurale, en particulier pour étudier de gros complexes ribonucléoprotéiques. En effet, des avancées significatives ont été réalisées dans ces méthodes de détermination de la structure 3D des acides nucléiques [44] à la fois à haute résolution par radiocristallographie [45, 46] ou RMN [47, 48] ou à plus faible résolution par microscopie cryo-électronique [49]. Toutefois, les ARN et leur(s) complexe(s) ARN/ligand restent des objets difficiles à étudier et dont la structure 3D est rarement déterminable dans des conditions standards. Dans ce contexte, une première étape vers la compréhension des liens structure/fonction des ARN passe donc souvent par une étude initiale biochimique et éventuellement l'acquisition de contraintes structurales par des méthodes spectroscopiques ou d'autres techniques. Lorsque la structure 3D expérimentale est connue, les méthodes de modélisation

peuvent permettre d'analyser les propriétés moléculaires des ARN et notamment les propriétés électrostatiques qui conditionnent énormément les interactions des ARN avec des métaux, des protéines, etc.

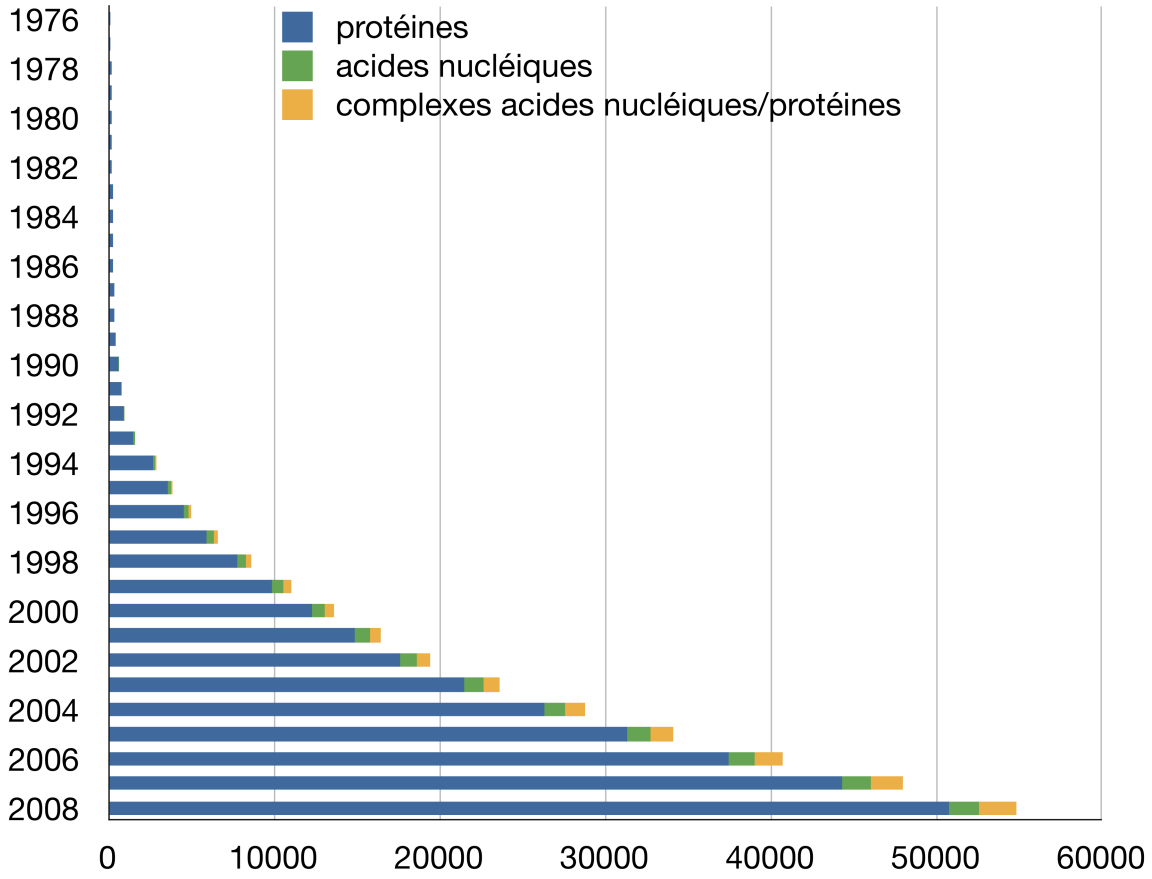


Figure 7. Statistiques du nombre de structures 3D expérimentales dans la banque RCSB-PDB. Les données sont exprimées en nombre cumulé d'entrées de la banque pour chaque année et par catégorie de macromolécules biologiques : protéines, acides nucléiques et complexes acides nucléiques/protéines. La catégorie "acides nucléiques" inclut ADN et ARN, sachant que la part ADN est assez conservée (bien qu'en légère diminution) au cours du temps avec une proportion d'environ les 2/3, les ARN représentant environ 1/3 des acides nucléiques. En 2008, la part des acides nucléiques représente 3,4% du total des entrées de la banque. Environ la moitié des entrées "acides nucléiques" correspondent à des complexes acides nucléiques/protéines.

La première structure 3D d'un ribozyme à tête de marteau biologiquement actif [11] fournissait les bases moléculaires pour comprendre la catalyse ARN. Cependant, si les variants conformationnels du ribozyme mis en évidence pouvaient expliquer les changements locaux dans la poche catalytique susceptibles de permettre l'attaque nucléophilique "in-line", ils laissaient beaucoup de zones d'ombre sur le rôle de nombreux résidus éloignés de la poche catalytique et qui pourtant sont essentiels pour la catalyse. En effet, ces résidus ne présentaient aucune interaction spécifique avec d'autres résidus ou un cation métallique. Plusieurs hypothèses pouvaient être avancées pour expliquer ces apparentes incohérences entre les données structurales et biochimiques reliées à leur implication dans des interactions dans d'autres conformations actives du ribozyme ou des conformations transitoires associées au repliement de l'ARN, ou dans des interactions avec des métaux [12]. Une tentative pour expliquer ces incohérences s'appuyait : sur l'interprétation de données de modifications chimiques de certaines bases nucléiques et sur l'analyse du profil électrostatique de l'ARN à des positions clés du ribozyme (Fig. 8) [50]. Le profil électrostatique a

révélé une région très électronégative à proximité des résidus A13, A14 et A 15.1 qui suggérait la présence d'un cation métallique [50]. La coordination d'un métal aux oxygènes non-liant des groupes phosphates des résidus A13, A14 et à l'azote N1 du résidu G8 est en mesure d'expliquer le rôle important de ces résidus. Des données biochimiques [51] et RMN [52] ont montré par la suite l'existence d'un site de fixation à cette position.

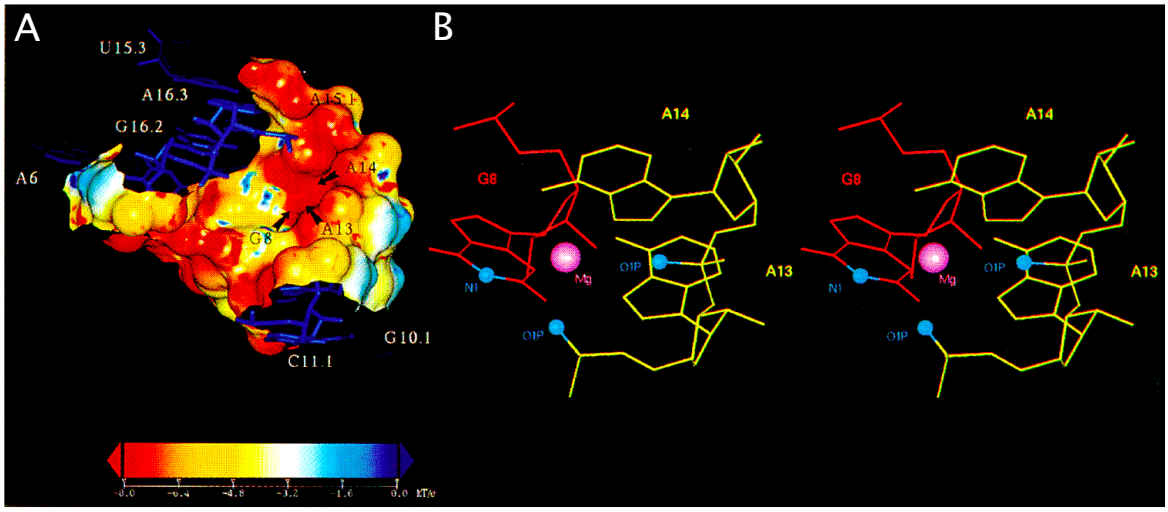


Figure 8. Prédiction d'un nouveau site de liaison pour un cation métallique du ribozyme à tête de marteau. A. Profil électrostatique dans la région du résidu A14 (calculé par le programme Delphi). B. Vue stéréo du site de fixation de Mg^{2+} prédit (coordinations avec les oxygènes non-liant des groupes phosphate de A13 et A14 et N1 de G8). D'après Chartrand *et al.*, 1997 [50].

Des données spectroscopiques (RMN ou fluorescence) peuvent être utilisées pour générer des modèles 3D en l'absence de données structurales à plus haute résolution. Des données de fluorescence ont ainsi été utilisées comme contraintes de distance pour modéliser la structure 3D du ribozyme à tête de marteau avant la détermination de sa structure 3D par radiocristallographie [10]. Le modèle prenait en compte les contraintes pour prédire l'orientation relative des 3 hélices de la jonction triple du motif "hammerhead". L'approche utilisée est essentiellement basée sur une modélisation manuelle. Ce type d'approche repose sur l'utilisation d'éléments de structure secondaire déjà identifiés dans des banques de données ou des nucléotides sont assemblés les uns aux autres en utilisant le graphisme moléculaire comme principal outil.

D'autres approches par modélisation automatique ont été développées. Le caractère automatique de ces méthodes de modélisation implique que la géométrie et la stéréochimie notamment, soient définies de façon formelle ainsi que les relations d'appariement entre nucléotides qu'ils soient canoniques (Watson-Crick ou Wobble) ou non canoniques. Ce type d'approche a été initialement popularisé par le programme Mc-Sym [53] qui a été largement utilisé avec différents types de contraintes de distance correspondant à des données de fluorescence [54], de RMN (NOE) [55], de pontages chimiques [56], etc. D'autres données expérimentales sur des covariations de séquence, obtenus par mutagénèse ou par des expériences de SELEX (" systematic evolution of ligands by exponential enrichment"), peuvent être utilisées pour déduire la présence d'appariements spécifiques qui constituent des contraintes de distance pour la modélisation 3D de structures d'ARN. Les données de SELEX permettent d'identifier des variants pouvant comporter des paires de bases isostériques qui renseignent sur le type d'appariement présent à des positions précises dans la séquence. L'information sur les paires isostériques constitue une contrainte structurale pouvant être exploitée pour la modélisation 3D [57]. L'ARN RBE est un motif tige-

boucle-tige qui constitue le site primaire de fixation de la protéine Rev du VIH-1 auquel cette approche a été appliquée avec succès [57, 58]. Bien que le modèle 3D diffère quelque peu de la structure 3D déterminée par RMN *a posteriori*, il montre une grande similarité et prédisait une ouverture du grand sillon de l'ARN du fait de la présence de deux paires non-canoniques empilées et de nucléotides non appariés apportant une flexibilité dans cette région de l'ARN insérée entre 2 tiges [57, 59]. La qualité du modèle 3D permettait de tester à nouveau les performances des méthodes de modélisation pour la prédiction de la structure 3D de complexes ARN/protéine et comprendre le lien entre le repliement spécifique de l'ARN et le mode de reconnaissance par Rev dans le grand sillon de l'ARN, habituellement peu accessible dans des régions strictement double-brin. L'utilisation d'une méthode de "docking" a permis proposer un modèle 3D d'interaction entre l'ARN RBE et le domaine basique de la protéine Rev [60]. La comparaison entre le modèle 3D et la structure RMN du complexe RBE/Rev (publiés séparément) a montré que le mode d'interaction de Rev était correctement prédit : le domaine basique de Rev (hélice-α) est dans la bonne orientation dans le grand sillon de l'ARN et la plupart des contacts établis entre résidus d'acides aminés et de nucléotides (en particulier ceux faisant intervenir les résidus essentiels) sont reproduits (Fig. 10) [59, 60].

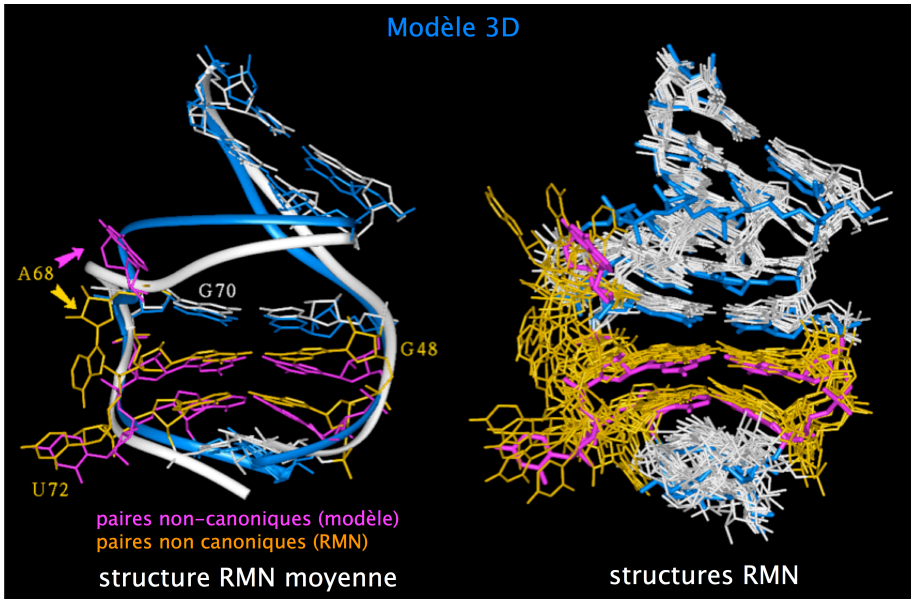


Figure 9. Comparaison entre la modèle 3D de l'ARN RBE généré par Mc-Sym et la structure 3D obtenue par RMN. Gauche : superposition du modèle 3D et de la structure 3D moyenne ; droite : superposition du modèle 3D et des 20 structures RMN. Modèle 3D : le squelette phosphodiester est représenté par un ruban (bleu), les paires par des bâtonnets pour les appariements canoniques (bleus) ou non canoniques (violets). Structure 3D : le squelette phosphodiester est représenté par un ruban (blanc), les paires par des bâtonnets pour les appariements canoniques (blancs) ou non canoniques (oranges). Les positions A68 et U72 correspondent à des nucléotides non appariés et non empilés.

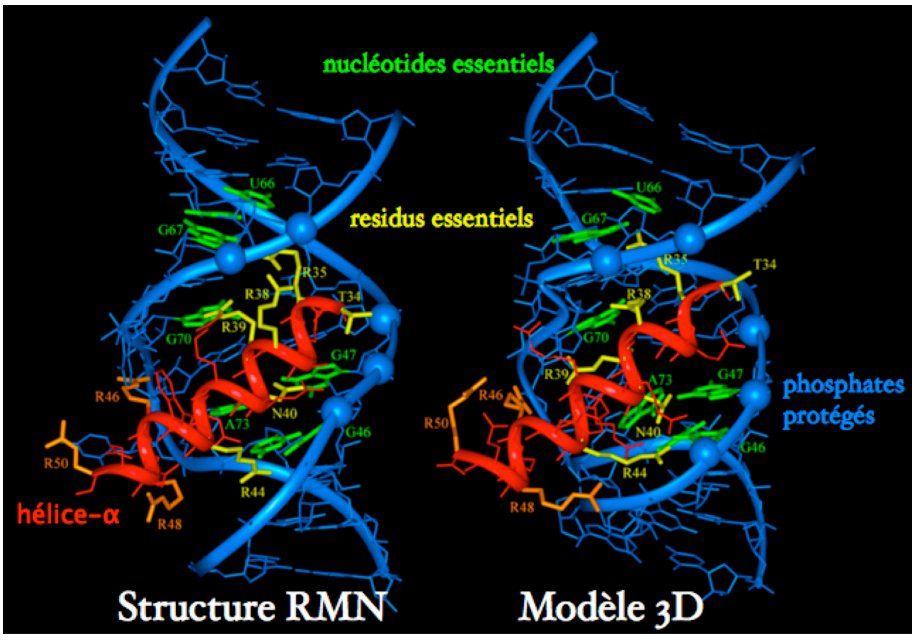


Figure 10. Comparaison entre la modèle 3D du complexe RBE/Rev généré par "docking" et la structure 3D obtenue par RMN. L'ARN (bleu) est représenté par un ruban pour le squelette phosphodiester et par des bâtonnets pour les nucléotides. Le peptide Rev (rouge), correspondant au domaine basique de la protéine, est également représenté par un ruban (hélice-α). Les résidus d'acides aminés (jaune) ou de nucléotides essentiels (vert) sont indiqués spécifiquement ; les nucléotides impliqués dans des contacts par le squelette phosphodiester sont indiqués par des sphères (5 résidus).

Chapitre 1

Les Méthodes MQ dans la Compréhension de la Réactivité des ARN Catalytiques

Sommaire

1.1 Résumé

La découverte des propriétés catalytiques des ARN a bouleversé nos conceptions en biologie moléculaire sur les rôles des ARN au cours de l'évolution, des origines de la vie jusqu'à maintenant. Du point de vue biochimique, on commence seulement à comprendre dans le détail la catalyse des ribozymes. L'enzymologie moléculaire offre des approches expérimentales pour étudier la catalyse par les ARN ; les méthodes de mécanique quantique permettent de modéliser les réactions qualitativement et quantitativement et d'essayer de comprendre l'origine du pouvoir catalytique des enzymes. C'est dans cette optique que nous avons modélisé la réaction chimique qui intervient dans le ribozyme à tête marteau qui est considéré comme un prototype dans la catalyse par les ARN car il est un des plus petits ribozymes naturels que l'on connaisse. Nous avons proposé le 1er modèle à 2 cations métalliques pour un ribozyme auto-clivable. A ce jour, le mécanisme réactionnel et les changements conformationnels associés à la catalyse ne sont pas complètement élucidés. Pour y contribuer, nous cherchons à comparer de façon théorique les catalyses de type "metallo-enzyme" et "nucléobase" ou hybrides.

1.2 Contexte

La partie de ce travail qui a donné lieu à publications a été réalisée en collaboration avec M. Karplus (ISIS-Université Louis Pasteur, Strasbourg) ainsi que dans le cadre d'un travail conjoint avec notamment Xavier Lopez (Euskal Herriko Unibertsitatea, Euskadi, Espagne), Annick Dejaegere (Université Louis Pasteur, Strasbourg) et Darrin York (University of Minnesota, USA). Les travaux en cours sont désormais développés en collaboration avec Z. Chval, Professeur à l'Université de Bohème du Sud dans le département de Physique Médicale et Biophysique (*Č*eské Budějovice, République Tchèque), qui a réalisé un stage post-doctoral d'un an et demi sous ma direction à Nancy. Cette thématique au sein de l'UMR 7567 se situe à l'interface des axes de recherche développés sur les ARN dans l'équipe de Maturation des ARN et sur la catalyse, dans l'équipe d'Enzymologie Moléculaire. Elle bénéficie du soutien l'IDRIS (Institut du Développement et des Ressources en Informatique Scientifique, Orsay) en terme de ressources de calcul.

1.3 Introduction

Jusqu'au début des années 2000, tous les ribozymes étaient considérés habituellement comme des métallo-enzymes en raison de la « pauvreté » du répertoire de groupes chimiques des bases nucléiques [61]. Elles étaient du coup jugées peu réactives et donc peu susceptibles de contribuer à la réactivité chimique. L'idée était donc que, dans ces métallo-enzymes, les cations métalliques jouent le rôle d'acide/base de Lewis et/ou d'acide/base générale en fonction de la géométrie du site actif et de son environnement [62]. Parmi les ARN naturels à activité catalytique, il existe 2 grandes familles : 1) les ARN auto-clivables, 2) les ARN auto-épissables ; outre leurs différences structurales, les premiers (ribozymes à tête de marteau, ribozyme du virus de l'hépatite D, ribozyme en hairpin) conduisent à des produits de réaction portant des extrémités 5'-OH et 2'-3'-cyclique (ensuite hydrolysé), alors que les seconds (ribozymes des introns de groupe I et de groupe II, spliceosome) génèrent des produits avec des extrémités 5'-phosphate et 3'-OH [5].

La découverte récente de l'implication de bases nucléiques dans la catalyse par des ribozymes auto-clivables [17, 63] et celle des métaux dans la catalyse par des ribozymes auto-épissables [64, 65] suggère une catalyse de type « nucléobase » pour les premiers et de type « métallo-enzyme » pour les seconds (Fig. 11). Bien que la catalyse de type « nucléobase » puisse apparaître comme plus efficace en ce qui concerne les étapes chimiques liées à l'activation de nucléophiles (par l'intervention d'une base nucléique dans des transferts de protons), aucune explication quantitative ne permet de comprendre exactement l'avantage d'une catalyse de type « nucléobase » par rapport à une catalyse par des métaux ou vice versa. La classification entre ribozymes auto-clivables à catalyse de type « nucléobase » (Fig. 11B-E) et ribozymes auto-épissables à catalyse de type « métallo-enzyme » (Fig. 11F) n'est pas non plus définitive ou absolue. En l'occurrence, les données expérimentales les plus récentes sur certains ribozymes à tête de marteau suggèrent l'intervention probable d'une base nucléique dans la catalyse [14] sans pour autant que les métaux ne soient totalement exclus de rôle(s) catalytique(s) [66].

1.4 Contribution des méthodes de chimie théorique à la compréhension des mécanismes d'action des ribozymes

Depuis la découverte des premiers ARN catalytiques il y a maintenant plus de 20 ans, les connaissances sur les liens structure/fonction des ribozymes se sont considérablement accrues

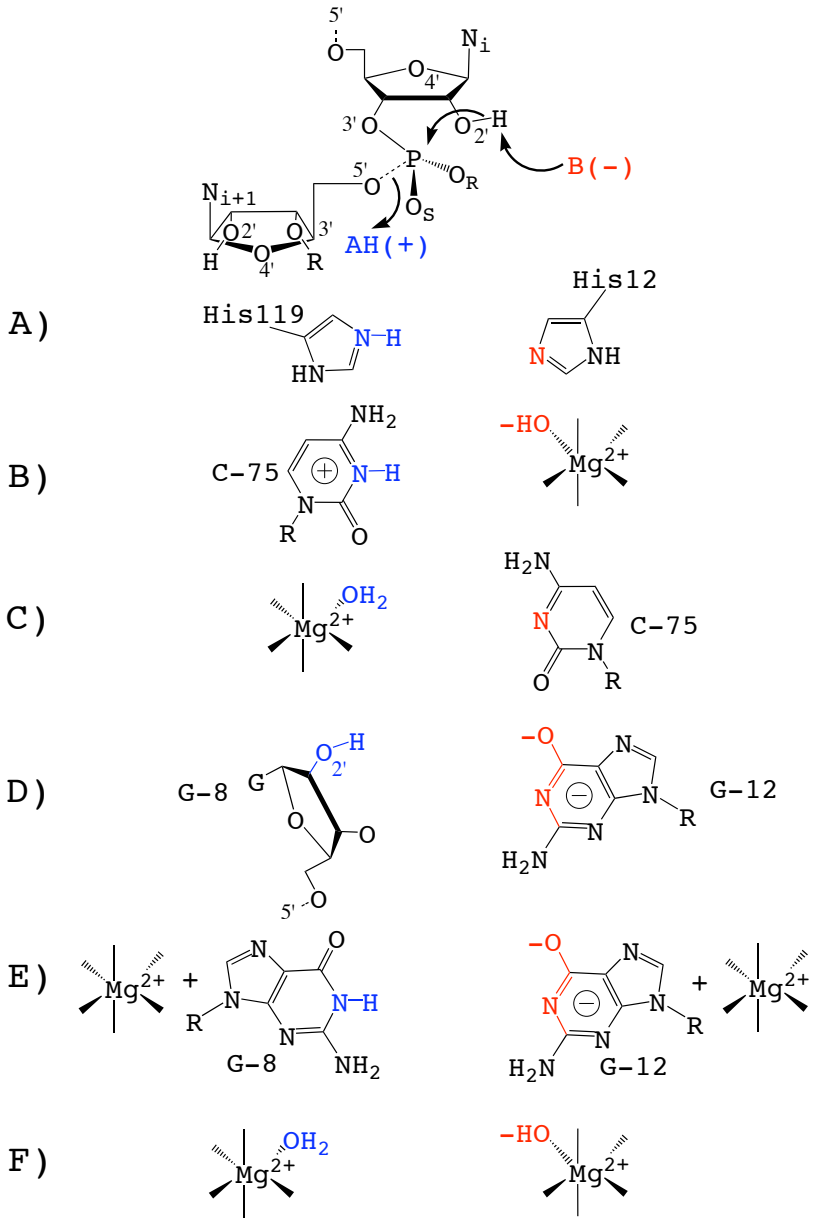


Figure 11. Stratégies catalytiques dans l'hydrolyse de groupes phosphates. A) Catalyse Acide-Base Générale dans la RNase A. Deux résidus histidine interviennent comme donneur et accepteur de protons pour activer le nucléophile (B-) et faciliter le départ du groupe partant (AH+). B) Catalyse acide générale dans le ribozyme HDV [67]. C) Catalyse base générale dans le ribozyme HDV [68]. D) Catalyse acide/base générale dans le ribozyme à tête de marteau [14]. E) Catalyse acid-base générale dans le ribozyme à tête de marteau [66]. Les cations métalliques contribuent à stabiliser les formes tautomériques et anioniques des bases nucléiques qui participent à la catalyse. F) Catalyse de type « métallo-enzyme » (mécanisme à 2 cations métalliques) dans ribozyme à tête de marteau [57, 69].

grâce à des approches expérimentales faisant appel à la biologie moléculaire, la chimie des acides nucléiques, l'enzymologie ou encore la biologie structurale. Elles ont montré en particulier l'importance des cations métalliques en permettant de localiser leurs sites de liaison à proximité des sites de clivage (rôle catalytique) ou au contraire éloignés (rôle structural) ; elles ont aussi révélé le rôle de bases nucléiques participant à la catalyse. Toutefois, le détail à l'échelle atomique des mécanismes catalytiques reste difficilement accessible par ces approches. Par contre, les approches théoriques (mécanique quantique, simulations Car-Parrinello, etc) offrent une bonne complémentarité en permettant de construire des modèles atomiques de mécanismes catalytiques et des chemins réactionnels complets du réactant au produit. Les méthodes dites « hybrides » combinant la mécanique moléculaire et la mécanique quantique permettent aussi de tenir compte des contributions aux barrières énergétiques provenant à la fois de la formation ou de la rupture de liaisons chimiques et des changements conformationnels qui jouent un rôle important dans les ARN et plus particulièrement les ribozymes. Les méthode théoriques restent toutefois très coûteuses en ressources de calcul, surtout pour des systèmes complexes qui incluent des cations métalliques solvatés. L'ensemble de ces approches ont permis de proposer des mécanismes réactionnels pour plusieurs ribozymes où les 2 rôles catalytiques principaux sont associés à l'activation du nucléophile (par une base) et au départ du groupe partant (facilité par un acide), (Fig. 11).

Une façon d'aborder le problème du rôle des cations métalliques dans la catalyse de type « métallo-enzyme » ou des bases nucléiques dans la catalyse de type « nucléobase » est d'utiliser des approches de chimie théorique. Notre objectif est donc d'utiliser ces approches pour proposer un modèle pour chaque type de catalyse pour les ribozymes à têtes de marteau et d'identifier les facteurs qui influencent les barrières d'énergie d'activation à l'échelle atomique (rôle des cations métalliques et des nucléobases) et à l'échelle moléculaire (changements conformationnels associés à la catalyse). La comparaison entre les 2 types de catalyse devrait donner des informations sur l'évolution moléculaire possible des ribozymes. La catalyse de type « métallo-enzyme » est plutôt considérée comme ancestrale ; l'évolution vers une catalyse de type « nucléobase » aurait pu se faire à une échelle locale atomique grâce à un abaissement de barrières énergétiques, ou à une échelle globale moléculaire par un réarrangement du site actif du ribozyme grâce à un repliement plus favorable de l'ARN (sans élévation des barrières énergétiques).

1.5 Modèles théoriques pour la catalyse de type « métallo-enzyme »

Plusieurs mécanismes hypothétiques sur la catalyse de type « métallo-enzyme » pour les ribozymes à tête de marteau ont été avancées dès les années 90 afin de rendre compte de résultats expérimentaux de sources diverses [11, 70]. Deux grands types de mécanismes sont proposés qui se distinguent par le nombre et le rôle des métaux impliqués dans la catalyse. Le premier type de mécanisme est largement inspiré de la catalyse acide/base générale de la RNase A, où un seul métal agit comme base pour activer le 2'-OH et éventuellement comme acide pour faciliter le départ du groupe partant [70]. Quant au second type, il reprend davantage les mécanismes d'exonucléase ou phosphatases ou encore des introns de groupe I ou de groupe II, où deux métaux interviennent comme acides de Lewis [71, 72]. On parle de mécanisme à un cation métallique ou de mécanisme à deux cations métalliques pour se référer à chacun des deux types de catalyse. Bien que ces propositions aient été disponibles pendant de nombreuses années, c'est seulement dans les années 2000 que sont publiés les premiers modèles théoriques pour une catalyse de type « métallo-enzyme ».

Le premier modèle publié de catalyse « métallo-enzyme » proposé correspond à un mécanisme à un cation métallique [73]. Bien que la barrière énergétique calculée dans ce modèle soit comparable à la barrière d'énergie libre déterminée expérimentalement, le modèle part d'une conformation initiale du substrat « activée », identifiée lors d'une simulation par dynamique moléculaire, qui néglige la barrière énergétique nécessaire à ce changement conformationnel. La conformation initiale du subtrat utilisée comme réactant correspond en fait à un minimum local moins stable d'environ 10 kcal/mol par rapport à la conformation « idéale » (Zdenek & Leclerc, en préparation). La barrière proposée dans ce modèle est donc vraisemblablement largement sous-estimée.

Le modèle à 2 cations métalliques que nous avons développé durant ces 4 ans [74], en utilisant une méthodologie similaire, a été publié à peu près en même temps qu'une comparaison entre les deux types de mécanismes établie par une approche différente (Car-Parinello) [69]. Cette dernière approche vise à reproduire la dynamique du processus catalytique plutôt que l'optimisation des points stationnaires situés sur le chemin réactionnel. Les conclusions de la comparaison des 2 types de mécanisme par simulations de dynamique moléculaire suggèrent que la mécanisme à 2 cations métalliques est plus favorable. Toutefois, les barrières énergétiques sont très élevées et peu réalistes par rapport aux données expérimentales, ce qui suggère que les chemins réactionnels proposés sont sous-optimaux. En revanche, dans le modèle que nous avons établi, les métaux établissent des coordinations directes avec l'oxygène 2' et/ou des oxygènes du groupe phosphate (en accord avec certaines données expérimentales), ce qui contribue à abaisser fortement les barrières d'énergie comme nous l'avons montré (Fig. 12). Notre modèle est d'ailleurs en bon accord avec les données thermodynamiques et cinétiques (calculs d'effets isotopiques) ; il suggère non seulement un rôle important des métaux dans l'abaissement des barrières d'énergie mais aussi une synergie entre les métaux aux 2 sites catalytiques à proximité du nucléophile et du groupe partant [7, 74].

1.6 Modèles théoriques pour la catalyse de type « nucléobase »

Le débat sur le mécanisme catalytique prédominant agissant dans les ribozymes à tête de marteau reste ouvert et encore sujet à polémique : à savoir s'il relève d'une catalyse de type « métallo-enzyme » ou au contraire de type « nucléobase ». Après avoir modélisé la catalyse de type « métallo-enzyme », il était intéressant d'essayer de proposer un modèle alternatif faisant appel à la catalyse de type « nucléobase » afin de comparer les deux stratégies catalytiques. Nous sommes partis de deux propositions de mécanismes catalytiques de type « nucléobase » qui ont été faites par 2 groupes différents pour les ribozymes à tête de marteau (Fig. 13). La première proposition correspond à une catalyse « nucléobase » dans laquelle les métaux n'ont pas de rôle catalytique dans le site actif du ribozyme [14] (Fig. 13A). La seconde proposition est également conforme à une catalyse « nucléobase » mais fait également intervenir un métal comme cofacteur dans la catalyse : elle correspond à une catalyse « nucléobase » hybride [66] (Fig. 13B).

Nous avons construit, en collaboration avec Z. Chval, un premier modèle théorique correspondant à une catalyse purement « nucléobase » (Fig. 13A). La barrière énergétique pour la première étape du mécanisme, l'activation du nucléophile, a pu être calculée. Une comparaison avec le modèle de catalyse « métallo-enzyme » dans l'étape d'activation montre que la catalyse « nucléobase » n'offre pas *a priori* d'avantage énergétique : la barrière d'énergie est d'environ 7 kcal/mol alors qu'elle n'est que de 2 kcal/mol ou même nulle dans les meilleurs cas de la catalyse « métallo-enzyme » (Fig. 14). Même si la catalyse « métallo-enzyme » exige une pré-activation

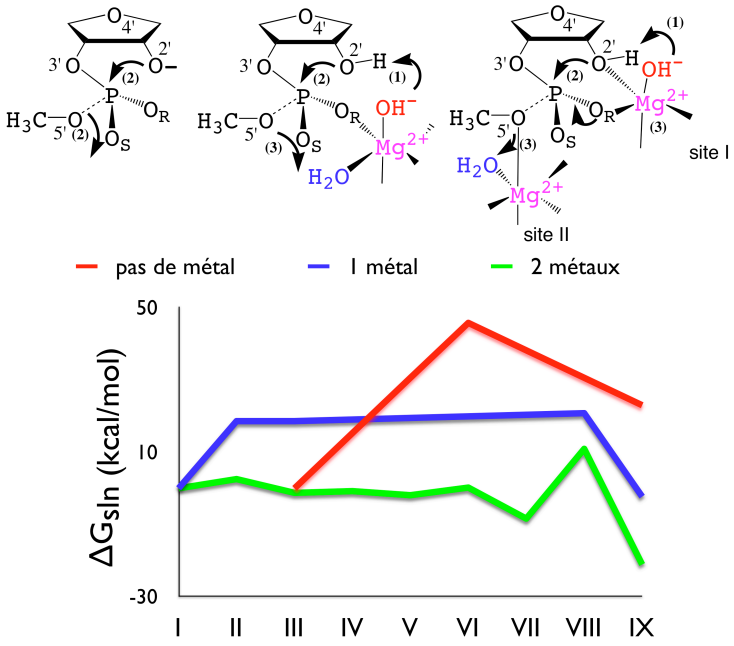


Figure 12. Rôle des métaux dans les mécanismes réactionnels à catalyse « métallo-enzyme » d'après des calculs de MQ. La réaction non catalysée ("pas de métal", rouge) présente une barrière d'énergie libre élevée [7]. La réaction catalysée en présence d'un métal ("1 métal", bleu) a une barrière d'énergie abaissée [73]. La réaction catalysée en présence de 2 métaux ("2 métaux", vert) présente la barrière d'énergie la plus faible [74]. Les étapes principales de la réaction sont numérotées : (1) activation du nucléophile, (2) attaque nucléophile, (3) départ du groupe partant.

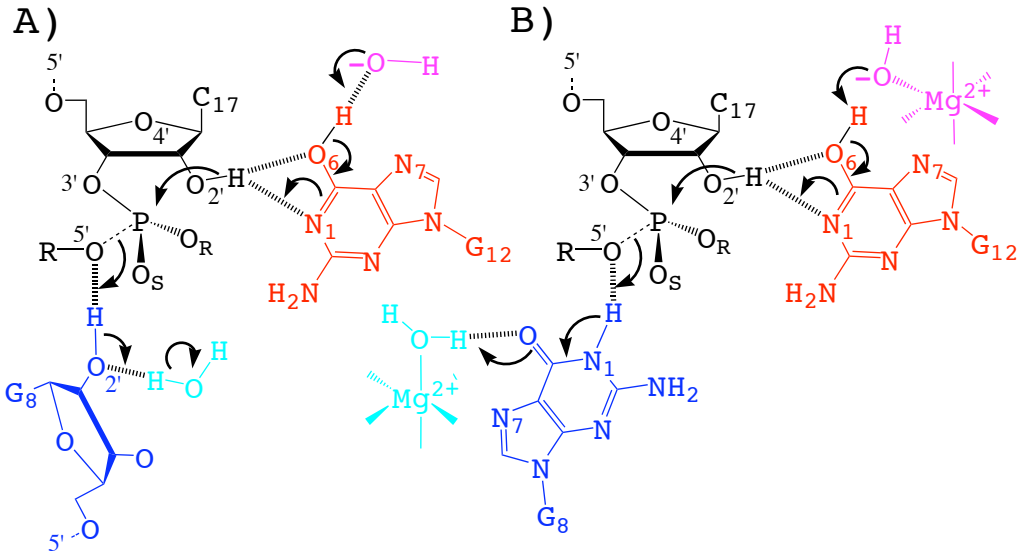


Figure 13. Modèles probables de catalyse de type « nucléobase » pour le ribozyme à tête de marteau. A) Mécanisme acide/base générale proposé par Martick & Scott (2006). G-12 est impliquée dans l'activation du nucléophile (2'-OH) ; G-8 facilite le départ du groupe partant en transférant un proton du 2'-OH sur l'oxygène 5'. B) Mécanisme acid-base générale inspiré de Roychowdhury-Saha & Burke (2006). G-12 est impliqué dans l'activation du 2'-OH, un cation métallique stabilise la forme tautomérique et facilite l'activation du 2'-OH. G-8 facilite le départ du groupe partant par le transfert du proton du (N1)H sur l'oxygène 5'.

dont la barrière est d'environ 10 kcal/mol, la barrière dans l'étape d'activation de la catalyse « nucléobase » ne tient pas compte d'une pré-activation due à un changement conformationnel lié au re-positionnement du résidu G12 dans le site actif du ribozyme. Il est donc raisonnable de penser que l'avantage d'une catalyse « nucléobase » ne se situe à l'étape d'activation mais au stade de l'attaque nucléophile ou du départ du groupe partant. Les calculs complémentaires qui sont en cours pour les autres étapes de la réaction devraient permettre d'identifier les facteurs susceptibles de favoriser davantage une catalyse de type « nucléobase ».

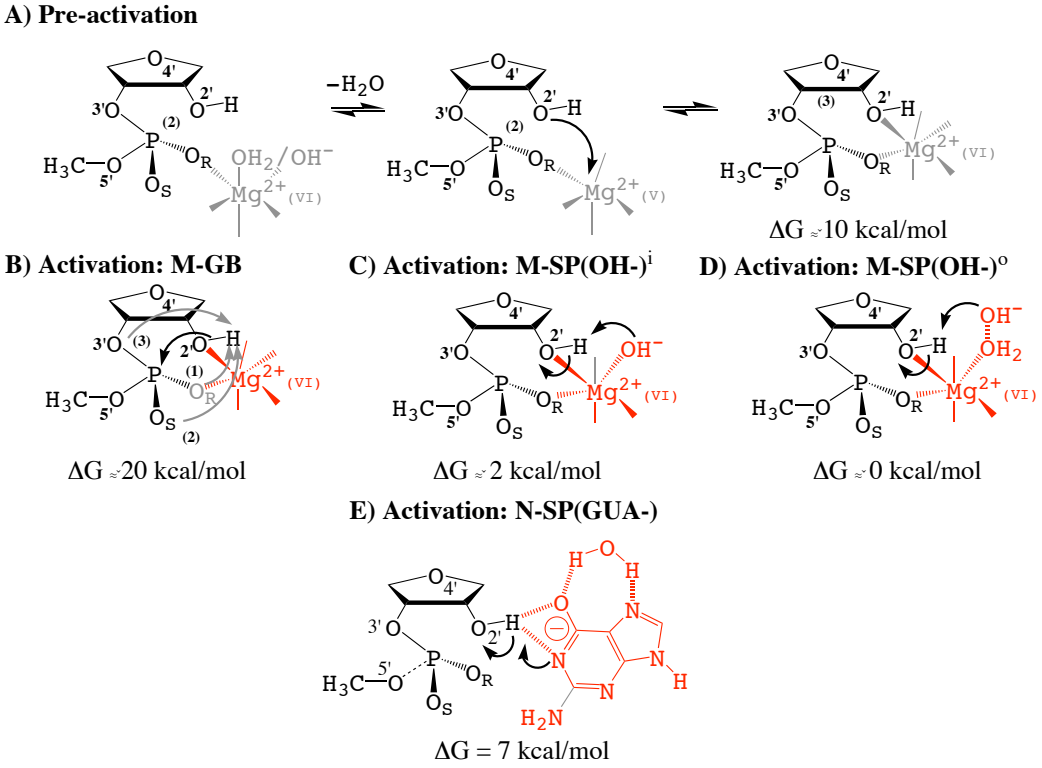


Figure 14. Mécanismes d'activation dans les catalyse de type « métallo-enzyme » et « nucléobase ». A. Les mécanisme de type « métallo-enzyme » réclament une étape de pré-activation qui permet de placer le métal dans une configuration favorable à l'activation du nucléophile : le groupe 2'-OH. B. Activation de type « métallo-enzyme » à base générale. C. Activation de type « métallo-enzyme » à base spécifique ; la base est directement coordinée au métal. D. Activation de type « métallo-enzyme » à base spécifique ; la base est coordinée au métal de façon indirecte. E. Activation de type « nucléobase » ; la base est une forme anionique tautomérique de la guanine (stabilisée par une molécule d'eau). Les barrières energétiques sont indiquées à un niveau de théorie MP2-COSMO/6-311+G(2d,2p)//B3LYP/6-31+G*.

1.7 Travaux publiés

1.7.1 "Two-Metal-Ion Mechanism for Hammerhead-Ribozyme Catalysis"

Two-Metal-Ion Mechanism for Hammerhead-Ribozyme Catalysis

Fabrice Leclerc*,†,‡,§ and Martin Karplus*,‡,§

Laboratoire de Maturation des ARN et Enzymologie Moléculaire, Université Henri Poincaré, Faculté des Sciences, B.P. 239, Bd. des Aiguillettes, 54506 Vandoeuvre-lès-Nancy, France, Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138, and Laboratoire de Chimie Biophysique, Université Louis Pasteur, Institut le Bel, 67000 Strasbourg, France

Received: July 12, 2005; In Final Form: November 2, 2005

The hammerhead ribozyme is one of the best studied ribozymes, but it still presents challenges for our understanding of RNA catalysis. It catalyzes a transesterification reaction that converts a 5′,3′ diester to a 2′,3′ cyclic phosphate diester via an S_N2 mechanism. Thus, the overall reaction corresponds to that catalyzed by bovine pancreatic ribonuclease. However, an essential distinguishing aspect is that metal ions are not involved in RNase catalysis but appear to be important in ribozymes. Although various techniques have been used to assign specific functions to metals in the hammerhead ribozyme, their number and roles in catalysis is not clear. Two recent theoretical studies on RNA catalysis examined the reaction mechanism of a single-metal-ion model. A two-metal-ion model, which is supported by experiment and based on ab initio and density functional theory calculations, is described here. The proposed mechanism of the reaction has four chemical steps with three intermediates and four transition states along the reaction pathway. Reaction profiles are calculated in the gas phase and in solution. The early steps of the reaction are found to be fast (with low activation barriers), and the last step, corresponding to the departure of the leaving group, is rate limiting. This two-metal-ion model differs from the models proposed previously in that the two metal ions function not only as Lewis acids but also as general acids/bases. Comparison with experiment shows good agreement with thermodynamic and kinetic data. A detailed analysis based on natural bond orbitals (NBOs) and natural energy decomposition (NEDA) provides insights into the role of metal ions and other factors important for catalysis.

1. Introduction

The discovery of catalytic RNA molecules (ribozymes) in the early 1980s,[1,2] at a time when proteins were thought to be the only enzymes, raised the fundamental question of how RNA enzymes work. Although ribozymes have been under intense study for the intervening years, no mechanism that provides a detailed description of the reaction is universally accepted for any ribozyme. Among the various known RNA enzymes, the best-characterized is the hammerhead ribozyme. It was the first ribozyme to be crystallized, and a series of X-ray structures corresponding to a biologically active ribozyme have been determined.[3−6] This ribozyme has also been the subject of numerous biochemical studies, yet questions remain regarding the reaction mechanism.[7−9] Like the RNA-catalyzed self-cleavage of other ribozymes,[10] the reaction catalyzed by the hammerhead ribozyme involves a transesterification step in the phosphate ester hydrolysis.[11] This step leads to isomerization from a 5′,3′ diester to a 2′,3′ cyclic phosphate diester. In a second step, the 2′,3′ cyclic phosphate is hydrolyzed to yield a 3′ phosphate and regenerate the 2′ OH group. The transesterification reaction has been shown to proceed via an S_N2(P) or "in-line" mechanism in which the attacking nucleophile (the 2′ oxygen) is aligned with the phosphorus atom and the 5′ oxygen atom of the phosphate group from the neighboring 3′ nucleotide.[12−14] Thus, the overall mechanism corresponds to that found in bovine pancreatic RNase,[15] although metal ions, which appear to play an essential role for the ribozymes, are not present in RNase. Models proposed for the reaction mechanism differ particularly with respect to the number of metal ions involved (single-metal-ion mechanisms[16−20] or two-metal-ion mechanisms[21,22]) and their specific role in the catalysis; that is, whether they act as a general acid/base, an electrophilic catalyst, or a Lewis acid[19] (Figure 1). When the metal is involved in the deprotonation of a nucleophile or in the protonation of a leaving group, it can function either as a Lewis acid or as a generalized acid/base. The metal acts as a Lewis acid if it stabilizes an anionic nucleophile or leaving group (by direct coordination of the metal to the oxygens of the phosphate group) but does not participate directly in the proton transfer, while it acts as a general acid or base when it is directly involved as a proton donor (from hydrated metal) or acceptor (by metal hydroxide). The metal functions as an electrophilic catalyst when it activates the electrophile (the phosphorus atom) by making it more susceptible to nucleophilic attack (by direct coordination of the metal to the nonbridging pro-R or pro-S oxygens). The single-metal-ion mechanisms are mostly based on a general acid/base model of catalysis (Figure 1B), while the two-metal-ion mechanisms are mostly based on a Lewis acid model of catalysis (Figure 1C and D). The experimental data, originally supporting a single-metal-ion mechanism (Figure 1A),[23−25] were shown subsequently to be more consistent with a two-metal-ion mechanism (Figure 1B and C).[26] Since then, additional experimental evidence has accumulated in favor of a two-metal-ion

* To whom correspondence should be addressed. E-mail: fabrice.leclerc@maem.uhp-nancy.fr (F.L.); marci@tammy.harvard.edu (M.K.).
† Université Henri Poincaré.
‡ Harvard University.
§ Université Louis Pasteur.

10.1021/jp053835a CCC: $33.50
Published on Web 02/01/2006

1.7.2 "Nucleophilic attack on phosphate diesters : a density functional study of in-line reactivity in dianionic, monoanionic, and neutral systems"

Nucleophilic Attack on Phosphate Diesters: A Density Functional Study of In−Line Reactivity in Dianionic, Monoanionic, and Neutral Systems

Xabier Lopez ,*,† Annick Dejaegere,‡ Fabrice Leclerc,§ Darrin M. York,|| and Martin Karplus*,⊥, #

Kimika Fakultatea, Euskal Herriko Unibertsitatea, P.K. 1072, 20080 Donostia, Euskadi, Spain, Laboratoire de Biologie et Génomique Structurales, Ecole Supérieure de Biotechnologie de Strasbourg, 674000 Illkirch, France, Laboratoire de Maturation des ARN et Enzymologie Moléculaire, CNRS-UHP Nancy I UMR 7567, Université Henri Poincaré, Faculté des Sciences, B.P. 239, 54506 Vandoeuvre-lès-Nancy, France, Department of Chemistry, University of Minnesota, 207 Pleasant St. SE, Minneapolis, Minnesota 55455-0431, Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, and Laboratoire de Chimie Biophysique, Institut le Bel, Université Louis Pasteur, 67000 Strasbourg, France

Received: January 19, 2006; In Final Form: April 18, 2006

A density functional study of the hydrolysis reaction of phosphodiesters with a series of attacking nucleophiles in the gas phase and in solution is presented. The nucleophiles HOH, HO^-, CH_3OH, and CH_3O^- were studied in reactions with ethylene phosphate, 2′3′-ribose cyclic phosphate and in their neutral (protonated) and monoanionic forms. Stationary-point geometries for the reactions were determined at the density functional B3LYP/6-31++G(d,p) level followed by energy refinement at the B3LYP/6-311++G(3df,2p) level. Solvation effects were estimated by using a dielectric approximation with the polarizable continuum model (PCM) at the gas-phase optimized geometries. This series of reactions characterizes factors that influence the intrinsic reactivity of the model phosphate compounds, including the effect of nucleophile, protonation state, cyclic structure, and solvent. The present study of the in-line mechanism for phosphodiester hydrolysis, a reaction of considerable biological importance, has implications for enzymatic mechanisms. The analysis generally supports the associative mechanism for phosphate ester hydrolysis. The results highlight the importance for the reaction barrier of charge neutralization resulting from the protonation of the nonbridging phosphoryl oxygens and the role of internal hydrogen transfer in the gas-phase mechanism. It also shows that solvent stabilization has a profound influence on the relative barrier heights for the dianionic, monoanionic, and neutral reactions. The calculations provide a comprehensive data set for the in-line hydrolysis mechanisms that can be used for the development of improved semiempirical quantum models for phosphate hydrolysis reactions.

1. Introduction

Phosphate diesters play a fundamental role in biology, including their role as the backbone of DNA and RNA.[1,2] The chemical properties and reactivity of phosphates determine how these biomolecules are formed and cleaved, and therefore, phosphate diesters have been the subject of numerous theoretical and experimental studies.[3−26] A first approximation to understand phosphate diester reactivity is to characterize the energetics of gas-phase model reactions for which a full quantum mechanical treatment is possible. Of particular interest are studies that cover nucleophilic attack of water or methanol on ethylene phosphate (EP), a model for the transphosphorylation and hydrolysis of RNA chains.

Previous computational studies on nonenzymatic phosphate hydrolysis reactions[3,5,9,11−13,17,23,24,27] have focused mainly on the so-called dianionic and monoanionic reaction mechanisms in which either a negatively charged nucleophile (XO^-; X = H, CH_3) or a neutral one (XOH; X = H, CH_3) attacks an unprotonated phosphate diester molecule (i.e., dimethyl phosphate or ethylene phosphate). Very few of these studies has addressed the effect of the sugar ring on the reaction. The pK_a values of phosphates (typically below 3) suggest that they are ionized in aqueous solution around neutral pH and, hence, are the most likely reactant species in nonenzymatic hydrolysis. However, the protonation state of the phosphate esters in enzymatic hydrolysis, especially of the phosphorane transition states and intermediates, are not clear. Recent experimental and theoretical results suggest that the phosphoranes exhibit significantly elevated pK_a[1] (e.g., a value of 7.9 has been suggested for ethylene phosphorane[27] and 8.6 for $P(OH)_5$[28]). Consequently, it is important also to characterize neutral reaction mechanisms because the charge state of the phosphorus species can play an important role in the enzyme-catalyzed reactions.

In this paper, we extend the scope of previous computational work[16,29] to consider neutral reaction mechanisms and to characterize the effect of the sugar ring at a high level of theory and basis set. We have studied by means of density functional theory the following set of reactions:

$$XO(H) + PO_2(H)(-O-Y-O) \rightarrow XO-PO_2(H) - (O-Y-O(H)) \quad (1)$$

where X can be either hydrogen or methyl, and Y is either C_2H_4

* Corresponding authors. E-mail: xabier.lopez@ehu.es (X.L.); marci@tammy.harvard.edu (M.K.).
† Kimika Fakultatea, Euskal Herriko Unibertsitatea.
‡ Laboratoire de Biologie et Génomique Structurales, Ecole Supérieure de Biotechnologie de Strasbourg.
§ Laboratoire de Maturation des ARN et Enzymologie Moléculaire, Université Henri Poincaré.
|| Department of Chemistry, University of Minnesota.
⊥ Department of Chemistry and Chemical Biology, Harvard University.
Laboratoire de Chimie Biophysique, Institut le Bel, Université Louis Pasteur.

10.1021/jp0603942 CCC: $33.50
Published on Web 05/19/2006

Perspectives : Etude théorique du rôle des cations métalliques dans la catalyse de type « nucléobase »

Nos modèles théoriques obtenus sur la catalyse « métallo-enzyme » suggèrent que les cations métalliques jouent un rôle important dans l'abaissement de la barrière d'énergie d'activation de la réaction par rapport à une réaction non catalysée. Ils montrent aussi qu'une contribution non négligeable à la barrière d'activation peut provenir d'une étape de pré-activation qui a été associée du point de vue expérimentale au changement de conformation entre une forme native mais inactive du point de vue biologique et une forme biologiquement active (avec un positionnement adéquat des atomes dans le site actif pour l'attaque nucléophile). Les premiers résultats obtenus sur un modèle de catalyse « nucléobase » suggère que cette stratégie catalytique n'offre pas d'avantage énergétique significatif dans l'étape d'activation des ribozymes à tête de marteau.

Les travaux en cours visent à identifier un chemin réactionnel complet dans la catalyse « nucléobase » à partir de l'intermédiaire formé après activation et déjà identifié. L'estimation de la barrière globale d'énergie libre sur ce modèle permettra d'évaluer en quoi la catalyse « nucléobase » peut être avantageuse et à quelle étape de la réaction. Toutefois, pour réaliser une comparaison précise, il est nécessaire de tenir compte d'éventuelles étapes de pré-activation. Dans le cas de la catalyse « métallo-enzyme », l'étape de pré-activation a été évaluée. Dans le cas de la catalyse « nucléobase », la pré-activation implique la simulation du changement conformationnel qui conduit au réarrangement du site actif dans une configuration biologiquement active. Grâce à des méthodes de dynamique moléculaire utilisant des contraintes géométriques, nous simulerons le changement de conformationnel qui permet de passer d'une forme inactive et à forme active du ribozyme à tête de marteau, les structures 3D des deux formes ayant été déterminées par diffraction des rayons-X. En complément, des méthodes hybrides alliant mécanique moléculaire et mécanique quantique permettront d'intégrer avec un seule méthodologie l'ensemble des contributions énergétiques qui proviennent à la fois de changements locaux à l'échelle atomique dans le site actif et de changements globaux à l'échelle de la molécule d'ARN.

La stratégie catalytique des ribozymes à tête de marteau fait peut-être appel, comme le suggèrent Roychowdhury-Saha & Burke (2006) [66], à une catalyse qui n'est purement « nucléobase » mais hybride entre les deux catalyses « nucléobase » et métallo-enzyme ». Nous tenterons également d'évaluer ce modèle hybride de catalyse et d'identifier l'avantage que peut représenter une telle stratégie par rapport à une catalyse purement « nucléobase » ou « métallo-enzyme » (Fig. 15).

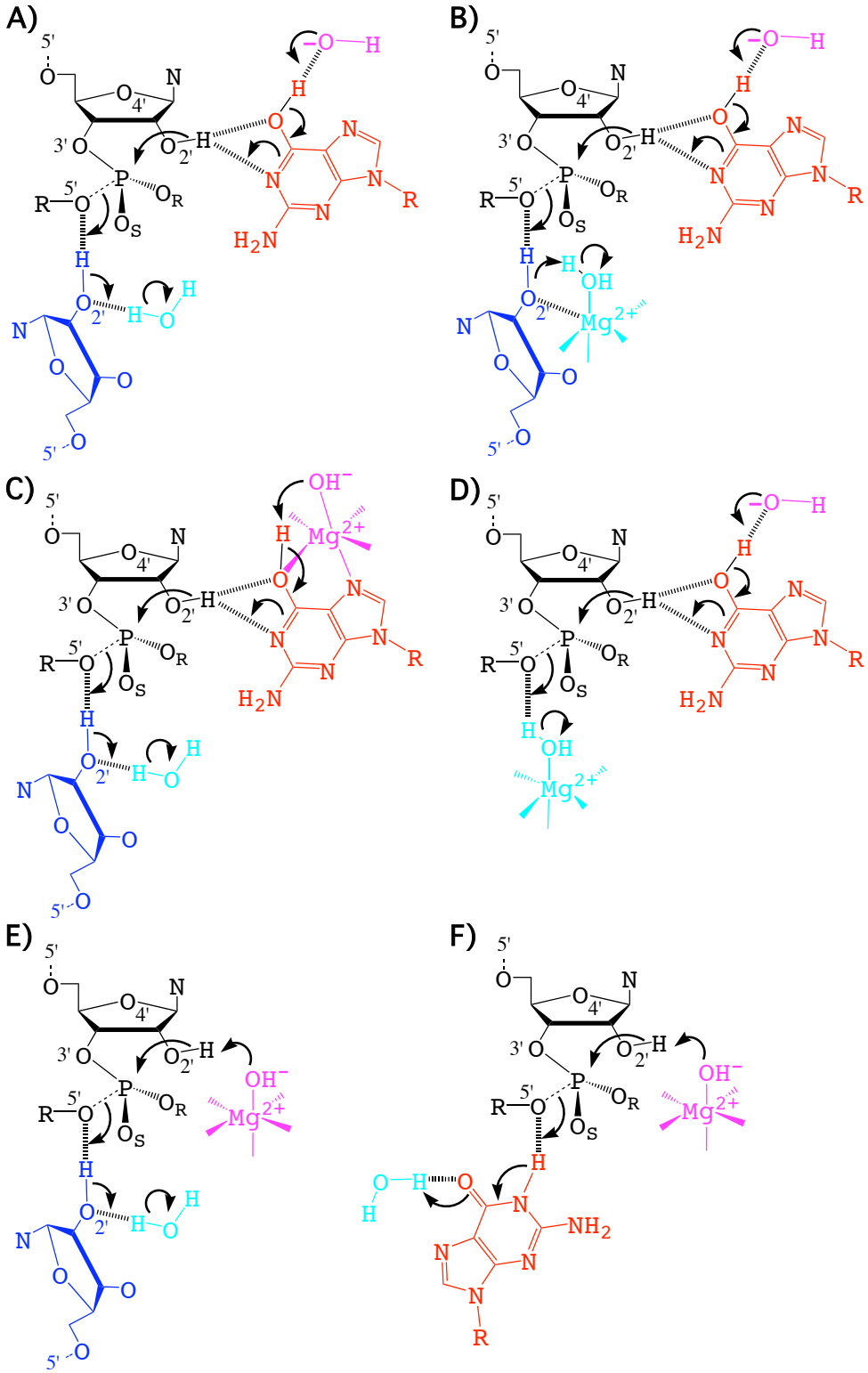


Figure 15. Modèles alternatifs de catalyses mixtes métallo-enzyme/nucléobase des ribozymes à tête de marteau. A. Modèle nucléobase proposé par Martick & Scott (2006) [14]. B. Modèle nucléobase avec départ du groupe partant co-assisté par un métal. C. Modèle nucléobase avec activation nucléophilique co-assistée par un métal. D. Modèle mixte avec départ du groupe partant assisté par un métal. E. Modèle mixte avec activation nucléophilique assistée par un métal. F. Modèle mixte avec activation nucléophilique assistée par un métal.

Chapitre 2

La Bioinformatique en Génomique Comparative : Application à la Recherche de Gènes d'ARNnc

Sommaire

2.1 Résumé

Des découvertes récentes ont mis en évidence l'existence de nombreux ARN non-codant impliqués dans de nombreuses et diverses fonctions biologiques. De plus, le fait que le génome humain soit transcrit de façon très étendue et qu'il existe de nombreux éléments phylogénétiquement conservés suggère l'existence de nombreux ARN fonctionnels dans les régions non codantes des génomes d'eucaryotes supérieurs et d'autres organismes. Dans cette thématique, nous nous

sommes focalisés sur un domaine du vivant encore peu étudié : les archaea. Ces organismes rassemblent des espèces qui se multiplient dans des conditions d'environnement très variées, que ce soit en termes de température, de pression, de salinité mais aussi dans des conditions plus standards et qui en font même parfois des hôtes dans le tractus digestif d'animaux et de l'homme en particulier. Ils ont l'avantage d'avoir de petits génomes sur lesquels il est possible de rechercher de façon rapide et efficace des gènes d'ARN non-codant. Nous avons développé et mis en œuvre une approche bio-informatique pour rechercher des gènes d'ARN non-codant dans ces génomes qui a été largement validée expérimentalement sur une classe particulière d'ARN non-codant que sont les ARN guides de modifications à boîtes H/ACA. Des développements permettant d'améliorer la sensibilité de la méthode sont prévus. D'autre part, les connaissances acquises sur les ARN à boîtes H/ACA ouvrent des perspectives dans la compréhension d'une maladie humaine (la dyskératose) reliée à un dysfonctionnement associée à cette famille d'ARN non-codant chez l'homme. Le prolongement de cette thématique abordée à l'aide d'outils bio-informatiques s'appuiera sur l'utilisation de méthodes de modélisation 3D développées par ailleurs.

2.2 Contexte

Ces travaux sur la recherche de gènes d'ARN non-codant ont été menés avec un étudiant (Sébastien Muller) qui a soutenu sa thèse le 27 novembre 2007, intitulée : « Développement et application de méthodes bioinformatiques pour la recherche de gènes de sRNA à boîtes H/ACA dans les génomes d'archaea et étude fonctionnelle des ARN et sRNP H/ACA correspondants ». Les nombreuses pistes de recherche ouvertes par ces travaux nous ont permis l'exploitation et la valorisation des résultats obtenus du point de vue expérimental grâce à une collaboration étroite avec d'autres membres de l'équipe (B. Charpentier, Pr. ; J-P. Fourmann, doctorant ; A. Urban, doctorant ; I. Motorine, Pr.). La thématique de recherche d'ARNnc a également donné lieu une collaboration ponctuelle, sur le plan méthodologique, avec D. Gautheret (Pr. à Paris11, Orsay), qui a permis de développer une approche bioinformatique plus particulièrement adaptée aux archaea. Les futurs développements méthodologiques s'inscrivent en partie dans le cadre d'un projet ANR auquel je suis associé : BRASERO (porteur : A. Denise, Pr. Paris11-Orsay). Ce projet vise à développer des outils bioinformatiques pour les ARN en exploitant notre expertise sur des systèmes biologiques bien caractérisés.

2.3 Introduction

Les gènes d'ARN non codant (ARNnc), par opposition aux gènes de protéines codent des ARN fonctionnels (ou non fonctionnels) qui ne contiennent aucune information codante ; on distinguera alors les ARN qui sont transcrits et traduits (ARN de classe I) de ceux qui sont transcrits mais non traduits, dépourvus de phase ouverte de lecture ou ORF (ARN de classe II), (Fig. 16). Ils sont étonnamment nombreux et peuvent être classés en deux grands groupes fonctionnels : (1) les ARNs de maintenance (ARNt, ARNr, snRNA, snoRNA, etc), (2) les ARNs de régulation (de la transcription, de la traduction, de la fonction ou de la distribution de protéines). Les ARNs de maintenance sont naturellement présents dans tous les génomes eucaryotes ou procaryotes et interviennent dans les fonctions cellulaires de base. De façon similaire, des ARN de régulation, bien que découverts plus récemment, sont retrouvés aussi bien dans les génomes d'eucaryotes que de procaryotes. Il s'agit notamment des microRNA (ou ARNi, siRNA, etc) dont la fonction de régulation peut être associée à la différenciation cellulaire des cellules du cœur [75], au cancer [76] et qui sont très présents (avec d'autres petits ARN) dans le système nerveux central

des mammifères [77]. Certains petits ARN pourraient d'ailleurs êtres associés à l'émergence de capacités cérébrales spécifiques chez l'homme et absentes chez d'autres primates [78].

Des gènes d'ARNnc ont été mis en évidence chez les bactéries, notamment chez *E. coli* par des approches expérimentales lourdes [79], ainsi que chez les eucaryotes [24]. Chez *E. coli*, certains de ces gènes sont impliqués dans le contrôle fin des réponses cellulaires aux changements de conditions de l'environnement. Chez les eucaryotes, ils peuvent aussi intervenir dans la stabilité des ARN messagers, dans des régulations traductionnelles (liés au développement, par exemple), dans la stabilité des protéines et leur sécrétion et agir à distance sur des cellules cibles via un système circulatoire (chez les plantes par exemple). L'explosion dans la découverte de petits ARN depuis les années 2000 est en partie liée à l'utilisation d'approches par génomique comparative faisant appel à des méthodes informatiques. Elles ont permis d'identifier beaucoup d'éléments conservés du point de vue évolutif qui ne se limitent pas à gènes potentiels d'ARNnc mais qui renseignent sur l'existence d'éléments fonctionnels ayant subi une pression de sélection et parfois conservés des insectes aux vertébrés [23]. Il est important de souligner aussi que la découverte d'un niveau extrêmement élevé de transcription des génomes d'eucaryotes supérieurs et en particulier du génome humain [21] suggère la présence de nombreux ARN cytosoliques (autres qu'ARNm) : plus de 50%, susceptibles de contenir de contenir des ARNnc.

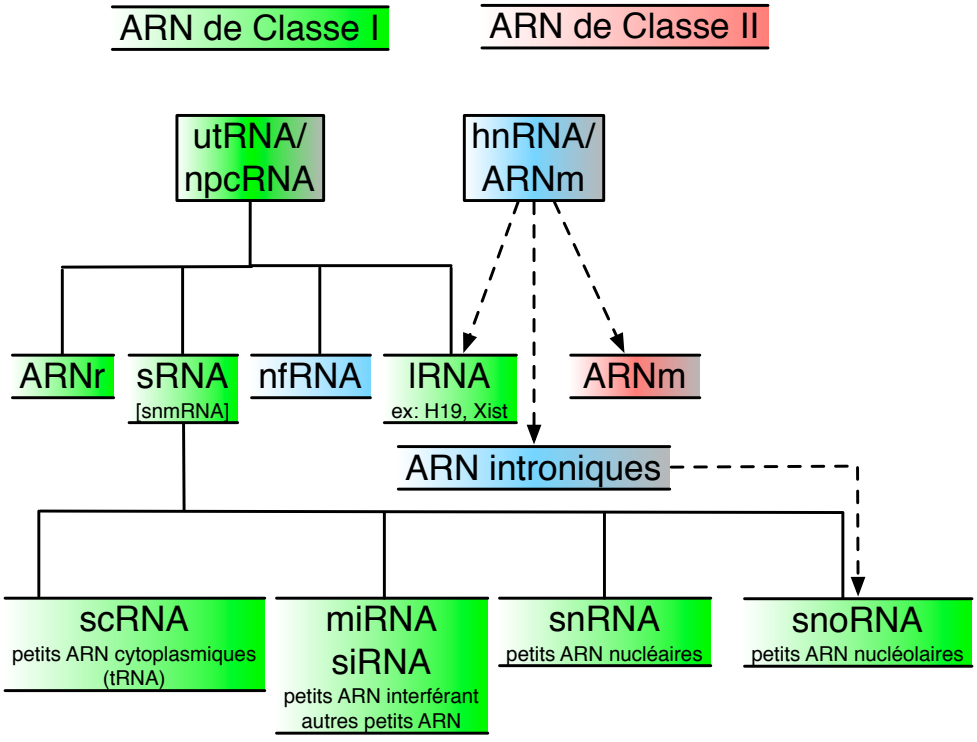


Figure 16. Classification fonctionnelle des ARN. Les ARN de classe I (vert) sont directement fonctionnels (après une éventuelle maturation) et ne contiennent pas d'ORF. Les ARN de classe II (rouge) sont traduits et contiennent donc une ORF qui est traduite. Certains ARN sont non fonctionnels ou non directement fonctionnels (en bleu) même si ils peuvent être à l'origine d'ARN de classe I ou de classe II. utRNAs ("untranslated RNAs") et npcRNAs ("non-protein-coding RNAs") sont des ARN non traduits et ne contenant pas d'information codante. Les utRNAs d'eucaryotes peuvent être générées par toutes les polymérases. Les petits ARN : sRNAs ou snmRNAs ("small non-messenger RNAs") comprennent les ARN cytoplasmiques tels que les ARNt, SRP RNA, etc. miRNA et siRNA correspondent à des ARN régulateurs de la famille des microRNA et ARN interférant ("short interfering RNAs"). Les snRNAs et snoRNAs correspondent aux petits ARN nucléaires et nucléolaires. Les pré-ARNm (ARN nucléaires transcrits par l'ARN polymérase II) donnent naissance aux ARNm matures, d'autres ARN nucléaires hétérogènes ("heterogeneous nuclear RNAs") peuvent aussi conduire à de longs ARN non traduits (lRNAs : "long RNAs"). Les régions introniques des pré-ARNm peuvent aussi renfermer des snoRNAs.

Jusqu'à maintenant, l'exploitation informatique des données brutes de séquence de génomes complets avait surtout été limitée à la recherche de séquences homologues dans les banques de données (BLAST, FASTA, etc) et à la recherche d'ORF potentielles. L'annotation systématique des gènes d'ARNnc est limitée aux ARNr dont la séquence primaire est suffisamment conservée pour être détectée par une simple recherche de similarité de séquences de type BLAST, ou à certains autres ARNnc tels que les ARNt qui possèdent une structure secondaire canonique bien définie et conservée. Les ARN fonctionnels peuvent avoir une séquence plus ou moins dégénérée d'un organisme à l'autre, mais leur structure secondaire est souvent bien définie pour un ARNnc donné (même s'ils ne sont pas forcément plus structurés que d'autres ARN [28]). Certains ARNnc présentent des motifs caractéristiques tels que ceux formés par les boîtes C/D ou H/ACA mis en évidence, au départ, dans les ARN nucléolaires (snoRNA) guides de 2'O-méthylations et de pseudo-uridylations des ARNr et qui sont parmi les ARNnc les mieux caractérisés (Fig. 17).

On peut distinguer deux types d'approches pour l'identification et l'annotation de nouveaux gènes d'ARNnc selon que l'on recherche des gènes appartenant à une famille déjà connue et bien caractérisée ou bien des gènes appartenant à une nouvelle famille ou une famille pour laquelle on dispose de peu de données. Dans le premier cas, les approches "RNomics" faisant appel aux méthodes de recherche d'éléments de structure ou motifs plus ou moins spécifiques des ARNnc recherchés peuvent être utilisées. Dans le second cas, une approche par génomique comparative peut permettre d'identifier des éléments conservés du point de vue évolutif et renfermant éventuellement des éléments de structure spécifiques.

Plusieurs programmes ont été développés pour rechercher des motifs d'ARN bien définis. Certains ont été conçus pour rechercher des motifs structuraux et/ou fonctionnels : des ARNt (tRNAScan-SE), des snoRNA (snoscan), ou d'autres ARN non-codant (ncrnascan), etc. D'autres programmes sont plus généralistes et permettent la recherche de motifs ARN en utilisant : un descripteur qui décrit l'enchaînement des éléments de structure primaire et secondaire (ou même tertiaire) de l'ARN (Rnamot, RNAMotif), ou un profil de séquences (ERPIN) qui correspond à un alignement de séquences d'ARN présentant le motif recherché. En génomique comparative, les méthodes classiques d'alignement et de recherche de similarités de séquence peuvent être mises à profit pour déceler des éléments conservés dans les génomes qui ne correspondent pas à des gènes de protéines déjà annotés. On pourra éventuellement distinguer les méthodes qui reposent sur des recherches de similarités locales à travers les génomes de celles qui font appel à des alignements globaux de génomes (lorsque peu de réarrangements génétiques ont eu lieu entre les génomes comparés).

Dans l'état actuel des connaissances, il apparaît que les gènes d'ARNnc sont présents, pour beaucoup, dans les régions situées entre les ORF identifiées et annotées dans les banques de séquences ou encore dans des régions introniques chez les eucaryotes. La présence de signaux d'expression comme un promoteur transcriptionnel en amont du gène, ou la forte structuration des ARNnc codés dans certaines régions du génome peuvent aussi être utilisés comme critères pour rechercher des gènes d'ARNnc. Toutefois ce type de critères intrinsèques à la séquence des ARNnc n'est pas nécessairement suffisant pour l'identification de ces gènes : les ARNnc ne possèdent pas tous un fort degré de structuration interne ; c'est vérifié pour un certain nombre d'organismes, et en particulier chez les archaea. Chez ces derniers, les informations sur les signaux de transcription sont quasiment inexistantes, ce qui limite encore les possibilités de détection de gènes d'ARNnc. Par contre, leur génome est de petite taille, de l'ordre de 2Mb, et il existe un certain nombre de génomes de ces organismes entièrement séquencés et proches entre eux du point de vue phylogénétique. Les archaea représentent donc des organismes de choix pour tester des approches bio-informatiques développées pour l'identification de gènes d'ARNnc.

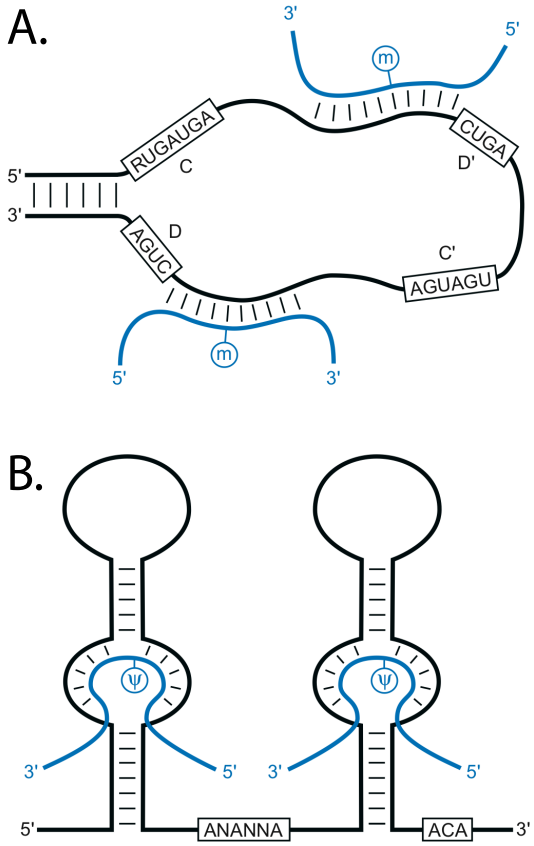


Figure 17. Structures des ARN guides de modification : snoRNAs à boîtes C/D et à boîtes H/ACA. A snoRNAs à boîtes C/D (noir) responsables de méthylations à des positions ciblées dans les ARNr (bleu). B snoRNAs à boîtes H/ACA (noir) responsables de pseudo-uridylation à des positions ciblées dans les ARNr (bleu).

2.4 Développement d'une approche bioinformatique pour la recherche d'ARNnc chez les Archaea

La recherche de gènes d'ARNnc s'est développée assez rapidement car ils sont ubiquitaires et impliqués dans la plupart des processus cellulaires de base (synthèse protéique via les ARNt, ARNr, snoRNA, etc) ou de régulation de la transcription. Toutefois, leur recherche reste ardue en raison de l'absence de biais de séquence liées à la présence d'informations codantes. En fait, la recherche de gènes d'ARNnc se trouve grandement facilitée lorsque les signaux transcriptionnels sont bien caractérisés et facilement identifiables comme c'est le cas pour les eubactéries. La distinction entre gène de protéine et gène d'ARNnc peut alors se faire sur la base des biais de séquence liés notamment à la présence de codons. Chez les archaea, les promoteurs transcriptionnels sont très mal connus, ce qui rend la recherche d'ARNnc chez ces organismes d'autant plus difficile. Outre les ARNr et ARNt, aucun autre ARNnc n'était annoté dans les génomes d'archaea au démarrage du projet, mis à part quelques ARN guides de 2'-O-méthylation : les sRNA à boîtes C/D annotés de façon non exhaustive. Chez les eucaryotes, l'existence de signaux transcriptionnels bien décrits avait permis dans le même temps d'annoter de façon beaucoup plus complète les gènes de sRNA à boîtes C/D [80].

Nous avons sélectionné plusieurs génomes d'archae complètement séquencés et proches du point de vue phylogénétique qui étaient susceptibles de présenter des éléments de séquence conser-

vés soumis à une pression de sélection et qui pourraient correspondre à des gènes d'ARNnc. Le genre *Pyrococcus* était alors le genre le mieux représenté avec les génomes séquencés de 3 espèces distinctes : *P. abyssi*, *P. furiosus* et *P. horikoshii* relativement proches du point de vue phylogénétique. L'approche par génomique comparative développée et mise en œuvre avec S. Muller consiste à rechercher ces régions conservées en dehors des régions codantes qui présentent un biais évident de conservation de séquence (Fig. 18). D'ailleurs, une analyse de la distribution de gènes d'ARNnc connus chez les archaea, les sRNA à boîtes C/D, nous avait permis d'observer que les gènes d'ARNnc semblent être présents de façon prépondérante dans les régions situées entre les ORF et parfois dans les régions chevauchantes [81]. La recherche de gènes d'ARNnc dans ces régions se déroule en plusieurs étapes. Dans un premier temps, les régions situées entre les régions codantes (régions IRC) sont extraites en utilisant l'annotation du génome ; les gènes d'ARNr et d'ARNt annotés sont également exclus (Fig. 18A). Des similarités de séquence sont ensuite recherchées avec des outils de type BLAST (ou FASTA) (Fig. 18B). Enfin, les éléments présentant les plus fortes similarités sont sélectionnés et réalignés par une méthode d'alignement global de type CLUSTALW permettant d'étendre les régions conservées ; un score permet d'apprécier le degré de similarité des régions conservées et de classer les gènes d'ARNnc candidats par ordre décroissant (Fig. 18C).

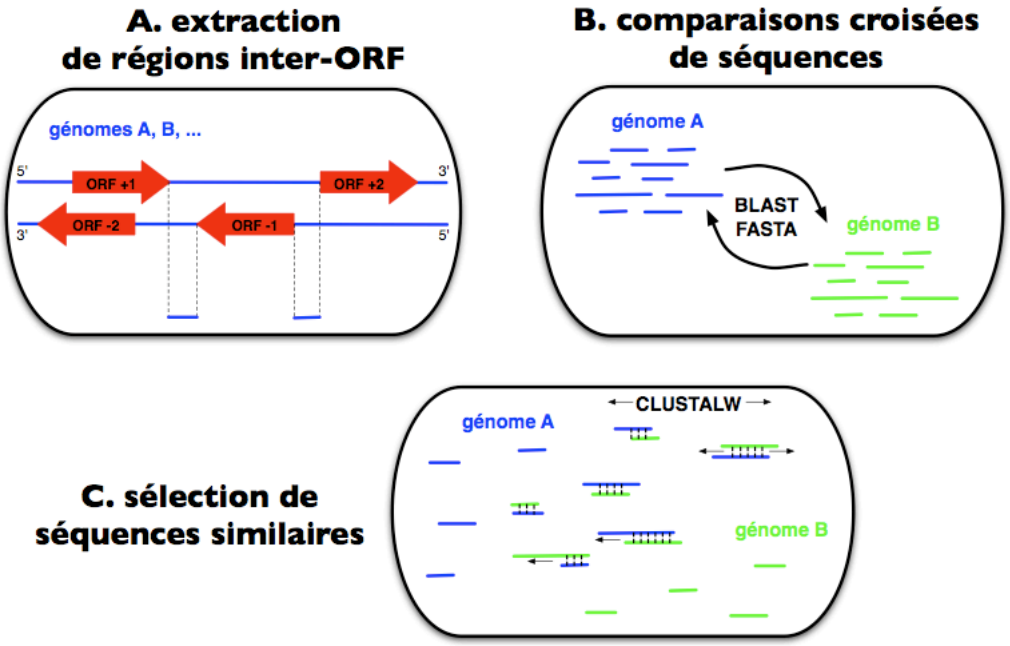


Figure 18. Principes d'une approche séquentielle de génomique comparative pour la recherche locale et non-spécifique de gènes d'ARNnc. A. Les régions IRC sont extraites en excluant les séquences des ORF sur chacun des brins ; une zone de recouvrement de 15 à 25 nucléotides entre les régions IRC avec les segments 5' et 3' des ORF est prise en compte dans les analyses afin de considérer les cas où une courte zone de chevauchement peut exister entre un gène d'ARNnc et une ORF. B. Les séquences des régions IRC sont comparées entre elles pour identifier des similarités locales ; La comparaison est faite avec BLAST ou FASTA. Le critère quantitatif utilisé pour sélectionner les candidats dans les régions IRC conservées correspond à la présence d'au moins 18 nucléotides consécutifs strictement conservés ou à un seuil minimum de signification statistique de 1e-04 pour la similarité calculée par BLAST C. Un alignement global est effectué par CLUSTAL à partir de la région de nucléation correspondant à l'alignement local. Les ensembles de séquences sont classées par un score qui varie en fonction de la qualité de l'alignement global ; une sélection peut être effectuée en définissant une valeur de score.

L'approche comparative conduit à l'identification de régions conservées qui ne correspondent pas forcément à des gènes d'ARNnc ni même à des régions transcrites mais pouvant contenir des

éléments de régulation au niveau ADN. Afin d'être plus sélectif vis-à-vis de régions conservées correspondant à des gènes potentiels d'ARNnc, nous avons mis en place deux filtres de sélection additionnels. Le premier est un filtre négatif qui vise simplement à éliminer les régions conservées correspondant à des ORF mal annotées. En effet, l'absence d'annotation de petites ORF (souvent mal reconnues par les programmes de prédiction d'ORF) ou une annotation imprécise sur le début ou la fin d'ORF conduit à inclure dans les régions IRC des séquences codantes. Entre les deux versions de l'annotation du génome de *P. abyssi* de juin 2001 et décembre 2005, plusieurs ORF ont été réannotées. Les régions IRC conservées qui ressemblent fortement, du point de vue statistique, à des ORF initialement mal identifiées sont alors éliminées des candidats (emploi du programme GenMarkS, [82, 83]). Le second filtre de sélection est un filtre positif destiné à identifier les ARNnc susceptibles de reconnaître une cible dans le génome sur la base d'une complémentarité de séquence. Ce filtre a été tout particulièrement utilisé pour la recherche d'ARN guides de pseudo-uridylation : les sRNA à boîtes H/ACA.

Cette approche de génomique comparative permet d'identifier un nombre réduit et pertinent de gènes d'ARNnc candidats (Fig. 19). Le filtre de sélection négatif a permis d'éliminer une région IRC conservée ressemblant à une ORF. A des fins de validation, nous avons vérifié que l'on retrouvait les gènes d'ARNnc connus chez ces 3 espèces : à savoir les sRNA à boîtes C/D (49 ARN au total avaient été identifiés et annotés dans les 3 génomes). Nous avons pu retrouver 45 de ces 49 ARNnc connus (92%). Les 4 gènes manquant n'ont pu être identifiés en raison de certaines limitations de la méthode : 1 gène est localisé dans une ORF. Les 3 autres ne présentent pas une similarité de séquence suffisante pour pouvoir être détectés (Fig. 18B). Il faut noter que plus la distance phylogénétique entre espèces comparées est élevée et plus la probabilité de divergence de séquence est grande et moins la probabilité d'identifier des ARNnc peu conservés est élevée. Une façon de repousser cette limitation est de rechercher des structures 2D conservées d'ARN plutôt que des séquences conservées, ce qui permet de rechercher des gènes d'ARNnc dont la séquence a beaucoup divergé.

2.5 Amélioration de l'approche par génomique comparative pour la recherche d'ARNnc structurés

Dans le cadre du stage de M1 de Cédric Bicep que j'ai encadré, j'ai proposé à la fois de faciliter l'utilisation de la méthode de recherche d'ARNnc (pour les non-experts) et d'améliorer sa sensibilité en ajoutant un volet de recherche de structures 2D. L'approche décrite ci-dessus a été initialement programmée en langage C-Shell en incluant des fonctions des outils UNIX sed et awk (Fig. 20). Avec C. Bicep, nous avons travaillé sur la programmation et le portage sous une forme informatiquement plus évoluée, en utilisant le langage Perl et les librairies spécifiques BioPerl pré-conçues pour l'analyse et la manipulation de séquences. Les améliorations majeures obtenues pour faciliter l'utilisation de la méthode sont : 1) l'automatisation complète de l'approche à l'aide d'outils Perl/BioPerl, 2) le calcul d'un score permettant de classer et de sélectionner les gènes d'ARNnc candidats en fonction du degré de similarité des IRC conservées. Quant à l'amélioration de sensibilité attendue, elle réside dans la possibilité de rechercher des ARN structurés conservés dont la séquence a pu diverger et qui ne peuvent pas être identifiés sur la seule base de la similarité de séquence.

Dans la nouvelle implémentation, la recherche de régions IRC conservées se fait de façon automatique avec la suite d'outils Perl/BioPerl. Par ailleurs, une fonction de score est calculée d'après le taux de conservation des séquences dans l'alignement global. Elle permet de classer les régions IRC homologues en fonction de leur degré de similarité. Un seuil minimum de conservation

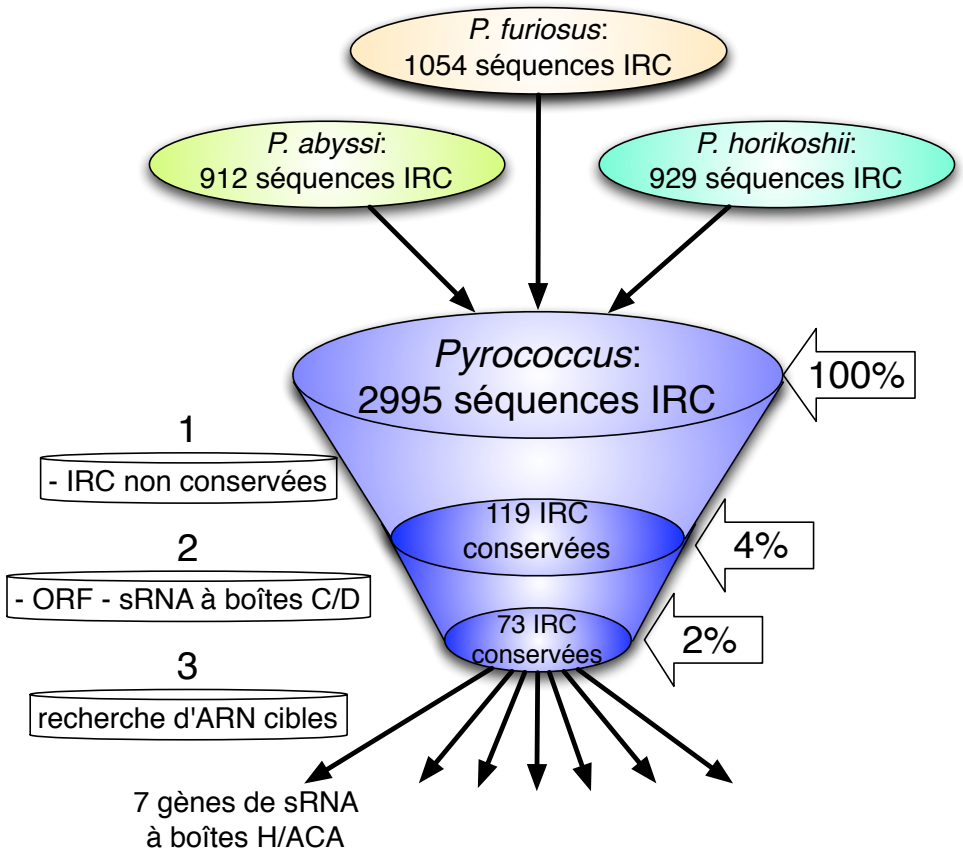


Figure 19. Filtrages des régions IRC pour la recherche de gènes d'ARNnc. Le premier filtre (positif) est la sélection de régions IRC conservées (similarité de séquence). Le deuxième filtre (négatif) est l'élimination de régions IRC mal annotées correspondant à des ORF ou de gènes d'ARNnc connus. Le troisième filtre (positif) est la recherche d'ARN cibles dans le génome.

de séquence peut être fixé. L'utilisateur peut réaliser une sélection d'après le score des différents candidats.

Un ensemble de méthodes peut être utilisé pour rechercher des ARN structurés (Fig. 20). On distingue alors 2 cas selon que le degré de conservation de séquence est fort (plus de 60%) ou faible (moins de 60%). Pour les ARN structurés qui présentent un fort degré de conservation de séquence, les méthodes telles que RNAz [84] ou RNAalifold [85] permettent de rechercher l'existence possible d'ARN dont les structures 2D potentielles sont conservées. Dans les autres cas, une méthode telle que CARNAC [86] permet de rechercher des structures 2D conservées sur des séquences qui ont fortement divergé et dont l'alignement risquerait de fausser les prédictions ; elle est alors plus performante que RNAz et RNAalifold.

Dans les prédictions effectuées par RNAz et RNAalifold, une structure 2D consensus est proposée à partir d'un alignement de type CLUSTAL (l'alignement sert de base à la recherche de régions complémentaires conservées entre les différentes séquences). La présence de mutations compensatoires dans les régions double-brin a une influence sur le score qui est calculée pour évaluer différentes structures 2D conservées possibles. Ce type de mutations permet donc d'obtenir des prédictions de meilleure qualité. Dans le cas où le degré de conservation dans l'IRC est plus faible, le programme CARNAC permet de proposer une structure 2D consensus en effectuant un repliement et un alignement des séquences simultanés sans alignement préalable. Ce programme

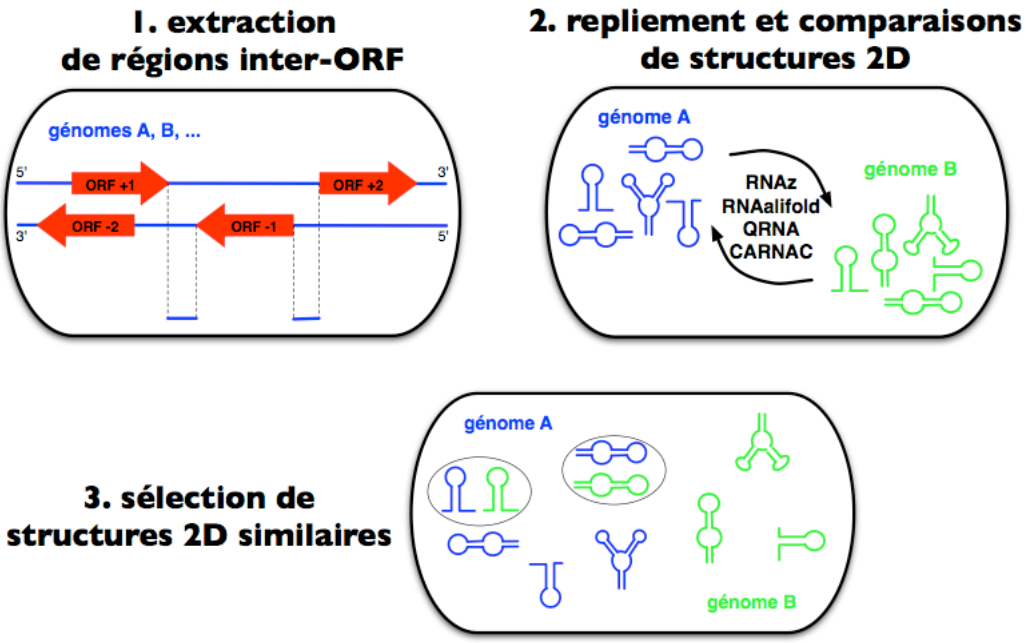


Figure 20. Principes d'une approche séquentielle de génomique comparative pour la recherche locale et non-spécifique de gènes d'ARNnc structurés. A. idem Fig. 18. B. Les structures 2D conservées sont soit déduites à partir d'un alignement CLUSTAL (RNAz, RNAalifold) ou sans alignement (repliement et alignement simultanés avec le programme CARNAC). C. Un score permet de classer les structures 2D les plus pertinentes.

est développé par l'équipe d'Hélène Touzet (LIFL, CNRS-INRIA Futurs, Lille1), qui travaille également sur le développement d'autres outils d'analyse de structures 2D d'ARN dans le cadre du projet ANR BRASERO auquel nous participons.

Cette version améliorée de l'approche initiale a aussi été testée sur les 3 génomes de *Pyrococcus* afin d'évaluer sa pertinence pour la détection de gènes de sRNA à boîtes C/D et à boîtes H/ACA. Comme leur séquence est fortement conservée, l'approche fait appel à un alignement préalable (CLUSTAL) tel qu'il était réalisé dans l'approche initiale. Cet alignement sert de base à la prédiction de structures 2D conservées par les programmes RNAz et RNAalifold. Ces programmes ont été développés en faisant deux hypothèses implicites. Premièrement, les séquences fournies dans l'alignement sont supposées correspondre aux brins d'ARN transcrits. Or, notre approche ne fait aucune distinction parmi les séquences conservées entre celles qui correspondent aux brins transcrits et celles qui sont sur le brin complémentaire (séquences « reverse complémentaire »), même si le score de l'alignement correspondant aux séquences du brin transcrit devrait être supérieur par rapport à l'autre cas. Deuxièmement, les séquences alignées sont repliées en supposant une structuration sur l'ensemble de la région alignée. Or, nos alignements des régions IRC conservées peuvent contenir des ARNnc structurés dans certaines parties de l'alignement d'une IRC mais pas nécessairement sur l'alignement complet (par exemple dans les régions non transcrites). La présence de parties non structurées dans l'alignement en 5' ou en 3' d'un ARNnc structuré peut donc conduire à favoriser des repliements alternatifs sous-optimaux.

Pour éviter le 1er écueil lié à l'orientation aléatoire des séquences dans alignement d'une région IRC conservée, la recherche de structures 2D conservées est réalisée à la fois sur les séquences originales et sur les séquences reverse complémentaires après les avoir préalablement réaligner. La stabilité de la structure 2D consensus proposée et le nombre de mutations compensatoires dans les régions double-brin de l'ARN permettent alors de distinguer le repliement correct. Dans

l'exemple donné ci-dessous (Fig. 21), un ARN partiellement structuré correspondant à un sRNA à boîtes C/D est retrouvé sous les 2 formes dans les régions IRC conservées. La structure 2D correspondant à l'ARN structuré est plus stable (Fig. 21A) que celle pour l'ARN correspondant aux séquences complémentaires inversées (Fig. 21B).

Pour éviter le second écueil lié à la présence de repliements sous-optimaux favorisés en considérant les régions IRC conservées totales, les séquences sont tronquées pour ne retenir que la partie de l'alignement la plus conservée (à au moins 75%) ; les séquences sont alors réalignées sur cette partie tronquée de l'alignement original et les structures 2D conservées sont prédites sur cette base. Dans le cas de motifs simples, correspondant par exemple à un gène de sRNA à boîtes H/ACA possédant un simple motif H/ACA, la prédiction sur l'alignement original non tronqué peut donner une bonne prédiction (Fig. 22). Dans le cas de gène de sRNA à boîtes H/ACA possédant un double ou triple motif, la troncature de l'alignement peut permettre d'identifier correctement tous les motifs H/ACA (données non montrées).

2.6 Approche bioinformatique pour la recherche ciblée de sRNA à boîtes H/ACA chez les Archaea

Une recherche des gènes codant potentiellement des sRNA à boîtes H/ACA a été mise au point par l'exploitation des connaissances disponibles sur les liens structure/fonction de ces ARN guides de modification [87]. Elles étaient limitées au début de notre étude ; seuls 3 sRNA à boîtes H/ACA avaient été identifiées chez *A. fulgidus*. Néanmoins, en se basant sur ce qui était connu chez les eucaryotes, les données existantes ont été suffisantes pour démarrer l'étude. Celles-ci ont permis de définir un motif caractéristique sur lequel les recherches pouvaient être basées : une structure tige-boucle contenant une large boucle interne, une boucle terminale et un élément simple-brin de type ANA (Fig. 23). Chez les archaea, ce motif H/ACA de base comporte en plus un motif en K-turn (ou K-loop) dans la partie apicale de la structure tige-boucle (Fig. 23A). La boucle interne délimite la poche de pseudo-uridylation où une uridine de l'ARN cible est convertie en pseudo-uridine. Dans les données connues sur les eucaryotes, l'ARN guide et l'ARN cible s'associent grâce à une complémentarité de séquence entre les éléments antisens 5' et 3' situés dans la boucle interne du motif H/ACA de l'ARN guide et l'ARN cible ; 2 résidus de l'ARN cible sont non-appariées, au centre le U du côté 5' est converti en pseudo-U (Fig. 23B). Un descripteur du motif H/ACA (Fig. 23A) a été défini à partir de ces connaissances à l'aide du programme Rnamot. La recherche a été réalisée sur les régions IRC des génomes de Pyrococcus. Le descripteur étant très général, une seconde étape importante dans le criblage a consisté à rechercher des cibles dans les ARNr, à l'aide de descripteurs spécifiques obtenus comme décrits ci-dessus, eux aussi construits avec le logiciel Rnamot (Fig. 23B) [29] . Pour chaque poche de pseudo-uridylation, plusieurs descripteurs ont été définis pour tenir compte de la variabilité de séquence et/ou de longueur des duplexes formés avec les éléments antisens 5' et 3' (Fig. 23B). L'existence *in cellulo* de candidats sRNA présentant une ou plusieurs cibles potentielles dans les ARNr a été validée expérimentalement par Sébastien Muller (Muller *et al.*, "Combined in silico and experimental identification of the Pyrococcus abyssi H/ACA sRNAs and their target sites in ribosomal RNAs", 2008) [88].

Nous avons ainsi identifié 5 gènes de sRNA à boîtes H/ACA correspondant à 9 motifs H/ACA dans le génome de *P. abyssi* : 2 gènes comportant un simple motif H/ACA (Pab-21, Pab-91), 2 comportant deux motifs (Pab-35 et Pab-105) et 1 comportant un triple motif (Pab-40) (Fig. 24). Avec S. Muller, nous avons ensuite utilisé les 5 sRNA H/ACA ainsi caractérisés afin de développer une méthode applicable à une recherche réalisée directement sur les génomes complets et plus

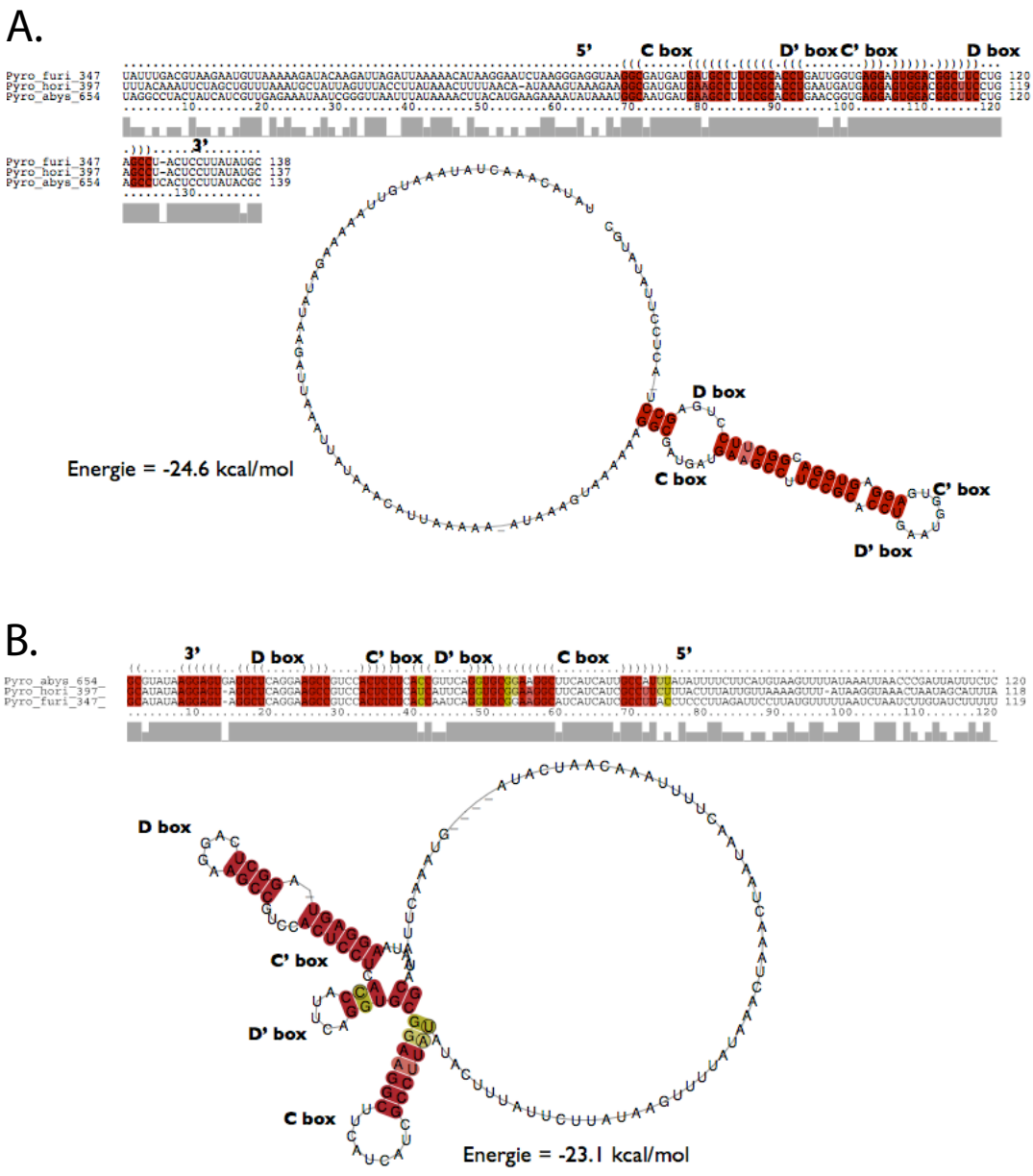


Figure 21. Recherche de gènes d'ARNnc structurés et prédiction de leur structure 2D conservée : le cas du sRNA à boîtes C/D sR1 des Pyrococcus (suite). A. Prédiction de la structure 2D conservée de l'ARN structuré. B. Prédiction de la structure 2D conservée pour les ARN correspondant aux séquences reverses complémentaires. Les positions en rouge indiquent l'absence de variation dans la structure, celles en ocre indiquent la présence de 2 types différents de paires.

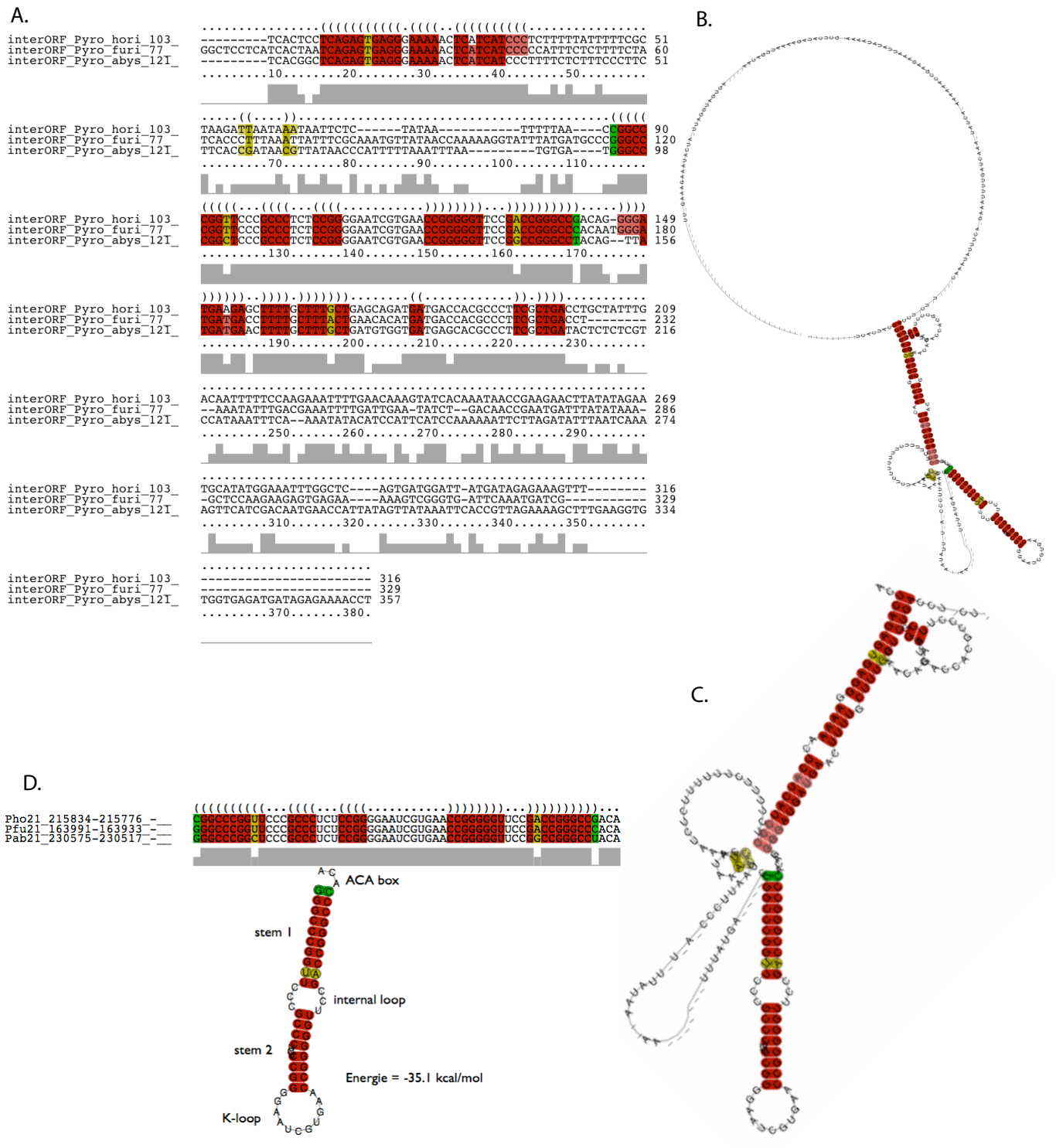


Figure 22. Recherche de gènes d'ARNnc structurés et prédiction de leur structure 2D conservée : le cas du sRNA à boîtes H/ACA Pab21 des *Pyrococcus*. A. Alignement CLUSTALW de régions inter-ORF homologues chez les *Pyrococcus*. B. Prédiction de la structure 2D conservée à partir de l'alignement original. C. Zoom sur la région structurée de la région inter-ORF conservée. D. Prédiction de la structure 2D conservée à partir d'un alignement tronqué (généré sur la base d'un pourcentage optimal d'identité de séquence). Les positions en rouge indiquent l'absence de variation dans la structure, celles en ocre et vert indiquent la présence respective de 2 ou 3 types différents de paires.

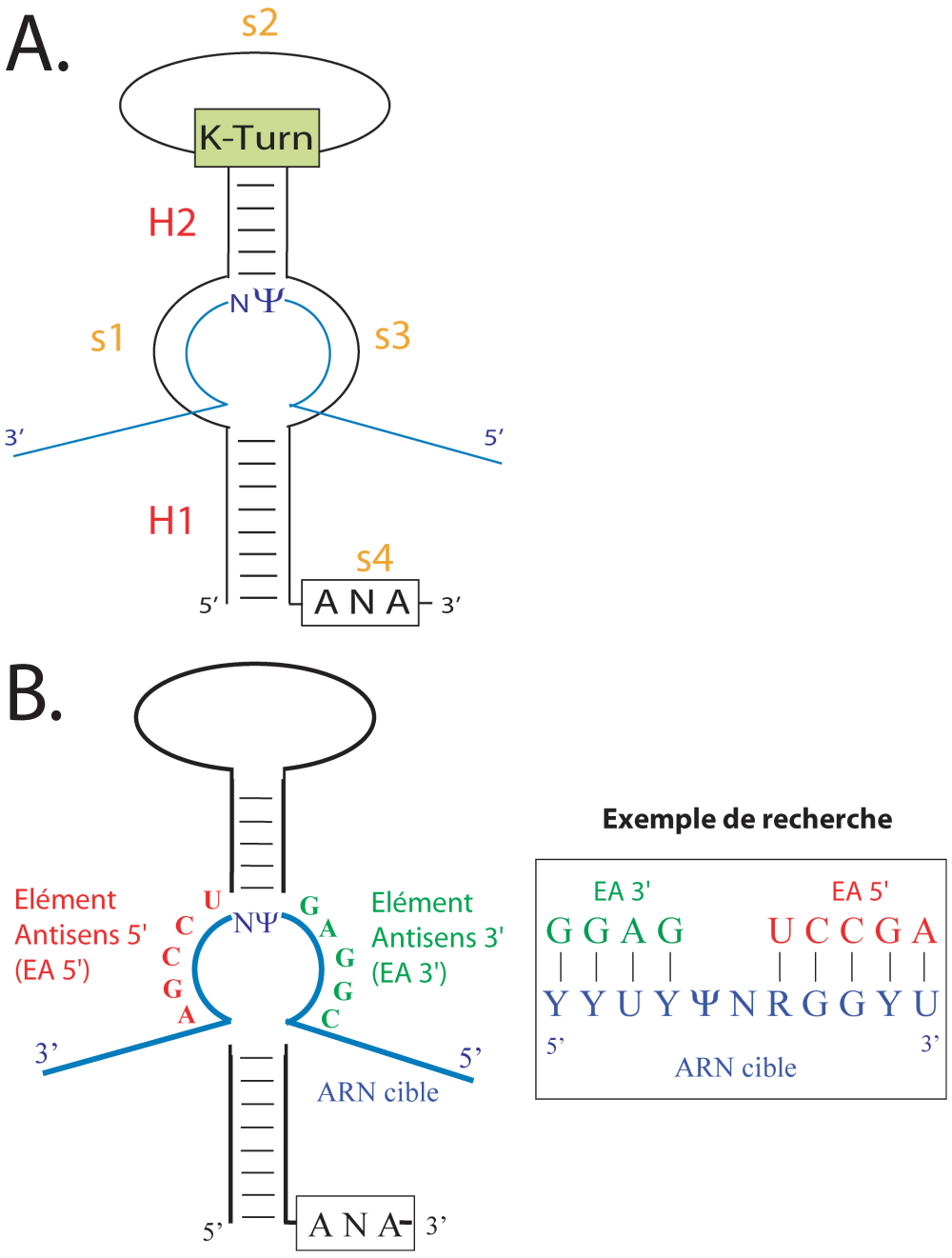


Figure 23. Relations structure/fonction des sRNA à boîtes H/ACA d'archaea utilisées dans la recherche bioinformatique. A. Description du motif H/ACA : les éléments « s » correspondent à des régions simple-brin, les éléments « H » à des régions double-brin. Le motif K-turn est défini de façon classique. B. Mode d'interaction guide/cible dans la poche de pseudo-uridylation. La position U convertie en pseudo-U est la 1ère position du dinucléotide non-apparié entre les éléments antisens 5' et 3'.

sensible. Pour cela, au lieu d'utiliser le descripteur Rnamot (comme décrit ci-dessus), nous avons utilisé le programme ERPIN qui est basé sur un alignement (ou profile) des séquences des ARN validés sur la base de leur structure 2D et des éléments de structure conservés (Fig. 25) [87]. Deux nouveaux sRNA à boîtes H/ACA potentiels renfermant un motif simple (Pab-19 et Pab-160) ont alors été détectés. La recherche sur les génomes complets a permis d'identifier un sRNA à boîte H/ACA supplémentaire (Pab-160) qui n'avait pas été identifié initialement car son gène se trouve localisé dans une région chevauchant une ORF [87]. D'autre part, l'utilisation de contraintes plus lâches dans certaines segments du motif H/ACA a permis d'identifier un second sRNA H/ACA supplémentaire qui comporte une boucle interne plus grande que celle des autres sRNA H/ACA (l'élément 5' de la boucle interne est plus long). L'expression de ces 2 sRNA H/ACA a aussi été validée expérimentalement par Sébastien Muller [88].

Sur la base des prédictions informatiques que nous avons faites, chacun des 11 motifs avait la possibilité de cibler un ou plusieurs séquences des ARNr. Ainsi, 23 couples ARN guide/ ARN cible potentiels ont été prédits (Fig. 26). Parmi les 23 prédictions faites, 15 se sont révélées correspondre à des cibles modifiées par l'ARN guide attendu. Sébastien Muller a alors employé 2 approches pour essayer de déterminer quelles séquences étaient réellement ciblées dans les ARNr et par lequel des motifs H/ACA. Tout d'abord avec I. Behm-Ansmant, il a identifié expérimentalement les positions pseudo-uridylées dans les ARNr en utilisant un réactif chimique spécifique des résidus Ψ. D'autre part, il a employé la méthode de reconstitution in vitro de particules sRNP H/ACA actives mise au point par Bruno Charpentier [89] et testé l'action de chaque motif sur chaque séquence cible proposée par l'approche informatique [87, 88].

Nos travaux ont largement contribué à mieux comprendre les liens structure/fonction des sRNA à boîtes H/ACA, c'est-à-dire à définir des règles de structuration des motifs H/ACA et d'association entre ARN guides et ARN cibles. Grâce à ces connaissances substantielles, une approche plus globale a été mise au point pour la recherche spécifique de gènes de sRNA à boîtes H/ACA dans les génomes d'archaea et de leur(s) cible(s). Elle permet une détection très sensible de motifs H/ACA dans les génomes. Contrairement au descripteur initial utilisé pour rechercher des motifs H/ACA dans les régions IRC conservées (Fig. 23A), le profile permet une recherche plus spécifique et sélective avec le programme ERPIN (Fig. 25) et nous autorise à ne faire aucune hypothèse *a priori* sur la localisation des motifs dans le génome.

L'approche a été utilisée spécifiquement pour la recherche de motifs H/ACA dans les génomes de *Pyrococcus*. Toutefois, grâce à quelques motifs H/ACA connus dans d'autres espèces d'archaea (notamment *A. fulgidus*), elle a pu être généralisée à la recherche de sRNA H/ACA dans de nombreux génomes d'archaea. Ainsi, des motifs H/ACA nouveaux qui n'étaient pas présents au départ dans la banque de motifs H/ACA connus ont pu être identifiés. L'avantage de la méthode est justement de permettre d'enrichir la banque de motifs H/ACA au fur et à mesure que de nouveaux motifs H/ACA sont identifiés et ainsi d'augmenter la spécificité et la sélectivité de la méthode.

En réunissant l'ensemble des données sur les motifs H/ACA d'archaea, la banque de motifs initiale utilisée par le programme ERPIN a compté 41 motifs différents correspondant à 26 sRNA à boîtes H/ACA identifiés parmi *Archaeglobus fulgidus*, *Methanocaldococcus jannaschii* et les espèces de *Pyrococcus* sur lesquelles nous avons plus spécifiquement travaillé. A chaque cycle de recherche de motifs H/ACA dans les génomes complets d'archaea, les nouveaux candidats identifiés ont été évalués par la recherche de cibles potentielles (Fig. 23B) et validés *in silico* lorsqu'un couple de motifs ARN guide/ARN cible pouvait être proposé. Les ARN guides alors validés ont été introduits dans la banque de motifs pour enrichir la profile ERPIN. Après plusieurs cycles successifs, le nombre de motifs H/ACA identifiés est resté invariable : la taille de la banque est actuellement 162 séquences alignées et annotées pour leur structure 2D [87]. Elle contient

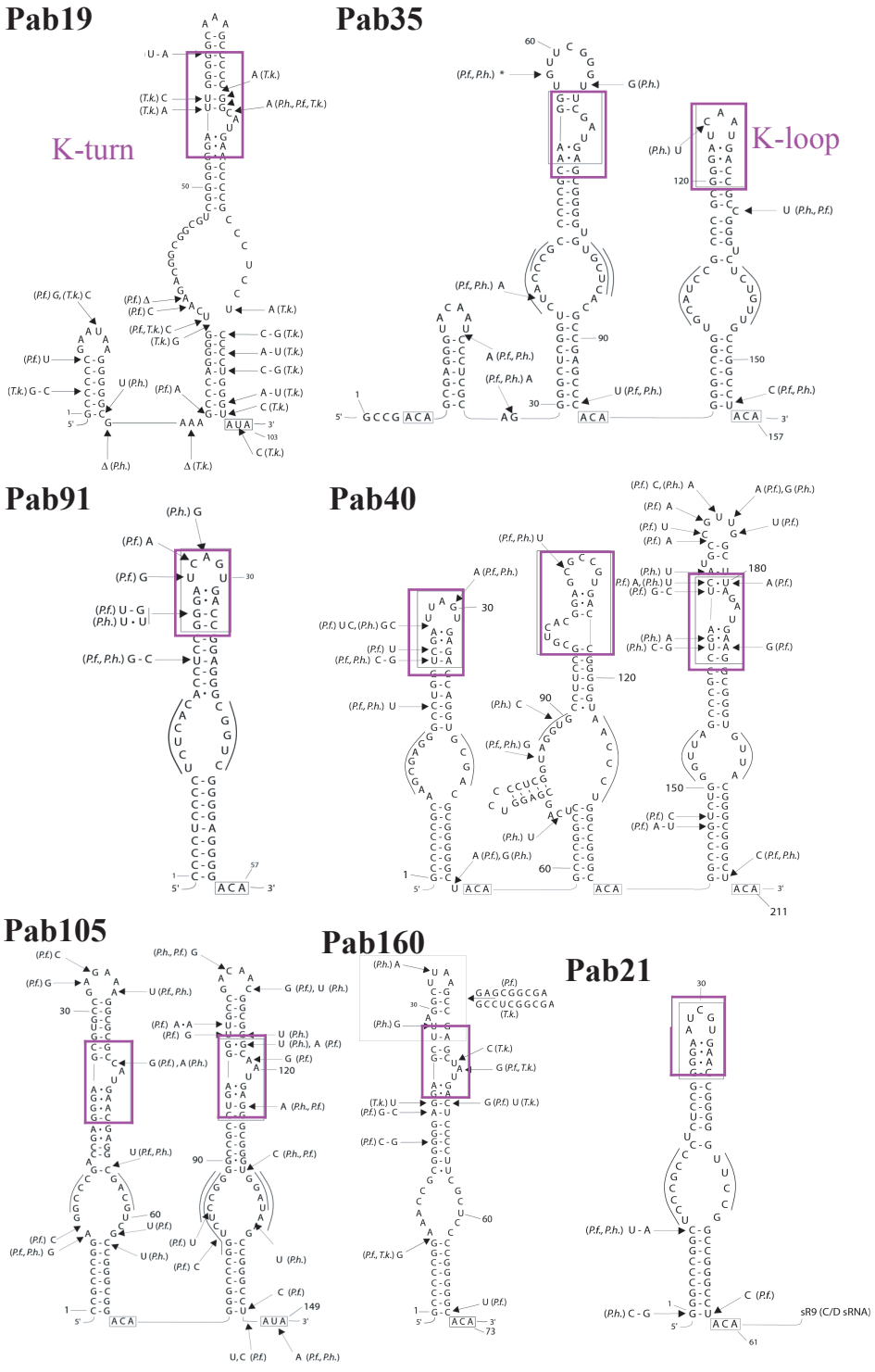
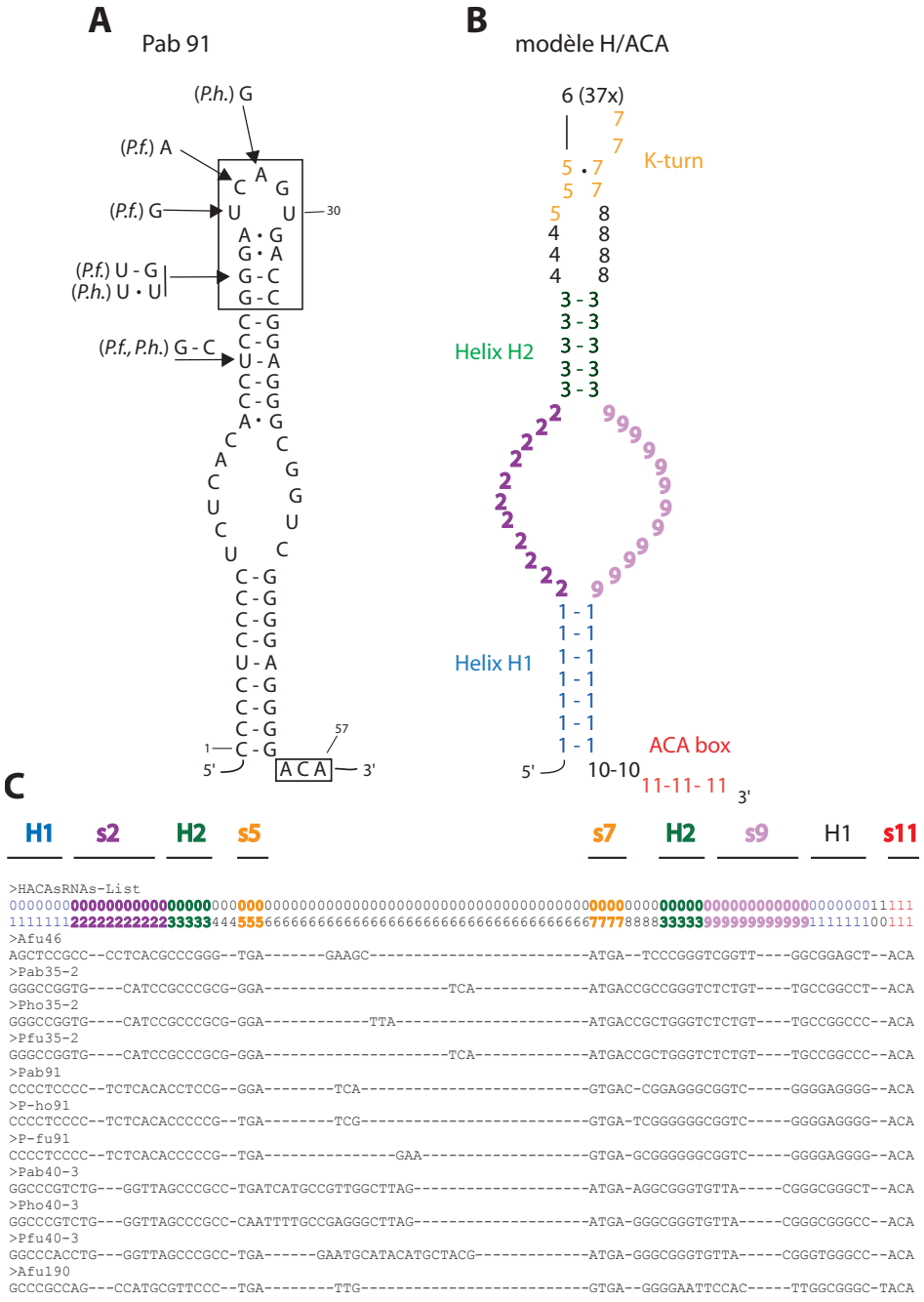


Figure 24. Structures secondaires des ARN H/ACA identifiés précédemment par l'approche informatique de recherche des gènes d'ARN non-codants dans les régions inter-ORF des génomes de *Pyrococcus*. Les structures des 7 ARN identifiés chez *P. abyssi* sont indiquées (Pab-19,Pab-35, Pab-21, Pab-91, Pab-105, Pab-40, Pab-160) ; seules les positions qui diffèrent dans la séquence de leur homologue correspondant chez *P. furiosus* (P. f.) et *P. horikoshii* (P. h.) et *T. kodakarensis* (T. k.) sont indiquées sur la structure 2D. Les boîtes ACA caractéristiques sont encadrées ainsi que les motifs K-turn ou K-loop (violet) dans la partie supérieure des tige-boucle. Les régions simple-brin des boucles internes qui sont soulignées par un trait correspondent à la séquence guide qui permet la reconnaissance de la séquence cible dans les ARNr 16S et 23S et la pseudo-uridylation.




```
>HACAsRNAs-List
>Afu46
AGCTCCGCC--CCTCACGCCCGGG--TGA-------GAAGC-------------------------ATGA--TCCCGGGTCGGTT----GGCGGAGCT--ACA
>Pab35-2
GGGCCGGTG----CATCCGCCCGCG-GGA---------------------TCA-------------ATGACCGCCGGGTCTCTGT----TGCCGGCCT--ACA
>Pho35-2
GGGCCGGTG----CATCCGCCCGCG-GGA------------TTA----------------------ATGACCGCTGGGTCTCTGT----TGCCGGCCC--ACA
>Pfu35-2
GGGCCGGTG----CATCCGCCCGCG-GGA---------------------TCA-------------ATGACCGCTGGGTCTCTGT----TGCCGGCCC--ACA
>Pab91
CCCCTCCCC--TCTCACACCTCCG--GGA--------TCA-------------------------GTGAC-CGGAGGGCGGTC-----GGGGAGGGG--ACA
>P-ho91
CCCCTCCCC--TCTCACACCCCCG--TGA--------TCG-------------------------GTGA-TCGGGGGGCGGTC-----GGGGAGGGG--ACA
>P-fu91
CCCCTCCCC--TCTCACACCCCCG--TGA---------------GAA-------------------GTGA-GCGGGGGGCGGTC-----GGGGAGGGG--ACA
>Pab40-3
GGCCCGTCTG---GGTTAGCCCGCC-TGATCATGCCGTTGGCTTAG--------------------ATGA-AGGCGGGTGTTA-----CGGGCGGGCT--ACA
>Pho40-3
GGCCCGTCTG---GGTTAGCCCGCC-CAATTTTGCCGAGGGCTTAG--------------------ATGA-GGGCGGGTGTTA-----CGGGCGGGCC--ACA
>Pfu40-3
GGCCCACCTG---GGTTAGCCCGCC-TGA------GAATGCATACATGCTACG-------------ATGA-GGGCGGGTGTTA-----CGGGTGGGCC--ACA
>Afu190
GCCCGCCAG---CCATGCGTTCCC--TGA--------TTG-------------------------GTGA--GGGGAATTCCAC-----TTGGCGGGC-TACA
```

Figure 25. Profile ERPIN pour la recherche de motifs H/ACA dans les génomes d'archaea. A. Exemple de motif tige-boucle H/ACA (Pab91) contenant une boucle interne centrale (poche de pseudo-uridylation) et un motif K-loop ou alternativement K-turn. B. Représentation encodée de la structure du motif H/ACA indiquant les différents éléments de structure 2D soit simple-brin (s) soit double-brin (H) qui définit un consensus structural sur lequel sont alignées les séquences des motifs H/ACA contenus dans la banque. C. Alignement structural des séquences des motifs H/ACA qui constitue le contenu tronqué du profile ERPIN utilisé pour la recherche de motifs H/ACA. Le profile ERPIN complet contient 162 séquences de motifs H/ACA spécifiques des archaea [87].

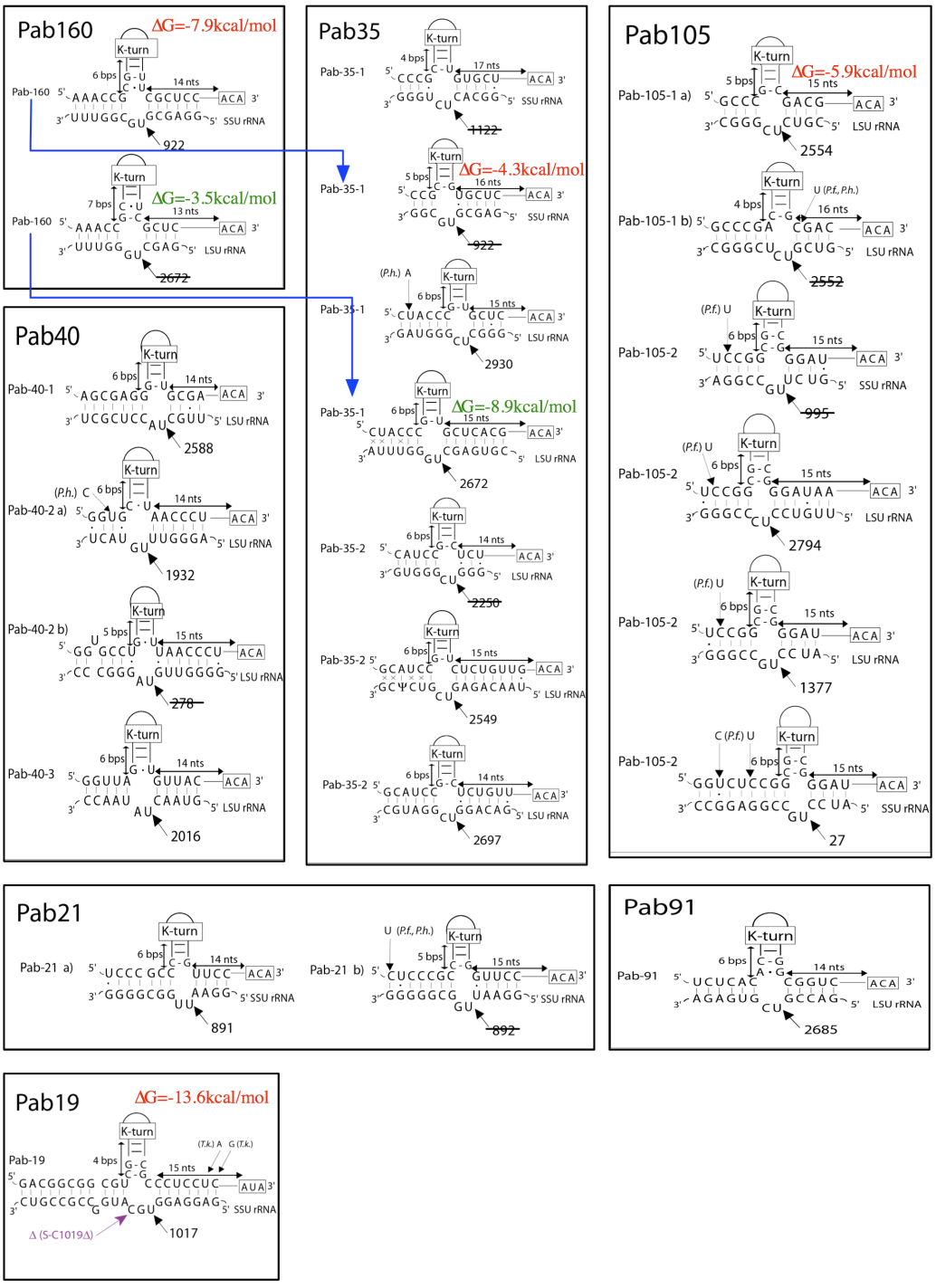


Figure 26. Structures secondaires des complexes ARN guide/ARN cible chez les *Pyrococcus*. La position modifiée dans l'ARN cible est indiquée par son numéro dans la séquence de l'ARNr 16S (SSU rRNA) ou 23S (LSU rRNA). Les numéros barrés correspondent à des couples de ARN guide/ARN cible qui satisfont les critères de base d'association (Fig. 23B) mais qui ne sont pas formés. Deux flèches (bleu) indiquent des postions redondantes susceptibles d'être modifiées par 2 ARN guides différents. Une valeur théorique de l'énergie libre d'interaction est calculée et indiquée sur quelques couples représentatifs. Les contraintes de distance entre la boîte H/ACA et le nucléotide modifé sont également indiquées.

les séquences de 73 gènes potentiels de sRNA à boîtes H/ACA réparties de façon non-homogène parmi 46 espèces différentes d'archaea mais qui couvrent l'ensemble du domaine des archaea (Fig. 27). Le calcul théorique de l'énergie libre d'interaction entre l'ARN guide et l'ARN cible à l'aide d'un modèle énergétique simple (implémentée dans le programme RNAup [90]) suggère que l'un des facteurs qui détermine la reconnaissance entre les 2 ARN partenaires est la stabilité thermodynamique du complexe (Fig. 26). Des facteurs structuraux interviennent probablement également, comme par exemple la distance, dans le complexe, entre la boîte ACA de l'ARN guide et le U modifié en pseudo-U, habituellement de 14 nucléotides.

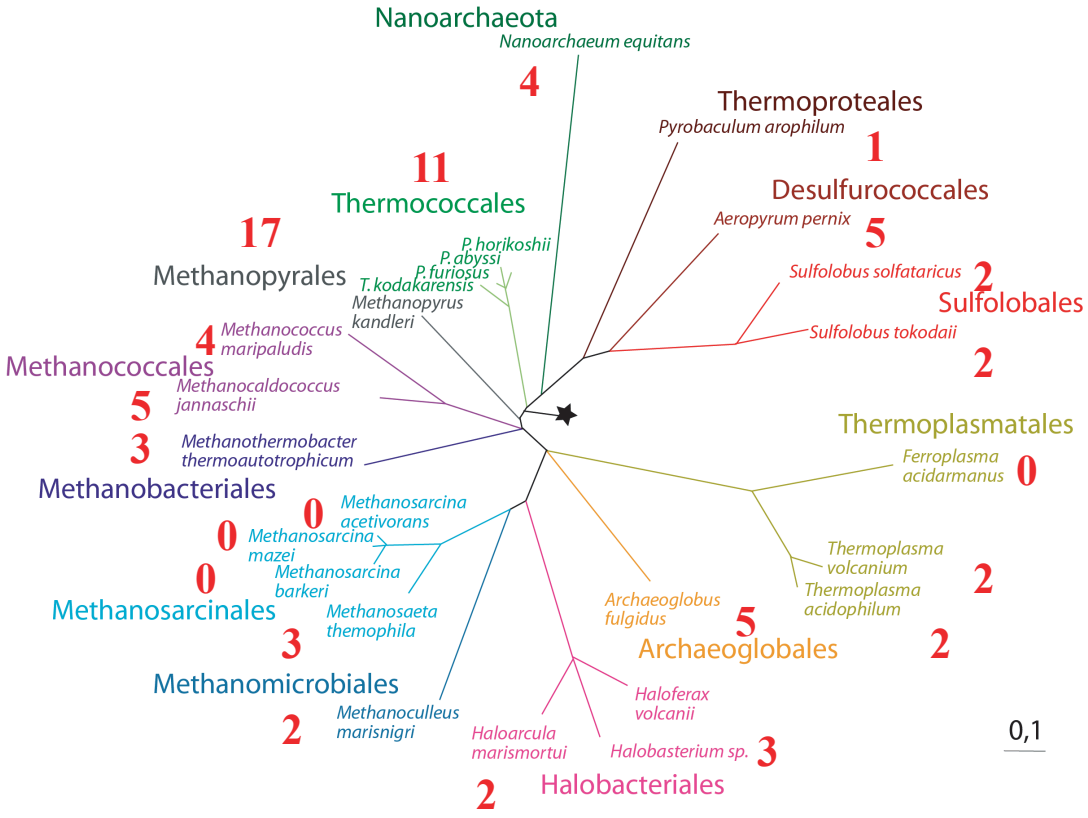


Figure 27. Répartition phylogénétique des gènes de sRNA à boîtes H/ACA chez les archaea. Les valeurs numériques indiquent le nombre de gène(s) pour une espèce (ex : 1 pour *Pyrobaculum arophilum*), ou un ordre (ex : 11 pour les Thermococcales incluant *P. abyssi*, *P. furiosus* et *P. horikoshii*). Au total, les 162 motifs H/ACA sont répartis dans 73 gènes au sein de 46 espèces différentes.

2.7 Exploitation des résultats pour la compréhension des liens structure/fonction des snRNP H/ACA

L'identification par notre approche bioinformatique de nombreux sRNA à boîtes H/ACA et de leur(s) cible(s) dans les ARNr chez les archaea et plus particulièrement chez les *Pyrococcus* a été à l'origine de nombreuses études expérimentales qui ont permis d'explorer plus avant les relations structure/fonction de ces ARN. Chez *P. abyssi*, 7 gènes de sRNA H/ACA ont été identifiés contenant un motif H/ACA : simple, double ou triple soit 11 motifs H/ACA au total (Fig. 24) pouvant chacun cibler une ou plusieurs positions dans les ARNr soit 23 positions potentielles (Fig. 25). Il était donc important de vérifier à la fois l'expression de ces gènes d'ARNnc prédits

mais aussi leur fonctionnalité *in vitro*. Le gène du sRNA H/ACA Pab91 qui cible potentiellement la position 2685 de l'ARNr 23S a été utilisé comme système modèle pour tester la fonctionnalité d'une particule snRNP reconstituée à modifier la cible attendue. C'était la première fois qu'une particule snRNP H/ACA active a été reconstituée *in vitro* (Charpentier *et al.*, 2007) [89].

L'analyse des autres sRNA à boîtes H/ACA chez *P. abyssi* et de leur fonctionnalité a également été réalisée. Cette autre étude a permis d'obtenir une image complète du rôle de ces ARN dans les pseudo-uridylations des ARNr. Elle a permis de montrer que la majorité des positions pseudo-uridylées sont modifiées par une snRNP H/ACA ; les exceptions concernent des positions pouvant être modifiées par les protéines de la particule mais sans ARN guide ou modifiée via un système Ψ-synthase. Sur les 23 positions de cibles prédites, 15 positions sont effectivement modifiées par l'ARN guide attendu dont l'activité *in vitro* a été démontré [88].

L'ensemble des données fonctionnelles a permis d'améliorer considérablement notre compréhension des liens structure/fonction de ces particules et permettront d'affiner à nouveau nos modèles (descripteurs, profiles) pour une détection plus spécifique et sélective de candidats possibles. Il est remarquable de noter que l'approche bioinformatique a également permis d'identifier une cible potentielle dans un ARN autre que les habituels ARNr puisque la position ciblée est présente dans un ARNt : c'est la première fois qu'il est montré qu'un sRNA H/ACA guide la formation de pseudo-uridine dans un ARNt. Il apparaît en effet que plusieurs espèces d'archaea (*S. solfataricus* et *A. pernix*) sont bien pourvues d'un sRNA à boîte H/ACA qui sert de guide pour la modification d'un pre-ARNt (Muller *et al.*, "Deficiency of the $tRNA^{Tyr}$:Ψ35-synthase aPus7 in Archaea of the Sulfolobales order might be rescued by the H/ACA sRNA-guided machinery", 2008) [91].

2.8 Travaux publiés

2.8.1 "The ERPIN server : an interface to profile-based RNA motif identification"

W160–W165 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh418

The ERPIN server: an interface to profile-based RNA motif identification

André Lambert, Jean-Fred Fontaine[1], Matthieu Legendre[1], Fabrice Leclerc[2], Emmanuelle Permal[3], François Major[3], Harald Putzer[4], Olivier Delfour[5], Bernard Michot[5] and Daniel Gautheret[1,*]

CNRS UMR 6207 and [1]INSERM ERM 206, Université de la Méditerranée, Luminy Case 906, 13288 Marseille, Cedex 09, France, [2]UMR 7567 CNRS-UHP, Université Henri Poincaré, 54506 Vandoeuvre-lès-Nancy cedex, France, [3]Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, CP 6128, Succ. Centre-Ville, Montréal, Québec, Canada H3C 3J7, [4]CNRS UPR 9073, 13 rue P. et M. Curie, 75005 Paris, France and [5]ACTiGenics, 10 avenue de l'Europe, 31525 Ramonville St Agne, France

Received February 13, 2004; Revised and Accepted April 5, 2004

ABSTRACT

ERPIN is an RNA motif identification program that takes an RNA sequence alignment as an input and identifies related sequences using a profile-based dynamic programming algorithm. ERPIN differs from other RNA motif search programs in its ability to capture subtle biases in the training set and produce highly specific and sensitive searches, while keeping CPU requirements at a practical level. In its latest version, ERPIN also computes *E*-values, which tell biologists how likely they are to encounter a specific sequence match by chance—a useful indication of biological significance. We present here the ERPIN online search interface (http://tagc.univ-mrs.fr/erpin/). This web server automatically performs ERPIN searches for different RNA genes or motifs, using predefined training sets and search parameters. With a couple of clicks, users can analyze an entire bacterial genome or a genomic segment of up to 5Mb for the presence of tRNAs, 5S rRNAs, SRP RNA, C/D box snoRNAs, hammerhead motifs, miRNAs and other motifs. Search results are displayed with sequence, score, position, *E*-value and secondary structure graphics. An example of a complete genome scan is provided, as well as an evaluation of run times and specificity/sensitivity information for all available motifs.

INTRODUCTION

The last few years have seen a continuous stream of novel non-coding RNA (ncRNA) genes and motifs being reported. Known RNA functions now display a wonderful diversity, ranging from genetic data storage to sensing, transport, targeting and even regulation of essential events such as cell differentiation or cell death. These discoveries have led biologists to undertake a systematic scrutiny of genomic sequences for functional ncRNAs in the form of either independent genes or structural motifs in the untranslated part of transcripts. This effort is carried out in part using experimental RNA amplification strategies. However, with the growth of sequence databases and the development of specific algorithms for RNA structure detection, bioinformatics is now emerging as an inexpensive yet efficient alternative. What tools are now available for computational RNA motif identification? Standard sequence alignment programs are generally not suited to ncRNA searches, because ncRNA is largely characterized by long-range base pair interactions and not by its linear sequence. Bioinformatics has addressed this problem in several ways, notably through descriptor-based systems in which the topology of base-paired regions is specified by the user (1–3), stochastic context free grammars (SCFGs), which use a complete statistical model of RNA elements (4,5), and secondary structure profiles, which are position weight matrices describing stems and single strands in the RNA motif, as implemented in the ERPIN program (6).

ERPIN takes an RNA sequence alignment and secondary structure annotation as input. From this 'training set', the

*To whom correspondence should be addressed. Tel: +33 491 828639; Fax: +33 491 828621; Email: gautheret@esil.univ-mrs.fr
Present address:
Jean-Fred Fontaine, INSERM EMI U 00.18, CHU d'Angers, 49033 Angers, France

2.8.2 "A Dedicated Computational Approach for the Identification of Archaeal H/ACA sRNAs"

CHAPTER FIFTEEN

A Dedicated Computational Approach for the Identification of Archaeal H/ACA sRNAs

Sébastien Muller, Bruno Charpentier, Christiane Branlant, *and* Fabrice Leclerc

Contents

Abstract

Whereas dedicated computational approaches have been developed for the search of C/D sRNAs and snoRNAs, as yet no dedicated computational approach has been developed for the search of archaeal H/ACA sRNAs. Here we describe a computational approach allowing a fast and selective identification of H/ACA sRNAs in archaeal genomes. It is easy to use, even for biologists having no special expertise in computational biology. This approach is a stepwise knowledge-based approach, combining the search for common structural features of H/ACA motifs and the search for their putative target sequences. The first step is based on the ERPIN software. It depends on the establishment of a secondary structure-based "profile." We explain how this profile is built and how to use ERPIN to optimize the search for H/ACA motifs. Several examples of applications are given to illustrate how powerful the method is, its limits, and how the results can be evaluated. Then, the possible target rRNA sequences corresponding to the identified H/ACA motifs are searched by use of a descriptor-based method (RNAMOT). The principles and the practical aspects of this method are also explained, and several examples are given here as well to help users in the interpretation of the results.

Laboratoire de Maturation des ARN et Enzymologie Moléculaire, Nancy Université, Faculté des Sciences et Techniques, Vandoeuvre-les-Nancy, France

Methods in Enzymology, Volume 425
ISSN 0076-6879, DOI: 10.1016/S0076-6879(07)25015-3

2.8.3 "Combined in silico and experimental identification of the Pyrococcus abyssi H/ACA sRNAs and their target sites in ribosomal RNAs"

Nucleic Acids Research Advance Access published February 27, 2008

Nucleic Acids Research, 2008, 1–17
doi:10.1093/nar/gkn077

Combined *in silico* and experimental identification of the *Pyrococcus abyssi* H/ACA sRNAs and their target sites in ribosomal RNAs

Sébastien Muller, Fabrice Leclerc, Isabelle Behm-Ansmant, Jean-Baptiste Fourmann, Bruno Charpentier and Christiane Branlant*

Laboratoire de Maturation des ARN et Enzymologie Moléculaire, UMR 7567 CNRS-UHP, Nancy Université, Faculté des Sciences et Techniques, 54506 Vandoeuvre-lès-Nancy, France

Received January 15, 2008; Revised February 7, 2008; Accepted February 8, 2008

ABSTRACT

How far do H/ACA sRNPs contribute to rRNA pseudouridylation in Archaea was still an open question. Hence here, by computational search in three *Pyrococcus* genomes, we identified seven H/ACA sRNAs and predicted their target sites in rRNAs. In parallel, we experimentally identified 17 Ψ residues in *P. abyssi* rRNAs. By *in vitro* reconstitution of H/ACA sRNPs, we assigned 15 out of the 17 Ψ residues to the 7 identified H/ACA sRNAs: one H/ACA motif can guide up to three distinct pseudouridylations. Interestingly, by using a 23S rRNA fragment as the substrate, one of the two remaining Ψ residues could be formed *in vitro* by the aCBF5/aNOP10/aGAR1 complex without guide sRNA. Our results shed light on structural constraints in archaeal H/ACA sRNPs: the length of helix H2 is of 5 or 6 bps, the distance between the ANA motif and the targeted U residue is of 14 or 15 nts, and the stability of the interaction formed by the substrate rRNA and the 3′-guide sequence is more important than that formed with the 5′-guide sequence. Surprisingly, we showed that a sRNA–rRNA interaction with the targeted uridine in a single-stranded 5′-UNN-3′ trinucleotide instead of the canonical 5′-UN-3′ dinucleotide is functional.

INTRODUCTION

Conversion of uridine into pseudouridine (Ψ) residues is one of the most abundant post-transcriptional modifications of tRNAs, rRNAs and UsnRNAs (1). Compared to U residues, Ψ residues can form an additional hydrogen bond at the N1–H position. Furthermore, the carbon–carbon link between the ribose and the base limits the flexibility of the ribose backbone of Ψ residues, which favours and stabilizes base-pair interactions (2). A role of Ψ residues in stabilization of the tRNA 3D structure has also been well documented (3), and the functional importance of Ψ residues in U2 snRNA for the activity of splicing complexes was demonstrated (4–7). Eukaryal rRNAs contain a large number of Ψ residues compared to bacterial rRNAs and they are concentrated in rRNA regions expected to play important functional roles, in particular in domains IV and V, which are directly involved in the peptidyl transferase activity (8–10). Taken individually, pseudouridylations in rRNAs are not essential for cell growth. However, the complete abolition of Ψ formation in rRNAs is deleterious for ribosome assembly and activity (10–12). Recent data suggest their possible involvement in: (i) subunit interaction (12–14), (ii) the translocation step (14,15), (iii) translation termination (12,16) and (iv) folding of 23S rRNA in an active form at the peptidyl transferase centre (PTC) (11,12,17). The large number of pseudouridylation sites in eukaryal rRNAs compared to bacterial rRNAs is explained by the use of different catalysts: stand-alone enzymes carrying both the RNA recognition capability and the catalytic activity are used in bacteria (10,18), whereas U to Ψ conversions are catalyzed by H/ACA snoRNPs in eukarya (19,20). The H/ACA snoRNPs contain four proteins and an H/ACA snoRNA that defines the targeted U residue by base-pair interaction with the rRNA. Recent data revealed a similar RNA-guided system in archaea (21–24). Most of the eukaryal H/ACA snoRNAs contain two characteristic stem-loop structures, with an internal loop (pseudouridylation pocket), that is complementary to two nucleotide stretches bordering the targeted U residue. Each of the stem-loops is flanked by a conserved motif (H and

*To whom correspondence should be addressed. Tel: +33 3 83 68 43 03; Fax: +33 3 83 68 43 07; Email: christiane.branlant@maem.uhp-nancy.fr

2.8.4 "Deficiency of the tRNA^{Tyr} :Ψ35-synthase aPus7 in Archaea of the Sulfolobales order might be rescued by the H/ACA sRNA-guided machinery"

1308–1322 *Nucleic Acids Research, 2009, Vol. 37, No. 4* *Published online 12 January 2009*
doi:10.1093/nar/gkn1037

Deficiency of the tRNA^{Tyr}:Ψ35-synthase aPus7 in Archaea of the Sulfolobales order might be rescued by the H/ACA sRNA-guided machinery

Sébastien Muller[1], Alan Urban[1], Arnaud Hecker[2], Fabrice Leclerc[1], Christiane Branlant[1,*] and Yuri Motorin[1]

[1]Laboratoire de Maturation des ARN et Enzymologie Moléculaire, UMR 7567 CNRS-UHP Nancy Université, BP 239, 54506 Vandoeuvre-les-Nancy Cedex and [2]Institut de Génétique et Microbiologie, Université Paris-Sud, IFR115, UMR8621-CNRS, 91405 Orsay, France

Received August 10, 2008; Revised December 11, 2008; Accepted December 12, 2008

ABSTRACT

Up to now, Ψ formation in tRNAs was found to be catalysed by stand-alone enzymes. By computational analysis of archaeal genomes we detected putative H/ACA sRNAs, in four Sulfolobales species and in *Aeropyrum pernix*, that might guide Ψ35 formation in pre-tRNA^{Tyr}(GUA). This modification is achieved by Pus7p in eukarya. The validity of the computational predictions was verified by *in vitro* reconstitution of H/ACA sRNPs using the identified *Sulfolobus solfataricus* H/ACA sRNA. Comparison of Pus7-like enzymes encoded by archaeal genomes revealed amino acid substitutions in motifs IIIa and II in Sulfolobales and *A. pernix* Pus7-like enzymes. These conserved RNA:Ψ-synthase-motifs are essential for catalysis. As expected, the recombinant *Pyrococcus abyssi* aPus7 was fully active and acted at positions 35 and 13 and other positions in tRNAs, while the recombinant *S. solfataricus* aPus7 was only found to have a poor activity at position 13. We showed that the presence of an A residue 3′ to the target U residue is required for *P. abyssi* aPus7 activity, and that this is not the case for the reconstituted *S. solfataricus* H/ACA sRNP. In agreement with the possible formation of Ψ35 in tRNA^{Tyr}(GUA) by aPus7 in *P. abyssi* and by an H/ACA sRNP in *S. solfataricus*, the A36G mutation in the *P. abyssi* tRNA^{Tyr}(GUA) abolished Ψ35 formation when using *P. abyssi* extract, whereas the A36G substitution in the *S. solfataricus* pre-tRNA^{Tyr} did not affect Ψ35 formation in this RNA when using an *S. solfataricus* extract.

INTRODUCTION

In all domains of life, pseudouridine residues (Ψ) are the most frequent post-transcriptionally modified residues in RNAs. They are universally found in ribosomal RNA (rRNAs) and in tRNAs (1,2). U to Ψ conversions are catalysed either by stand-alone enzymes [specific RNA:Ψ-synthases, (3)] or by small ribonucleoprotein particles [H/ACA snoRNPs or H/ACA scaRNPs in eukarya, and H/ACA sRNPs in Archaea, (4)]. H/ACA RNPs contain a small RNA that defines the targeted U residue by base-pair interaction with the RNA substrate (5–8). Archaeal and eukaryal H/ACA RNPs contain 4 proteins: Nop10, Gar1, L7ae/Nhp2p and Cbf5/Dyskerin (9). CBF5 belongs to the TruB family of RNA:Ψ-synthases. Whereas aCBF5 alone has no activity on rRNAs (5–8), recent data showed an *in vitro* activity of aCBF5 at position 55 in tRNAs. This activity does not require the presence of a guide RNA (10–12).

The additional free N_1-H of the Ψ nucleobase allows the formation of an additional hydrogen bond, either in *cis*, within the RNA molecule, or in *trans*, with another RNA molecule or a protein. For instance, residue Ψ35 in the eukaryal cytoplasmic tRNA^{Tyr}(GUA) allows the formation of an hydrogen bond with the O2′ of residue U33, which stabilizes the anticodon loop structure (13,14). Furthermore, substitution of the C-N bond between the ribose and the nucleobase by a C-C bond limits the flexibility of the ribose-phosphate backbone and favours RNA-RNA base-pair interactions (13–15).

In all organisms, pseudouridylations were found to occur at numerous positions in tRNAs (seven in *Escherichia coli*, at least 15 in *Saccharomyces cerevisiae* and even more in higher eukaryotes) (1). Formation of residue Ψ55 in the TΨC loop is the most frequent

*To whom correspondence should be addressed. Tel: +33 3 83 68 43 03; Fax: +33 3 83 68 43 07; Email: christiane.branlant@maem.uhp-nancy.fr

Perspectives : Recherche et modélisation des liens structure/fonction d'ARN non-codants

Extension de la recherche de sRNA à boîtes H/ACA chez les archaea

Les perspectives futures dans ce projet sont d'améliorer la spécificité et la sélectivité des méthodes de détection *in silico* à la fois des sRNA à boîtes H/ACA et de leur(s) cibles(s) afin d'étendre notre compréhension des liens structure/fonction de ces ARN guides de modification. L'augmentation de la spécificité et sélectivité de détection de motifs H/ACA qui pourraient différer de façon notable de ceux déjà identifiés demande un changement de stratégie. En effet, nous avons développé des outils de plus en plus spécifiques qui ont permis d'identifier un grand nombre de sRNA à boîtes H/ACA dans tous les génomes disponibles d'archaea (Fig. 27). Toutefois, ces sRNA à boîtes H/ACA peuvent ne représenter qu'une sous-classe d'ARN guides de modification par rapport aux sRNA à boîtes H/ACA ou aux autres sRNA et qu'une petite partie de l'ensemble des ARN non-codant présents dans ces organismes. L'enrichissement graduel du profile ERPIN à partir de nouveaux candidats a permis d'augmenter la diversité des séquences représentées dans la banque de motifs H/ACA et donc de continuer à identifier de nouveaux motifs. Cette approche semble maintenant avoir atteint ses limites puisqu'aucun nouveau candidat n'a pu être identifié à partir de la banque actuelle qui compte 162 motifs H/ACA.

Afin d'identifier de nouveaux sRNA à boîtes H/ACA ou d'autres ARN non-codant, un relâchement supplémentaire des contraintes de structuration des motifs H/ACA sera testé. Un retour à une stratégie plus générale par génomique comparative (Fig. 18 & Fig. 20) sera aussi effectuée, afin d'identifier des ARN potentiellement structurés dans les régions IRC conservées qui correspondraient à des motifs H/ACA différents de la forme canonique.

Les règles d'association entre ARN guides et ARN cibles semblent plus flexibles que celles supposées initialement, en particulier en ce qui concerne le nombre de nucléotides non appariés entre les 2 éléments de séquence complémentaires aux éléments antisens de la poche de pseudouridylation (voir Pab19, Fig. 26). D'après des résultats de calculs théoriques d'énergie libre de liaison effectués sur les couples possibles ARN guide/ ARN cible (Fig. 26), il apparaît qu'une modèle reposant sur la stabilité du complexe ARN/ARN et des règles structurales liées à la position de la cible par rapport à la boîte ACA pourrait améliorer la prédiction de cibles. Des contraintes moins sévères quant au nombre d'appariements dans les séquences complémentaires ou au nombre de nucléotides non appariés entre ARN guide et ARN cible combinées à une évaluation de l'énergie libre de liaison sont susceptibles d'améliorer la prédiction de cibles. Un nouveau descripteur pour les cibles des sRNA à boîtes H/ACA sera défini afin de tenir compte de

modes d'appariement entre ARN guide / ARN cible faisant intervenir plus de 2 nucléotides non appariés entre les régions complémentaires des 2 ARN. Davantage de cibles potentielles pourront être considérées au départ mais elles seront sélectionnées ensuite selon les règles structurales établies de façon empirique (Fig. 26) et d'après les données structurales (Fig. 28).

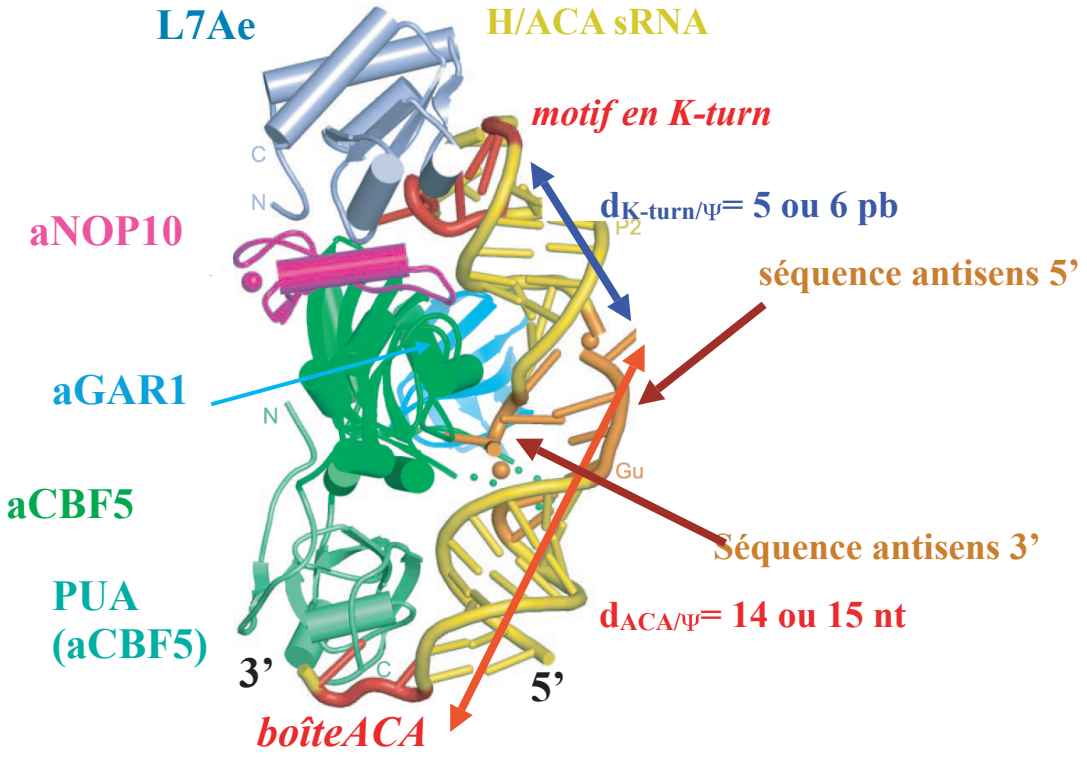


Figure 28. Structure 3D d'un complexe ribonucléoprotéique sRNA H/ACA-aCBF5-aNOP10-aGAR1-L7AE déterminée par radiocristallographie. Les contraintes empiriques déduites de l'analyse des règles d'association entre ARN guide et ARN cible sont reportées sur la structure 3D.

L'approche pourra être appliquée notamment aux Sulfolobales (3 génomes : *S. tokodaii*, *S. solfataricus* et *S. acidocaldarius*) et aux Desulfurococcales (2 génomes : *H. butylicus* et *A. pernix*). Ce sont les 5 génomes les plus proches du point phylogénétique qui soient complètement séquencés. L'approche basée sur ERPIN n'a permis d'identifier que 2 gènes de sRNA à boîtes H/ACA chez les *Sulfolobus*. Il est possible que d'autres positions habituellement modifiées chez les archaea grâce à des ARN guides le soient par un système enzymatique de type Ψ-synthase sachant que l'enzyme présente chez les Sulfolobales a une spécificité différente de celle de *P. abyssi* [91]. L'autre possibilité est qu'on ne les ait pas identifié car la structuration du sRNA H/ACA dévie de la forme canonique (Fig. 23A), ou bien que l'association ARN guide / ARN cible dévie trop des règles classiques (Fig. 23B). Par exemple, l'identification de l'ARN cible du gène Pab19 (Fig. 26), montre que les règles d'association peuvent faire intervenir un trinucléotide non-apparié au lieu du classique dinucléotide non-apparié.

De données structurales récentes obtenues sur un complexe ribonucléoprotéique comprenant un ARN guide à boîte H/ACA et sa cible permettent de mieux comprendre les règles structurales que nous avions déduites de façon empirique sur les distances entre la position pseudo-uridylée d'une part et la boîte ACA (de 14 ou 15 nucléotides) ou le motif K-turn (5 ou 6 nucléotides) d'autre part (Fig. 26). La structure 3D déterminée par Li & Ye en 2006 [92] éclaire de façon inté-

ressante ces contraintes empiriques qui permettent de discriminer parmi plusieurs cibles potentielles celles qui sont les plus pertinentes et qui correspondent à des contraintes structurales fortes pour la reconnaissance des cibles (Fig. 28). D'autres éléments de relations structure/fonction sont étudiés au laboratoire : il a été montré expérimentalement que la séquence de l'hélice H1 a un impact sur le taux de fixation de certaines protéines sur le complexe ARN/ARN. A partir des données structurales existantes, l'influence de différents éléments de structures 3D pourra être étudiée de façon théorique par modélisation moléculaire afin d'essayer de comprendre le rôle des hélices, de la poche de pseudo-uridylation, ou du K-turn dans les différentes étapes d'assemblage et de la catalyse.

Modèles 2D et 3D pour la co-évolution et la reconnaissance ARN guide/ARN cible

Les données moléculaires sur la reconnaissance entre ARN guide/ARN cible des sRNP H/ACA obtenues à partir des gènes candidats de sRNA à boîtes H/ACA identifiés par génomique comparative ont permis de valider expérimentalement un certain nombre de couples guide/cible(s). L'absence de contraintes fortes imposées au départ dans la recherche de cibles potentielles des sRNA H/ACA a permis d'identifier une assez grande diversité de cibles et de modes d'interaction (longueurs des éléments antisens 5' ou 3' appariés, nombre de mésappariements, positionnement dans la structure par rapport aux motif K-turn et ACA). L'existence de couples fonctionnels et d'autres non fonctionnels (non validés expérimentalement) a permis d'édicter un certain nombre de règles d'association qui se sont révélées être en accord avec les données structurales disponibles depuis peu sur un complexe sRNP H/ACA (Fig. 28). L'ensemble de ces données et modèles permettent d'envisager d'étendre et de rationaliser les résultats obtenus dans 2 directions : la première est d'essayer de comprendre l'évolution moléculaire des gènes de sRNA et des motifs H/ACA chez les *Pyrococcus* dans un premier temps et ensuite chez l'ensemble des archaea pour lesquelles des gènes de ce type ont été identifiés (le fait, par exemple, qu'il existe des gènes comportant des motifs H/ACA triples, une caractéristique spécifique des archaea par rapport aux eucaryotes). La seconde direction est d'exploiter les données structurales désormais disponibles sur deux complexes sRNP H/ACA avec et sans substrat [92, 93] et les données moléculaires obtenues notamment au laboratoire (sur les motifs naturels et mutants pour leur affinité avec les protéines du complexe RNP, leur cinétique d'association et leur structuration étudiée par dichroïsme circulaire, etc) pour développer des modèles 3D permettant de comprendre les liens structure/fonction. On cherchera en particulier à comprendre le rôle de déterminants structuraux connus ou à identifier de nouveaux déterminants structuraux des ARN et des protéines liées à la fonction en modélisant la structure 3D de complexes sRNP H/ACA naturels ou mutants artificiels ou associés à une maladie.

Les gènes de sRNA à boîtes H/ACA d'archaea sont organisés sous la forme de motifs H/ACA simples (structure tige-boucle), de motifs doubles ou triples. La similarité de séquence et de structures de ces motifs suggèrent que ces gènes ont pu évolué de façon modulaire par des transferts conservant le nombre de motifs H/ACA dans un gène entre différentes espèces d'archaea ou au contraire en scindant certains gènes à motifs multiples ou encore en fusionnant des motifs pour générer un gène à motifs multiples. Notre approche combinant une recherche basée sur le profil ERPIN des motifs H/ACA et de leur(s) cible(s) a permis d'identifier 162 candidats de motifs à boîtes H/ACA identifiés chez les archaea. Des données sur des variants des motifs H/ACA chez les *Pyrococcus* et la comparaison des séquences et structures 2D des 162 motifs suggèrent également qu'il y a pu avoir des co-évolutions entre le motif H/ACA (au niveau du site de pseudo-

uridylation) et sa ou ses cibles. Par exemple, des variations d'un ou quelques nucléotides dans la boucle interne du motif H/ACA ont pu conduire à un changement de spécificité de reconnaissance de sa cible sur l'ARNr ou d'autres cibles. Nous tenterons d'identifier les mécanismes qui peuvent expliquer ce type de co-évolution entre ARN guide et ARN cible par la mise en œuvre de méthodes telle que celle développée par Dutheil et al. [94] pour l'identification de positions covariantes entre ARN faisant appel à un mode de reconnaissance de type sens/anti-sens.

Les résultats préliminaires obtenus sur les 33 motifs identifiés chez les *Pyrococcus* ont permis de construire un arbre de similarités de structures 2D suggérant l'existence d'un ancêtre commun à l'ensemble des motifs H/ACA de *Pyrococcus* et dont les plus proches descendants correspondent au motif Pab40.2 et ses homologues chez *P. horikoshii* et *P. furiosus* (Fig. 29). Il est intéressant de noter que ces 3 motifs (Pab4.2, Ph40.2 et Pfu7.2) ont la possibilité de se replier sous 2 formes alternatives en raison de la présence d'une région auto-complémentaire dans la boucle interne : le repliement atypique de Pab40.2 semble fonctionnel puisqu'il permet *a priori* de cibler la position L1932 (position 1932 dans la grande sous-unité du ribosome) alors que le repliement canonique similaire à celui rencontré dans tous les autres motifs pourrait ne pas avoir de cible. Etant donné la similarité relativement forte entre les motifs homologues de Pab40.3 avec le Pab40.2, cela suggère un remaniement génétique impliquant une duplication et juxtaposition du motif ancestral Pab40.2. Le motif Pab40.1 semble avoir divergé plus fortement par rapport au motif ancestral et aurait pu être intégré dans le gène Pab40 (comportant le motif triple) moyennant un remaniement génétique avec un déplacement important au sein du génome. Les autres gènes comportant des motifs doubles (Pab105 et Pab35) pourraient avoir évolué à la suite de duplications assez récentes.

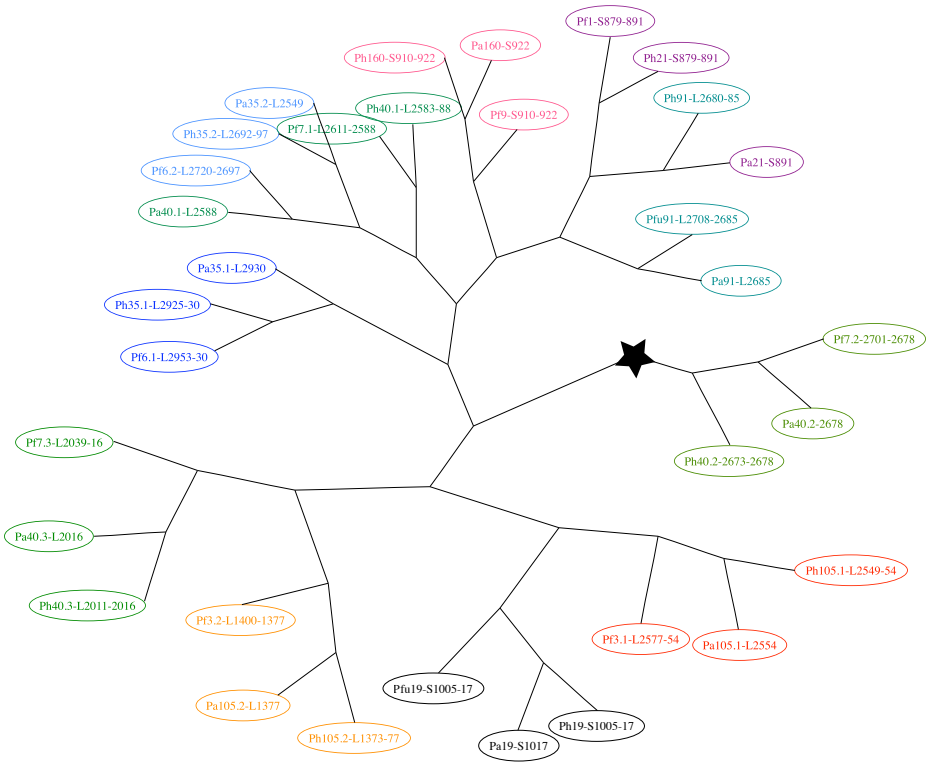
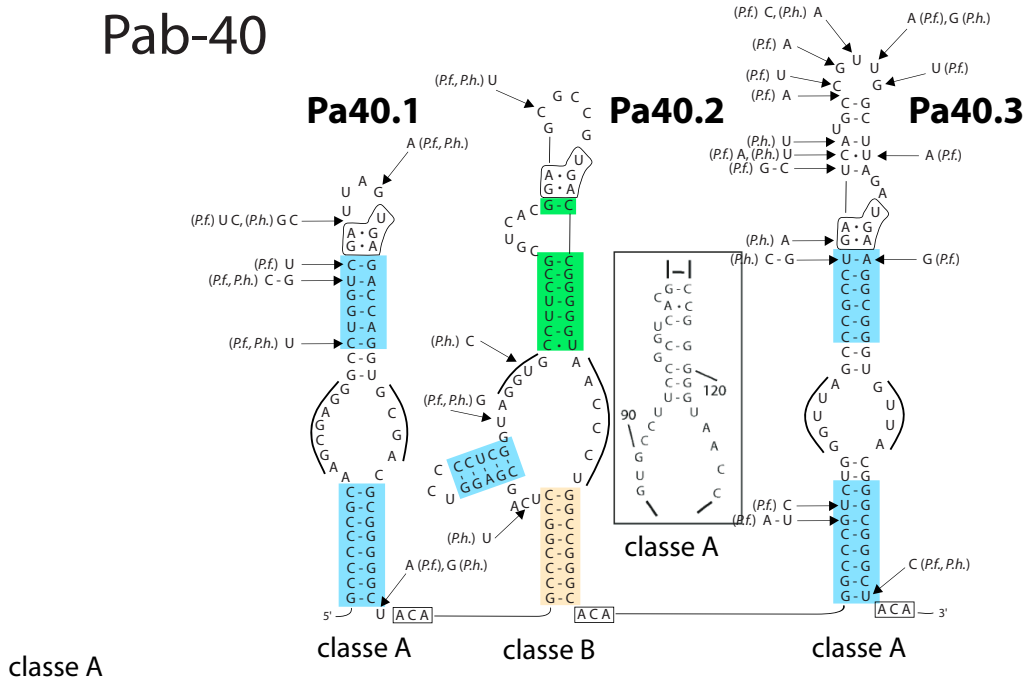


Figure 29. Arbre phylogénétique basée sur une mesure de similarité entre structures 2D des sRNA H/ACA proposés chez les *Pyrococcus*. Les 33 motifs H/ACA sont classés en fonction de la similarité de leur structure 2D calculée par le programme RNAforester (Vienna package).

Nos travaux ont conduit à proposer 2 repliements alternatifs pour les motifs H/ACA tels qu'ils ont été déposés dans la banque de données Rfam (http://www.sanger.ac.uk/Software/Rfam) et identifiés comme appartenant à la classe A pour le repliement canonique et la classe B pour le repliement non canonique (Fig. 30). Le fait que l'ancêtre commun des motifs H/ACA ait pu adopter 2 repliements alternatifs concomitants pourrait indiqué une plus grande versatilité potentielle des motifs ancestraux pour la reconnaissance de cibles potentielles. Les motifs H/ACA auraient ensuite évolué et co-évolué avec la séquence de leur(s) cible(s) en adoptant un repliement plus canonique. Un exemple hypothétique de changement de spécificité de cible à partir du motif ancestral Pa40.2 est donné pour illustrer l'effet possible de simples mutations ponctuelles ; la mutation à la position 85 de U en G entre la Pa40.2 et le Ph40.2 ou vice versa aurait permis un changement de spécificité de la position ciblée de L1932 vers L2678 chez *P. horikoshii* (Fig. 31).

Une étude structurale des relations structure/fonction des sRNA à boîtes H/ACA a été amorcée en utilisant une approche par modélisation 3D. Il a été montré expérimentalement au laboratoire que le motif H/ACA du Pab21 a une meilleure affinité pour aCBF5 que le motif H/ACA de Pab91. Les résultats expérimentaux suggèrent aussi que cette différence d'affinité provient de la séquence de l'hélice H1. Pab91 a la particularité de posséder une hélice H1 composée d'une longue série de pyrimidines en 5' et de purines en 3' alors que celle de Pab21 fait alterner purines et pyrimidines sur le même brin de l'hélice. Des simulations effectuées par dynamique moléculaire sur une forme tronquée des ARN Pab21 et Pab91 (comportant l'hélice H1 et la boucle interne) font apparaître de légères altérations des régions hélicoïdales en ce qui concerne l'élévation et l'inclinaison de l'hélice H1. Un léger étirement de l'hélice H1 est observée pour l'ARN Pab91 (h-Rise moyen de 2.83Å) par rapport à une hélice standard d'un ARN de forme A (h-Rise moyen de 2.73Å) même s'il n'est pas significatif par rapport à l'élévation moyenne d'une hélice d'ARN correspondant à un duplex poly(A)-poly(U) (h-Rise moyen de 2.81Å). Par contre, l'élévation moyenne pour l'ARN Pab21 est conforme à celle d'une hélice standard d'ARN. Sur les 9 paires de base de l'hélice H1, la différence moyenne d'élévation correspond à un étirement de plus d'1Åpour l'hélice H1 de Pab91 (Fig. 32). En revanche, l'inclinaison de l'hélice H1 est plus marquée dans le cas de l'ARN Pab21 que dans le cas de l'ARN Pab21 (Fig. 32). Ces légères différences structurales suggère qu'un positionnement légèrement différent de la poche de pseudo-uridylation de Pab91 dans la ribonucléoprotéine pourrait expliquer la différence d'affinité entre les 2 ARN.



classe A

```
# STOCKHOLM 1.0

Ph40.2-2673-2678    GCCCGGCc.ttagcgaggtc.....CCCTCgG..GAggcgccttccgcgtcacggagtgccGTGA.ccGGGGG....taaccct.gGCCGGGC..ACA
Pa40.2-2678         GCCCGGCc.tcagcgaggtc.....CCCTCgG..TAggtgccttccgcgtcacggagcgccGTGA.ccGGGGG....taaccct.gGCCGGGC..ACA
Pf7.2-2701-2678     GCCCGGCc.tcagcgaggtc.....CCCTCgG..GAggtgccttccgcgtcacggagtgccGTGA.ccGGGGG....taaccct.gGCCGGGC..ACA
Pf3.1-L2577-54      CCGCCCGGG...CGGCCCGA........CCGAGGGA...GCGTGCCGGCAATGGCGCGCG.ATGAA.CGAGG.....TGACGTCTCCGGGCGG..ACA
Pa105.1-L2554       CCGCCCGG....AGGCCCGA........CCGAGGGA...GCGTGCCGAGAAAGGCGCGCC.ATGAA.CGAGG.....CGACGTCGCCGGGCGG..ACA
Ph105.1-L2549-54    CCGCCCGG....GGGCCCGA........CCGAGGGA...GCGTGCCGAGAATGGCGCGCA.ATGAA.CGAGG.....TGACGTCGTCGGGCGG..ACA
Pfu19-S1005-17      ACCCAGGGG..cccgacggcggcgT...CGGGgGGA...ttgggggcaaagccccccgga..ATGAacCCCGC......cctcctCCCCTGGGT..ATA
Pa19-S1017          GCCCAGGGG.tcaagacggcggcGT...CGGggGGA...ttgggggcaaagccccccggc..ATGAacCCCGC......cctcctCCCCTGGGT..ATA
Ph19-S1005-17       GCCCAGGGGt.caagacggcggcGT...CGGggGGA...ttgggggcaaagccccccgga..ATGAacCCCGC......cctcctCCCCTGGGT..ATA
Pf3.2-L1400-1377    GGGCCCGGC....CTTCGG......GGCCGCCT.GA...GGGAGCCGAGAAGGGCAGACG.ATGA..AGGCGGGC...GGATAAGCCGGGCCCTCAAA
Pa105.2-L1377       GGGCCCGGT....CTCCGG......GGCCGCCT.GA...GGTTGCCGACAACGGCGGGCA.ATGA..GGGCGGGT...GGATAAGCCGGGCCT..ATA
Ph105.2-L1373-77    GGGCCCGGT....CTCCGG......GGCCGCCT.GA...GGTTGCCGAGAATGGCGTTCA.ATGA..AGGCGGGC...GGATAATCCGGGCCT..AAA
Pf7.3-L2039-16      GGCCCACCTG....GGTTA......GCCCGCCT.GA....GAATGCATACATGCTACG...ATGA..GGGCGGGT...GTTAC.GGGTGGGCC..ACA
Pa40.3-L2016        GGCCCGTCTG....GGTTA......GCCCGCCT.GA....TCATGCCGTTGGCTTAG....ATGA..AGGCGGGT...GTTAC.GGGCGGGCT..ACA
Ph40.3-L2011-2016   GGCCCGTCTG....GGTTA......GCCCGCCC.AA....TTTTGCCGAGGGCTTAG....ATGA..GGGCGGGT...GTTAC.GGGCGGGCC..ACA
Pa35.1-L2930        GGGCTCGGT.....CTACCCGC.....CCCCGC.AA......GGTGTTCGGGTTCG.....ATGA.GCGGGG...TGTGCTCACGCCGAGCCC..ACA
Ph35.1-L2925-30     GGGCTCGGT.....CAACCCGC.....CCCCGC.AA......GGTTTCGGGG.TCG.....ATGA.GCGGGG...TGTGCTCACGCCGAGCCT..ACA
Pf6.1-L2953-30      GGGCTCGGT.....CAACCCGC.....CCCCGC.AA......GGTTTCGGGT.TCG.....ATGA.GCGGGG...TGTGCTCACGCCGAGCCT..ACA
Pf9-S910-922        GCCCCCGGG.....AAACCGC......GGCGGGGGA...GCGAGCGG......CGAGCT..GTGAG.CCCGCT....TCGCTCC.CCGGGGGT..ACA
Pa160-S922          GCCCCCGG......AAACCGC.....GGGGGAG.GA...GCTTAGGCTTAAGCCGAGCT..ATGA..CTCCCCT...TCGCTCC.CCGGGGGC..ACA
Ph160-S910-922      GCCCCCGG......AAACCGC.....GGGGGAG.GA...GCTGAGGCTAAAGCCGAGCT..ATGA..CTCCCCT...TCGCTCC.CCGGGGGC..ACA
Pa21-S891           GGGCCCGGC.....TCCCGCC....CTCTCCGGGGA...........ATC...........GTGAACCGGGGG.....TTCC.GGCCGGGCCT..ACA
Ph91-L2680-85       CCCCTCCCC.....TCTCACA.....CCCCCGTGA...........TCG...........GTGAT.CGGGGGG....CGGTC.GGGGAGGGG..ACA
Ph21-S879-891       CGGCCCGGT.....TCCCGCC....CTCTCCGGGGA...........ATC...........GTGAACCGGGGG.....TTCC.GACCGGGCCG..ACA
Pf1-S879-891        GGGCCCGGT.....TCCCGCC....CTCTCCGGGGA...........ATC...........GTGAACCGGGGG.....TTCC.GACCGGGCCC..ACA
Pa91-L2685          CCCCTCCCC.....TCTCACA.....CCTCCGG.GA...........TCA...........GTGA..CCGGAGGG...CGGTC.GGGGAGGGG..ACA
Pfu91-L2708-2685    CCCCTCCCC.....TCTCACA.....CCCCCGT.GA...........GAA...........GTGA..GCGGGGGG...CGGTC.GGGGAGGGG..ACA
Ph40.1-L2583-88     GCCCCCGC.....AAGCGAGG....GCTTGGCC.GA.........GCTTA...........ATGA.GCCAGGT.....GCGAC..GCGGGGGCG.ACA
Pf7.1-L2611-2588    GCCCCCGC.....AAGCGAGG....GCTTGGCT.GA.........TCTTA...........ATGA.GCCAGGT....GCGAC..GCGGGGGCA.ACA
Pf6.2-L2720-2697    GGGCCCGT.....GCATCC.....GCCCCGCGG.GA...........TCA...........ATGA.CCGCTGGGT..CTCTGTT.GCCGGCCC..ACA
Pa40.1-L2588        GCCCCCGC.....AAGCGAGG....GCCTGGTC.GA...........TTA...........GTGA.GACCAGGT....GCGAC..GCGGGGGCT.ACA
Pa35.2-L2549        GGGCCGGT.....GCATCC.....GCCCCGCGG.GA...........TCA...........ATGA.CCGCCGGGT...CTCTGTTGCCGGCCT..ACA
Ph35.2-L2692-97     GGGCCGGT......GCATCC.....GCCCGCGG.GA...........TTA...........ATGA.CCGCTGGGT..CTCTGTT.GCCGGCCC..ACA
#=GC SS_cons        <<<<<<.....................<<<<<<.<<.........................>>..>>>>>>.............>>>>>>.....
//
```

classe B

```
# STOCKHOLM 1.0

Pa40.2-1932         GCCCGGCcucagcgaggguCCCCUcgGUAggugccuuccgcgucacgGAgcgccGUGAccGGGGGuaacccugGCCGGGCACA
Ph40.2-2673-78      GCCCGGCcuuagcgaggcCCCCUcgGGAggcgccuuccgcgucacggagugccGUGAccGGGGGuaacccugGCCGGGCACA
Pf7.2-2701-2678     GCCCGGCcucagcgaggcCCCCUcgGGAggugccuuccgcgucacggagugccGUGAccGGGGGuaacccugGCCGGGCACA
#=GC SS_cons        <<<<<<<<....<<<<<...>>>>>.......<<.<<<<......<<<.......>>>>>>>>>.......>>>>>>>...
//
```

Figure 30. Repliements alternatifs de classe A et classe B des motifs H/ACA chez les Archaea. Les 2 repliements alternatifs sont illustrés sur la structure 2D du triple motif Pab40 (les 2 motifs extrêmes sont de classe A, le motif central de classe B ou de classe A). Les alignements de l'ensemble des séquences des motifs H/ACA de *Pyrococcus* de classe A et de classe B sont donnés sous le format Stockholm.

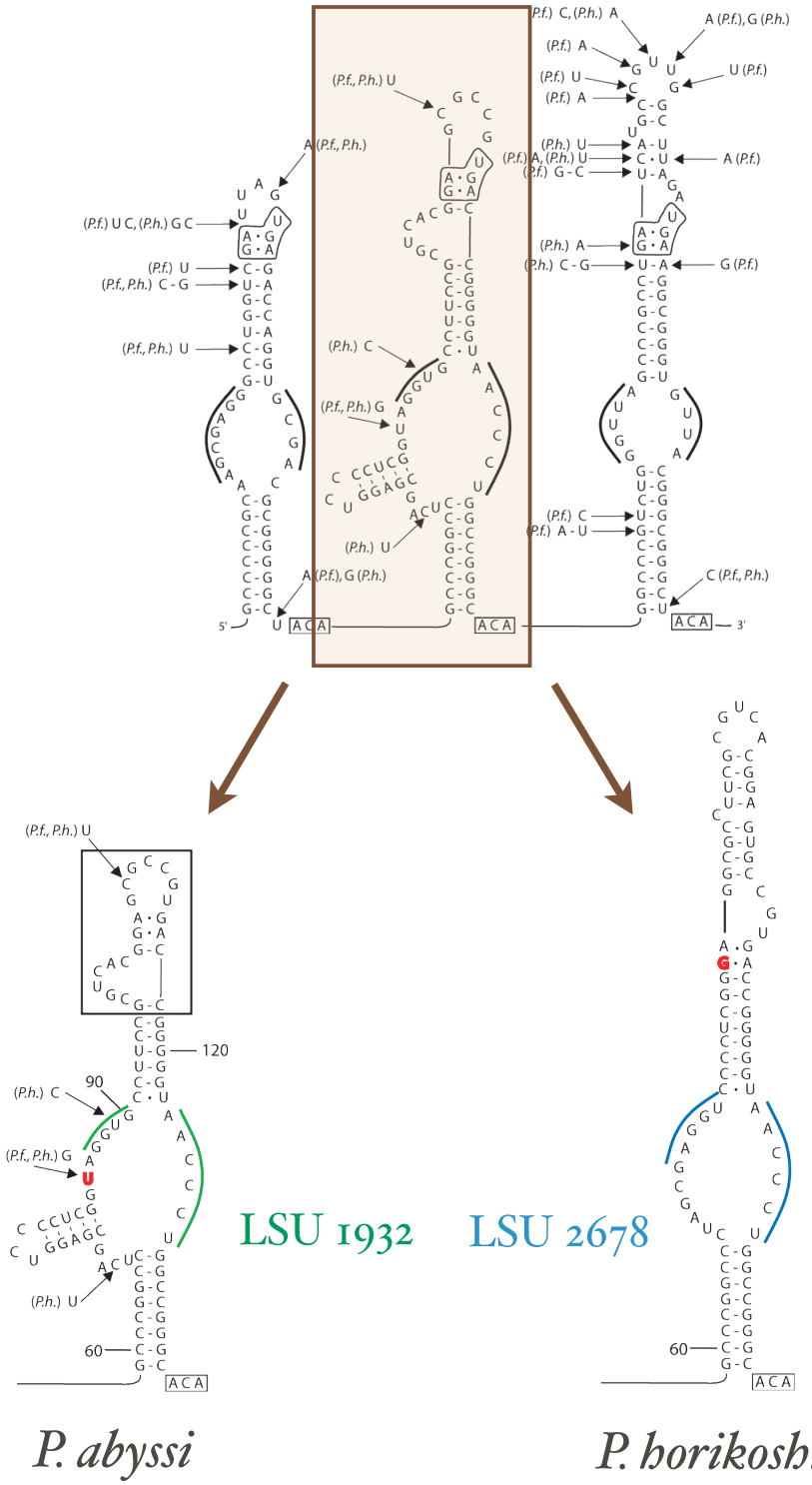


Figure 31. Effet possible de mutations ponctuelles sur la spécificité de reconnaissance des cibles d'ARN guides à boîtes H/ACA. Une mutation à la position 85 dans la boucle interne du motif Pa40.2 de *P. abyssi* aurait conduit à un changement de spécificité de l'ARN guide pour reconnaître la position L1932 au lieu de la position L2678 ciblée par le Ph40.2.

Grâce aux données structurales récentes obtenues sur 2 particules ribonucléoprotéiques contenant un sRNA à boîte H/ACA, il devient envisageable de modéliser la structure complète de la ribonucléoprotéine pour les complexes correspondant à Pab21 et Pab91. Une modélisation des différentes étapes d'assemblage de l'ARN guide avec sa cible et les protéines de la particule pourra être effectuée en exploitant les données expérimentales obtenues au laboratoire sur les affinités de liaison, les cinétiques d'association et autres données moléculaires relatives à ces 2 complexes. Le rôle d'autres déterminants fonctionnels potentiels sur les ARN pourront être étudiés par une approche similaire (taille de la poche de pseudo-urydilation, présence de nucléotides non appariés dans les régions double-brin, etc).

Du point de vue méthodologique, les approches de modélisation et plus spécifiquement de "docking" utilisés pour modéliser la structure 3D des complexes RNP sont décrites en détail dans le chapitre suivant. Les données structurales obtenues par Li & Ye (2006) [92] correspondent à l'association de l'ensemble des protéines du complexe RNP (Cbf5, Nop10, Gar1 et L7ae) et un ARN guide artificiel dérivé en partie d'une sRNA de *P. furiosus*. Ces données ont été utilisées pour construire 2 modèles initiaux correspondant aux complexes sRNP H/ACA avec les ARN guides homologues : Pab21 et Pfu1 (Fig 33). Le complexe avec Pfu1 a été construit par modélisation de l'ARN guide dans le site de fixation de l'ARN dans le complexe RNP en lieu et place de l'ARN guide artificiel. Le complexe avec Pab21 a été construit de la même façon pour l'ARN (qui diffère par quelques mutations avec Pfu1) et en substituant les protéines de *P. furiosus* par celles de *P. abyssi* dont les structures 3D ont été déterminées au laboratoire par l'équipe structurale (Xavier Manival et Christophe Charron) [95, 96].

Outre les déterminants fonctionnels sur l'ARN guide, des déterminants fonctionnels protéiques sont connus : une mutation dans le gène codant la protéine Cbf5 est associée par exemple à une maladie, la dyskeratose [97]. La modélisation 3D des complexes sRNP H/ACA avec des mutants protéiques, cette fois, sera également utilisée pour essayer de comprendre les bases moléculaires de la fonction de ces RNP dans un contexte fonctionnel et non fonctionnel lorsque la mutation abolit ou altère fortement la fonction. Une des mutations ponctuelles associées à la dyskeratose est celle qui affecte la position 34 de Cbf5 où une arginine est convertie en tryptophane (R34W). A l'aide de modèles 3D des structures des complexes ribonucléoprotéiques, nous essaierons de comprendre l'effet de cette mutation sur la structure et la dynamique en solution des complexes par des simulations de "docking" (optimisation de la structure de départ des complexes) et de dynamique moléculaire (étude de la flexibilité) (Fig. 34).

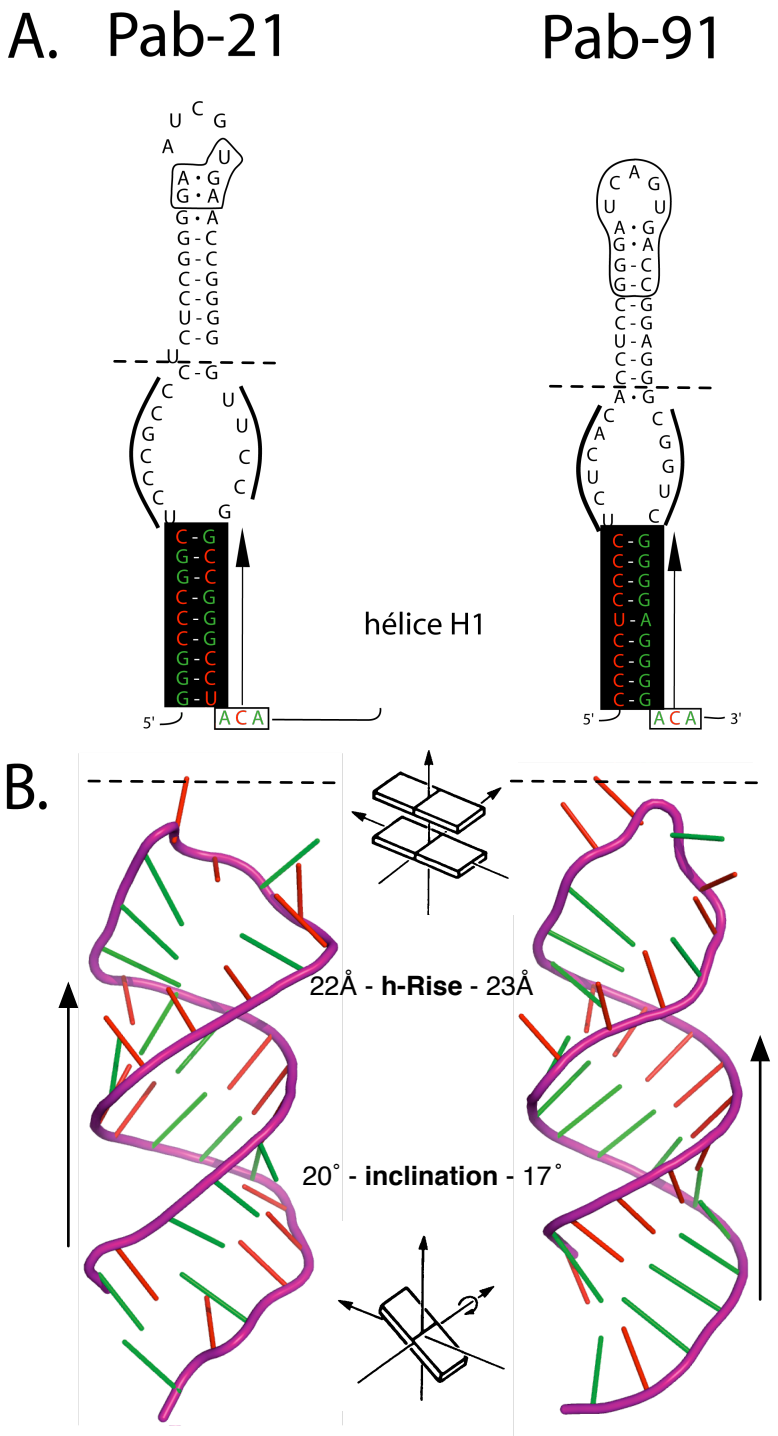


Figure 32. Modèles des sRNA H/ACA de *Pyrococcus abyssi* : Pab21 et Pab91. A. Modèles de structures 2D des sRNA Pab 21 et Pab91. B. Modèles 3D des formes tronquées sRNA Pab 21 et Pab91 obtenus par dynamique moléculaire. Les sRNA comportent les parties basales incluant l'hélice H1, la boucle interne et la boîte ACA. Les modèles 3D ont été construits avec des paramètres d'hélice standard (Mc-Sym) ; la boucle terminale (assimilée à la boucle interne de la forme non tronquée) a été extraite d'une structure 3D correspondant à une dodecaloop de l'ARNr 23S (PDB ID : 1J5A) choisie pour sa similarité de séquence et proportion purines/pyrimidines. Une simulation par dynamique moléculaire (CHARMM) de 5ns (en plus de l'équilibration du système : 500ps) a été effectuée pour chaque ARN immergé dans une boîte d'eau contenant des contre-ions (ions Na^+). Les mesures d'élévation et d'inclinaison de l'hélice (h-Rise, inclination) ont été réalisées sur la structure "moyenne" (coordonnées moyennes sur l'ensemble de la simulation et optimisées par minimisation d'énergie). Les valeurs d'élévation et d'inclinaison hélicoïdales ont été calculées sur les 9 paires de base de l'hélice (X3DNA). Dans les 2 modèles, les purines apparaissent en vert et les pyrimidines en rouge.

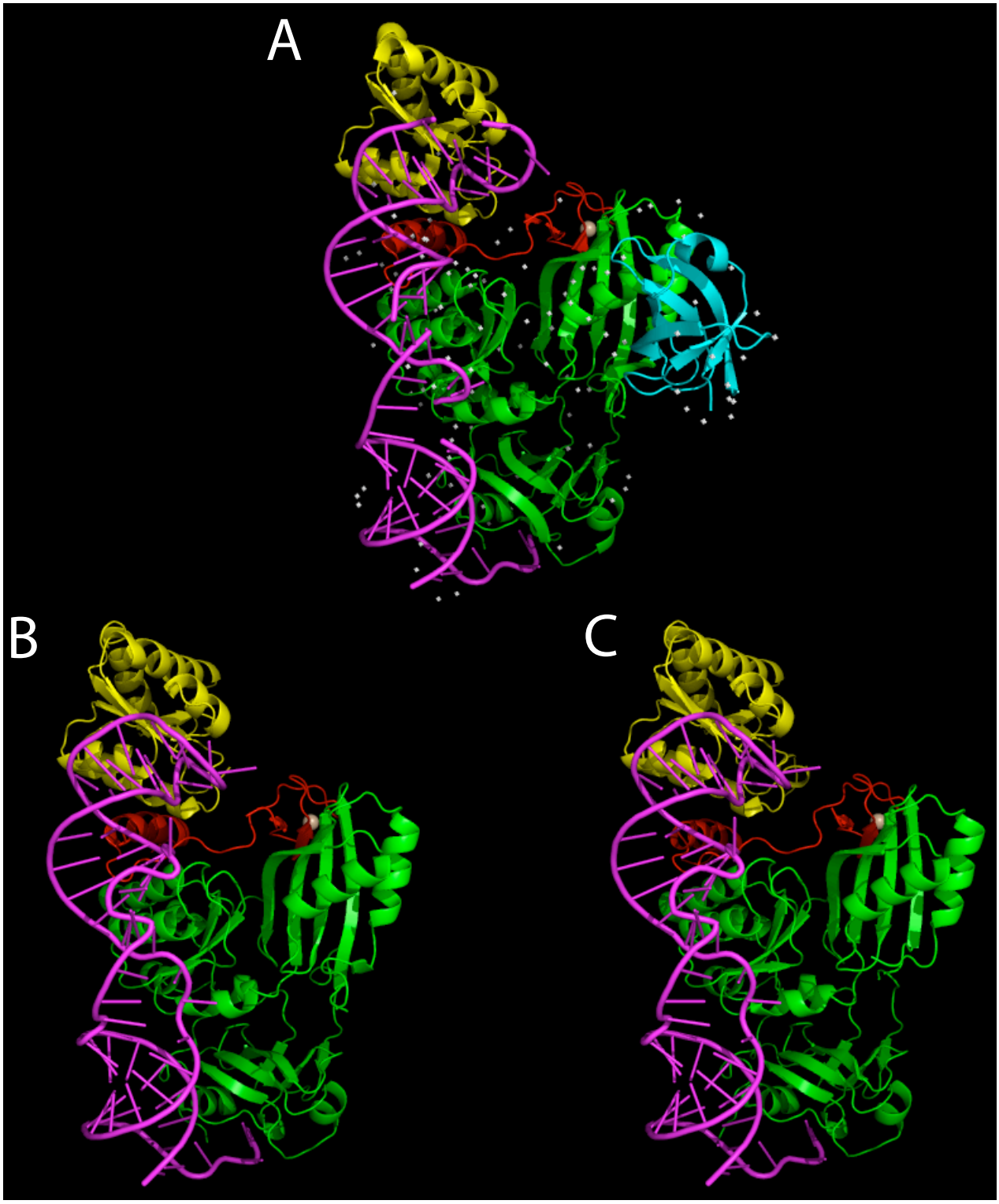


Figure 33. Modèles 3D des complexes sRNP H/ACA de *P. abyssi* et *P. furiosus* construits à partir des données structurales de Li & Ye (2006). A Structure 3D du complexe sRNP déterminée par radio-cristallographie (PDB ID : 2HVY). B Modèle 3D du complexe sRNP de *P. abyssi* avec l'ARN guide Pab21. C Modèle 3D du complexe sRNP de *P. furiosus* avec l'ARN guide Pfu1. L'ARN (magenta) et les protéines (L7ae : jaune ; Cbf5 : vert ; Nop10 : rouge ; Gar1 : cyan) sont représentés de façon schématique ("cartoon").

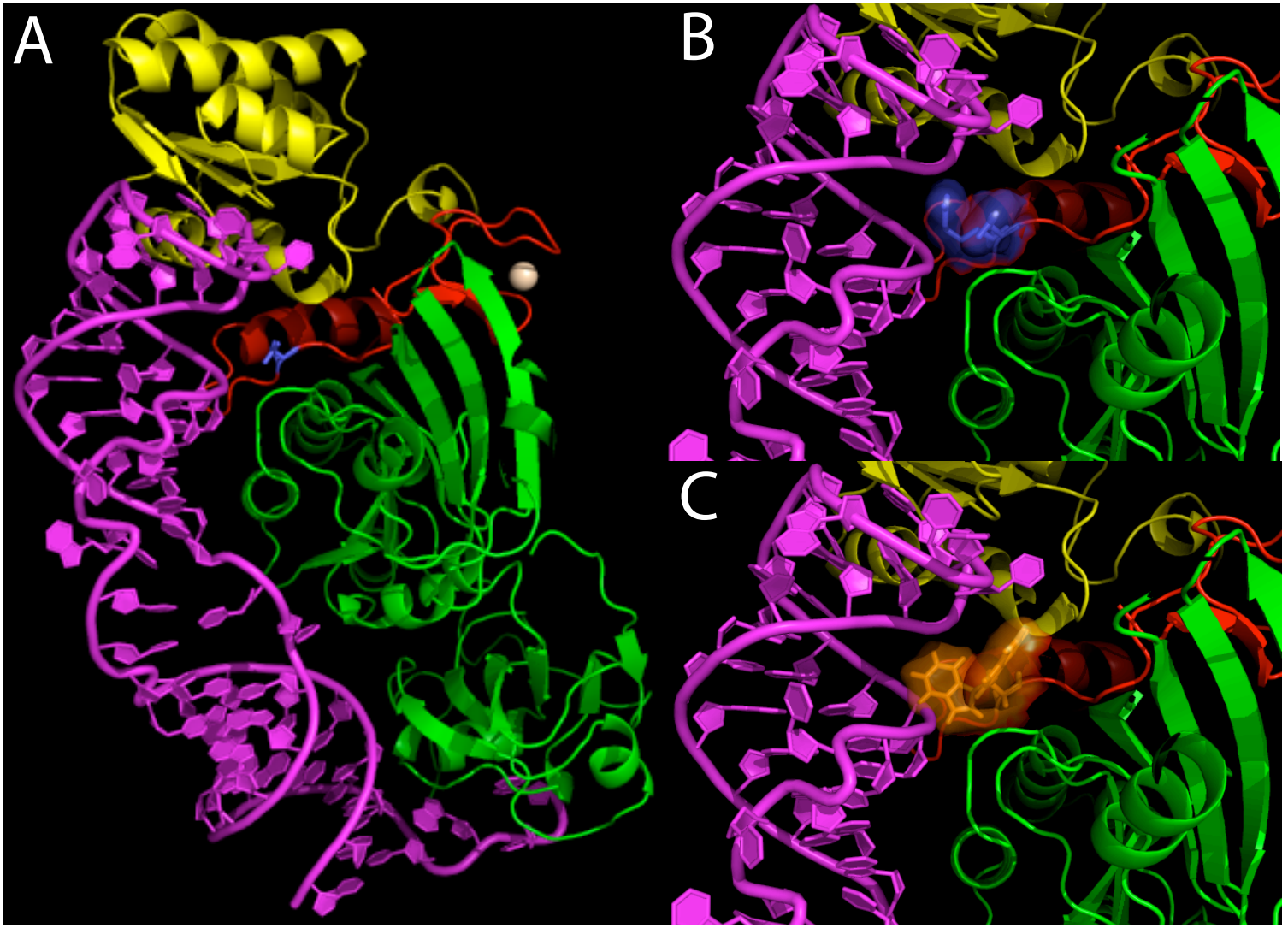


Figure 34. Modèles 3D du complexe snRNP Pab21 (*P. abyssi*) sauvage (R34) et mutant (W34). A Modèle 3D de la nRNP Pab21 sauvage. La position R34 est indiquée en bleu : elle se trouve à l'interface ARN/protéine proche du motif K-loop ou de l'hélice distale (H2) en fonction du rotamère de l'arginine (surface en bleu). B. La position W34 où l'arginine a été substituée par un tryptophane positionne également le résidu du motif K-loop ou de l'hélice distale (H2) en fonction du rotamère (surface en orange).

Chapitre 3

Approches pour la Modélisation d'Interactions avec des Cibles ou Ligands ARN

Sommaire

3.1 Résumé

Les travaux menés sur les ribozymes et l'essor de la RNomics démontrent la versatilité de fonction des ARN ; celle-ci repose sur la diversité de structure des ARN et leur capacité à interagir avec d'autres molécules (ARN, protéines, antibiotiques, etc). La biologie structurale a permis d'acquérir beaucoup de connaissances sur la structuration des ARN et leurs interactions avec d'autres molécules même si les ARN restent des objets d'étude difficiles. La conception assistée par ordinateur de ligands dirigée contre des ARN comme cibles thérapeutiques est dès lors possible. La détermination de la structure 3D de certains ARN et complexes ARN/ligand peut toutefois rester inaccessible ou les structures déterminées peuvent ne pas correspondre à une forme biologiquement active. La modélisation 3D de macromolécules offre une approche possible d'étude des liens structure/fonction en l'absence de données structurales mais aussi d'étude théorique (par simulation) de la flexibilité conformationnelle qui joue un rôle primordial dans le repliement des ARN et leur interaction avec des ligands. Les travaux présentés ici s'inscrivent dans le cadre d'un développement d'approches de modélisation de complexes ARN/protéines et ARN/ligand adaptées au mode d'interaction étudié, selon que l'ARN est essentiellement structuré

sous forme double-brin ou bien sous forme simple-brin. Dans le premier cas, une approche de docking classique est utilisée où l'ARN peut être considéré comme une cible pour un ligand protéique ou autre. Notre participation au défi CAPRI2008 a permis de démontrer la performance de notre approche de docking (HEX+MCDOCK) dans ce cas de figure. Dans le second cas, une approche de docking "par fragment" (SELEX *in silico*) inspirée des méthodes de conception de drogues est utilisée pour prédire la reconnaissance par des protéines de liaison à l'ARN et éventuellement concevoir d'autres ligands ARN. La méthode est en cours de validation et sera appliquée à des cibles d'intérêt impliquées dans les dystrophies myotoniques.

3.2 Contexte

Le développement des méthodes de docking et de conception de ligands sont développées dans le cadre du CPER 2008 MISN : "Modélisation, information, systèmes numériques" (thème MBI : "Modélisation des biomolécules et de leurs interactions") en collaboration avec Bernard Maigret de l'équipe ORPAILLEUR du LORIA et avec Dave Ritchie (Université d'Aberdeen, Ecosse) qui rejoindra prochainement le LORIA pour occuper une chaire d'excellence (projet ANR) et renforcer l'axe de recherche. L'approche SELEX *in silico* a été développée avec Manuel Simoes (posdoctorant de 2006 à 2008 au MAEM), dans le cadre du programme Décrypthon financée conjointement par l'AFM, IBM-France et les départements STIC et SDV du CNRS (projet : "Développement et utilisation d'approches informatiques et théoriques pour l'analyse des liens existant entre défauts d'épissage et maladies génétiques"). Le projet s'inscrivait dans une thématique phare du laboratoire impliquant doctorant et ingénieurs pour la partie expérimentale (Audrey Vautrin : doctorante, Nathalie Marmier-Gourrier : ingénieur). Le prolongement du projet est focalisé sur la problématique des dystrophies myotoniques et soutenu par un financement AFM (Denis Furling, Institut de Myologie, Paris) avec la participation d'autres chercheurs du MAEM (Isabelle Behm-Ansmant et Xavier Manival). Le projet bénéficie également du soutien du CINES pour l'accès à des ressources de calcul.

3.3 Introduction

La banque RCSB-PDB (http ://www.rcsb.org/pdb/) contient l'ensemble des structures 3D expérimentales de macromolécules biologiques déterminées depuis 1976 (Fig. 7). En 2008, elle recensait 54 825 structures 3D dont les protéines représentent 92,3% et les acides nucléiques 3,4% (1163 structures d'ADN et 685 d'ARN). Les complexes acide nucléique/protéine (2225 structures) représentent une part équivalente à celle des acides nucléiques non complexés (1848 structures). Le nombre de structures 3D de complexes ARN/protéine reste donc encore limitée malgré le nombre important de structures issues de la détermination récente de la structure du ribosome [98]. Près de la moitié des structures de complexes correspondent à de complexes ribonucléoprotéiques de grosse taille dans lesquels on retrouve aussi, outre les protéines ribosomiques, les tRNA synthétases, les polymérases ou encore des composants du spliceosome. Pourtant, de plus en plus de structures 3D de protéines de liaison à l'ARN sont déterminées. On compte actuellement plus de 6700 structures de protéines de liaison à l'ARN dans la banque PDB pour lesquelles la structure 3D du complexe avec l'ARN n'est pas connue. D'autre part, le nombre de modes de liaison ARN/protéine apparaît élevé, des protéines de la même famille pouvant se lier de façon différente à l'ARN (protéines à domaines RRM ou protéines à doigts de zinc). La détermination de la structure 3D d'un complexe pour une famille donnée de protéines ne garantit donc pas que les autres membres de la famille protéique se lient de la même façon à l'ARN. Le

développement de méthodes de modélisation permet de proposer des modèles 3D de complexes ARN/protéines. Couplées à des techniques expérimentales d'étude d'interactions ARN/protéine, ces méthodes offrent une alternative dans l'étude des relations structure/fonction des ARN. De plus, elles offrent aussi la possibilité de simuler les propriétés dynamiques de ces complexes et l'influence de la solvatation ou des contre-ions sur leur stabilité.

Depuis quelques années, la communauté scientifique propose, à des équipes du domaine, de prédire en aveugle la structure 3D de cibles choisies correspondant à des complexes entre macromolécules (dont les coordonnées sont gardées secrètes) ; il s'agit des compétitions CAPRI (http ://www.ebi.ac.uk/msd-srv/capri/) qui confrontent des équipes développant des méthodes dites de "docking" pour la modélisation et la prédiction de structures 3D de complexes entre macromolécules. Les méthodes de "docking" disposent normalement d'un outil de recherche conformationnelle et d'une fonction de score afin d'explorer l'espace des configurations possibles en terme de modes d'interaction possibles et d'évaluer quantitativement la force de l'interaction ou le degré de complémentarité en termes d'énergie (enthalpie ou énergie libre). L'exploration conformationnelle se fait généralement par des méthodes classiques de simulations moléculaires (dynamique moléculaire ou Monte-Carlo) qui peuvent être couplées à des méthodes d'optimisation (algorithmes génétiques, recuit simulé) ou à des banques de données (banque de rotamères, de contacts, etc) ; la fonction de score est généralement une somme de termes correspondant à différentes contributions énergétiques (enthalpiques et entropiques) définies de façon plus ou moins empiriques. Outre les performances variables que peuvent avoir telle ou telle méthode de recherche conformationnelle et de fonction de score, la performance globale d'une méthode de "docking" peut varier beaucoup en fonction de la cible car tous les complexes ne sont pas équivalents du point de vue du niveau de difficulté de prédiction. L'historique des compétitions CAPRI montre en effet que les performances peuvent varier d'une cible à une autre. Un des facteurs clé qui explique en partie ces différences de performance est lié à la flexibilité conformationnelle des partenaires dont la conformation peut changer dramatiquement entre la forme liée et la forme "libre" (non complexée) qui met fortement à l'épreuve la méthode de recherche conformationnelle et la fonction de score.

Dans le cas des complexes ARN/protéines, il est courant que l'ARN (et parfois aussi la protéine) subisse un changement conformationnel important au cours de l'association. D'autre part, les interactions électrostatiques mises en jeu dans les complexes ARN/protéine demandent beaucoup de ressources de calcul pour une évaluation précise. On peut donc distinguer les complexes ARN/protéine en fonction des niveaux de difficulté différents en terme de prédiction de la structure 3D d'après ce critère de flexibilité ; une classification naturelle est de séparer les complexes qui contiennent des ARN intrinsèquement flexibles de ceux qui contiennent des ARN *a priori* plus rigides (Fig. 35). Les ARN dont l'interface avec la protéine est non structurée, c'est-à-dire sous forme simple-brin (Fig. 35A-C) , seront a priori les plus difficiles à prédire car le nombre de degrés de liberté est très élevé (conformation de la base nucléique, du sucre et du squelette phosphodiester). En revanche, les ARN très structurés et dont l'interface avec la protéine est sous forme double-brin (Fig. 35D) comportent beaucoup moins de degrés de liberté et leur structure 3D liée est donc *a priori* plus facile à prédire. En raison des difficultés à évaluer la flexibilité conformationnelle des ARN et les interactions électrostatiques, peu de méthodes ont été développées pour traiter spécifiquement le "docking" ARN/protéine. Toutefois, l'intérêt grandissant pour les ARN qui se trouvent impliquées dans de nombreuses fonctions biologiques (chapitre précédent) et qui constituent des cibles thérapeutiques attractives est une nouvelle motivation pour le développement de méthodes de "docking" performantes pour la prédiction de la structure 3D de complexes ribonucléoprotéiques. Pour la première fois, l'édition CAPRI2008 proposait une cible correspondant à un complexe ARN/protéine.

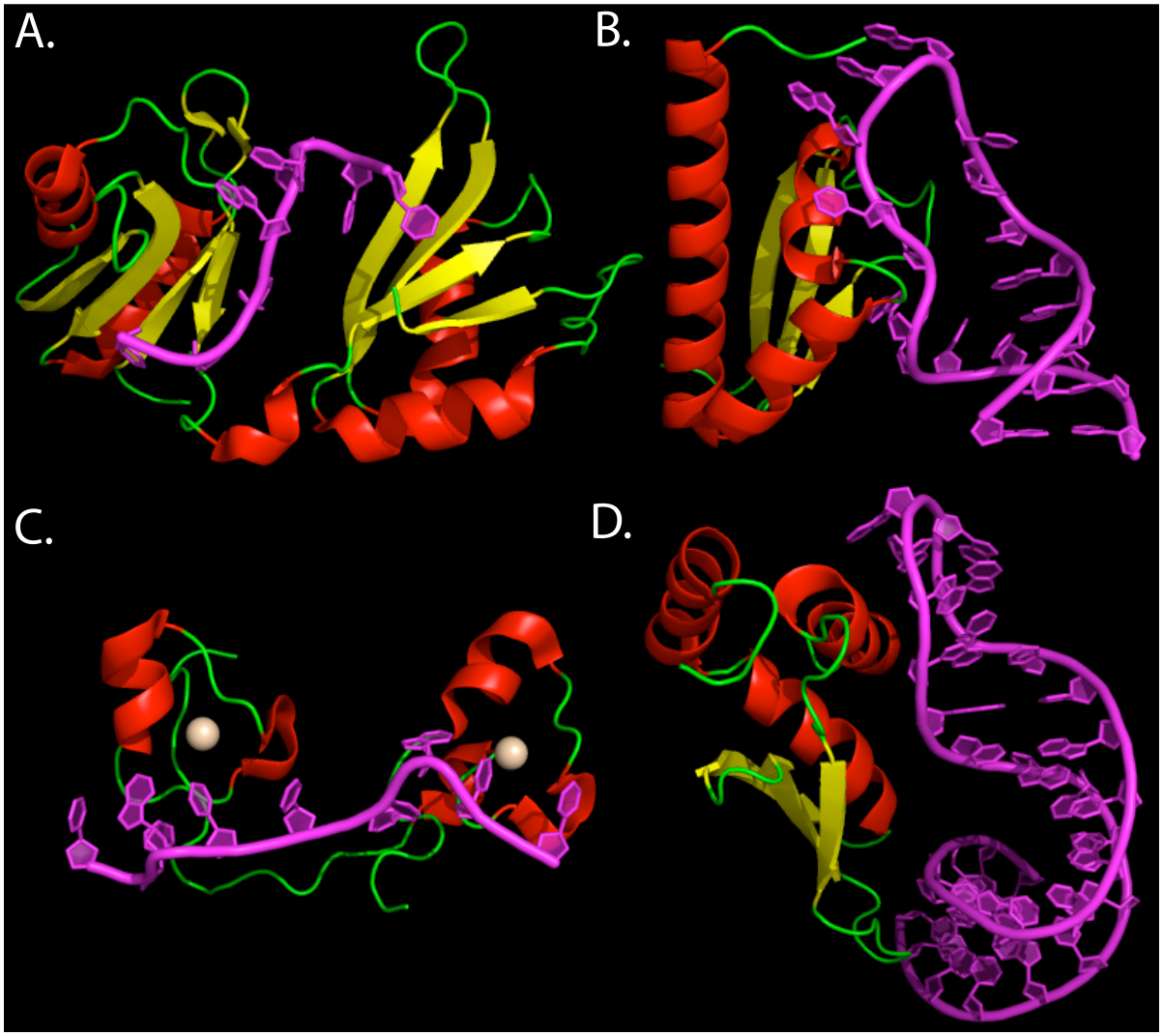


Figure 35. Exemples représentatifs de complexes ARN/protéine avec des domaines de liaison à de l'ARN simple-brin (A, B, C) ou double-brin (D). A Complexe formé entre la protéine HuD (stabilisation des extrémités 3'UTR des messagers) à domaines RRM (RRM1 et RRM2) et un ARN simple-brin (riche en U et A ; PDB ID : 1FXL). B Complexe formé entre la protéine Nova-2 (régulation du métabolisme des neurones) à domaines KH (KH3) et un ARN partiellement structuré lié à la protéine par sa partie simple-brin (boucle terminale ; PDB ID : 1EC6). C Complexe formé entre la protéine TIS11d (dégradation des messagers par leur l'extrémité 3'UTR) à doigts de zinc (Zn1 et Zn2) et un ARN simple-brin (riche en U et A ; PDB ID : 1RGO). D Complexe formé entre la protéine Rnt1 à domaines RBD (RNase III de levure) et un ARN double-brin (précurseur du snR47 ; PDB ID : 1T4L).

La maturation des précurseurs des ARNm (hnRNA) est un processus biologique complexe qui fait appel à une machinerie, le spliceosome, constituée de différents composants ribonucléo-protéiques (snRNP). Ce processus fait intervenir de nombreuses interactions ARN/protéine à la fois dans la biogénèse des snRNP et dans la reconnaissance des signaux d'épissage (présents dans les séquences introniques et aux jonctions intron/exon) qui permet l'élimination de séquences introniques au sein des transcrits (hnRNA), ainsi que la ligation des exons par le spliceosome. Par ailleurs, d'autres interactions ARN/protéine interviennent aussi dans la reconnaissance de signaux d'épissage non constitutifs reconnus par des facteurs protéiques de régulation de l'épissage. Ces interactions jouent un rôle majeur dans la régulation de l'épissage alternatif qui permet de générer une grande diversité de transcrits matures (ARNm) à partir du même gène ; on estime que 60% des gènes humains produisent au moins deux ARNm différents par ce processus qui joue

un rôle majeur notamment dans le développement et la différenciation cellulaire.

Les mutations dans les signaux d'épissage constitutifs ou alternatifs peuvent conduire à des maladies génétiques. Il est actuellement estimé qu'environ 30% des maladies d'origine génétique sont liées à la génération de défauts d'épissage (proportion sans doute sous-estimée). Ces défauts ont été répertoriés en différentes classes selon que la mutation, qui peut n'être qu'une mutation ponctuelle, altère les signaux d'épissage (constitutifs ou alternatifs), les composants du spliceosome (snRNP) ou des facteurs protéiques de régulation de l'épissage [99, 100, 101, 102]. Une classe d'altération de l'épissage alternatif échappe pourtant à cette classification et concerne des mutations de plus grande ampleur qui induisent un "gain de fonction" associé à l'amplification de séquences répétées du génome humain [103, 104, 105]. Ce "gain de fonction", qui correspond à un effet toxique dû à la présence des séquences répétées dans les transcrits, intervient dans environ une quinzaine de maladies neurologiques humaines [106, 107]. Dans le cas des amplifications de triplets CTG et de quadruplets CCTG, le "gain de fonction" se traduit par le développement plus ou moins tôt dans la vie de dystrophies myotoniques (DM) de type 1 (répétitions CTG) et de type 2 (répétitions CCTG) [108]. Ces séquences répétées, présentes au niveau des transcrits ARN (répétitions CUG et CCUG), ont un "effet toxique" en modifiant la disponibilité respective de facteurs impliqués dans la régulation de l'épissage alternatif : les protéines des familles Muscleblind (dont MBNL1) et CELF (dont CUG-BP1). Le "gain de fonction" a pour effet de favoriser un épissage alternatif de plusieurs transcrits (codant un canal chlore, le récepteur de l'insuline, etc) sous une forme anormale plutôt que sous leur forme adulte attendue et conduit au développement des symptômes de la maladie. Du point de vue moléculaire, les répétitions CUG et CCUG forment de longues structures ARN tige-boucles [109, 110, 106] à l'origine de la toxicité ARN associé aux DM. MBNL1 serait séquestrée de façon spécifique sur les longues répétitions $(CUG)_n$ ou $(CCUG)_n$ (n $\geq$ 37 pour la DM1 et n $\geq$ 75 pour la DM2) présentes dans des ARNm de patients atteints de DM1 ou de DM2 [111]. La séquestration de MBNL1 romprait l'équilibre avec un autre facteur de régulation de l'épissage antagoniste, CUG-BP1, ce qui conduirait aux dérégulations de l'épissage observées pour plusieurs gènes chez les patients atteints de DM1 ou de DM2 (Fig. 36) [112, 113].

L'émergence des ARN comme cibles thérapeutiques potentielles pour de petites molécules offre de nouvelles opportunités de valoriser les données génomiques [114, 115, 116, 117]. En dehors des cibles habituelles des antibiotiques que représentent les ARN ribosomiques bactériens, l'acquisition de nouvelles données structurales sur les ARN ouvre la voie à la conception de ligands dirigés contre des cibles ARN par les méthodes traditionnelles de conception basées sur l'utilisation de la structure 3D. Ces méthodes ont été largement utilisées depuis une dizaine d'années contre des cibles protéiques et font désormais parti intégrante des outils utilisés en conception de drogues. Les ARN représentent donc des cibles de choix pour la conception d'antiviraux par exemple mais aussi de façon plus générale pour interférer avec la fonction associée à la reconnaissance de la cible ARN par une ou plusieurs protéines, un autre ARN, ou d'autres molécules biologiques. La détermination récente de la structure 3D d'un petit duplex d'ARN qui mime les longues répétitions CUG associées à la DM1 [118] ouvre des perspectives pour la conception des ligands contre une cible ARN à l'aide d'approches théoriques. Par ailleurs, les protéines de la famille CELF et les complexes formés par MBNL1 et les séquences répétées représentent aussi des cibles thérapeutiques potentielles pour le développement de traitements de la DM1 et la DM2. Acquérir des connaissances sur le mode d'interaction des protéines CUG-BP1 et MBNL1 avec leurs cibles normales dans les ARN pré-messagers d'une part (contexte non-pathologique) ou avec leurs cibles "toxiques" (contexte pathologique) est la première étape pour pouvoir envisager une stratégie thérapeutique rationnelle. L'une de ces stratégies serait, par exemple, de pouvoir bloquer l'interaction de MBNL1 avec les séquences répétées en altérant le moins possible son

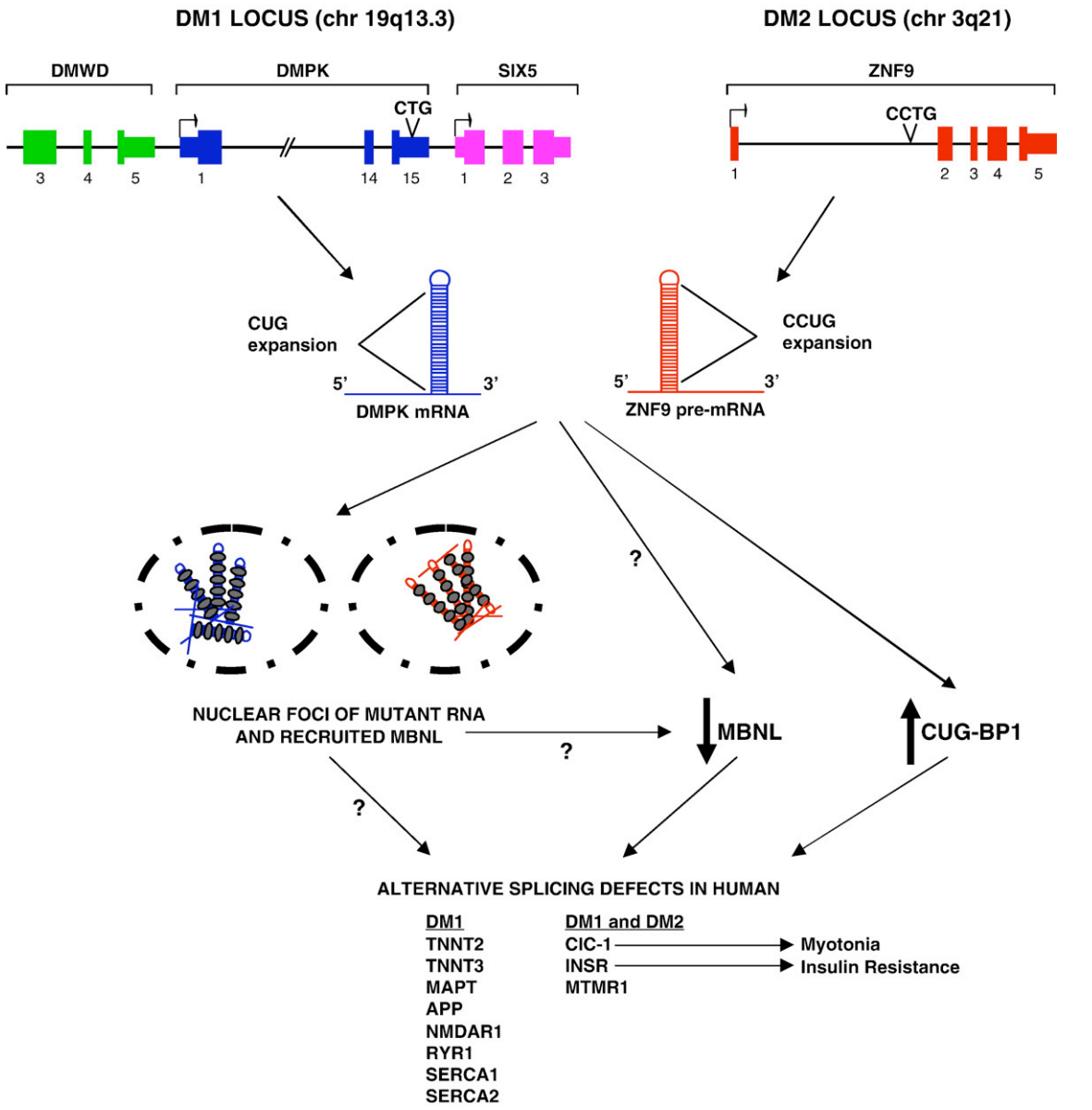


Figure 36. Modèle "gain de fonction" et toxicité des ARN à séquence répétées (CUG et CCUG) dans les dystrophies myotoniques. Les dystrophies myotoniques sont associées avec des expansions de répétitions CTG dans la région 3'UTR du gène DMPK ou CCTG dans l'intron 1 du gène codant la protéine ZNF9. Les transcrits ARN forment des longues structures double-brin auxquelles se lie MBNL1. Les protéines MBNL co-localisent dans les inclusions ribonucléaires formées par les mutants ARN. Il n'est pas encore clair si la séquestration de MBNL1 conduit à une baisse de son activité normale ou bien si les expansions CUG et CCUG redirigent MBNL vers un chemin alternatif qui inhibe sa fonction. Le niveau de CUG-BP1 est accru dans les cellules DM1, indépendamment de la régulation par MBNL. Un accroissement de l'activité de CUG-BP1 ou la perte de fonction de MBNL1 peuvent conduire à des événements aberrants d'épissage associés avec les DM1 et DM2. Il a été montré que la myotonie et la résistance à l'insuline peut être attribué à des défauts d'épissage des transcrits du canal chlore 1 (CIC-1) et du récepteur de l'insuline dans les DM1 et DM2. D'autres altérations de l'épissage dans les tissus DM impliquent notamment : la troponine T cardiaque (TNNT2), la troponine T du muscle squelettique (TNNT3), la protéine Tau associée aux microtubules (MAPT), le précurseur de la protéine amyloïde beta (APP), le récepteur de la N-méthyl-D-aspartate (NMDAR1), le récepteur 1 de la ryanodine (RyR1), l'ATPase calcium-dépendant du réticulum sarco et endoplasmique (SERCA), la protéine 1 myotubularine-dépendante (MTMR1). D'après Cho & Tapscott (2007) [113].

interaction avec ses cibles naturelles (cible ARN). Une autre possibilité pourrait être de limiter l'interaction de CUG-BP1 avec ses cibles naturelles dans les cas où MBNL1 est séquestrée (cible protéique).

Du point de vue méthodologique, les approches "par fragment" ("fragment-based") sont celles qui ont remporté les succès les plus retentissants car elles offrent de fortes potentialités pour la conception de ligands forts à partir de petits groupements chimiques qui n'ont qu'une affinité relativement faible (de l'ordre du μM ou mM) pris séparément. Le concept de ces approches est fondé sur l'idée qu'une cible thérapeutique, en général une macromolécule biologique (protéine, ADN, ARN) présente des cavités qui constituent autant de sites potentiels de fixation pour des groupes chimiques de caractéristiques différentes en termes de complémentarité de forme, d'hydrophobicité, de charge, etc. La cible peut ainsi être considérée comme une succession de cavités avec des propriétés spécifiques ; la conception de drogues est ainsi assimilée à la recherche de groupes chimiques qui présentent la meilleure complémentarité possible dans une cavité donnée. Une drogue potentielle ou ligand peut alors être conçu en reliant entre eux plusieurs de ces groupes chimiques présentant une affinité de liaison minimale dans leur cavité respective et situés à proximité les uns des autres. Bien que la dernière étape de cette approche qui est de relier entre eux des groupes pouvant avoir des propriétés chimiques très différentes soit la plus délicate (car elle pose des problèmes parfois insolubles du point de vue de la synthèse organique des ligands), elle offre aussi les meilleures perspectives d'optimisation en termes d'affinité de liaison en permettant d'augmenter l'affinité globale du ligand ainsi généré de plusieurs ordres de grandeur par rapport à l'affinité des groupes chimiques pris de façon isolée (Fig. 37). La mise en œuvre de ces approches a été réalisée à la fois du point de vue expérimental et théorique. Le concept et son application ont été magnifiquement illustrés par le travail réalisé au sein des Laboratoires Pharmaceutiques Abbott sur la conception de ligands dirigés contre la protéine de liaison de l'immunosuppresseur FK506 (FKBP) [119]. Dans ce cas, il s'agit d'une approche expérimentale où 2 groupes chimiques différents ont été identifié et localisés à proximité l'un de l'autre par RMN. En dernier lieu, le ligand “fusionné” a été synthétisé et son activité biologique confirmée. Une approche par fragment *in silico* a également permis de proposer des inhibiteurs peptidiques dirigés contre la protéine Ras afin d'empêcher la formation du complexe Ras-Raf-GTP qui joue un rôle dans le processus d'oncogénèse [120]. Une validation expérimentale *in vitro* a démontré la pertinence des inhibiteurs proposés par cette approche [121].

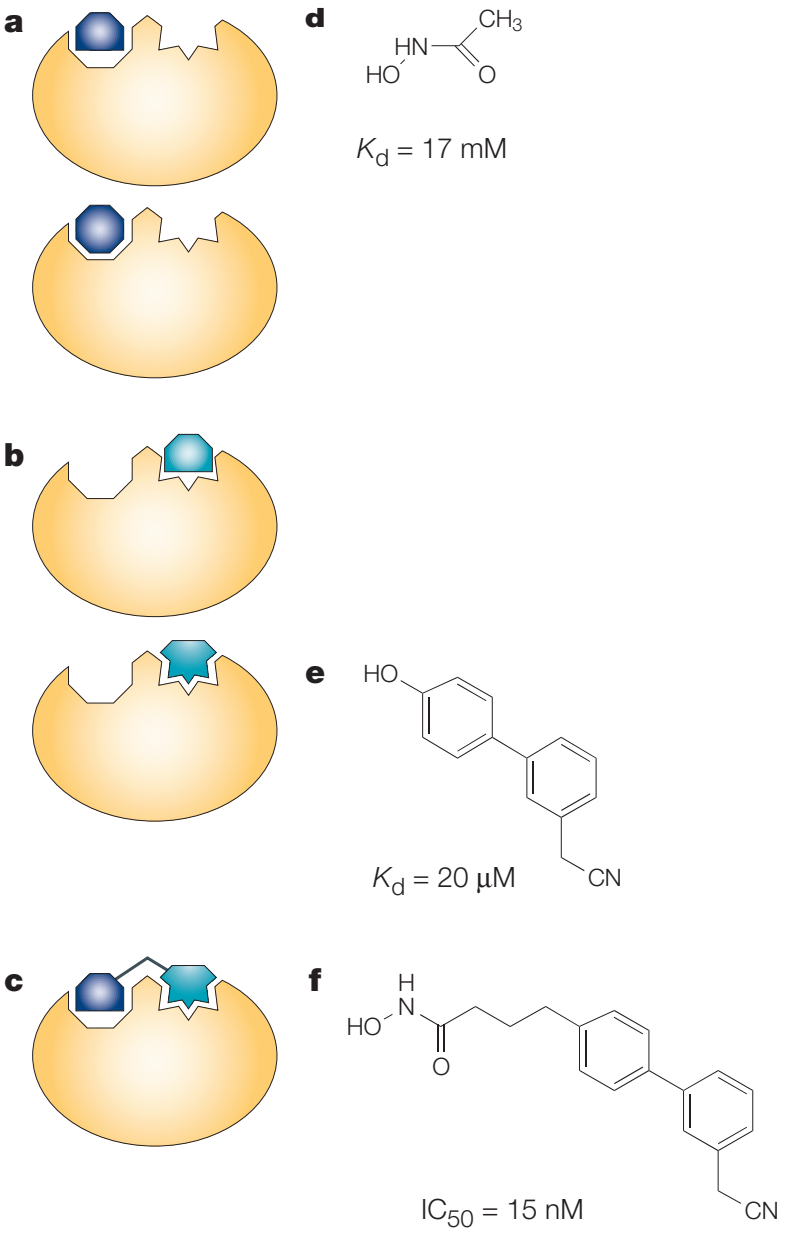


Figure 37. Exemple de conception de ligands en utilisant une approche "par fragment". Les 2 cavités à la surface de la cible peuvent accueillir des groupes chimiques complémentaires à des degrés divers (forme, électrostatique, hydrophobicité, etc) aux sites en (a) et (b). La connexion par une liaison chimique entre 2 groupes chimiques aux sites (a) et (b) conduit à un ligand (c) ayant une affinité globale (f) supérieure de plusieurs ordres de grandeur à celles des groupes chimiques pris séparément en (d) et (e). Les structures chimiques correspondent à un inhibiteur de la liaison de le protéine E2 du papillomavirus humain (cible) à l'ADN (Hajduk et al., 1997) [122]. D'après Pellecchia *et al.* (2002) [123].

3.4 Approches par modélisation pour la compréhension des bases moléculaires des DM

Une meilleure compréhension des maladies dues à des défauts d'épissage implique une étude structurale des interactions ARN/ARN et ARN/protéines. Ainsi, dans le cas des mutations correspondant à des amplifications de triplets de nucléotides et qui conduisent aux dystrophies myotoniques, l'influence des séquences répétées CUG et CCUG sur les dérégulations de l'épissage est encore mal comprise. Initialement, la séquestration de différentes protéines de la famille CELF sur les séquences répétées avait été associée à l'origine des anomalies observées, mais les bases moléculaires de la reconnaissance entre ces protéines et leur(s) cible(s) ARN sont encore inconnues. Des données récentes ont montré que c'est la protéine MBNL1 qui est séquestrée sur les séquences répétées CUG et non CUG-BP1 dans un contexte pathologique [124, 125] ; CUG-BP1 intervient elle comme antagoniste de MBNL1 mais dans un contexte non pathologique. Initialement, la protéine CUG-BP1 avait été baptisée ainsi pour son affinité pour des séquences contenant des triplets CUG ; toutefois, elle semble ne se lier qu'à de courtes séquences répétées (CUG)n (avec n $\leq$ 8) ne présentant pas de structuration en double-brin contrairement aux longues répétitions CUG. De plus, il a été montré ultérieurement de façon expérimentale au sein du MAEM que CUG-BP1 n'avait pas d'affinité particulière in vitro pour de longues répétitions $(CUG)_n$ (avec n $\geq$ 16). L'absence de données structurales sur la protéine MBNL1 ou sur une protéine homologue qui aurait pu être utilisée pour la modélisation de sa structure 3D ne permettait pas d'envisager la modélisation des complexes entre MBNL1 et les ARN double-brin à séquences répétées (CUG)n. Par contre, l'existence de données structurales abondantes sur les protéines à domaines RRM (famille à laquelle appartient CUG-BP1) et leur complexe avec des ARN simple-brin non structurés, nous a conduit à développer une nouvelle approche. En effet, l'absence de structuration des ARN simple-brin liés aux protéines à domaines RRM ne permet pas d'utiliser d'emblée une approche de "docking" classique.

Dans une approche de "docking" classique, on dispose d'une structure initiale définie pour chacun des 2 partenaires à partir desquelles la procédure de "docking" est amorcée, même si la conformation de l'un ou l'autre des partenaires est susceptible de changer de façon conséquente au cours de la simulation. Or, la grande flexibilité des ARN simple-brin ne permet pas de modéliser sa structure séparée de celle de la protéine puisque l'ARN se structure essentiellement au contact de la protéine ; nous avons donc développé une nouvelle approche par modélisation baptisée "SELEX *in silico*" (inspirée de la méthode expérimentale SELEX) conçue et adaptée pour traiter le cas de complexes avec un partenaire ARN très flexible. L'approche "SELEX *in silico*" a été développée dans l'optique de pouvoir prédire des interactions ARN/protéine et concevoir des ligands ARN ayant une affinité optimale pour une protéine cible donnée. En effet, la méthode SELEX expérimentale a déjà montré qu'il est possible de générer des ligands ARN ayant une meilleure affinité que le ligand ARN naturel. Il devient alors possible d'utiliser des ligands ARN générés artificiellement pouvant interférer avec le ligand naturel, une protéine par exemple. Dans le même temps et dans la continuité de travaux antérieurs, une approche de "docking" classique a été développée avec l'objectif de pouvoir modéliser les interactions entre un ARN de type (CUG)n (pour lequel nous disposons de données structurales partielles [118]) et la protéine MBNL1 dont nous espérons obtenir une structure 3D complète au laboratoire (Xavier Manival et Christophe Charron). Pour l'instant, on dispose de données structurales très parcellaires sur les seuls domaines en doigts à zinc de MBNL1 [126].

Dans le cas des DM, les protéines CUG-BP1 et MBNL1 agissent comme antagonistes dans la régulation d'un certain nombre de gènes ; interférer avec l'une ou l'autre des 2 protéines dans son

interaction avec son ligand naturel peut représenter une stratégie thérapeutique. Par exemple, lorsque MBNL1 est séquestrée sur les longues répétitions ARN $(CUG)_n$ dans un contexte pathologique, CUG-BP1 se trouve en excès par rapport à MBNL1 et bloquer sa fonction par un ligand ARN peut atténuer la dérégulation de l'épissage lié au déséquilibre entre ces 2 facteurs de régulation de l'épissage. Interférer avec MBNL1 par un ligand ARN empêchant qu'elle soit séquestrée sur les répétitions ARN $(CUG)_n$ peut atténuer les dérégulations de l'épissage dans les DM. Nous nous sommes donc focalisé sur les interactions entre ARN simple-brin et protéines à domaines RRM dans un premier temps en vue de modéliser les interactions entre CUG-BP1 et ses ligands ARN naturels. Afin de valider la méthode SELEX *in silico*, un système test a été utilisé pour reproduire le complexe formé entre une protéine à domaines RRM, la protéine HuD, et un de ses ligands ARN riches en U et A (Wang et al., 2001 ; PDB ID : 1FXL). La détermination récente de la structure 3D des domaines en doigts à zinc de MBNL1 nous a également conduit à tester notre méthode sur un autre complexe de structure 3D connue formée entre la protéine à doigts à zinc TIS11d et son ligand ARN, l'extrémité 3'UTR d'un ARN messager (PDB ID : 1RGO [127]). De même, la détermination récente de la structure 3D d'un court duplex $r(CUG)_6$:$r(CUG)_6$ [118] nous permet d'envisager la modélisation des structures tige-boucle des longues répétitions $(CUG)_n$ et la recherche de ligands potentiels contre cette cible ARN.

3.5 SELEX *in silico* : Modélisation et Conception de Ligands ARN

Une nouvelle méthode appelée SELEX *in silico* a été développée pour modéliser des complexes protéine/ARN où l'ARN est non structuré et essentiellement sous forme simple-brin. La méthode repose sur une approche dite "par fragment" en faisant l'hypothèse que l'ARN simple-brin peut être considéré comme un ligand que l'on peut naturellement décomposé en groupes individualisés que sont les nucléotides. Plusieurs méthodes théoriques ont été développées dans la philosophie d'une approche "par fragment" (MCSS [128] ; Ludi [129] ; HOOK [130] ; DLD [131] ; etc) mais aucune n'apporte réellement de solution globale et très intégrée entre les étapes de recherche de groupes chimiques à la surface de la cible (MCSS), de construction de ligands à partir des groupes chimiques (HOOK, Ludi) et d'optimisation des liens chimiques servant de connecteurs entre les groupes chimiques (DLD).

La démarche "par fragment" pour modéliser la structure des complexes protéine/ARN consiste donc à : 1. rechercher des sites d'interaction favorables de nucléotides à la surface de la protéine cible (programme MCSS), 2. reconstruire des chaînes d'ARN à partir de nucléotides proches les uns des autres et susceptibles de former un lien phosphodiester (programme BUILD), 3. optimiser la structure globale de la chaîne d'ARN ainsi construite en interaction avec la protéine cible (programme MCDOCK), (Fig. 38). A la différence de la plupart des approches "par fragment", l'approche SELEX *in silico* développée dans ce sens intègre l'ensemble des étapes qui vont de la recherche de sites de fixation de nucléotides, à la construction de chaînes d'ARN et à leur optimisation liées à la protéine cible. Elle a été nommée par analogie avec la technique SELEX expérimentale car elle doit permettre soit de faire de la "prédiction" de structure 3D d'ARN simple brin liée à leur protéine cible en imposant la séquence, soit de faire de la "conception" d'ARN simple brin sans contraintes de séquence (Fig. 38). Le choix de faire simplement de la prédiction est particulièrement pertinent dans le cadre du développement et de la validation de la méthode. Il s'agit en effet de tester rapidement si la méthode est en mesure de reproduire la structure 3D de complexes ARN/protéine connus ; ce qui est beaucoup moins coûteux en termes de recherche et temps de calcul que de sélectionner parmi un espace de 4n séquences, la ou les

plus favorables en terme d'affinité théorique calculée pour sa liaison avec la protéine. D'ailleurs, il a été démontré par la technique SELEX que les partenaires générés ainsi artificiellement peuvent avoir une affinité améliorée de façon significative par rapport au partenaire naturel. Par conséquent, une approche "design" ne garantit pas non plus forcément que l'on retrouve, pour une cible donnée, la séquence ARN correspondant au ligand naturel tel qu'il est lié à la protéine. Dans un souci premier de validation, nous avons donc testé la capacité de la méthode à reproduire les 2 complexes ARN/protéine utilisés comme systèmes tests faisant intervenir la protéine HuD, protéine à domaines RRM comme CUG-BP1, et la protéine TIS11d, protéine à doigts de zinc comme MBNL1. Les séquences des ARN reconnus par chacune de ces 2 protéines étant uniquement composés de 2 nucléotides : A ou U, les calculs MCSS s'en trouvaient également simplifiés.

Les résultats obtenus sur la protéine HuD (Fig. 39) comme cible ont révélé un des écueils majeurs de l'approche SELEX *in silico*, qui est intrinsèque aux approches par fragment. Les sites d'interaction identifiés pour les nucléotides A ou U en utilisant des filtres standards (valeurs seuil d'énergie pour considérer qu'une interaction avec un nucléotide à un site précis est "optimale") sont des sites optimaux alors que la position des nucléotides dans le ligand ARN naturel ne correspond pas toujours à un site optimal. En effet, en raison des contraintes géométriques imposées par la liaison des nucléotides successifs dans la chaîne, la liaison de l'ARN à sa protéine cible apparaît comme une succession de sites où peuvent alterner sites optimaux et sites sous-optimaux (Fig. 39A). C'est ce que révèle une analyse par décomposition de la contribution de chaque nucléotide, dans la chaîne ARN liée à HuD, à l'énergie globale d'interaction ARN/protéine (Fig. 39B). La méthode MCSS, utilisée dans la première étape de l'approche SELEX *in silico*, se révèle donc très performante pour identifier les sites d'interaction optimaux (Fig. 39C). Malheureusement, dans l'optique de l'approche "prédiction" (où l'on cherche à reproduire la structure 3D d'un complexe avec une chaîne d'ARN de séquence donnée) la construction de chaînes ARN à partir de sites nucléotidiques isolés (dans la 2ème étape) demande de connecter des sites optimaux à des sites sous-optimaux. Un cas typique de site sous-optimal dans le complexe HuD est celui occupé par le résidu U6 qui interagit en fait très peu avec la protéine : intercalé entre les résidus U5 et A7, il est essentiellement stabilisé par une interaction ARN/ARN et non ARN/protéine par son empilement sur A7 (Fig. 39A).

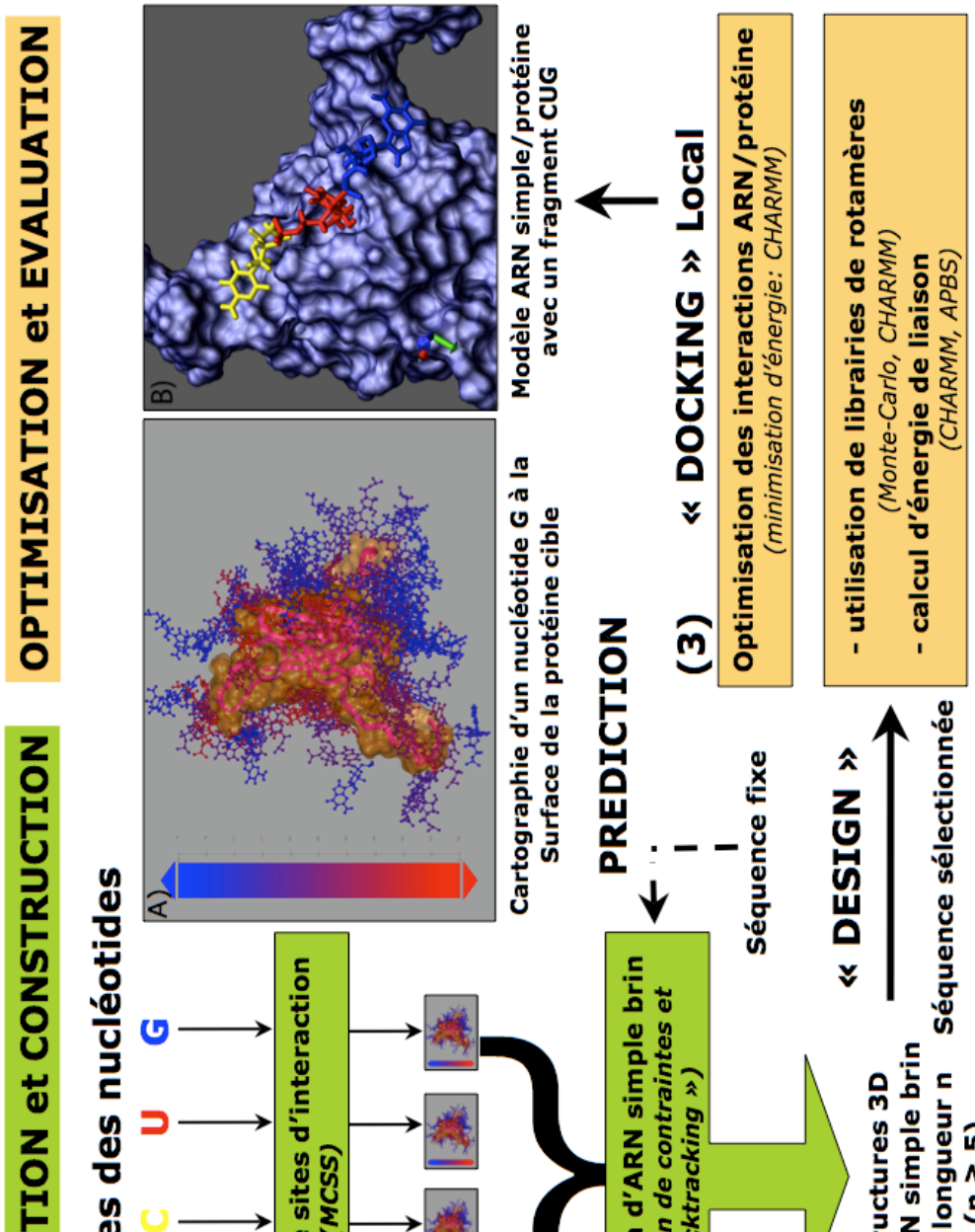


Figure 38. Approche SELEX *silico* in silico pour la modélisation et la conception de chaînes d'ARN simple brin comme ligands de protéines. L'approche SELEX in silico en 3 étapes consiste d'abord : 1) à rechercher, à la surface de la protéine, des sites d'interaction privilégiés pour des nucléotides, ensuite : 2) à reconstruire une chaîne complète d'ARN à partir de ces sites, enfin : 3) à optimiser la chaîne d'ARN dans son entier en interaction avec la protéine. Les étapes 1 et 2 sont focalisées sur l'identification de sites de liaison nucléotidiques et la construction de chaînes d'ARN. L'étape 3 est l'optimisation de l'interface ARN/protéine et le calcul d'une estimation de l'énergie de liaison. L'interface entre les différentes méthodes est développée en langage Python (version 2.4 et ultérieure).

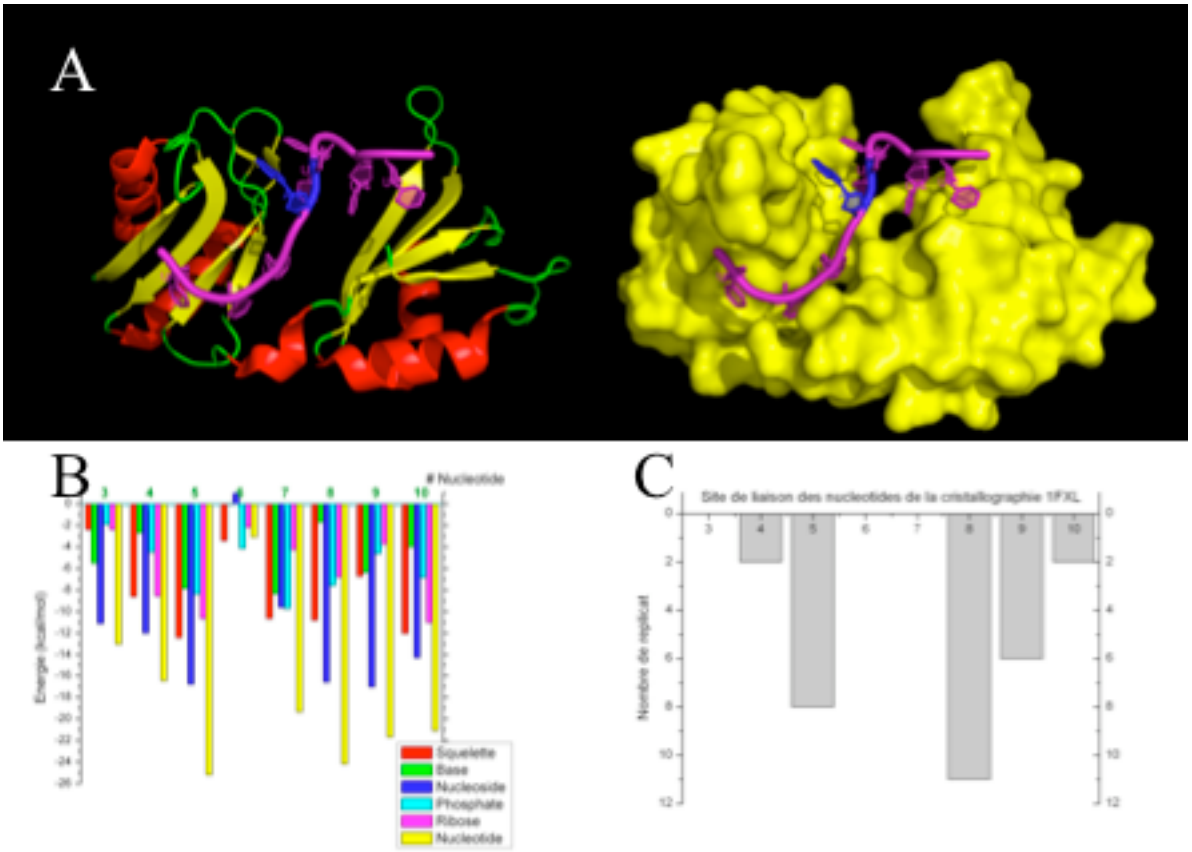


Figure 39. Relation entre la force de liaison des résidus de l'ARN et la précision de prédiction MCSS. A. La protéine cible utilisée est la protéine HuD à domaines RRM (RRM1 + RRM2) de topologie $\alpha\beta$ qui reconnaît des ARN simple-brin riches en U et A. L'ARN cristallisé avec HuD a pour séquence 5'-3UUUUAUUU10-3' (le 1er résidu est numéroté 3, le résidu terminal 10). B. L'énergie de liaison (contribution enthalpique) de l'ARN à HuD a été décomposée par résidu et par groupes chimiques à chaque position dans la séquence de l'ARN (3 à 10). C. L'histogramme indique le nombre de prédictions MCSS des résidus d'ARN (U et A) en accord avec la structure 3D expérimentale (RMSD $\leq$ 2.0Å). Les sites le plus peulplés sont considérés comme optimaux (4, 5, 8, 9 et 10), les autres comme sous-optimaux.

Plusieurs stratégies ont été envisagées et testées sur le complexe HuD pour remédier au problème de la sous-représentation de sites sous-optimaux dans l'approche "prédiction". La première stratégie consiste à abaisser les seuils d'énergie utilisés pour identifier les sites favorables de telle sorte à ce que MCSS donne des positions de nucléotides à des sites optimaux mais aussi à des sites sous-optimaux (stratégie "force brute", Fig. 40). En contrepartie, le nombre de sites à considérer peut devenir très important et les ressources en temps de calcul et mémoire deviennent alors limitantes dans l'étape de construction de chaînes ARN. La deuxième stratégie est d'effectuer une "recherche locale" et non plus une recherche globale sur la surface entière de la protéine cible : la probabilité de trouver des résidus dans une conformation adéquate est accrue. La recherche locale est alors ciblée dans des régions de la protéine à proximité de sites optimaux afin d'accroître la possibilité de construire des chaînes ARN incluant des sites sous-optimaux. La troisième stratégie dite "des groupes étendus", déjà utilisée dans un travail antérieur [132], consiste à utiliser des groupes chimiques plus volumineux susceptibles de se loger à deux sites proximaux différents : l'un optimal et l'autre sous-optimal. En l'occurrence, il s'agit de rechercher des sites d'interaction pour des dinucléotides au lieu de sites pour des mono-nucléotides. Dans le cas des résidus U6 et A7, l'utilisation de dinucléotides UA permet en effet d'identifier un résidu A au site optimal

de A7 et un résidu U au site sous-optimal de U6 (Fig. 41). Ces trois stratégies ont contribué à améliorer les résultats obtenus sur la prédiction du mode de liaison du ligand ARN à la protéine HuD.

La mise en œuvre des trois stratégies décrites a nécessité la modification du programme MCSS original, utilisé initialement dans sa version 2.5 académique. Une nouvelle version beta académique (2.6) a été développée et intègre l'ensemble des modifications qui permettent : la recherche et l'identification d'un grand nombre de sites optimaux et sous-optimaux, l'utilisation de nucléotides comme groupes chimiques, l'utilisation de groupes poly-nucléotidiques (dinucléotides ou autres poly-nucléotides). Cette dernière fonctionnalité pourra d'ailleurs être étendue à d'autres groupes chimiques pour la recherche, par exemple, de ligands peptidiques ou oligosaccharidiques.

L'approche "prédiction" de la méthode SELEX *in silico* a été appliquée à la protéine TIS11d qui appartient à la famille CCCH de protéines à doigts de zinc. La recherche d'un grand nombre de sites optimaux et sous-optimaux a été effectuée en utilisant les 2 mono-nucléotides A et U qui composent la séquence du ligand ARN ainsi que les 3 dinucléotides UU, AU, UA rencontrés dans la séquence. Plus de 40 000 sites nucléotidiques potentiels ont ainsi été identifiés, distribués autour de la protéine sur ces 2 faces majeures : la face qui représente la zone de fixation de l'ARN et la face opposée (Fig. 42). Pour traiter un nombre aussi importants de sites, nous avons défini une procédure de filtre des résultats MCSS et des résultats BUILD (2ème étape de la méthode SELEX *in silico*) afin de ne retenir dans les chaînes ARN possibles que les plus pertinentes. Du point de vue thermodynamique, la liaison d'un ligand à sa cible est pénalisée par la perte de degrés de liberté du ligand (contribution entropique) lorsqu'il est lié par rapport à sa forme libre. Or, les sites identifiés par MCSS sont classés seulement en fonction de l'énergie d'interaction (contribution enthalpique). Un filtre entropique a été ajouté en considérant le nombre de groupes présents à un site donné. D'autre part, nous avons introduit une technique de classification hiérarchique ("clustering") au programme BUILD afin d'éliminer les chaînes trop similaires. Lorsque plusieurs chaînes ARN sont proches les unes des autres, on les regroupe entre elles ("cluster") et l'on ne retient qu'une chaîne parmi n possibles : celle qui est la plus représentative au sein du cluster. Ces 2 filtres ont permis d'identifier et de sélectionner 14 familles de chaînes correspondant à la séquence du ligand ARN naturel. L'ensemble des 14 familles sont localisées sur la face de la protéine qui correspond à la zone de fixation de l'ARN (Fig. 42). De plus, elles recouvrent aussi toutes le ligand naturel.

Le meilleur résultat, la chaîne avec le meilleur score prédite par SELEX *in silico*, est aussi celle qui donne la plus faible déviation par rapport à la structure 3D expérimentale (Fig. 43). Elle donne une déviation RMSD de 3,1Å et lorsque le 1er résidu de la chaîne (le plus mobile dans la structure déterminée par RMN) est exclu, la RMSD est inférieure à 3,0Å (Fig. 43). La dernière étape est d'optimiser les chaînes ARN sélectionnées et classées. Cette étape n'est pas détaillée du point de vue méthodologique car elle correspond essentiellement à l'utilisation de MCDOCK (voir section suivante). Les calculs MCDOCK sont susceptibles d'améliorer les résultats avec une plus petite déviation entre les chaînes reconstruites et la structure 3D expérimentale ; les calculs sont actuellement en cours. Bien que la dernière étape du SELEX *in silico* ne soit pas totalement achevée, on peut considérer que la méthode a été en grande partie validée, MCDOCK ayant été en partie validé lors du défi CAPRI2008.

3.6 MCDOCK : docking de complexes ARN/protéine

MCDOCK a été développée sur les principes de la méthode de "docking" Monte-Carlo d'Amadéo Caflisch [133]. La procédure de docking repose sur une exploration conformationnelle à la

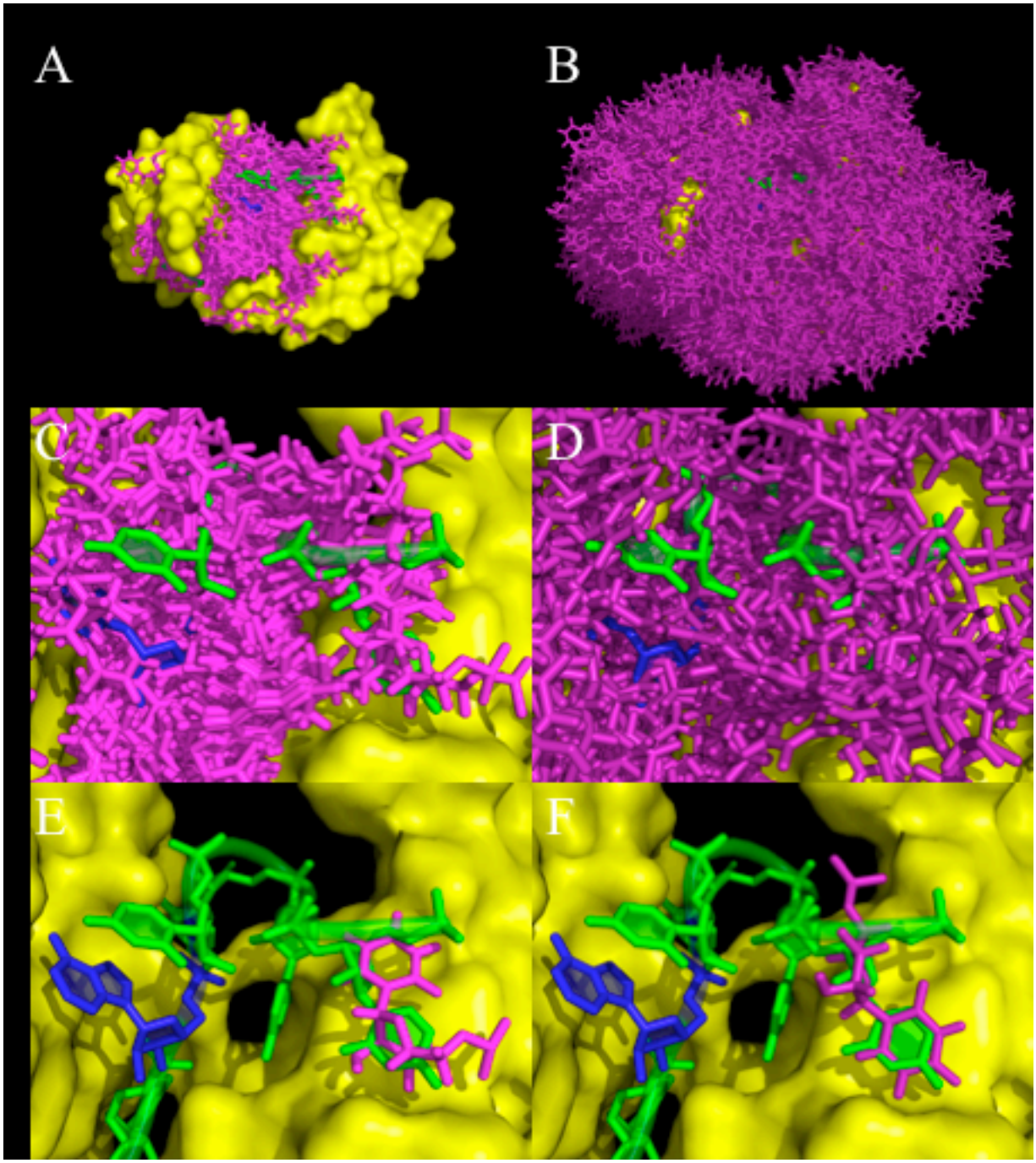


Figure 40. Stratégie "force brute" pour l'identification de sites sous-optimaux par MCSS. Le site ciblé est celui du résidu U3 (sous-optimal). La stratégie standard de MCSS (A, E, F) ne permet pas d'identifier des nucléotides se liant à cette position qui correspond à un site sous-optimal. La stratégie "force brute" (B, D, F) modifie les seuils d'énergie et la densité de sites identifiés. Un zoom des sites identifiés proches du résidu U3 sont montrés en C) et E) et révèlent l'absence de candidats proches de la structure 3D expérimentale. Dans la stratégie "force brute", la densité de sites est plus élevée (D), et des sites sous-optimaux proches de la structure 3D expérimentale sont identifiés (F).

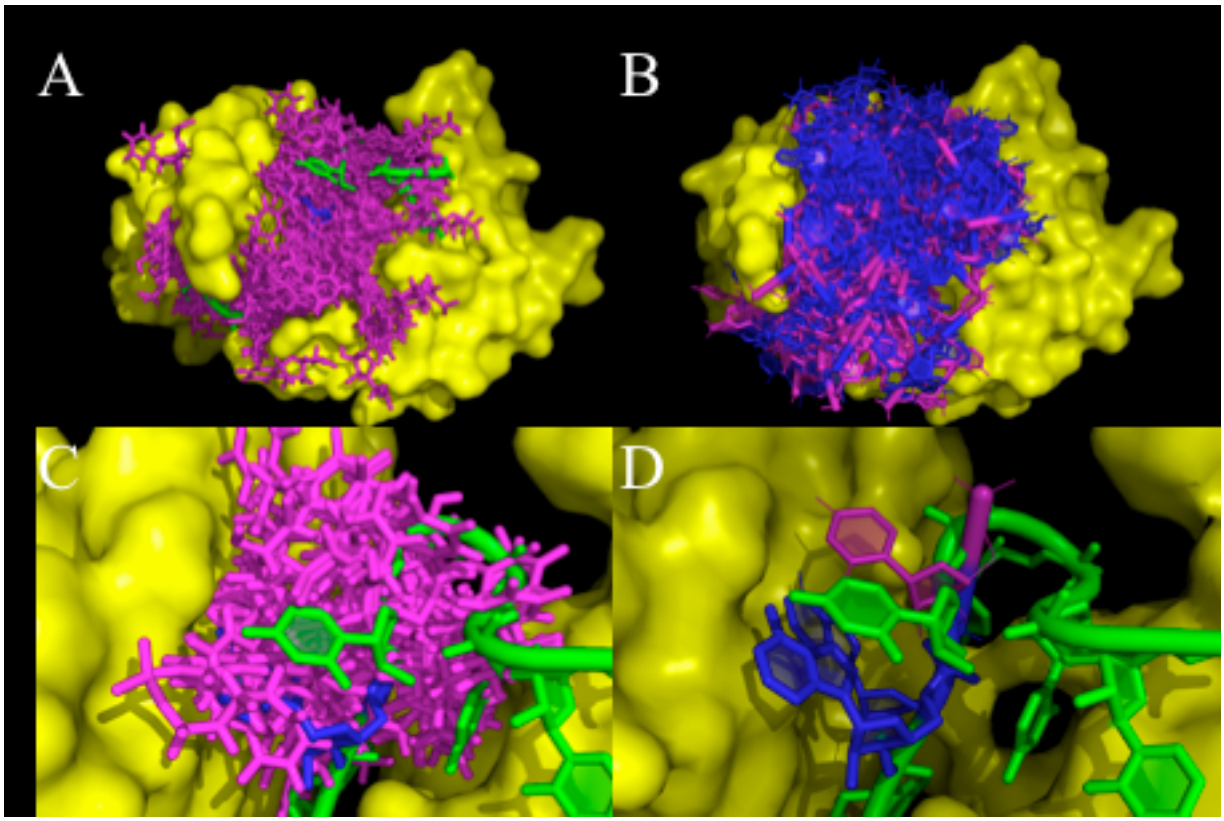


Figure 41. Stratégie "groupes étendus" pour l'identification de sites sous-optimaux par MCSS. Les site ciblé est celui du résidu U6 (sous-optimal). La stratégie standard de MCSS (A, C) ne permet pas d'identifier des nucléotides se liant à cette position qui correspond à un site sous-optimal (C). La stratégie des "groupes étendus" utilisant un groupe MCSS composé d'un dinucléotide UA (B) permet d'identifier un site de U6 proche de la structure 3D expérimentale (D).

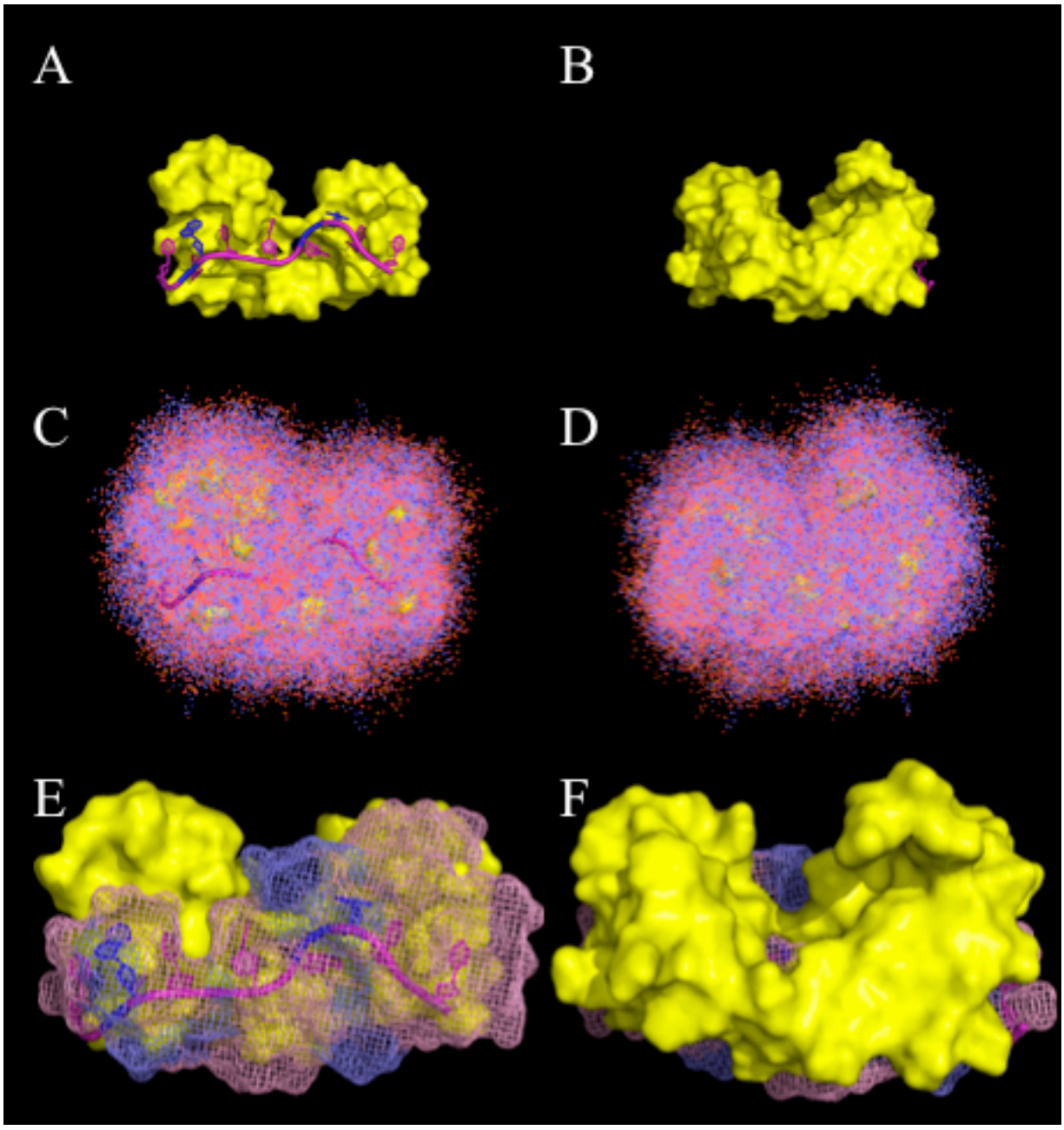


Figure 42. Application de l'approche "prédiction" de SELEX in silico à la protéine TIS11d : identification de zones de liaison à l'ARN. Les différentes stratégies MCSS testées sur la protéine HuD ont été appliquées à la protéine TIS11d, protéine appartenant à la famille de doigts à zinc (CCCH) de MBNL1. La face de liaison de la protéine à l'ARN (A) comme la face opposée (B) comportent de sites mono- et di-nucléotidiques potentiels identifiés par MCSS (C et D). Ces sites ont permis de reconstruire plusieurs familles de chaînes ARN représentées par une surface "maillée" (les positions correspondant aux sites pour les U sont indiquées en rose, celles pour les résidus A en bleu). Les chaînes reconstruites par SELEX in silico (de la même taille que celle du ligand ARN) sont toutes localisées sur la face de liaison à l'ARN.

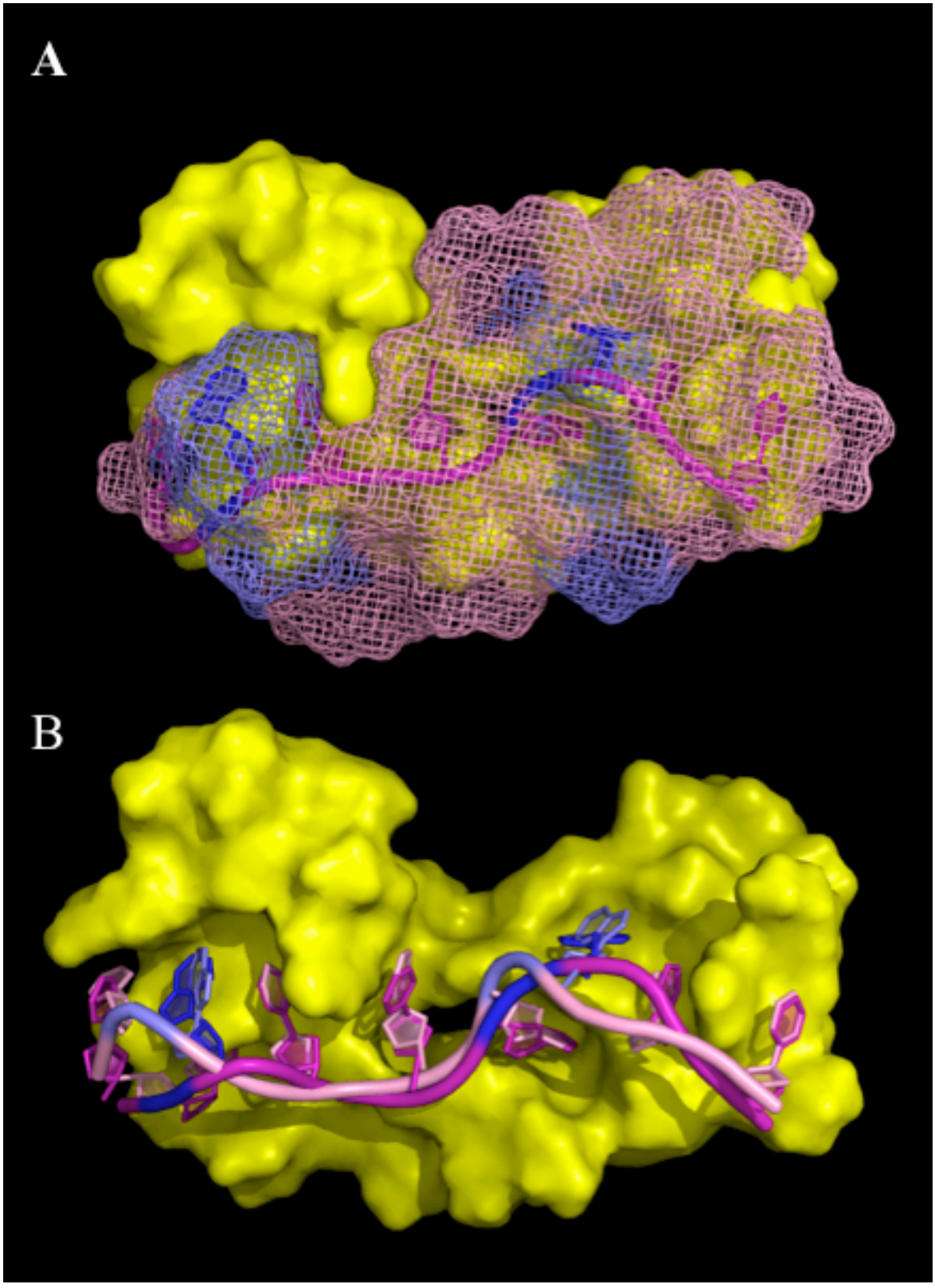


Figure 43. Application de l'approche "prédiction" de SELEX in silico à la protéine TIS11d : identification et sélection de chaînes d'ARN. Parmi les chaînes obtenues (A), la chaîne d'ARN avec le meilleur score (B) est celle qui présente la plus faible déviation par rapport à la structure 3D expérimentale (RMSD=3,1Å). L'ARN dans la structure expérimentale du complexe est représentée par son squelette phosphodiester et la position des nucléotides (U en pourpre, A en bleu), ainsi que l'ARN de la chaîne prédite par SELEX *in silico* (U en rose, A en bleu clair).

fois globale et locale (Monte-Carlo) et une évaluation d'énergie en 2 étapes permettant d'utiliser au mieux les ressources de calcul (Fig. 44). La recherche conformationnelle tient compte des degrés de liberté des partenaires en termes de translation et rotation globales et de variations des angles de torsion des chaînes latérales des résidus d'acides aminés situés à l'interface ARN/protéine et éventuellement des nucléotides. L'énergie du système est évalué de façon approximative après une minimisation d'énergie afin d'éliminer des configurations trop défavorables *a priori* et non pertinentes pour un calcul coûteux d'énergie libre (Fig. 44). Le développement et une validation partielle de l'approche de "docking" MCDOCK ont été réalisés sur un complexe ARN/protéine utilisé comme système test (le complexe formé entre le snRNA U4 et la protéine 15.5kDa humaine). Plusieurs complexes de départ (5 au total en plus du complexe natif) sont générés aléatoirement en modifiant les positions et orientations relatives des 2 partenaires ARN et protéine ; ensuite une simulation par complexe est réalisée. Les résultats obtenus sur les 5 complexes générés aléatoirement sont comparés au résultat obtenu sur le complexe natif : la fonction de score permet de départager les complexes a priori les plus favorables. Les résultats obtenus sur le système test plaçaient le complexe natif avec le meilleur score (données non publiées).

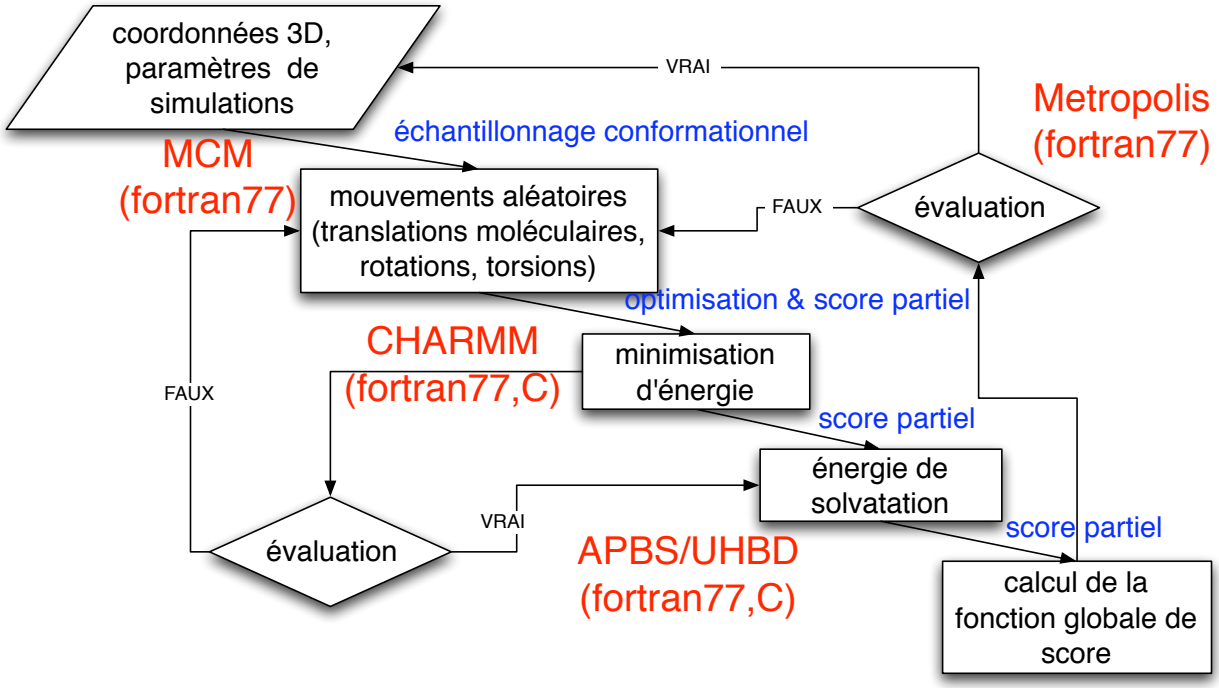


Figure 44. Schéma de la procédure de docking par MCDOCK. La recherche conformationnelle est effectuée par simulations Monte-Carlo à haute température (600K) avec le programme CHARMM. Un terme d'interaction non-polaire est également calculé avec CHARMM à partir de la différence de surface accessible entre les molécules dans le complexe et les molécules libres. Les interactions électrostatiques sont calculées par la résolution de la forme non-linéaire de l'équation de Poisson-Boltzmann, par les programmes UHBD [134] ou APBS [135]. Les programmes appelés à différentes étapes sont indiqués en rouge. Les tâches accomplies sont indiquées en bleu. Les indications "vrai/faux" correspondent aux décisions qui sont prises en fonction du résultat des évaluations correspondantes.

Le développement de MCDOCK a été poursuivie et elle a été utilisée récemment dans le cadre de l'édition 2008 du défi annuel international CAPRI auquel nous avons répondu comme équipe participante en collaboration avec l'équipe ORPAILLEUR du LORIA (M. Chavent et B. Maigret) et Dave Ritchie, Professeur à l'Université d'Aberdeen (Ecosse, RU). Parmi les cibles proposées, cette dernière édition du défi CAPRI (http://www.ebi.ac.uk/msd-srv/capri/round15/round15.html)

proposait en effet, pour le première fois, de prédire la structure d'un complexe ARN/protéine (cible T34). Il s'agit du complexe formé entre la méthyltransférase RlmA de classe II qui modifie certaines positions de l'ARN 23S bactérien et notamment la position G748 qui se situe dans le domaine de l'ARN 23S formant une jonction triple entre les hélices 33 à 35 [136]. Sur les 44 équipes participant à la prédiction en aveugle (les structures 3D expérimentales du complexe et de la protéine seule étant inconnues), les prédictions que nous avons réalisées avec MCDOCK (Fig. 44) nous placent au 7ème rang des meilleures équipes participantes pour cette cible (Table 2). La méthodologie reposait sur la génération de 15 complexes de départ générés par le programme de docking HEX, développé par D. Ritchie, et qui permet d'identifier les complexes entre molécules présentant la meilleure complémentarité de surface de contact [137]. La structure 3D de l'ARN était donnée aux participants alors que la structure 3D de la protéine, inconnue, a été modélisée par homologie par B. Maigret. Ensuite, les 15 complexes initiaux obtenus par HEX étaient soumis à une approche par dynamique moléculaire (M. Chavent & B. Maigret) ou à MCDOCK (M. Simoes & F. Leclerc). La comparaison des résultats obtenus avec HEX et les 2 approches par simulation : dynamique moléculaire (MD) et Monte-Carlo (MCDOCK) montrent que les approches par simulation améliorent la précision des prédictions, les prédictions obtenues par MD étant légèrement meilleures (Table 3). La courte durée des simulations qui ont limité fortement l'exploration conformationnelle des modes d'interaction ARN/protéine laisse espérer encore de meilleures résultats avec les méthodes combinées HEX+MD ou HEX+MCDOCK (Fig. 45) dans des conditions plus idéales. Un travail complémentaire est en cours pour valider plus avant les méthodes (Chavent et al., "Improving rigid body docking predictions for the CAPRI Round 15 targets using short molecular dynamics and Monte Carlo simulations and bioinformatics databases", en préparation).

Nous avons montré, dans le cadre du défi international CAPRI2008, que les méthodes combinées telles que HEX+MCDOCK sont performantes pour prédire la structure 3D de complexes ARN/protéine y compris avec un modèle 3D pour la structure de la protéine, même si la structure 3D expérimentale de l'ARN lié était elle connue. Des modifications seront apportées afin d'améliorer la fonction de score de MCDOCK qui n'a pas permis de classer la meilleure prédiction (le plus forte similarité avec la structure expérimentale) avec le meilleur score. En incorporant ces modifications, la méthode pourra raisonnablement être appliquée avec des chances de succès pour prédire les complexes formés entre les répétitions $(CUG)_n$ et ses partenaires protéiques sous réserve que les structures 3D des protéines soient disponibles ou à défaut celle d'un homologue proche.

Les résultats de nos prédictions lors de CAPRI2008 suggéraient des difficultés à bien évaluer les meilleures modes d'interaction par la fonction de score de MCDOCK. Une série test de complexes ARN/protéine représentatifs des principaux domaines de liaison à l'ARN (RRM, KH, dsRBD, ZnF-CCCH, Zn-CCHH, etc) [138] sera construit et utilisé pour ajuster le paramétrage de la fonction de score. On tiendra également compte de l'importance des interactions électrostatiques dans le choix des complexes à tester. Des résultats préliminaires suggèrent que la mauvaise évaluation par la fonction de score est en partie reliée à la faible exploration conformationnelle durant le simulations effectuées sur la cible T34 de CAPRI2008 correspondant au complexe ARN/protéine (Fig. 45). En effet, de plus longues simulations permettent d'obtenir un classement des 10 modèles proposés en meilleur accord avec la déviation de ceux-ci par rapport à la structure 3D expérimentale du complexe.

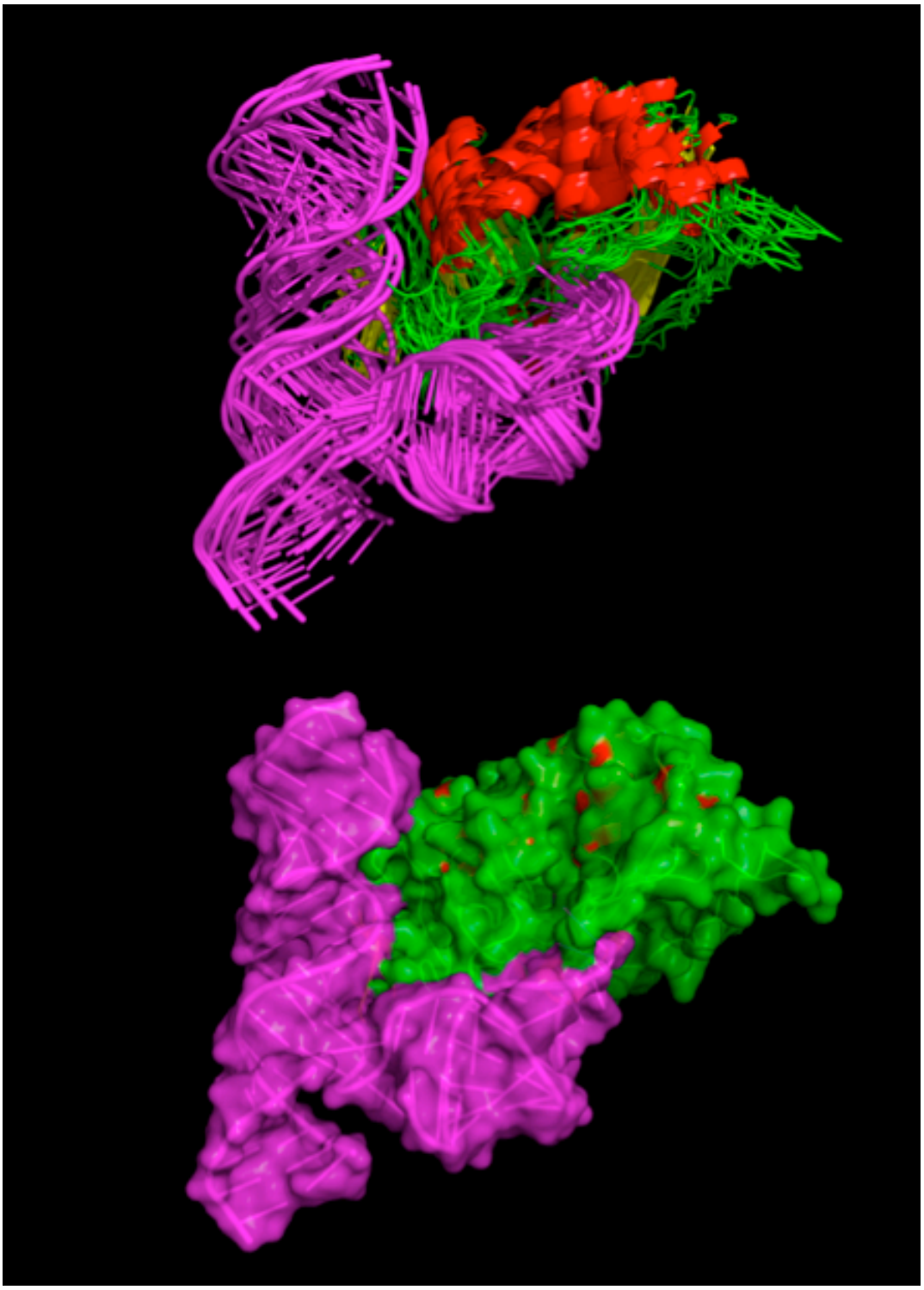


Figure 45. Modèles 3D du complexe ARN/protéine méthyltransférease RlmA/ARNr 23S (CAPRI2008). Les 10 modèles générés par HEX+MCDOCK sont montrés superposés les uns par rapport aux autres (haut) ; le meilleur modèle dans le classement CAPRI est montré représenté par une surface accessible (ARN en pourpre, protéine en vert).

Table 2 : Résultats CAPRI 2008 sur la cible T34

prédiction	distance	clashes	L_rmsd	M_RMSD	Classification
capri_t34_xray	0.000	22	0.000	0.000	high
1. T34_P37.M01	0.837	17	1.675	1.442	medium
2. T34_P37.M03	1.504	15	2.344	1.528	medium
3. T34_P61.M03	1.055	15	2.381	1.887	medium
4. T34_P24.M03	1.073	129	2.610	1.358	medium
5. T34_P29.M01	1.168	48	2.853	1.343	medium
6. T34_P46.M03	1.510	18	3.189	1.450	medium
7. T34_P63.M06	1.635	15	3.223	1.231	medium
16. T34_P39.M02	1.936	128	4.18	1.221	medium
26. T34_P25.M05	3.626	13	4.081	1.716	acceptable
66. T34_P25.M05	-	-	-	-	incorrect
164. T34_P25.M05	-	-	-	-	removed

P61 : HEX+MD (Chavent & Maigret) ; P63 : HEX+MC (Simoes & Leclerc) ; HEX (Ritchie) ; les critères L_rmsd et M_RMSD sont deux critères utilisés par mesurer la déviation entre un modèle donné et la structure 3D expérimentale ; "clashes" indique le nombre de contacts entre atomes anormalement proches dans le complexe ; la classification va de "high" pour les meilleures prédictions à "medium", "acceptable" et "incorrect". La Table complète est disponible en suivant le lien : http ://www.ebi.ac.uk/msd-srv/capri/round15/R15_T34/index.html.

Table 3 : Performance des méthodes de docking

Modèle HEX	Classement	HEX + MD	Classement	HEX + MCDOCK	classement
1	removed	5	acceptable	-	-
2	medium	9	medium	2	medium
3	acceptable	-	-	4	acceptable
4	medium	7	medium	5	acceptable
5	acceptable	-	-	6	acceptable
6	acceptable	-	-	3	acceptable
7	removed	4	acceptable	-	acceptable
8	removed	3	medium	10	acceptable
9	incorrect	10	incorrect	9	acceptable
10	acceptable	6	acceptable	7	incorrect
11	?	8	acceptable	1	medium
12	?	-	- 8	acceptable	
13	?	1	medium	?	?
14	?	2	medium	?	?

pour chaque méthode utilisée (HEX seule, HEX+MD, HEX+MCDOCK), le classement CAPRI est indiqué ; les lignes en gras indiquent les cas où les méthodes de simulation ont conduit à une amélioration directe de la prédiction à partir du même modèle de départ.

Dans MCDOCK, la fonction de score est calculée comme la somme de 3 contributions principales qui donne une approximation de l'énergie libre de liaison :

$$\Delta G_{binding} = \Delta E_{el+vdw} + \Delta G_{el,solv} + \Delta G_{np,solv} \tag{3.1}$$

où les différentes contributions pour soluté (complexe ARN+protéine) représentent : l'énergie interne (contributions électrostatique et van der Waals) : ΔE_{el+vdw} calculée par le programme CHARMM, la contribution électrostatique de l'énergie de solvatation du soluté : $\Delta G_{el,solv}$ calculée par la résolution de la forme non-linéaire de l'équation de Poisson-Boltzmann (programme UHBD ou APBS), la contribution non polaire de l'énergie de solvatation du soluté : $\Delta G_{np,solv}$ calculée par le programme CHARMM (comme la variation de surface accessible entre le complexe et les partenaires pris séparément). La fonction de score pourra être ajustée à partir des calculs effectués sur la série test de complexes ARN/protéine représentatifs en ajoutant un coefficient multiplicateur à chacune des 3 contributions dans l'équation (3.1). L'équation prend alors la forme suivante :

$$\Delta G_{binding} = \alpha \Delta E_{el+vdw} + \beta \Delta G_{el,solv} + \gamma \Delta G_{np,solv} \tag{3.2}$$

Une méthode statistique permettra ensuite de calculer les coefficients α, β et γ permettant d'obtenir le meilleur accord possible entre la fonction de score et la déviation des modèles générés par rapport à la structure 3D expérimentale.

3.7 Travaux publiés

3.7.1 "DNA Polymorphism : A Comparison of Force Fields for Nucleic Acids"

DNA Polymorphism: A Comparison of Force Fields for Nucleic Acids

Swarnalatha Y. Reddy,* Fabrice Leclerc,*†‡ and Martin Karplus*‡

*Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138 USA; †Laboratoire de Maturation des ARN et Enzymologie Moléculaire, CNRS-UHP Nancy I UMR 7567, Université Henri Poincaré, Faculté des Sciences, B.P. 239, 54506 Vandoeuvre-lès-Nancy, France; and ‡Laboratoire de Chimie Biophysique, ISIS, Université Louis Pasteur, 4 rue Blaise Pascal, 67000, Strasbourg, France

ABSTRACT The improvements of the force fields and the more accurate treatment of long-range interactions are providing more reliable molecular dynamics simulations of nucleic acids. The abilities of certain nucleic acid force fields to represent the structural and conformational properties of nucleic acids in solution are compared. The force fields are AMBER 4.1, BMS, CHARMM22, and CHARMM27; the comparison of the latter two is the primary focus of this paper. The performance of each force field is evaluated first on its ability to reproduce the B-DNA decamer d(CGATTAATCG)$_2$ in solution with simulations in which the long-range electrostatics were treated by the particle mesh Ewald method; the crystal structure determined by Quintana et al. (1992) is used as the starting point for all simulations. A detailed analysis of the structural and solvation properties shows how well the different force fields can reproduce sequence-specific features. The results are compared with data from experimental and previous theoretical studies.

INTRODUCTION

Nucleic acids can adopt different conformations in solution depending on the base composition (Hunter, 1993) and the environment (for example pH and temperature, Kumar and Maiti, 1994), including the nature of the solvent (Fang et al., 1999), the counterions (Minasov et al., 1999), their concentration (Ali and Ali, 1997), and interactions with proteins (Jones et al., 1999), or small molecules (Reinert, 1999). Even a given sequence of DNA or RNA can exhibit multiple conformations (Kielkopf et al., 2000). In living systems, the conformational flexibility of DNA resides primarily in the polymorphs of the DNA double helix (including right-handed and left-handed double-helical DNA) that occur under various experimental conditions (Gupta et al., 1980). By contrast, double-stranded helical RNA is confined to two very similar polymorphs of the A form (A and A′), and the wide range of single-stranded nonhelical RNA folds introduces the essential structural variability.

Significant progress in the development of empirical force fields and molecular dynamics (MD) simulation methods has lead to a more reliable description of the structure, energetics, and dynamics of nucleic acids (Auffinger and Westhof, 1998; Beveridge and McConnell, 2000; Cheatham and Kollman, 2000; Cheatham and Young, 2001). However, some limitations related to the improper treatment of the equilibrium between the A and B forms of DNA (Feig and Pettitt, 1997, 1998) and the deviations of helicoidal parameters from canonical B values (Cheatham and Kollman, 1996) have been reported. The overstabilization of the A form relative to the B form of DNA (Yang and Pettitt, 1996; MacKerell, 1997; Feig and Pettitt, 1997) with the CHARMM22 force field (MacKerell et al., 1995) has been addressed in a recent reoptimization of the CHARMM22 all-atom nucleic acid force field. The new nucleic acid force field, called CHARMM27, has small but important changes in both the internal and interaction parameters relative to CHARMM22 (Foloppe and MacKerell, 2000) and appears to treat well the equilibrium between the A and B forms of DNA and the influence of the environment, such as the water activity (MacKerell and Banavali, 2000). A revised and improved version of the AMBER4.1 force field has also been presented that shows better agreement with experimental data as a result of the adjustment of internal force field parameters (Cheatham et al., 1999). An alternative nucleic acid force field, which we refer to as the Bristol-Myers Squibb (BMS) force field, has been developed by Langley (Langley, 1998). Both the CHARMM27 and AMBER force field parameters are based on the reproduction of experimental results for nucleic acid oligomers (e.g., condensed phase structural properties of DNA and RNA) and consistency with small molecule results obtained from quantum mechanical calculations and experimental data. The BMS force field was developed, in part, by adaptation of the CHARMM22 (MacKerell et al., 1995), QUANTA and AMBER force fields (Cornell et al. 1995. The backbone angle and dihedral parameters were derived from quantum mechanical calculations with refinements based on a series of MD simulations. All the force fields used condensed-phase MD simulations in the final stage of the parameter optimization. The CHARMM27 force field has also been applied to model compounds to evaluate the contributions from the individual moieties to the overall conformational properties of DNA and RNA (Foloppe and MacKerell, 2000).

Recent simulations of nucleic acids using an explicit solvent representation and an ionic environment have led to high structural stability on the nanosecond time scale (Beveridge and McConnell, 2000). This accuracy was

Submitted March 15, 2002, and accepted for publication August 6, 2002.

Address reprint requests to Martin Karplus, Dept. of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138 USA. Tel.: 617-495-1768; Fax: 617-496-3204; E-mail: marci@tammy.harvard.edu.

0006-3495/03/03/1421/29 $2.00

3.7.2 "MCSS-based predictions of RNA binding sites"

Theor Chem Acc (1999) 101:131–137
DOI 10.1007/s002149800m85

Theoretical
Chemistry Accounts

Regular article

MCSS-based predictions of RNA binding sites*

Fabrice Leclerc[1], **Martin Karplus**[1,2]

[1] Harvard University, Department of Chemistry and Chemical Biology, 12 Oxford Street, Cambridge, MA 02138, USA
[2] Université de Strasbourg I, Institut le Bel, Laboratoire de Chimie Biophysique, F-67000 Strasbourg, France

Received: 24 April 1998 / Accepted: 4 August 1998 / Published online: 7 December 1998

Abstract. The diversity of RNA tertiary structures provides the basis for specific recognition by proteins or small molecules. To investigate the structural basis and the energetics which control RNA-ligand interactions, favorable RNA binding sites are identified using the MCSS method, which has been employed previously only for protein receptors. Two different RNAs for which the structures have been determined by NMR spectroscopy were examined: two structures of the TAR RNA which contains an arginine binding site, and the structure of the 16S rRNA which contains an aminoglycoside binding site (paromomycin). In accord with the MCSS methodology, the functional groups representing the entire ligand or only part of it (one residue in the case of the aminoglycosides) are first replicated and distributed with random positions and orientations around the target and then energy minimized in the force field of the target RNA. The Coulombic term and the dielectric constant of the force field are adjusted to approximate the effects of solvent-screening and counterions. Optimal force field parameters are determined to reproduce the binding mode of arginine to the TAR RNA. The more favorable binding sites for each residue of the aminoglycoside ligands are then calculated and compared with the binding sites observed experimentally. The predictability of the method is evaluated and refinements are proposed to improve its accuracy.

Key words: MCSS – Ligand Design – RNA – Binding

1 Introduction

Nucleic acids make logical targets for drug design, since all enzymes and receptor proteins depend on RNA for their synthesis. Unlike DNA, RNA may fold into complex and diverse molecular shapes which constitute attractive targets for specific and selective binding. A number of X-ray or NMR structures of RNA and its complexes with drugs, peptides, or proteins have been recently determined and provide the opportunity to better understand RNA-ligand interactions [6]. Nevertheless, only a few studies on modeling RNA-ligand interactions have been published [7, 8].

A primary step towards the design of drugs against macromolecule receptors is the identification of potential binding sites for ligand fragments (functional groups). Such an approach has been applied in computational and laboratory combinatorial ligand design for protein receptors [2, 9–11]. The TAR RNA of HIV-1 and the bacterial 16S rRNA, for which the structures have been determined by NMR spectroscopy [3–5], represent two interesting targets because of their biological role in the regulation of the HIV cycle and in the protein synthesis of pathogenic bacteria, respectively. They are used here as macromolecule targets with the MCSS method to predict favorable binding sites for arginine in the case of the TAR RNA and for aminoglycoside moieties in the case of the 16S rRNA. The MCSS method [1] is divided into two steps. The first step involves the replication of pre-defined functional groups and their distribution in random positions and orientations in a binding region delimited by a spherical or rectangular boundary. The second step involves their energy minimization in the force field of the receptor [1], and the selection of the local minima based on an energy cutoff. The search for minima is performed by an iterative process. Details of the method are given by Miranker and Karplus [1].

2 Specific approach

Two sets of force field parameters were used to identify and select the MCSS minima. In the first set (Model 1), the non-bonded interactions between the replicated functional groups (replica) and the receptor were calculated based on the CHARMM force field using the recent parameters for nucleic acids [12]. Since these parameters are designed for use with explicit solvents

* Contribution to the Proceedings of Computational Chemistry and the Living World, April 20–24, 1998, Chambery, France

Correspondence to: M. Karplus
e-mail: marci@brel.u-shasbg.fr

Perspectives : Modélisation d'interactions ARN/ligands : Application aux macromolécules biologiques associées aux DM

Application de l'approche SELEX *in silico* aux cibles des DM

La méthode SELEX *in silico* est actuellement appliquée aux cibles d'intérêt, à savoir les protéines CUG-BP1 et MBNL1. CUG-BP1 comporte trois motifs RRM : RRM1, RRM2 et RRM3 ; elle est produite dans l'UMR 7567 sous différentes formes entière et tronquées. La structure 3D des domaines RRM1 + RRM2 d'une part (PDB ID : 2DHS) et celle du domaine RRM3 d'autre part (PDB ID : 2CPZ) sont disponibles (Fig. 46A). Une très petite partie de la structure de MBNL1 est disponible : il s'agit des domaines en doigts de zinc (PDB ID : 2E5S ; Fig. 46B). L'application du SELEX *in silico* est utilisée pour tenter de répondre à 2 objectifs : le premier est de prédire la structure de complexes formés entre les cibles et un certain nombre de ligands naturels ou artificiels (obtenus par SELEX expérimentale) d'après des données connues ou obtenus au sein de l'UMR 7567 ; on utilisera alors l'approche "prédiction" de la méthode. Le second objectif est de concevoir des ligands ARN potentiels susceptibles d'interférer efficacement avec les ligands naturels ; pour cela, on utilisera l'approche "conception". L'écueil que représente la sur-représentation de sites optimaux (ou la sous-représentation de sites sous-optimaux) dans l'approche "prédiction" devrait être un avantage dans l'approche "conception". Dans le SELEX *in silico*, les calculs les plus coûteux étant les calculs MCSS, l'ensemble des calculs MCSS sont préparés sur l'ensemble des cibles en considérant tous les groupes possibles pour pouvoir utiliser les 2 approches "prédiction" et "conception". Pour cela, les calculs MCSS sont effectués avec les 4 groupes correspondant aux 4 mono-nucléotides (A, U, C, G) et les 16 di-nucléotides possibles, donc 20 groupes au total. La version beta (2.6) de MCSS a été migrée sur la plateforme IBM-SP4 du CINES qui est utilisée pour réaliser les calculs.

Application d'approches de criblage et de conception de ligands contre les répétitions $(CUG)_n$

Un modèle 3D atomique d'ARN $r(CUG)_{17}$ a été généré en utilisant la structure 3D du duplex $r(CUG)_6$: $r(CUG)_6$ obtenue par diffraction des rayons-X (PDB ID : 1ZEV, Mooers

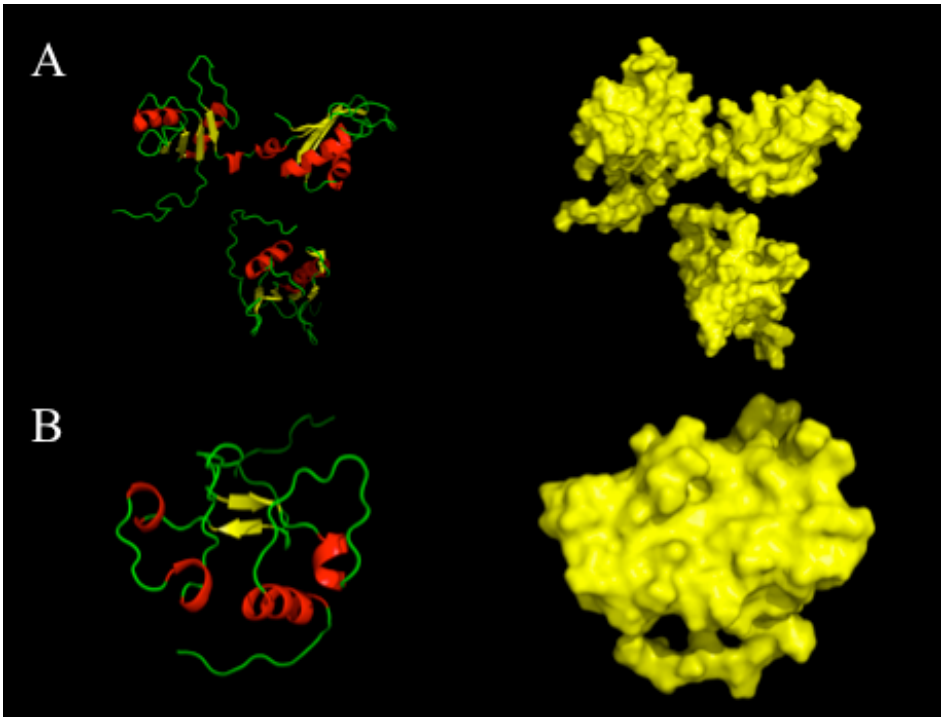


Figure 46. Protéines cibles impliquées dans les dystrophies myotoniques (DM). A. Représentations de CUG-BP1 : "cartoon" et "surface". Les domaines RRM1 et RRM2 correspondent à la structure PDB 2DHS, le domaine RRM3 à la structure PDB 2CPZ. B. Représentations de MBNL1 : "cartoon" et "surface". Les domaines en doigts de zinc correspondent à la structure PDB 2E5S.

et al., 2005). A la différence de la structure 3D expérimentale, le modèle correspond à une chaîne d'ARN monocaténaire obtenue par réplication symétrique du duplex pour la partie double-brin (8 répétitions en double-brin) et auquel nous avons ajouté une boucle terminale qui ferme la tige : la "triloop" CUG (Fig. 47) ; r$(CUG)_{17}$ correspond à l'une des molécules à répétitions CUG utilisée expérimentalement au laboratoire pour étudier *in vitro* les interactions avec des partenaires protéiques.

Avant la détermination de la structure 3D du duplex, un petit modèle 3D à trois répétitions r$(GG(CUG)_3CC)$ avait été construit afin d'étudier la flexibilité d'une tige boucle comportant un mésappariement U :U (pour la méthodologie, voir Reddy *et al.*, 2000 [139]). Les résultats de simulations par dynamique moléculaire ont montré que l'ARN semble très flexible dans la partie de la boucle terminale. Le mésappariement U :U correspond aussi à une région de l'ARN elle aussi très dynamique, où les deux résidus U peuvent glisser l'un par rapport à l'autre tout en maintenant un appariement par leur face Watson-Crick (Fig. 48). Les deux U peuvent passer d'un appariement où le U en 5' est déplacé vers le petit sillon (Fig. 48A) à un appariement où le U en 3' est déplacé vers le petit sillon (Fig. 48B), ceci par référence aux appariements Watson-Crick de l'hélice (Fig. 48). Dans la structure 3D du duplex, les mésappariements U :U correspondent à un déplacement du U en 3' vers le petit sillon (Fig. 48C). La transition entre le 1er type d'appariement (Fig. 48A) et le second (Fig. 48B) est observée en quelques nanosecondes à l'échelle de la simulation. Les paires U :U constituent vraisemblablement une zone de flexibilité dans la partie double-brin des répétitions $(CUG)_n$.

Les résultats obtenus en dynamique moléculaire sur le petit modèle à 3 répétitions se sont révélés transposables au modèle à 17 répétitions. Les simulations montrent une grande flexibilité de la boucle terminale en accord avec l'instabilité conformationnelle de la partie apicale observée

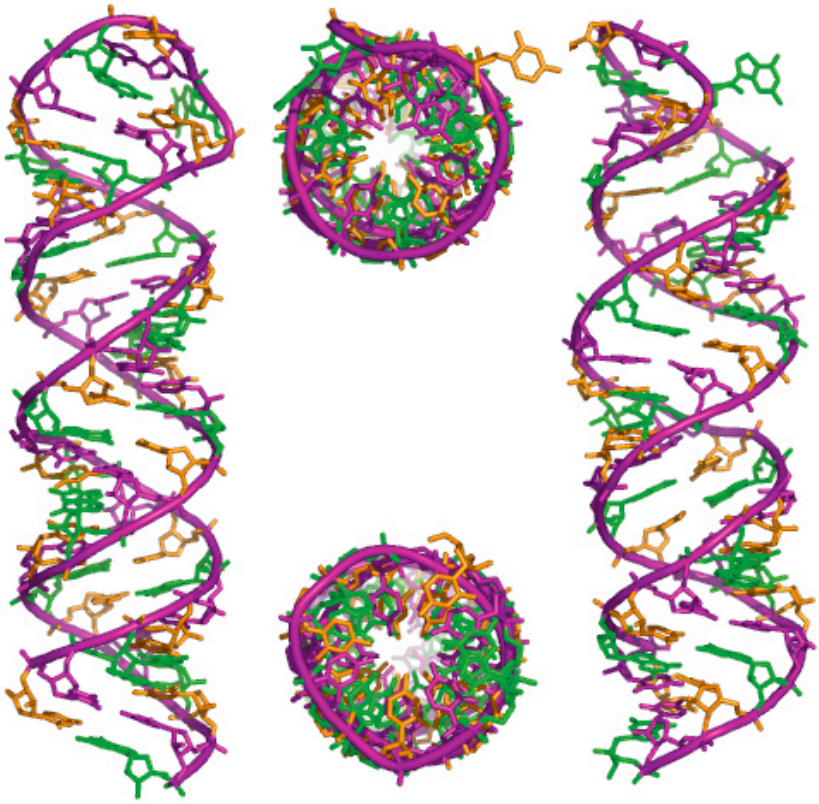


Figure 47. Modèle 3D d'un ARN à séquence répétée : r(CUG)$_{17}$. Les 2 vues avec de la structure tige-boucle représentées longitudinalement montrent les sillons et la plus grande accessibilité du petit sillon. Les 2 vues axiales montrent une structure tige-boucle très rectiligne un peu moins régulière qu'une double hélice standard d'ARN. La région double-brin de l'ARN a été construite à partir des coordonnées atomiques de la structure 3D d'un duplex r(CUG)$_6$:r(CUG)$_6$ (PDB ID : 1ZEV).

expérimentalement par la technique d'empreintes enzymatiques (Fig. 49). Les paires U :U sont également flexibles et pourraient représenter une zone déformable au contact MBNL1 ou d'autres ligands. Une des perspectives de ce travail sur l'ARN r(CUG)$_{17}$ est d'utiliser des conformations représentatives obtenues en dynamique moléculaire comme cibles pour la conception de ligands susceptibles d'interférer dans l'interaction de MBNL1 avec les répétitions $(CUG)_n$. La région des paires U :U apparaît comme un site privilégié à cibler pour la conception des ligands. Deux approches de philosophies différentes seront utilisés pour rechercher des ligands potentiels : la première entre dans la catégorie des méthodes de recherche et filtrage virtuel de molécules actives alors que la seconde est une approche par fragment pour la conception *de novo* de ligands. Dans le premier cas, l'approche consiste à rechercher des ligands potentiels parmi des molécules biologiques "actives" connus à une site particulier de la cible alors que la seconde, plus risquée, vise à proposer des ligands à partir de groupes chimiques susceptibles de se lier à différents sous-sites.

La première approche entre dans le cadre d'une collaboration avec Bernard Maigret pour l'adaptation du programme VSM-G ("Virtual Screening Manager for the Grid" [140]) pour le criblage virtuel de banque de données 3D de ligands ayant une complémentarité potentielle pour les ARN à séquence répétée CUG (Fig. 50). Une première banque de ligands connus comme inhibiteurs de protéines kinases a été sélectionnée ; comme il s'agit de molécules ayant des similarités moléculaires avec les nucléotides, ces inhibiteurs sont susceptibles de mimer des interactions

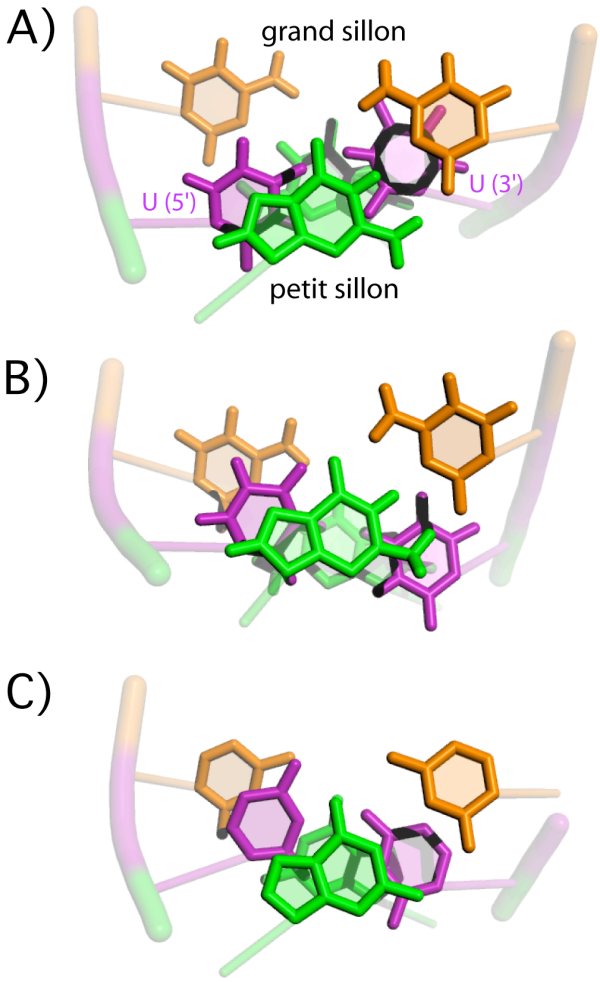


Figure 48. Appariements et empilements de paires U :U dans un petit modèle ARN à séquence répétée CUG. L'appariement U :U (violet) est représenté dans le contexte d'un duplex formé par des répétitions CUG avec les 2 paires Watson-Crick C :G et G :C (C : orange, G : vert) qui l'encadrent. A) Appariement U :U observé en simulation dans le cas du petit modèle d'ARN r(GG(CUG)$_3$CC : le U en 5' (côté gauche) est déplacé vers le petit sillon. B) Appariement U :U observé en simulation dans le cas du petit modèle d'ARN r(GG(CUG)$_3$CC : le U en 3' (côté droit) déplacé vers le petit sillon. C) Appariement U :U, observé dans la structure expérimentale du duplex d'ARN r(CUG)$_6$: r(CUG)$_6$ (PDB ID : 1ZEV [118]) : le U en 3' (côté droit) déplacé vers le petit sillon.

ARN/ARN avec les répétitions CUG. La recherche de molécules actives pourra ensuite être étendue à d'autres banques de molécules bioactives. La seconde approche correspond à une extension du programme MCSS pour l'identification de sites d'interaction favorables de groupes chimiques ayant une affinité potentielle pour les ARN et s'inscrit dans la continuité de travaux antérieurs menés dans la laboratoire de Martin Karplus. Une banque de groupes chimiques MCSS a été construite à partir de données extraites de la littérature sur des ligands connus pour se lier aux acides nucléiques (ADN ou ARN). Les groupes existant, qui ont été paramètrés dans le champ de force CHARMM, sont classés en 4 familles : les composés cycliques (pouvant se lier dans les sillons ou s'intercaler), les glycosides correspondant aux résidus qui des antibiotiques de la classe des aminoglycosides (pouvant se lier dans les sillons), les groupes amino chargés (extraits de composés polycationiques se liant à l'ADN), les métaux ou complexes avec des métaux (Fig. 51).

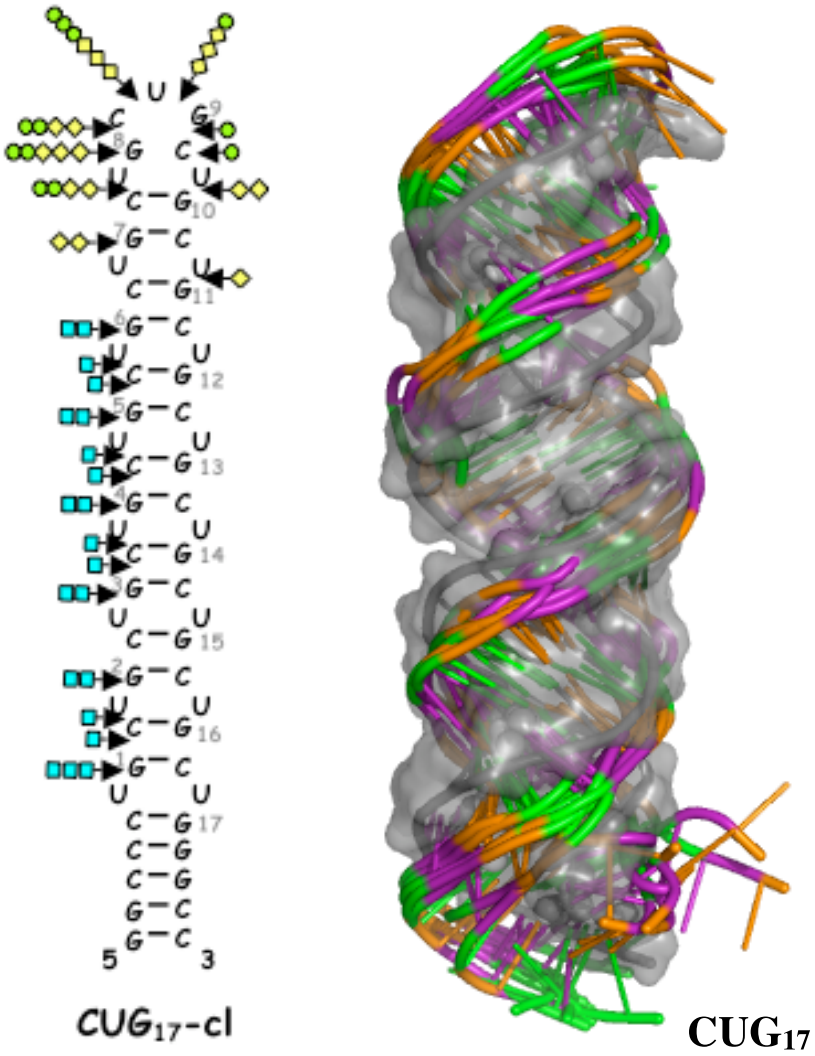


Figure 49. Structures 2D et 3D proposée pour l'ARN r(CUG)$_{17}$. Structure 2D proposée sur une base expérimentale : les clivages par les RNases V1, T1 et T2 sont respectivement représentés par des carrés bleus (V1), des losanges jaunes (T1) et des cercles verts (T2). Le nombre (1, 2 ou 3) de carrés, losanges ou cercles représente l'intensité du clivage (faible, moyen, fort, respectivement). Les carrés bleus indiquent les régions en double-brin, les losanges jaunes et cercles verts celles en simple-brin. Les appariemments Watson-Crick sont indiqués par un trait. Structure 3D proposée sur une base théorique : 10 conformations représentatives de la structure modélisée et soumise à une simulation par dynamique moléculaire sont superposées. Les résidus sont colorés par type : U en pourpre, C en orange, G en vert. La conformation initiale de la structure modélisée est représentée par une surface accessible afin de souligner les régions flexibles (les extrémités : la base de l'hélice et la boucle terminale, les paires U :U de la tige).

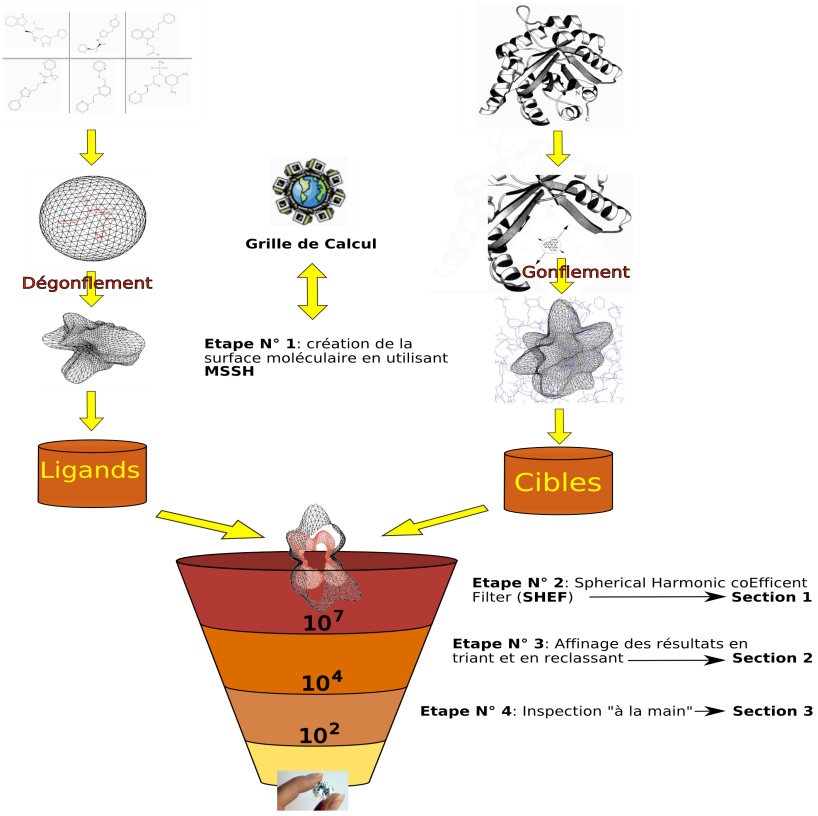


Figure 50. Schéma du principe de fonctionnement VSM-G. Les différentes étapes permettant le criblage de molécules sur une cible donnée sont indiquées.

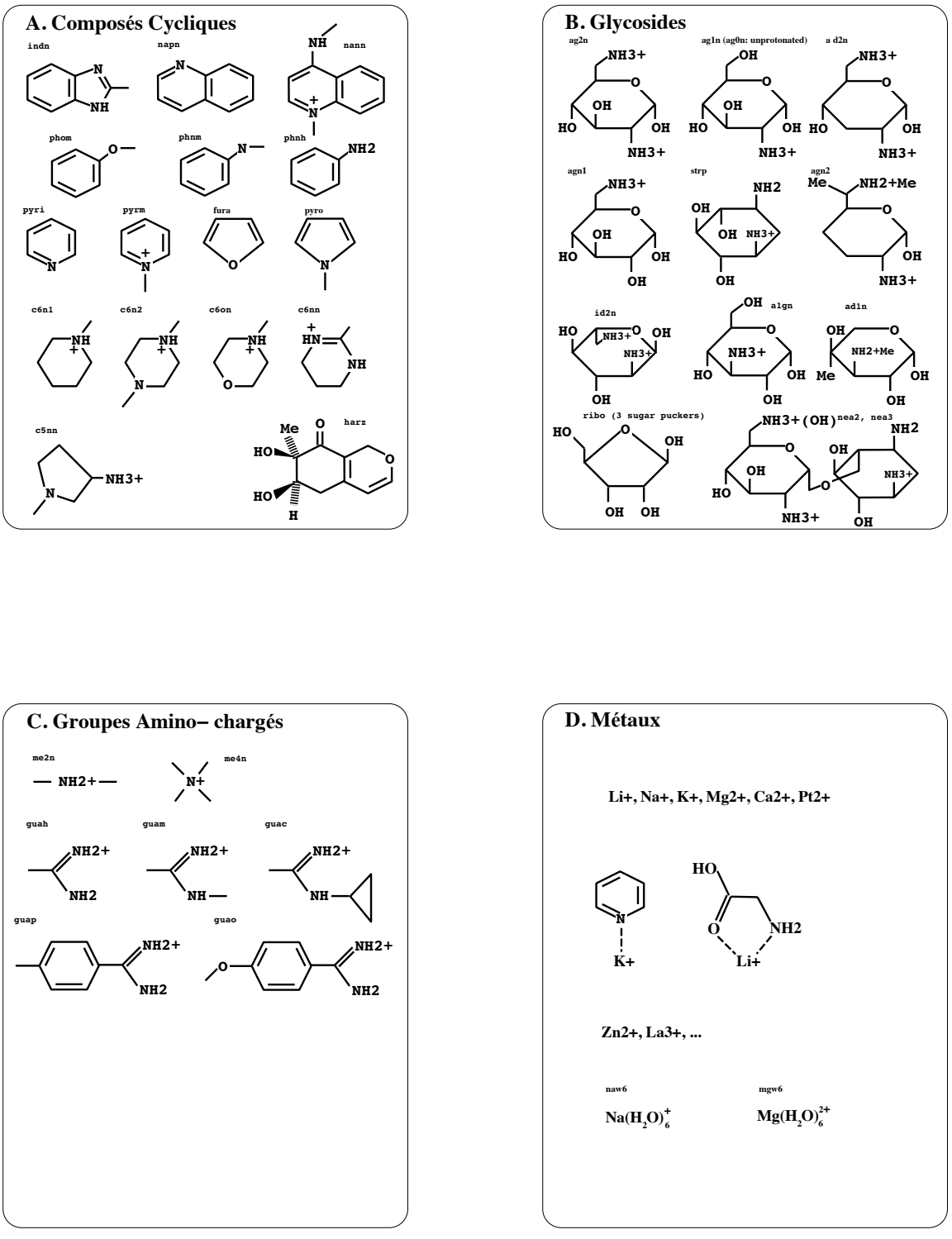


Figure 51. Groupes MCSS extraits de ligands des acides nucléiques. A. Groupes cycliques correspondant à des intercalants, à des ligands spécifiques des sillons ou à des ligands mixtes. B. Glycosides extraits des résidus des aminoglycosides. C. Groupes amino chargés extraits de composés polycationiques se liant dans les sillons de l'ADN. D. Métaux et complexes métalliques.

Annexe A

Publications

Bibliographie personnelle

[1] **F Leclerc**, R Cedergren, and A D Ellington. A three-dimensional model of the Rev-binding element of HIV-1 derived from analyses of aptamers. *Nat Struct Biol*, 1(5) :293–300, May 1994.

[2] A D Ellington, **F Leclerc**, and R Cedergren. An RNA groove. *Nat Struct Biol*, 3(12) :981–4, Dec 1996.

[3] J Srinivasan, **F Leclerc**, W Xu, A D Ellington, and R Cedergren. A docking and modelling strategy for peptide-RNA complexes : applications to BIV Tat-TAR and HIV Rev-RBE. *Folding & design*, 1(6) :463–72, Jan 1996.

[4] B Llorente, **F Leclerc**, and R Cedergren. Using SAR and QSAR analysis to model the activity and structure of the quinolone-DNA complex. *Bioorg Med Chem*, Jan 1996.

[5] P Chartrand, **F Leclerc**, and R Cedergren. Relating conformation, Mg2+ binding, and functional group modification in the hammerhead ribozyme. *RNA*, 3(7) :692–6, Jul 1997.

[6] **F Leclerc**, J Srinivasan, and R Cedergren. Predicting RNA structures : the model of the RNA element binding Rev meets the NMR structure. *Folding & design*, 2(2) :141–7, Jan 1997.

[7] S Steinberg, **F Leclerc**, and R Cedergren. Structural rules and conformational compensations in the tRNA L-form. *J Mol Biol*, 266(2) :269–82, Feb 1997.

[8] B Cousineau, **F Leclerc**, and R Cedergren. On the origin of protein synthesis factors : a gene duplication/fusion model. *J Mol Evol*, 45(6) :661–70, Dec 1997.

[9] **F Leclerc** and R. Cedergren. Modeling ligand interactions with nucleic acids using structure/activity relationships. *Encyclopedia of Computational Chemistry*, 1998.

[10] **F Leclerc** and R. Cedergren. Modeling RNA-ligand interactions : the Rev-binding element RNA-aminoglycoside complex. *J Med Chem*, 41(2) : 175–182, Jan 1998.

[11] **F Leclerc** and M Karplus. MCSS-based predictions of RNA binding sites. *Theoretical Chemistry Accounts : Theory, Computation, and Modeling (Theoretica Chimica Acta)*, 101(1) :131–137, Feb 1999.

[12] **F Leclerc**, B Llorente, and R Cedergren. Structure-function relationships of RNA : a modeling approach. *Meth Enzymol*, 317 :457–70, Jan 2000.

[13] M Schaefer, C Bartels, **F Leclerc**, and M Karplus. Effective Atom Volumes for Implicit Solvent Models : Comparison between Voronoi Volumes and Minimum Fluctuation Volumes. *J Comp Chem*, 22(15) :1857–1879, Nov 2001.

[14] N Marmier-Gourrier, A Cléry, V Senty-Ségault, B Charpentier, F Schlotter, **F Leclerc**, R Fournier, and C Branlant. A structural, phylogenetic, and functional study of 15.5-kD/Snu13 protein binding on U3 small nucleolar RNA. *RNA*, 9(7) :821–38, Jul 2003.

[15] S Y Reddy, **F Leclerc**, and M Karplus. DNA Polymorphism : A Comparison of Force Fields for Nucleic Acids. *Biophys J*, 84(3) :1421–49, Mar 2003.

[16] A Lambert, J-F Fontaine, M Legendre, **F Leclerc**, E Permal, F Major, H Putzer, O Delfour, B Michot, and D Gautheret. The ERPIN server : an interface to profile-based RNA motif identification. *Nucleic Acids Res*, 32(Web Server issue) :W160–5, Jul 2004.

[17] **F Leclerc** and M Karplus. Two-Metal-Ion Mechanism for Hammerhead-Ribozyme Catalysis. *J Phys Chem B, Condensed matter, materials, surfaces, interfaces & biophysical*, 110(7) :3395–409, Feb 2006.

[18] X Lopez, A Dejaegere, **F Leclerc**, D M York, and M Karplus. Nucleophilic Attack on Phosphate Diesters : A Density Functional Study of In-Line Reactivity in Dianionic, Monoanionic, and Neutral Systems. *J Phys Chem B, Condensed matter, materials, surfaces, interfaces & biophysical*, 110(23) :11525–39, Jun 2006.

[19] S Muller, B Charpentier, C Branlant, and **F Leclerc**. A Dedicated Computational Approach for the Identification of Archaeal H/ACA sRNAs. *Meth Enzymol*, 425 :355–87, Jan 2007.

[20] A Cléry, V Senty-Ségault, **F Leclerc**, H A Raué, and C Branlant. Analysis of Sequence and Structural Features that Identify the B/C Motif of U3 Small Nucleolar RNA as the Recognition Site for the Snu13p-Rrp9p Protein Pair. *Mol Cell Biol*, 27(4) :1191–206, Feb 2007.

[21] S Muller, **F Leclerc**, I Behm-Ansmant, J Fourmann, B Charpentier, and C Branlant. Combined *in silico* and experimental identification of the *Pyrococcus abyssi* H/ACA sRNAs and their target sites in ribosomal RNAs. *Nucleic Acids Res*, Feb 2008.

[22] S Muller, A Urban, A Hecker, **F Leclerc**, C Branlant, and Y Motorin. Deficiency of the $tRNA^{Tyr}$:Ψ35-synthase aPus7 in Archaea of the Sulfolobales order might be rescued by the H/ACA sRNA-guided machinery. *Nucleic Acids Res*, in press, 2008.

Annexe B

Publications significatives

Sommaire

Theor Chem Acc (1999) 101:131–137
DOI 10.1007/s002149800m85

Theoretical
Chemistry Accounts

Regular article

MCSS-based predictions of RNA binding sites*

Fabrice Leclerc[1], **Martin Karplus**[1,2]

[1] Harvard University, Department of Chemistry and Chemical Biology, 12 Oxford Street, Cambridge, MA 02138, USA
[2] Université de Strasbourg I, Institut le Bel, Laboratoire de Chimie Biophysique, F-67000 Strasbourg, France

Received: 24 April 1998 / Accepted: 4 August 1998 / Published online: 7 December 1998

Abstract. The diversity of RNA tertiary structures provides the basis for specific recognition by proteins or small molecules. To investigate the structural basis and the energetics which control RNA-ligand interactions, favorable RNA binding sites are identified using the MCSS method, which has been employed previously only for protein receptors. Two different RNAs for which the structures have been determined by NMR spectroscopy were examined: two structures of the TAR RNA which contains an arginine binding site, and the structure of the 16S rRNA which contains an aminoglycoside binding site (paromomycin). In accord with the MCSS methodology, the functional groups representing the entire ligand or only part of it (one residue in the case of the aminoglycosides) are first replicated and distributed with random positions and orientations around the target and then energy minimized in the force field of the target RNA. The Coulombic term and the dielectric constant of the force field are adjusted to approximate the effects of solvent-screening and counterions. Optimal force field parameters are determined to reproduce the binding mode of arginine to the TAR RNA. The more favorable binding sites for each residue of the aminoglycoside ligands are then calculated and compared with the binding sites observed experimentally. The predictability of the method is evaluated and refinements are proposed to improve its accuracy.

Key words: MCSS – Ligand Design – RNA – Binding

1 Introduction

Nucleic acids make logical targets for drug design, since all enzymes and receptor proteins depend on RNA for their synthesis. Unlike DNA, RNA may fold into complex and diverse molecular shapes which constitute attractive targets for specific and selective binding. A number of X-ray or NMR structures of RNA and its complexes with drugs, peptides, or proteins have been recently determined and provide the opportunity to better understand RNA-ligand interactions [6]. Nevertheless, only a few studies on modeling RNA-ligand interactions have been published [7, 8].

A primary step towards the design of drugs against macromolecule receptors is the identification of potential binding sites for ligand fragments (functional groups). Such an approach has been applied in computational and laboratory combinatorial ligand design for protein receptors [2, 9–11]. The TAR RNA of HIV-1 and the bacterial 16S rRNA, for which the structures have been determined by NMR spectroscopy [3–5], represent two interesting targets because of their biological role in the regulation of the HIV cycle and in the protein synthesis of pathogenic bacteria, respectively. They are used here as macromolecule targets with the MCSS method to predict favorable binding sites for arginine in the case of the TAR RNA and for aminoglycoside moieties in the case of the 16S rRNA. The MCSS method [1] is divided into two steps. The first step involves the replication of pre-defined functional groups and their distribution in random positions and orientations in a binding region delimited by a spherical or rectangular boundary. The second step involves their energy minimization in the force field of the receptor [1], and the selection of the local minima based on an energy cutoff. The search for minima is performed by an iterative process. Details of the method are given by Miranker and Karplus [1].

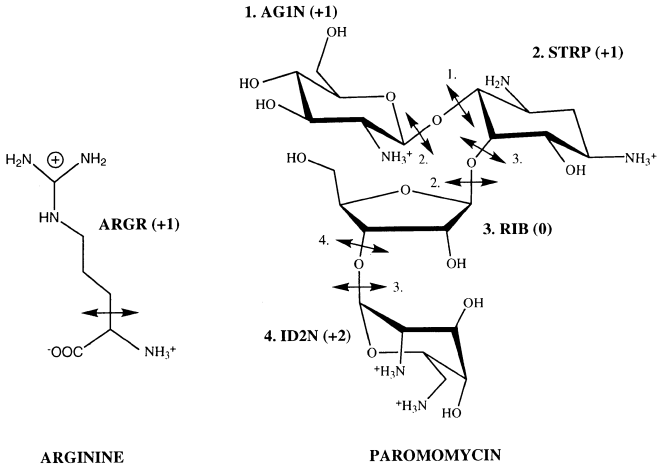
2 Specific approach

Two sets of force field parameters were used to identify and select the MCSS minima. In the first set (Model 1), the non-bonded interactions between the replicated functional groups (replica) and the receptor were calculated based on the CHARMM force field using the recent parameters for nucleic acids [12]. Since these parameters are designed for use with explicit solvents

* Contribution to the Proceedings of Computational Chemistry and the Living World, April 20–24, 1998, Chambery, France

Correspondence to: M. Karplus
e-mail: marci@brel.u-shasbg.fr

and the calculations are done in vacuo for the sake of efficiency, a modified set (Model 2) was derived from the first one by scaling down the atomic charges on the phosphate groups to mimic the presence of counterions [13]. In both sets, the solvent screening is modeled using a distance-dependent dielectric $\epsilon(r) = (c \cdot r)$ where c varies from 1 to 4. The search for MCSS minima is first carried out in a restricted binding region defined by a sphere of 12 Å of radius centered on the position of the bound arginine residue. A more extensive search is then performed in a box including the whole RNA. In the case of the 16S rRNA, the first search is carried in a box centered on the aminoglycoside binding site (8250 $Å^3$). The MCSS functional group for the arginine side-chain is represented by a polar hydrogen model [1, 14]. The four MCSS functional groups corresponding to the individual residues of paromomycin are represented with the all-hydrogen model. The chemical structures of the RNA ligands, arginine, and paromomycin are shown in Scheme 1; the functional groups are separated by arrows (the number attached to the arrow indicates where the bond is broken for the corresponding residue) and the net charge per group is indicated in parentheses.



Scheme 1.

The minima are sorted based on the energy of interaction, which is defined by the sum of the van der Waals and electrostatic contributions. The improved MCSS program (version 2.1) developed by Erik Evensen (unpublished), a complete reimplementation of the original MCSS method with increased efficiency, flexibility, and ease-of-use, and the CHARMM program [15] are used for the calculations.

3 Results

The strong arginine binding site of the TAR RNA is used to calibrate the method by adjusting the force field parameters to account for the implicit presence of counterions and/or the screening effect of solvent. The two classes of model were tested to reproduce the position of the arginine side-chain in the TAR RNA binding site. For Model 1 the MCSS minima are closer to the phosphate backbone, while they are closer to the nucleic acid bases for the Model 2. In both models, the increase of the dielectric constant produces more favorable MCSS minima in the known arginine binding site. Model 2, which takes into account both the implicit presence of counterions and the screening effect of solvent ($c = 3$), gave the minima with the best score and the lowest root mean square deviation (RMSD) with respect to the position of the arginine residue in the NMR structure (see Table 1). In what follows, we describe the results for Model 2.

The results are presented in Fig. 1: the MCSS minima identified in a restricted searched region around the arginine binding site (Fig. 1A) and on the entire RNA surface (Fig. 1B). The minima exhibit three different binding modes: the first one involves arginine fork-like structures with nucleic acid bases (guanine) [4], the second one corresponds to non-specific interactions with phosphate groups, and the last one to stacking interactions with the base of bulge nucleotides

Table 1. Prediction accuracy of RNA binding sites[a]

RNA	Ligand	MCSS functional group	Model 1			Model 2		
			RMSD	Score	Occupancy[b]	RMSD	Score	Occupancy
HIV-1 TAR[c]	Arginine	Arginine side-chain	0.6 Å	100	7%	0.6 Å	100	8%
HIV-1 TAR[d]	Arginine	Arginine side-chain	1.1 Å	100	6%	1.0 Å	100	15%
16S rRNA	Paromomycin	1st residue (AG1N)	1.0 Å	65	1%	1.0 Å	76	3%
	Paromomycin	2nd residue (STRP)	≥4	≤1	–	2.8 Å	1.2	–
	Paromomycin	3rd residue (RIB)	≥4	≤1	–	≥4	≤1	–
	Paromomycin	4th residue (ID2N)	0.3 Å	74	2%	0.3 Å	100	4%
	Paromomycin	Residues 1 and 2 (NEA2)	–	–	–	0.3 Å/1.5 Å	34/92	3%

[a] The prediction accuracy is evaluated by two criteria: the root mean square deviation (RMSD) and the score. The RMSD is measured for all non-hydrogen atoms between the positions of the MCSS minima and the ligands observed experimentally. The score (min = 0, max = 100) is given by the following expression: $100(1 - |(E_{max} - E_i)/(E_{max} - E_{min})|)$, where E_{max} and E_{min} are the maximum and minimum energies of interaction, respectively, and E_i the energy of interaction for the corresponding MCSS minimum. A score of 100 indicates that the minimum has the best energy of interaction

[b] The occupancy is defined as the proportion of MCSS minima similar to the ligand observed experimentally (RMSD ≤ 1.5 Å)

[c] HIV-1 TAR RNA structure determined by Puglisi et al. [4]

[d] HIV-1 TAR RNA structure determined by Aboul-ela et al. [3]

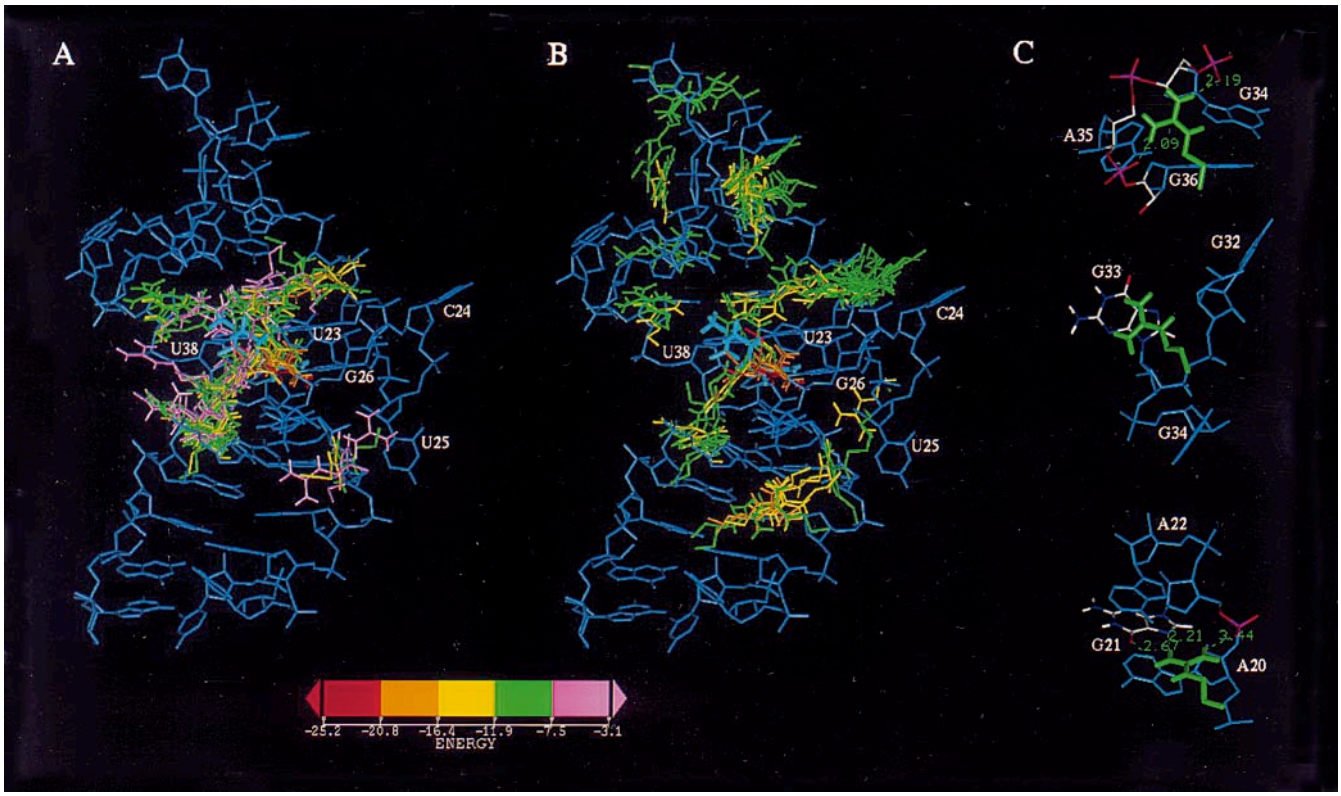


Fig. 1A–C. Distribution of MCSS minima at the surface of the HIV-1 TAR RNA [3]. **A** Minima identified within a sphere of 12 Å centered on the structure of the arginine residue. **B** Minima identified on the entire surface of the RNA. The RNA target, colored in *blue*, corresponds to the bound TAR conformation; its bound arginine is colored in *light blue*. The score of each minima is given by the color scheme at the bottom representing the energy of interaction. **C** Detail of the three arginine binding modes (*top*: interactions with the phosphate backbone; *middle*: stacking interactions with the bulge nucleotide G33; *bottom*: arginine fork motif). The hydrogen bonds are represented by *dashed lines*. The program InsightII was used for the graphical representation (MSI, San Diego)

(Fig. 1C). The minima with high scores largely overlap with the arginine residue in the NMR structure of the complex. They also correspond to the more favorable binding sites on the entire surface of the RNA. The details of the interaction between the minima with the best score and the RNA binding site are represented on Fig. 2. The best minima (with the 10 highest scores) reproduce accurately the position of the arginine side-chain in the two NMR structures of the arginine-TAR RNA complexes, but with a lower RMSD in the case of the structure determined by Puglisi et al. [3, 4] (see Table 1 and Fig. 2).

The distribution of the minima corresponding to the four paromomycin residues are shown in Fig. 3A and 3B. The global distribution for the AG1N (first residue), STRP (second residue), RIB (third residue), and ID2N (fourth residue) groups is similar: for all the residues (Fig. 3C), independently of the charge and shape, three different favorable binding regions are found. The minima with high scores are concentrated mostly in one region (Fig. 3B). The position of the best AG1N and ID2N minima are in good agreement with that of the first and fourth residues of paromomycin (Fig. 4); the minima for the STRP and RIB groups, which share some overlap with the actual positions of the corresponding paromomycin residues, have very low scores. These latter could be due to the fact that these residues occupy a sub-optimal position in the binding site as part of the paromomycin molecule. To test this hypothesis, an additional search was performed using a larger functional group (NEA2) including both the first and second residues in order to determine the more favorable positions of the STRP group when merged to AG1N (see Scheme 1). The results demonstrate that the binding site of the second residues can be better predicted in the context of a larger functional group (see Table 1). In the case of aminoglycoside moieties, the computational cost of the MCSS search remains linearly dependent on the size of the functional group (on a DEC 3000 AXP 500, around 4 h of computer time for the AG1N and STRP groups and around 10 h of computer time for NEA2). A general view of the paromomycin binding site and the positions of some minima are shown in Fig. 5. The accuracy of the predicted binding sites for each functional group is summarized in Table 1.

4 Discussion

We show that a simple model that makes use of the MCSS method with a modified CHARMM force field (reduced phosphate charges) can simulate RNA-ligand interactions and predict RNA binding sites. The predic-

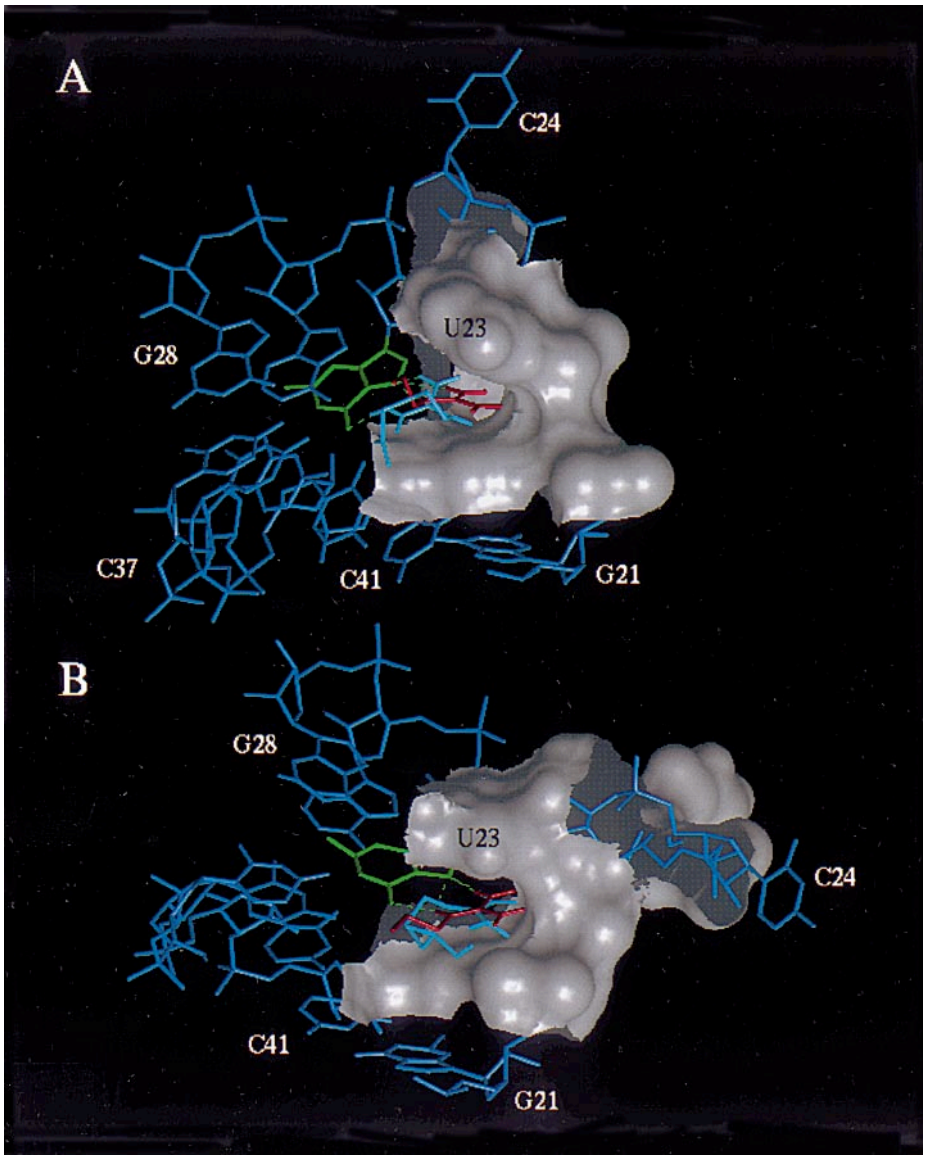


Fig. 2A, B. Close-up view of the MCSS minima within the TAR arginine binding site. The structure of the arginine-RNA complex is colored as in Fig. 1 (RNA in *blue*, arginine in *black*). The binding site is represented by a solvent accessible surface. **A** NMR structure by Aboul-ela et al. [3]. **B** NMR structure by Puglisi et al. [4]. The MCSS minimum with the highest score is shown in *red*, the guanine forming the arginine fork motif is in *green*

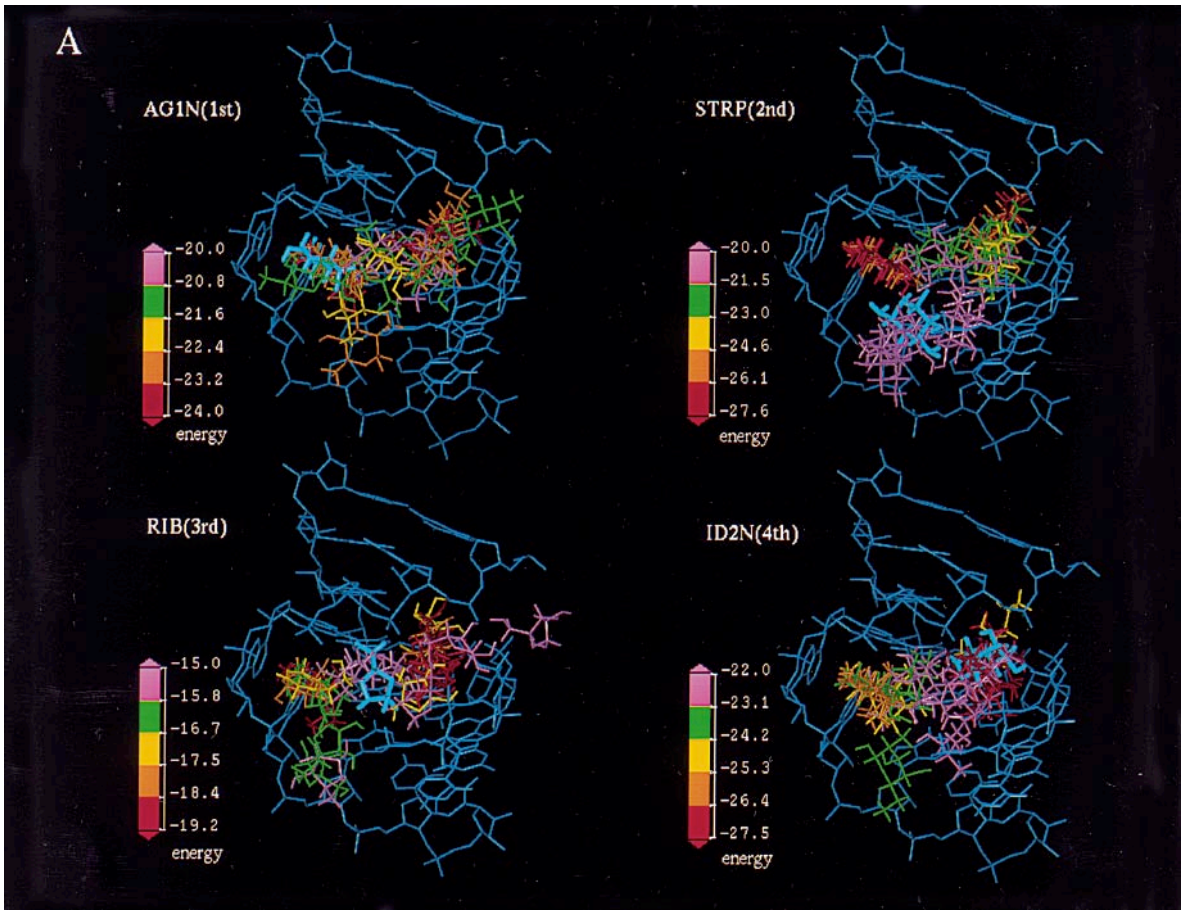
tion accuracy is dominated by the geometry and electrostatic properties of the RNA binding site.

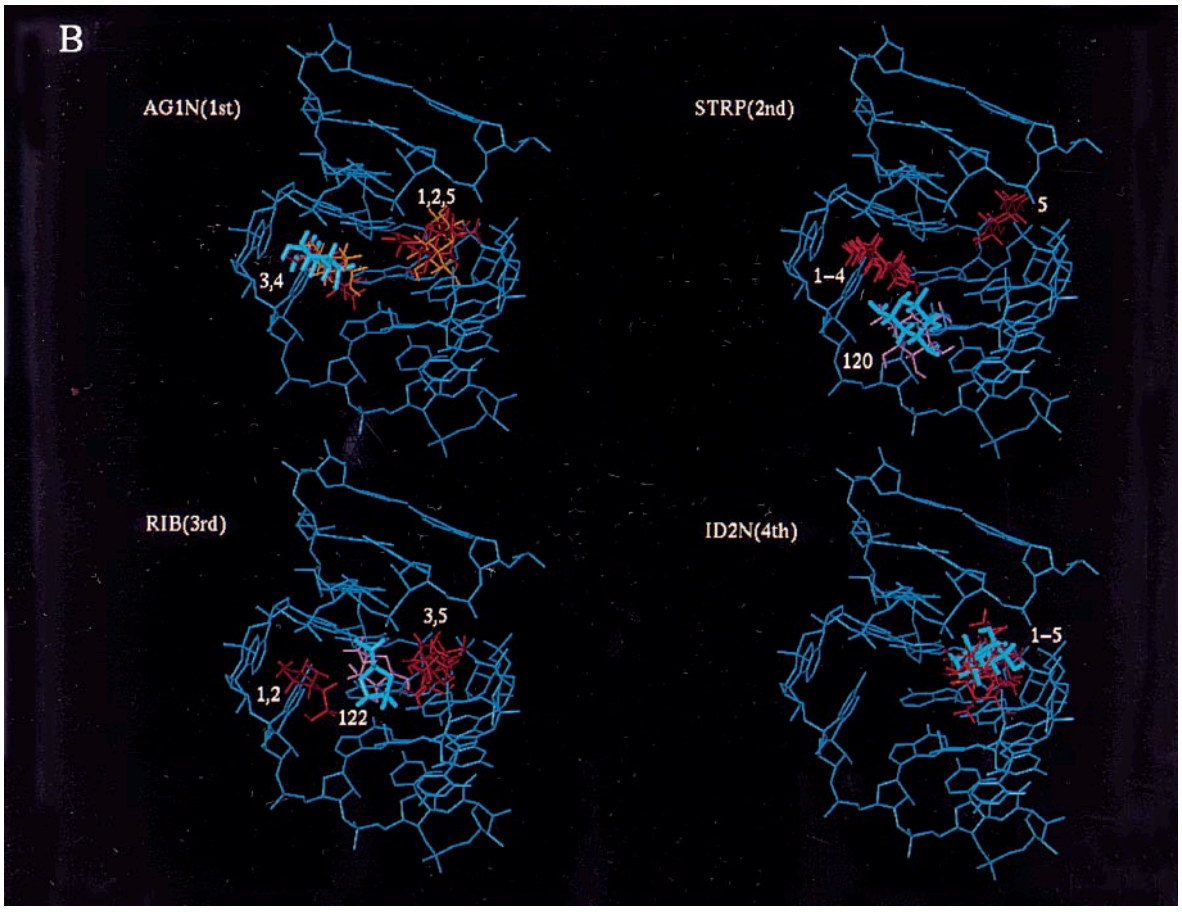
In the case of the TAR RNA, the accuracy is sensitive to the NMR structure used as the target (see differences in RMSD and score in Table 1). In the first TAR RNA structure [4], the bulge nucleotide U23 forms a base-triple with the Watson-Crick base pair A27 · U38, a structural feature absent from the second structure [3]. The presence of the base triple restrains the positions of the MCSS minima to a well-defined cavity where the guanidinium group of the arginine residue is perfectly stacked over the bulge nucleotide U23 and paired to the guanine G26 according to the arginine fork motif [4] (see Fig. 2). In the second structure [3], the MCSS minima tend to stack over U23 close to the phosphate backbone and are more poorly paired to G26.

In the case of the 16S rRNA, the predictions depend strongly on the geometry and electrostatic properties of the RNA binding site. Independently of the functional group used to represent one of the four residues of paromomycin, two regions of the RNA appear to be very favorable for binding. One binding region is a well-defined cavity in which the first residue of paromomycin fits by stacking interactions with the nucleotide A1492 (see Fig. 5). This region is also an attractive pole for monocationic residues like the first and second residue of paromomycin (see Fig. 3A, B). The second region is also

►

Fig. 3. **A** Distribution of MCSS minima at the surface of the 16S rRNA. The four panels represent the distribution for each of the functional groups corresponding to the four paromomycin residues (first residue: AG1N; second residue: STRP; third residue: RIB; fourth residue: ID2N) colored in *light blue*. A color scheme indicates the score of the minima for each functional group. **B** Distribution of high score MCSS minima. The five minima with the highest scores are shown for each functional group. The minima closer to the corresponding paromomycin residue are minimum 4 for the first residue (AG1N), minimum 122 for the second residue (STRP), minimum 122 for the third residue (RIB), and minima 1 and 2 for the fourth residue (ID2N)


A
AG1N(1st)
-20.0
-20.8
-21.6
-22.4
-23.2
-24.0
energy
STRP(2nd)
-20.0
-21.5
-23.0
-24.6
-26.1
-27.6
energy
RIB(3rd)
-15.0
-15.8
-16.7
-17.5
-18.4
-19.2
energy
ID2N(4th)
-22.0
-23.1
-24.2
-25.3
-26.4
-27.5
energy


B
AG1N(1st)
1,2,5
3,4
STRP(2nd)
5
1–4
120
RIB(3rd)
3,5
1,2
122
ID2N(4th)
1–5

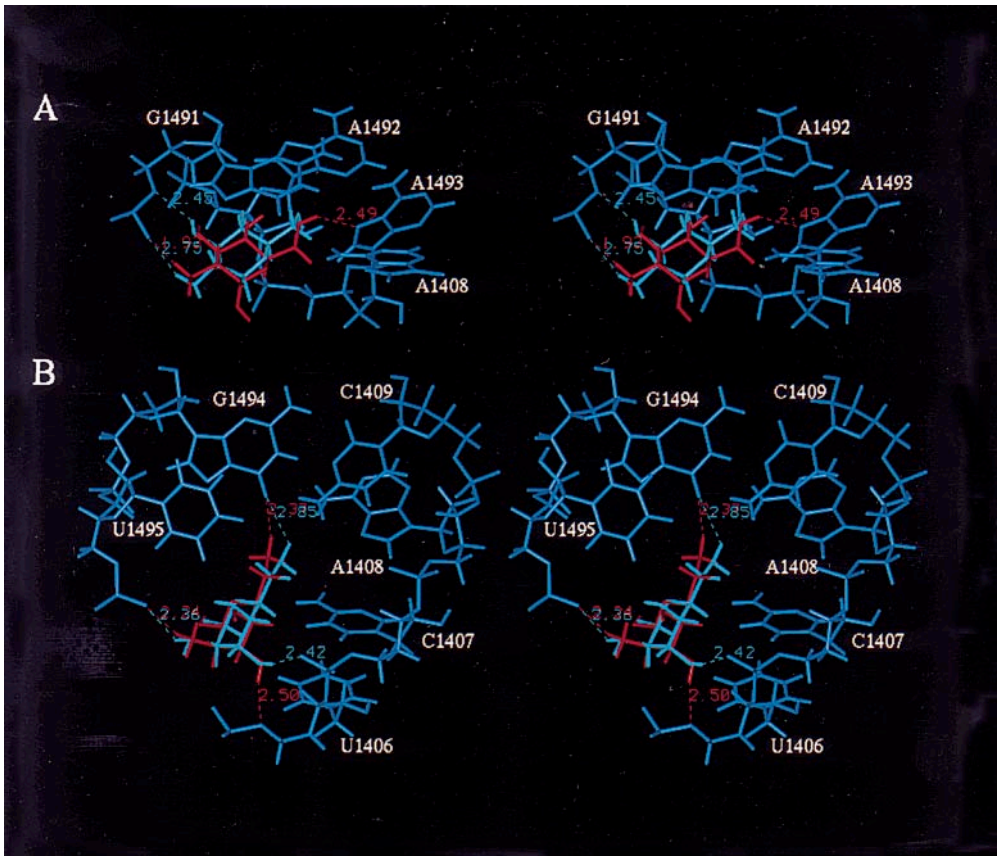
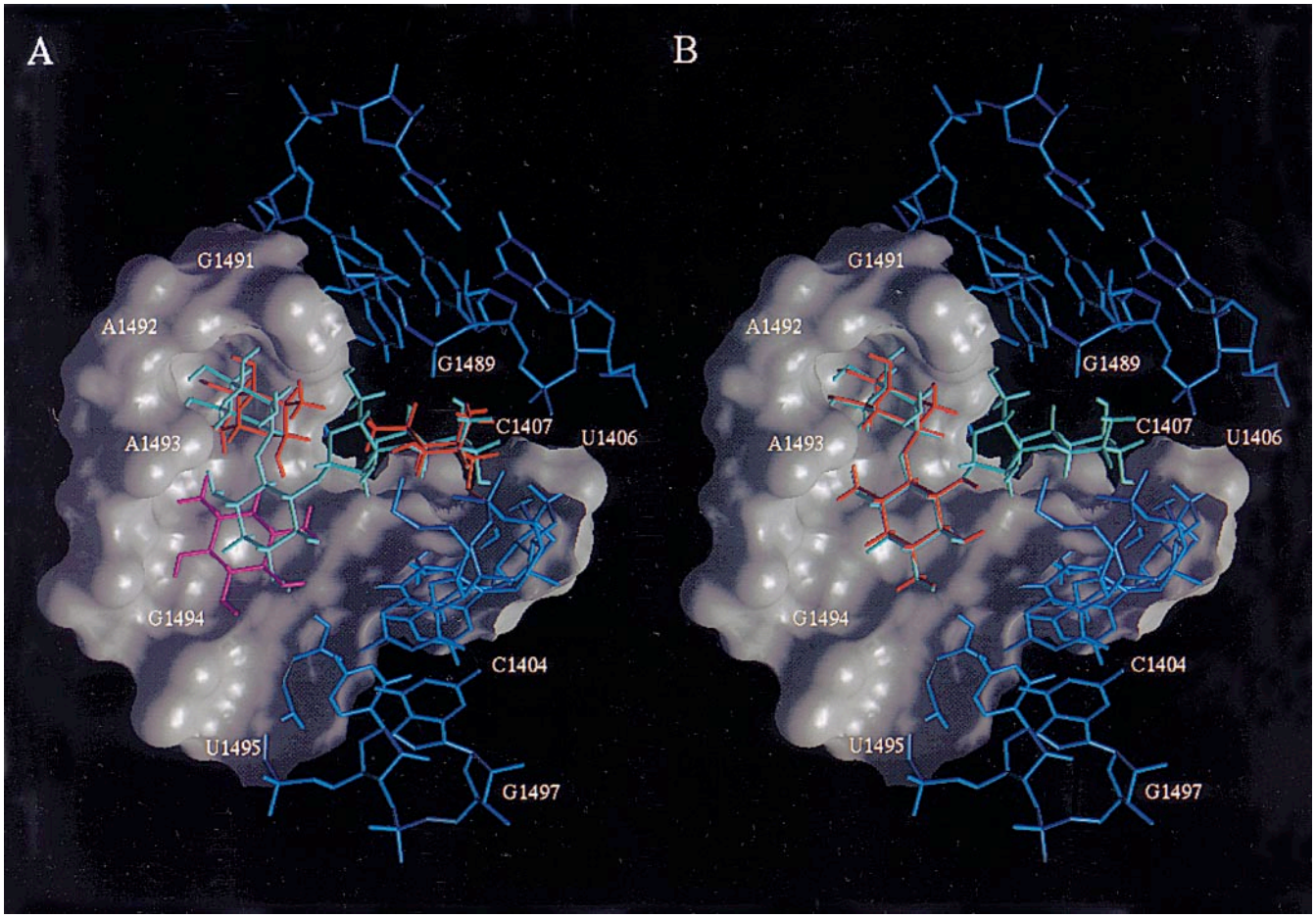


Fig. 4A, B. Close-up view of the MCSS minima for the AG1N and ID2N groups in the paromomycin binding site. **A** MCSS minimum for the AG1N group. **B** MCSS minimum for the ID2N group. The MCSS minima are colored in *red*, the corresponding residues in the RNA-aminoglycoside complex in *light blue*. The hydrogen bonds between the RNA and the ligands (distance X-H $\leq$3.0 Å where X = O, N) are indicated by *dashed lines* according to the same color scheme. The MCSS minima represented correspond to those described in Table 1



a very attractive electrostatic pole for mono- or dicationic groups because of the proximity of two portions of the RNA backbone (the phosphates of nucleotides G1489 and C1407 in Fig. 5). The decomposition of the binding free energy on a per residue basis indicates that the group for which the binding position and conformations are predicted best corresponds to the residue that contributes the most to the free energy of binding (results not shown), in agreement with recent experimental data [16].

5 Conclusions

A preliminary study has been made which indicates that the MCSS method can be used to investigate functional group binding to nucleic acids. To improve the reliability of the predictions, models which give a more accurate description of the solvent and polyelectrolyte effects on RNA-ligand interactions are being evaluated. Reliable predictions of known RNA binding sites will open new perspectives for the application of combinatorial structure-based drug design to these molecules, which are emerging as attractive drug targets.

Acknowledgements. This work was supported in part by a grant from the Department of Energy to Harvard University.

◀

Fig. 5A, B. General view of the MCSS minima within the paromomycin binding site. The structure of the paromomycin-16S rRNA complex is colored as in Fig. 3 (RNA in *blue*, paromomycin in *light blue*). The binding site is represented by a solvent accessible surface. **A** The minima 1 and 4 (AG1N and ID2N) corresponding to the first and fourth residues of paromomycin are shown in *red*; the minimum 122 (STRP) corresponding to the second residue of paromomycin is shown in *pink* according to the score. **B** The minimum with the lowest RMSD (see Table1) corresponding to the residues 1 and 2 (NEA2) is shown in *red*

References

1. Miranker A, Karplus M (1991) Proteins 11:29
2. Caflisch A, Miranker A, Karplus M (1993) J Med Chem 36:2142
3. Aboul-ela F, Karn J, Varani G (1995) J Mol Biol 253:313
4. Puglisi JD, Tan R, Calnan BJ, Frankel AD, Williamson JR (1992) Science 257:76
5. Fourmy D, Recht MI, Blanchard SC, Puglisi JD (1996) Science 274:1367
6. Ramos A, Gubser CC, Varani G (1997) Curr Opin Struct Biol 7:317
7. Veal JM, Wilson WD (1991) J Biomol Struct Dyn 8:1119
8. Leclerc F, Cedergren R (1998) J Med Chem 41:175
9. Gordon EM, Barrett RW, Dower WJ, Fodor SP, Gallop MA (1994) J Med Chem 37:1385
10. Shuker SB, Hajduk PJ, Meadows RP, Fesik SW (1996) Science 274:1531
11. Caflisch A (1996) J Comput Aided Mol Des 10:372
12. MacKerell AD Jr, Wiorkiewicz-Juczera J, Karplus M (1995) J Am Chem Soc 117:11946
13. Tidor B, Irikura KK, Brooks BR, Karplus M (1983) J Biomol Struct Dyn 1:231
14. Neria E, Fischer S, Karplus M (1996) J Chem Phys 105:1902
15. Brooks BR, Bruccoleri RE, Olafson D, States DJ, Swaminathan A, Karplus M (1983) J Comput Chem 4:187
16. Alper PB, Hendrix M, Sears P, Wong C-H (1998) J Am Chem Soc 120:1965

DNA Polymorphism: A Comparison of Force Fields for Nucleic Acids

Swarnalatha Y. Reddy,* Fabrice Leclerc,*†‡ and Martin Karplus*‡

*Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138 USA; †Laboratoire de Maturation des ARN et Enzymologie Moléculaire, CNRS-UHP Nancy I UMR 7567, Université Henri Poincaré, Faculté des Sciences, B.P. 239, 54506 Vandoeuvre-lès-Nancy, France; and ‡Laboratoire de Chimie Biophysique, ISIS, Université Louis Pasteur, 4 rue Blaise Pascal, 67000, Strasbourg, France

ABSTRACT The improvements of the force fields and the more accurate treatment of long-range interactions are providing more reliable molecular dynamics simulations of nucleic acids. The abilities of certain nucleic acid force fields to represent the structural and conformational properties of nucleic acids in solution are compared. The force fields are AMBER 4.1, BMS, CHARMM22, and CHARMM27; the comparison of the latter two is the primary focus of this paper. The performance of each force field is evaluated first on its ability to reproduce the B-DNA decamer d(CGATTAATCG)$_2$ in solution with simulations in which the long-range electrostatics were treated by the particle mesh Ewald method; the crystal structure determined by Quintana et al. (1992) is used as the starting point for all simulations. A detailed analysis of the structural and solvation properties shows how well the different force fields can reproduce sequence-specific features. The results are compared with data from experimental and previous theoretical studies.

INTRODUCTION

Nucleic acids can adopt different conformations in solution depending on the base composition (Hunter, 1993) and the environment (for example pH and temperature, Kumar and Maiti, 1994), including the nature of the solvent (Fang et al., 1999), the counterions (Minasov et al., 1999), their concentration (Ali and Ali, 1997), and interactions with proteins (Jones et al., 1999), or small molecules (Reinert, 1999). Even a given sequence of DNA or RNA can exhibit multiple conformations (Kielkopf et al., 2000). In living systems, the conformational flexibility of DNA resides primarily in the polymorphs of the DNA double helix (including right-handed and left-handed double-helical DNA) that occur under various experimental conditions (Gupta et al., 1980). By contrast, double-stranded helical RNA is confined to two very similar polymorphs of the A form (A and A′), and the wide range of single-stranded nonhelical RNA folds introduces the essential structural variability.

Significant progress in the development of empirical force fields and molecular dynamics (MD) simulation methods has lead to a more reliable description of the structure, energetics, and dynamics of nucleic acids (Auffinger and Westhof, 1998; Beveridge and McConnell, 2000; Cheatham and Kollman, 2000; Cheatham and Young, 2001). However, some limitations related to the improper treatment of the equilibrium between the A and B forms of DNA (Feig and Pettitt, 1997, 1998) and the deviations of helicoidal parameters from canonical B values (Cheatham and Kollman, 1996) have been reported. The overstabilization of the A form relative to the B form of DNA (Yang and Pettitt, 1996; MacKerell, 1997; Feig and Pettitt, 1997) with the CHARMM22 force field (MacKerell et al., 1995) has been addressed in a recent reoptimization of the CHARMM22 all-atom nucleic acid force field. The new nucleic acid force field, called CHARMM27, has small but important changes in both the internal and interaction parameters relative to CHARMM22 (Foloppe and MacKerell, 2000) and appears to treat well the equilibrium between the A and B forms of DNA and the influence of the environment, such as the water activity (MacKerell and Banavali, 2000). A revised and improved version of the AMBER4.1 force field has also been presented that shows better agreement with experimental data as a result of the adjustment of internal force field parameters (Cheatham et al., 1999). An alternative nucleic acid force field, which we refer to as the Bristol-Myers Squibb (BMS) force field, has been developed by Langley (Langley, 1998). Both the CHARMM27 and AMBER force field parameters are based on the reproduction of experimental results for nucleic acid oligomers (e.g., condensed phase structural properties of DNA and RNA) and consistency with small molecule results obtained from quantum mechanical calculations and experimental data. The BMS force field was developed, in part, by adaptation of the CHARMM22 (MacKerell et al., 1995), QUANTA and AMBER force fields (Cornell et al. 1995. The backbone angle and dihedral parameters were derived from quantum mechanical calculations with refinements based on a series of MD simulations. All the force fields used condensed-phase MD simulations in the final stage of the parameter optimization. The CHARMM27 force field has also been applied to model compounds to evaluate the contributions from the individual moieties to the overall conformational properties of DNA and RNA (Foloppe and MacKerell, 2000).

Recent simulations of nucleic acids using an explicit solvent representation and an ionic environment have led to high structural stability on the nanosecond time scale (Beveridge and McConnell, 2000). This accuracy was

Submitted March 15, 2002, and accepted for publication August 6, 2002.

Address reprint requests to Martin Karplus, Dept. of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138 USA. Tel.: 617-495-1768; Fax: 617-496-3204; E-mail: marci@tammy.harvard.edu.

0006-3495/03/03/1421/29 $2.00

achieved due to improvements of the force fields, as described above, the inclusion of an appropriate number of counterions, and the use of the Ewald method for the long-range electrostatic interactions after 1995 (Cheatham et al., 1995; Lee et al., 1995; York et al., 1995). The implementation of the particle-mesh-Ewald (PME) method (Feller et al., 1996), which is faster than standard Ewald method, allows accurate treatment of long-range electrostatic interactions while preserving a reasonable simulation time. It is becoming a standard in nucleic acid simulations, although some simulations with the CHARMM27 force field suggest that long-range electrostatic interactions can also be treated using cutoffs (Norberg and Nilsson, 1996). In addition to overall stability, force-field-based simulations should also reproduce the sequence-dependent structure variations in DNA, as manifested by the local backbone conformation and basepair geometry at different basepair steps, and the hydration patterns associated with these structural variations. In this paper, we present the results of molecular dynamics simulations that assess the ability of the CHARMM force fields, CHARMM22 (MacKerell et al., 1995) and CHARMM27 (Foloppe and MacKerell, 2000), to address the structure and dynamics of DNA in aqueous solution. For comparison, the performance of two additional force fields, the AMBER4.1 (Cornell et al., 1995) and BMS (Langley, 1998) force fields, is also examined. Our goal is to evaluate the ability of these force fields to address sequence-dependent aspects of the structure of DNA duplexes and DNA hydration. Simulations are carried out for a B-DNA decamer with a central TpA step, d(CGATTAATCG)$_2$ (Quintana et al., 1992). This sequence is of particular interest because its structure, obtained at very high resolution (1.5 Å), shows certain twisting and bending properties. The minor groove is wide at the central TpA step rather than narrow, and the twist angle of the TpA step is small rather than large, contrary to other sequences with a central TTAA tetramer. The presence of a Mg^{2+} cation bound to DNA at the TpT step probably contributes to this local widening of the minor groove (Fig. 1). These properties appear to confer a greater possibility of deformation that could be exploited for sequence recognition by drugs and by proteins (Quintana et al., 1992; Goodsell et al., 1994).

The DNA structures generated during the simulations are analyzed in terms of global structural parameters, such as the DNA form and the size of the major and minor grooves, and local structural parameters such as sugar pucker, phosphate backbone conformation, and basepair geometry. The relative conformational flexibility arising from the different force fields is determined by the fluctuations of these parameters during the dynamics. The DNA hydration and distribution of counterions is analyzed and compared with high resolution x-ray data (Egli et al., 1998; Tereshko et al., 1999). In the manuscript, (in preparation) on the structural and hydration changes in the transition from A- to B-DNA, it is shown that the DNA decamer, starting from the canonical A-DNA form

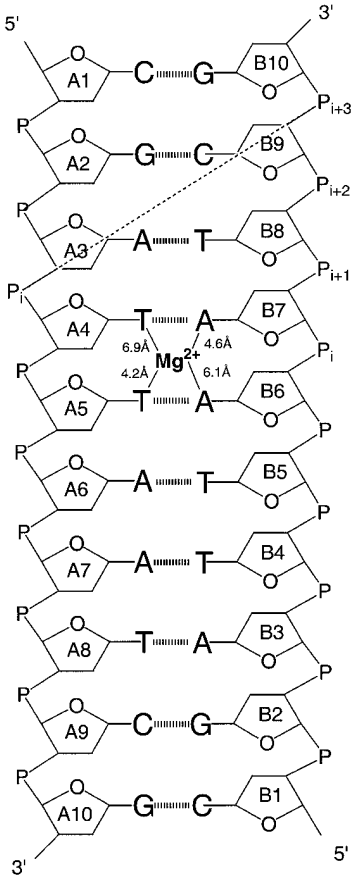


FIGURE 1 Secondary structure and numbering of the DNA decamer d(CGATTAATCG)$_2$ (Quintana et al., 1992). The position of the Mg^{2+} binding site, in the minor groove, is indicated with the distances to O2 and N3 atoms of thymines and adenines, respectively. A dashed line joins the P_i and P_{i+3} phosphorus atoms on opposing strands. The corresponding distances are used as measure of the minor groove width.

with the CHARMM27 force field, can undergo a quick transition to the B form (in a simulation time as short as 1.4 ns). This contrasts with the CHARMM22 force field, which overstabilizes the A form. The A to B transition is analyzed to obtain an understanding of the contributing factors. It is shown that the internal motions and hydration of DNA are both involved in the transition.

METHOD

Molecular simulations

Simulations were performed with the CHARMM program (Brooks et al., 1983), in the constant NVT ensemble at 298 K. The leapfrog Verlet (Verlet, 1967) integration scheme was used with a 2-fs time step and SHAKE applied to all covalent bonds involving hydrogens (Ryckaert et al., 1977). Images were generated using the CRYSTAL module (Field and Karplus, 1992) in CHARMM. The different force fields used are all-atom force fields (both polar and nonpolar hydrogen are included) implemented in the CHARMM program. In all the MD simulations, electrostatic interactions are treated with the Ewald method (Ewald, 1921) as implemented in the PME formulation (Darden et al., 1993; Petersen, 1995); the latter was implemented in the

CHARMM program by Feller et al. (1996). PME calculations were performed using a real space cutoff of 9 Å with Lennard-Jones interactions truncated at the same distance. A dielectric constant of unity was used. A convergence parameter (κ) of 0.32 $Å^{-1}$ and a fifth degree B-spline interpolation was employed with the PME method. In the case of the AMBER force field, explicit scaling of 1-4 electrostatic function was applied, as recommended for nucleic acid simulations (Cheatham et al., 1995).

To neutralize the DNA molecule, 18 Na^+ ions were introduced with each one initially placed at a distance of 6.0 Å from the phosphorous atom on the perpendicular bisector defined by the phosphorous and nonbridging oxygen atoms; the ions were fixed in the first stage of the simulation to allow the water molecules to relax around the DNA and counterions. The model was minimized for 1000 steps with the Adopted Basis Newton Raphson (ABNR) method. Periodic boundary conditions were defined using an orthorhombic box (36.0 Å × 40.0 Å × 46.5 Å) filled with TIP3P model water molecules (Jorgensen et al., 1983) so that the minimum thickness of the solvation shell around the DNA and counterions centered in the box was 5 Å. The water molecules in the box were minimized for 400 steps of the Steepest Descent (SD) method. The box was then overlaid onto the system of the DNA decamer model with the sodium ions. Solvent molecules with the oxygen within 2.7 Å of any DNA nonhydrogen atom or any sodium ion were deleted. The solvent was minimized for 100 steps of SD followed by 1000 steps of ABNR, keeping the DNA and ions fixed. After that, the entire model was minimized for 2000 steps with the ABNR method before starting the simulations.

During the equilibration, the structure was relaxed in stages, so that the most strained parts of the system could adjust without introducing artifacts. Harmonic constraints with a force constant of 1 kcal/mol-Å were used. Initially, the DNA and Na^+ ions were constrained and only water molecules were allowed to move. The water molecules were simulated at 298 K for 40 ps. The constraints on the ions were released and the water and ions were reheated gradually from 0 to 298 K at increments of 100 K, each for 20 ps for a total of 100 ps. These stages were carried out in the NPT ensemble ($T =$ 298 K, $P = 1$ atm), so that the water box could equilibrate in accord with the number of 1947 water molecules included in the simulation system. The dimensions of water box were allowed to vary only along the z-axis, because the DNA molecule is oriented in that direction. During the course of the100 ps simulation, the box dimension fluctuated by 2–3 Å from the initial value. Then, the constraints on DNA were removed, and it was allowed to move along with the ions and water molecules, except that NOE-like distance constraints on the terminal basepairs of DNA between the heavy atoms involved in hydrogen bonds were introduced. The distance constraints applied with a force constant of 10 kcal/mol-Å correspond to the distances observed in the x-ray structure with an allowed deviation of ~0.2 Å. For subsequent simulations, the NVT ensemble was used, as it provides more stable trajectories (Brown and Clarke, 1984). The system was reequilibrated by heating the entire model at increments of 50 K for 20 ps each, from 0 to 298 K. Then a 30-ps simulation was run at 298 K to equilibrate the entire system at this temperature. The heating and equilibration phase of dynamics thus lasted 250 ps for each of the MD simulations. The production simulation was then started and continued for an additional 950 ps at an average temperature 298 K yielding a total simulation time of 1200 ps with each force field. The distance constraints on the terminal basepairs were present during the production simulation. In previous simulations, distance constraints have been used in some cases (Auffinger and Westhof, 1997) and not in others (Cheatham and Kollman, 1996); in the latter there are usually significant distortions in these basepairs and they are not included in the analysis. The simulations with the CHARMM27 force field were extended to 2000 ps, with both B- and A-type DNA starting structures; the B-DNA starting structure corresponded to the x-ray structure (Quintana et al., 1992) and the A-DNA to the canonical A form (Arnott and Hukins, 1972).

Structural analysis

IUPAC-IUB and EMBO nomenclature (IUPAC-IUB joint commission on Biochemical Nomenclature, 1983; EMBO Workshop, 1989) for nucleic acids were followed for the representation of conformational and helicoidal parameters of DNA, respectively. The antiparallel chains of DNA were specified as strand A with nucleotide residues from A1 to A10 and in strand B from B1 to B10 (Fig. 1). The conformational and helicoidal parameters of the double helix for the analysis of DNA structures were calculated, excluding the terminal residues/basepairs which exhibit larger fluctuations than internal basepairs. In the case of the calculated average DNA structures, a minimization of 500 steps of SD was performed before analysis of the structural parameters (average values and standard deviations) to remove unfavorable steric interactions; these arise, in particular, from hydrogen atoms of rotatable methyl groups on the thymines. Root-mean-square deviations (RMSD) were calculated between the DNA structures from the simulation and the minimized x-ray structure, and the canonical A-DNA or B-DNA structures. The RMSD was evaluated after least-square fitting of all the DNA heavy atoms, except for the terminal basepairs. Base helicoidal parameters were evaluated using the program FREEHELIX (Dickerson, 1998). Conformational and helicoidal parameters of canonical A-type DNA (Arnott and Hukins, 1972) and B-type DNA (Arnott and Hukins, 1973), designated hereafter as A and B, were evaluated from DNA duplexes generated with the program InsightII (Molecular Simulation Inc., San Diego, CA, version 98.0).

A cluster analysis was performed to identify a number of representative conformations from each simulation. Root mean square deviations calculated along the simulation with respect to the average structure show that five clusters can be generated using a cluster radius between 1.2 Å (BMS) and 1.9 Å (AMBER and CHARMM). A two-dimensional matrix of the RMSD between 1200 sets of DNA coordinates from the simulation (one every 1 ps) and the x-ray structure was built using the CHARMM program. A RMSD threshold value was chosen so that the 1200 DNA coordinates are distributed in five different clusters for all the force fields, each member of the cluster being more similar to all members of the same cluster and more dissimilar to any member of the four other clusters. The coordinate sets were then organized into five subsets based on the RMSD threshold values, using the program QUANTA (Molecular Simulation Inc., version 98). All the conformers in one subset have RMSDs of less than 1.8 Å from each other. The clusters contain between 96 and 546 coordinate sets. Average coordinates for each cluster were calculated and the conformer having the lowest RMSD with respect to the average structure was chosen as the cluster representative. The five cluster representatives were used for the detailed analysis of the conformational and helicoidal parameters. The RMSD between the five cluster representatives generally varies from 1.2 Å to 1.9 Å (e.g., between 1.3 Å and 1.9 Å for the simulation with the CHARMM27 force field).

Water and ion distributions were computed from the trajectories using the program Surfnet, which generates three-dimensional density distributions of data points and which has been widely used to analyze intermolecular interactions (Laskowski, 1995). Before generation of the distributions, a root mean square fit was carried out between all the DNA conformers obtained from the simulation and the average structure used as reference. The superimposition of all the DNA conformers on the reference structure leads to a distribution of data points corresponding to the water oxygens and sodium ions. A map of these points was written by counting the number of oxygen or sodium atoms on a grid with (144 × 160 × 179) grid points in the (x,y,z) directions, corresponding to a grid separation of 0.25 Å. No correction was applied to account for the nearest periodic images of water or sodium atoms because they lead to only small underestimations of the density at the corners of the box. The grid maps were then contoured using the program InsightII to visualize high-density regions and identify specific hydration patterns.

RESULTS AND DISCUSSION

General features and average structural parameters

The results for the RMSD versus time with respect to the x-ray, canonical B, and canonical A structures are shown in

Fig. 2. The MD structure simulated with CHARMM22 exhibits significant deviations from both the experimental B structure (3.8 Å) and canonical B-DNA structure (4.2 Å). After 600 ps, the structure is close to a canonical A-DNA with a RMSD of less than 2.0 Å. This corresponds to a B to A transition. CHARMM27 yields a stable B-DNA structure with a small RMSD, both from the x-ray structure (1.8 Å) and canonical B-DNA (2.2 Å); i.e., it remains slightly closer to the x-ray structure. The AMBER and BMS force fields give a MD structure that deviates more or less equally from the x-ray structure and the canonical B-DNA. The BMS force field has the best agreement with the x-ray structure with an average RMSD of ~1.5 Å. In all cases, there is a clear anticorrelation between the RMSD with respect to the canonical A-DNA and B-DNA forms. For CHARMM27, in a simulation of two B-DNA duplexes, MacKerell and Banavali (2000) found that as the MD structure deviates from canonical B-DNA, it moves toward canonical A-DNA and vice versa. They demonstrated that this behavior is specific to DNA (and not to RNA). They also showed that CHARMM27 can address conformational changes involved in the A to B or B to A transitions, as influenced, for example, by the ionic environment. The intersection of the curves representing the RMSD versus time with respect to canonical A- and B-DNA in the case of the AMBER and CHARMM27 force fields (e.g. at 500 ps for CHARMM27) reflects a significant sampling of the conformational space between the A and B forms. This highly anticorrelated behavior is common to the CHARMM force fields; the correlation coefficient for the A versus B RMSD is −0.96 for both CHARMM22 and CHARMM27. For the AMBER force field, the correlation coefficient is −0.80. For the BMS force field, the simulated DNA remains close to B-DNA during the entire simulation; the correlation coefficient is −0.68.

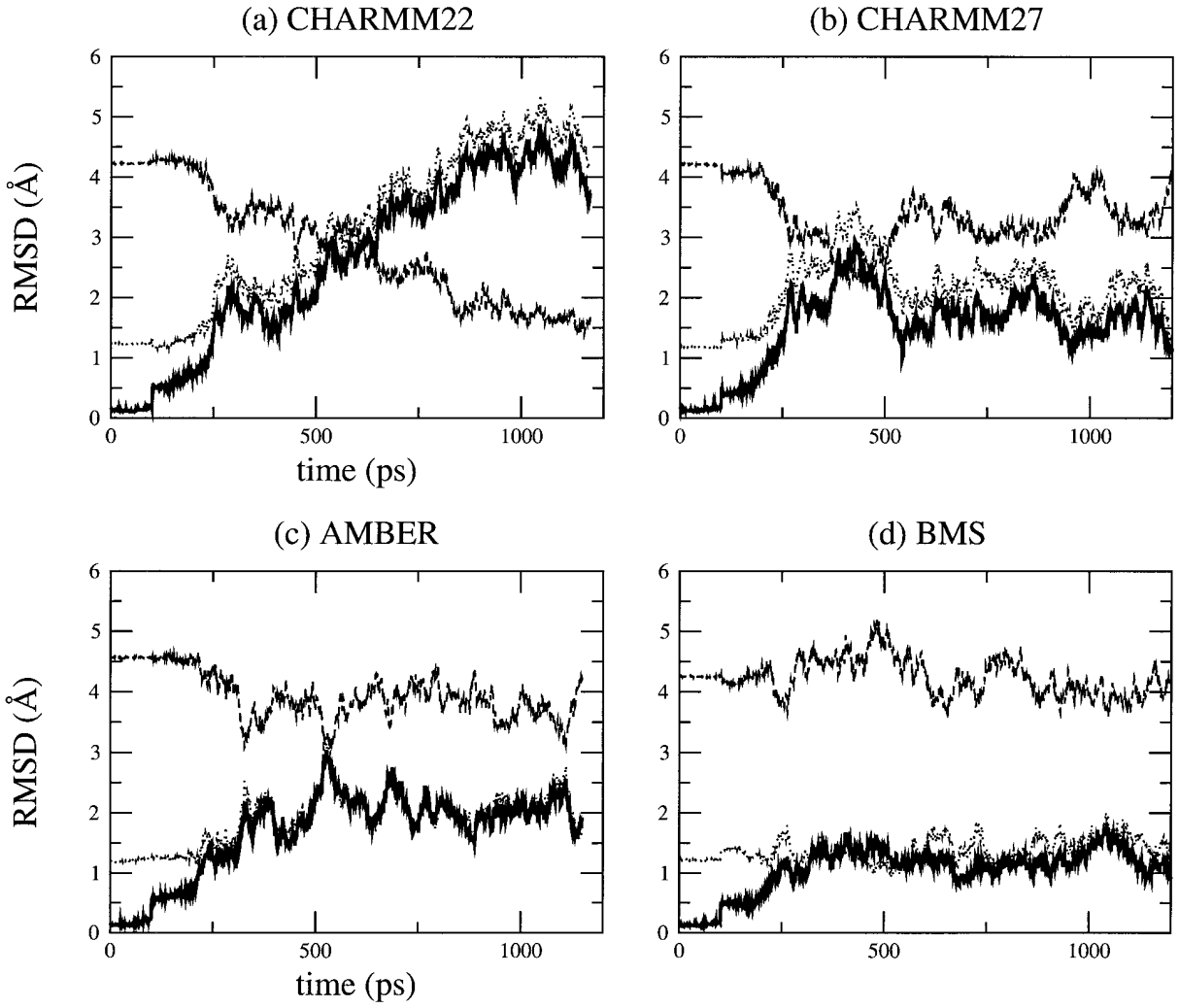


FIGURE 2 Root mean square deviation of all atoms, excluding the end basepairs of the DNA duplex, d(CGATTAATCG)$_2$, versus simulation time using different force fields: (*a*) CHARMM22, (*b*) CHARMM27, (*c*) AMBER, (*d*) BMS. Deviations from the minimized crystal, canonical B-DNA, and canonical A-DNA structures are indicated by the solid, dotted and dashed lines, respectively.

Table 1 gives the details of the conformational and helicoidal parameters and minor groove distances of the simulated duplexes averaged for the period from 601 to 1200 ps; the corresponding values for canonical A-DNA and B-DNA and the x-ray structure are also listed.

A-DNA and B-DNA conformations are differentiated by certain features related to the conformation of the phosphodiester backbone and the geometry of the nucleic acid bases (base stacking and basepairing). The sugar pucker is partly responsible for the DNA form adopted; other backbone movements can vary without changing the backbone conformation due to crankshaft-type displacements. The pucker, characterized by the pseudorotation angle, tends to have values corresponding to C3′-endo/C2′-exo (P from −36° to +36°) and C3′-exo/C2′-endo (P from 137° to 194°) in A-DNA and B-DNA respectively. The base stacking is primarily determined by the basepair separation and in particular the rise, which significantly differs between the A form (2.56 Å) and the B form (3.37 Å). It is also influenced by the relative orientation of adjacent basepairs (twist, slide). Helical twist varies considerably in B-type helices (35°–45°); in A-DNA the variation of twist is less (30°–33°). The large variations of twist in B-DNA and the small difference between the A and B forms makes the identification difficult in terms of this parameter. The slide, corresponding to the relative displacement between two adjacent basepairs along their longitudinal axis, can be used to distinguish the A (−1.7 Å) and B forms (+0.4 Å) because of the small magnitude of its variation and its strong correlation with helical twist. Basepairs are also significantly displaced from the helix axis (X-displacement) into the major groove by −4.4 to −4.9 Å in A-DNA and into the minor groove (0.2–1.8 Å) in B-DNA. This affects the macroscopic structure of the grooves of the helix, resulting in a narrow minor groove in B-DNA and a wide major groove in A-DNA, as described above. The inclination of basepairs to the helix axis is positive (10°–20°) in A-type, whereas it is negative (−6°–16.5°) in B-type.

TABLE 1 Comparison of conformational and helical parameters and minor groove distance of d(CGATTAATCG)$_2$ DNA duplexes and MD simulated structures obtained with different force fields

				Average simulated structure*			
	Canonical A-DNA	Canonical B-DNA	Crystal structure	CHARMM22	CHARMM27	BMS	AMBER
α (°)	276	313.2	303.9(13.5)	289.3(14.4)	304.1(16.3)	298.1(7.0)	297.5(11.4)
β (°)	208	214	179.1 (2.8)	167.3(27.3)	159.4(25.4)	171.2(24.0)	160.3(28.5)
γ (°)	45.5	36.4	47.8(10.3)	57.7(4.9)	54.7(12.4)	52.2(5.7)	52.8(6.5)
δ (°)	84.3	156.4	127.1(15.9)	94.1(16.6)	128.2(14.0)	133.7(19.3)	126.8(18.6)
ε (°)	179.5	155	184.2(25.6)	205.1(10.3)	191.9(6.7)	201.0(25.6)	211.0(29.8)
ζ (°)	311	264.8	260.1(28.9)	282.9(7.5)	244.7(17.8)	242.1(40.8)	237.3(46.7)
χ (°)	206	262	245.4(16.2)	203.4(14.9)	247.8(15.4)	249.1(17.8)	246.6(18.3)
Pucker (°)	13.3	191.8	139.4(38.1)	45.8(47.7)	136.4(39.1)	138.1(39.1)	137.6(25.1)
Amplitude (°)	40.2	37.5	36.5(7.5)	37.5(9.6)	39.1(5.5)	39.6(7.9)	40.1(7.1)
Base step parameters:							
Tilt (°)	0	0	0.1(1.9)	−1.4(2.5)	−1.5(4.4)	0.1(2.5)	−0.2(4.2)
Slide (Å)	−1.53	−0.16	−0.10(0.38)	−1.59(0.43)	−0.18(0.23)	0.09(0.55)	−0.08(0.59)
Roll (°)	10.7(2.2)	−3.6	0.0(5.1)	6.4(7.5)	3.9(8.8)	1.3(4.4)	1.7(5.8)
Shift (Å)	0	0	0.03(0.27)	0.16(0.56)	−0.17(0.69)	0.07(0.75)	0.03(0.78)
Twist (°)	32.7	36	36.9(3.7)	28.5(5.1)	35.1(2.2)	36.1(3.6)	33.9(3.9)
Rise (Å)	2.56	3.37	3.20(0.21)	3.39(0.39)	3.4(0.29)	3.20(0.20)	3.39(0.17)
Basepair parameters:							
Tip (°)	0	0	−0.8(3.4)	2.3(7.8)	0.9(6.1)	0.2(4.1)	0.1(5.1)
Prop. twist (°)	11.7	4.3	−14.2(3.1)	−10.6(6.8)	−6.2(11.9)	−7.2(5.0)	−6.4(7.4)
Buckle (°)	0	0	−2.5(6.3)	−3.2(14.6)	3.7(8.4)	−0.7(7.4)	1.1(4.3)
Inclination (°)	19.9	−5.7	0.3(1.2)	16.9(3.4)	7.3(2.7)	3.5(2.2)	5.2(1.7)
X-disp. (Å)	−4.49	0.23	−0.27(0.42)	−4.92(0.66)	−1.29(0.42)	−0.49(0.93)	−0.87(0.58)
Y-disp (Å)	0	0	0.22(0.60)	−0.51(0.49)	0.14(0.65)	0.53(0.65)	−0.36(0.47)
Minor groove distances (Å)†:							
$PA_4 \ldots PB_{10}$	16.95	13.17	10.48	15.33	14.57	14.47	12.91
$PA_5 \ldots PB_9$	16.95	13.17	11.49	15.25	14.81	14.46	11.87
$PA_6 \ldots PB_8$	16.95	13.17	13.15	14.87	13.31	15.01	13.05
$PA_7 \ldots PB_7$	16.95	13.17	13.68	15.09	13.41	13.07	13.19
$PA_8 \ldots PB_6$	16.95	13.17	12.68	15.31	12.71	11.72	13.24
$PA_9 \ldots PB_5$	16.95	13.17	11.85	15.86	13.86	12.69	13.48
$PA_{10} \ldots PB_4$	16.95	13.17	11.09	14.65	15.82	15.12	13.62

Averages of conformational and helical parameters exclude the end basepairs of DNA. Values in parentheses correspond to standard deviations.

*The values for the torsions, base step and basepair parameters are averaged over the simulation time.

†The minor groove distances are calculated between the phosphorus atoms P_i on one strand and P_{i+3} on the opposing strand.

The comparison of the torsions and parameters related to the geometry of the basepairs confirms the previous trend: CHARMM22 yields an A-like DNA whereas CHARMM27 yields a B-like geometry in good agreement with the x-ray structure (Table 1). Most of the torsion angles generated with CHARMM22 correspond to an A-like geometry for the phosphate backbone (α, ε, ζ) including the sugar pucker (δ corresponding to C3′-endo). The same holds true for the basepair geometry of the nucleic acid bases (χ). The particular orientation of the nucleic acid base ($\chi \approx -156°$) corresponds to an inclination of $\sim$20° of the basepair planes with respect to the plane perpendicular to the helical axis observed in A-DNA, whereas there is almost no inclination of the basepairs in B-DNA. The X-displacement is another basepair parameter that can be used to discriminate between A and B-DNA because it gives a measure of the depth of the major groove. Again, the average structure obtained with CHARMM22 is close to the A form. The relative orientations of successive basepairs, defined by the base step parameters, differ between A and B-DNA particularly in the slide (relative translation of the basepairs about the long axis of the base step), roll (relative rotation of the basepairs about the long axis of the base step), and twist (relative rotation of the basepairs about an axis perpendicular to the plane of the base step). Based on these parameters, the geometry of the double helix in the CHARMM22 simulation corresponds to that of A-DNA, even if the average rise is very close to that of a standard B-DNA (Table 1).

The average structure obtained with CHARMM27 exhibits torsion angles corresponding to the B form and close to those of the x-ray structure. The CHARMM27 force field correctly reproduces the average ε and ζ torsion angles, related to the conformation of the phosphate groups. The AMBER and BMS force fields, give values that deviate slightly more from those of the x-ray structure. Regarding the geometry of the basepairs, the roll and inclination are systematically larger in comparison with those of the x-ray structure. They are also somewhat too large, as well as with the AMBER force field, whereas the helical twist is smaller than the standard values for B-DNA and for the x-ray structure. A decrease in the average value of propeller twist of basepairs relative to the x-ray structure is found in all the MD structures. The geometry of the basepairs that differs least from the x-ray structure is obtained with the BMS force field.

Because the CHARMM22 force field leads to an A-like DNA, the minor groove is wide and shallow (whereas it is deep and narrow in the B-DNA structure); the distance between phosphorous atoms across the minor groove (P_i on one strand and P_{i+3} on the opposing strand, Fig. 1) are all more than 1.4 Å over 13.17 Å, a distance that corresponds to the minor groove width for a canonical B-DNA (Table 1). Although the minor groove is slightly wider than in B-DNA in the structure simulated with CHARMM27, it is considerably narrower and B-DNA like.

The minor groove is locally wider at the terminal basepairs (PA_4-PB_{10}, PA_5-PB_9, PA_9-PB_5, PA_{10}-PB_4) and narrower at the center, in particular at the ApA (A_6A_7) step (PA_8-PB_6, Table 1). It is also relatively narrow at the TpT and TpA basepair steps (PA_6-PB_8, PA_7-PB_7). It differs from the x-ray structure where the minor groove morphology is reversed: a local widening at the central basepair steps (PA_6-PB_8, PA_7-PB_7, PA_8-PB_6), and narrowing at the terminal basepair steps. The BMS force field gives a more similar minor groove morphology than CHARMM27: a local narrowing at the central basepair steps (TpA and ApA) but shifted toward the ApT step (A_7A_8). By contrast, the AMBER force field leads to a local widening at the three first terminal basepair steps GpA, ApT, and TpT. None of the force field reproduce a local widening at the center of the double helix. These discrepencies between the x-ray structure and the simulated structures will be discussed later.

The dynamic behavior of structural parameters specific to A- and B-DNA

As the dynamic changes in sugar pucker influence the backbone conformation, the variations of the pucker during the MD trajectories are of interest (Fig. 3). Most of the nucleotides adopt a C3′-endo conformation after the B to A transition observed with CHARMM22 at 600 ps (Fig. 3 *a*); the sugar conversion from C2′-endo to C3′-endo occur earlier (375 ps) at some positions (B2, B3, B7), simultaneously (A2, A5, A6), or later at other positions (A3, A4, B4, B5, B8, B7, B9). These results are consistent with the average parameters (Table 1), showing that the CHARMM22 force field favors A-type over B-type DNA, and contrast with those obtained with the CHARMM27 force field for which the preferred sugar pucker is C2′-endo (Fig. 3 *b*). Some C2′-endo to C3′-endo transitions with a short lifetime for the C3′-endo pucker (less than 80 ps) are observed with CHARMM27. They occur at a few positions via a O4′-endo intermediate ($P \approx 90°$) at A3, A7, B4. Some very short transitions C2′-endo to C3′-endo and back to C2′-endo are also observed at positions A4 and A8 (Fig. 3 *a*). Nevertheless, the more stable sugar pucker is C2′-endo at all the positions, except for B7 which conserves a C3′-endo conformation during the entire simulation. For the other force fields, AMBER and BMS (Fig. 3, *c* and *d*,), C2′-endo to C3′-endo transitions, and back to C2′-endo via a O4′-endo intermediate, are also observed at various positions. The deviations from the standard pucker forms (C2′-endo or C3′-endo) tend to be larger with the AMBER and BMS force fields. The strong tendency for B7 to adopt a C3′-endo conformation with CHARMM27, AMBER, and BMS is consistent with the x-ray data (Quintana et al., 1992).

Although the average torsion and pseudorotation angles provide a description of the general conformational features related to A or B-DNA, discrepancies are observed at particular positions; B7, already mentioned above, is one

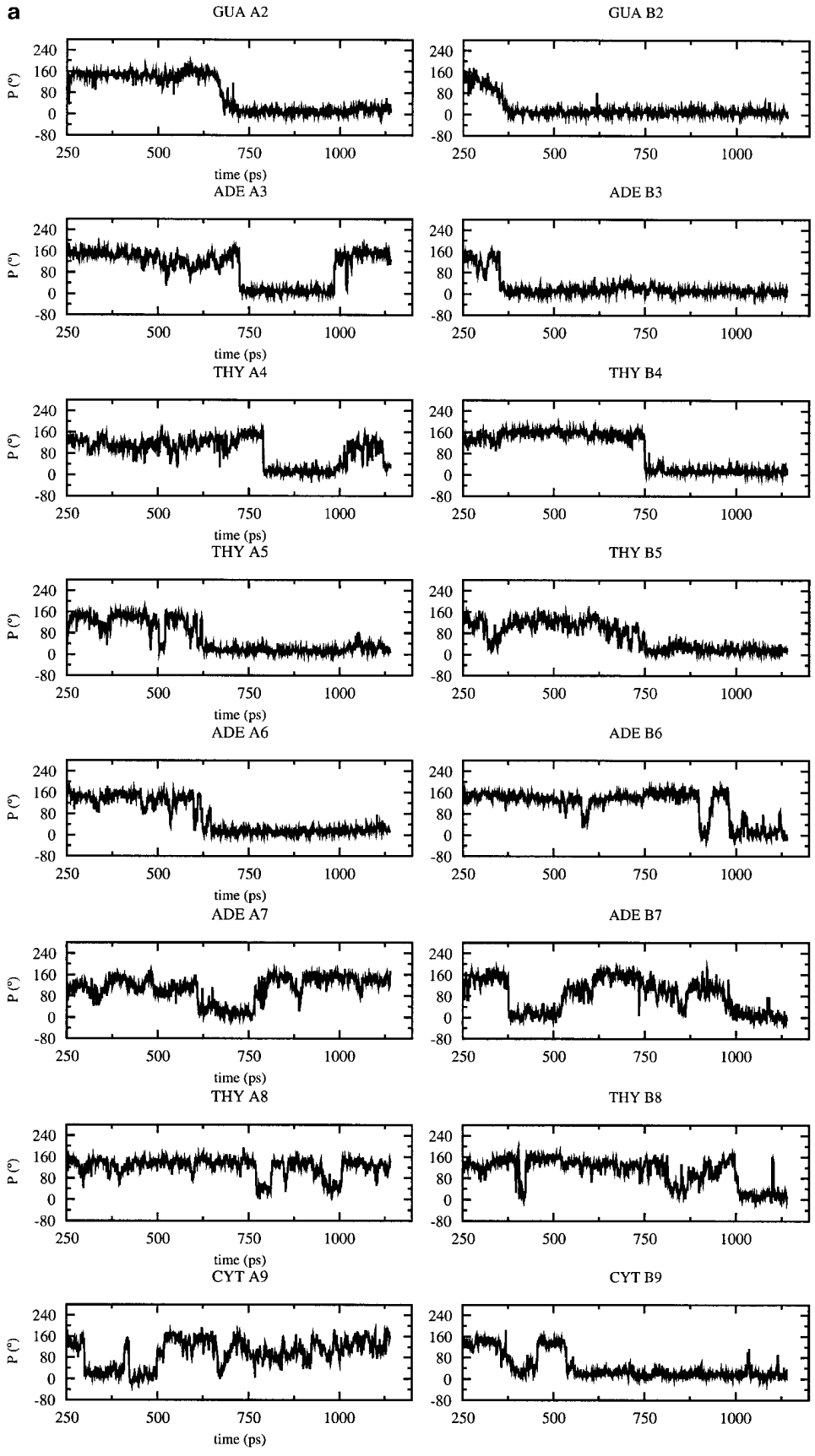


FIGURE 3 Time dependence of the sugar pseudorotation angle at the all nucleotide positions of the DNA decamer except at the terminal base-pairs using different force fields: (*a*) CHARMM22, (*b*) CHARMM27, (*c*) AMBER, (*d*) BMS.

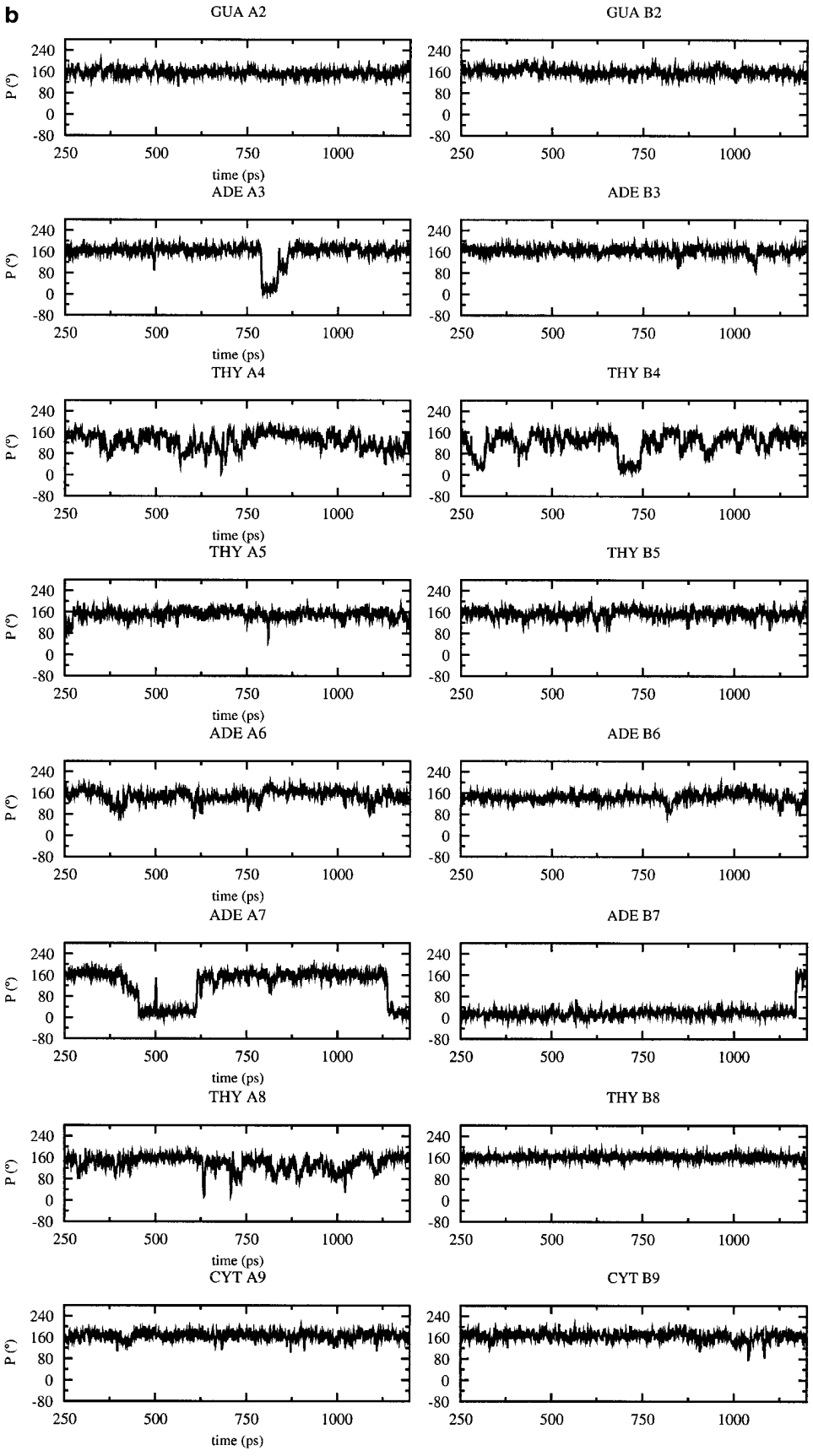


FIGURE 3 Continued

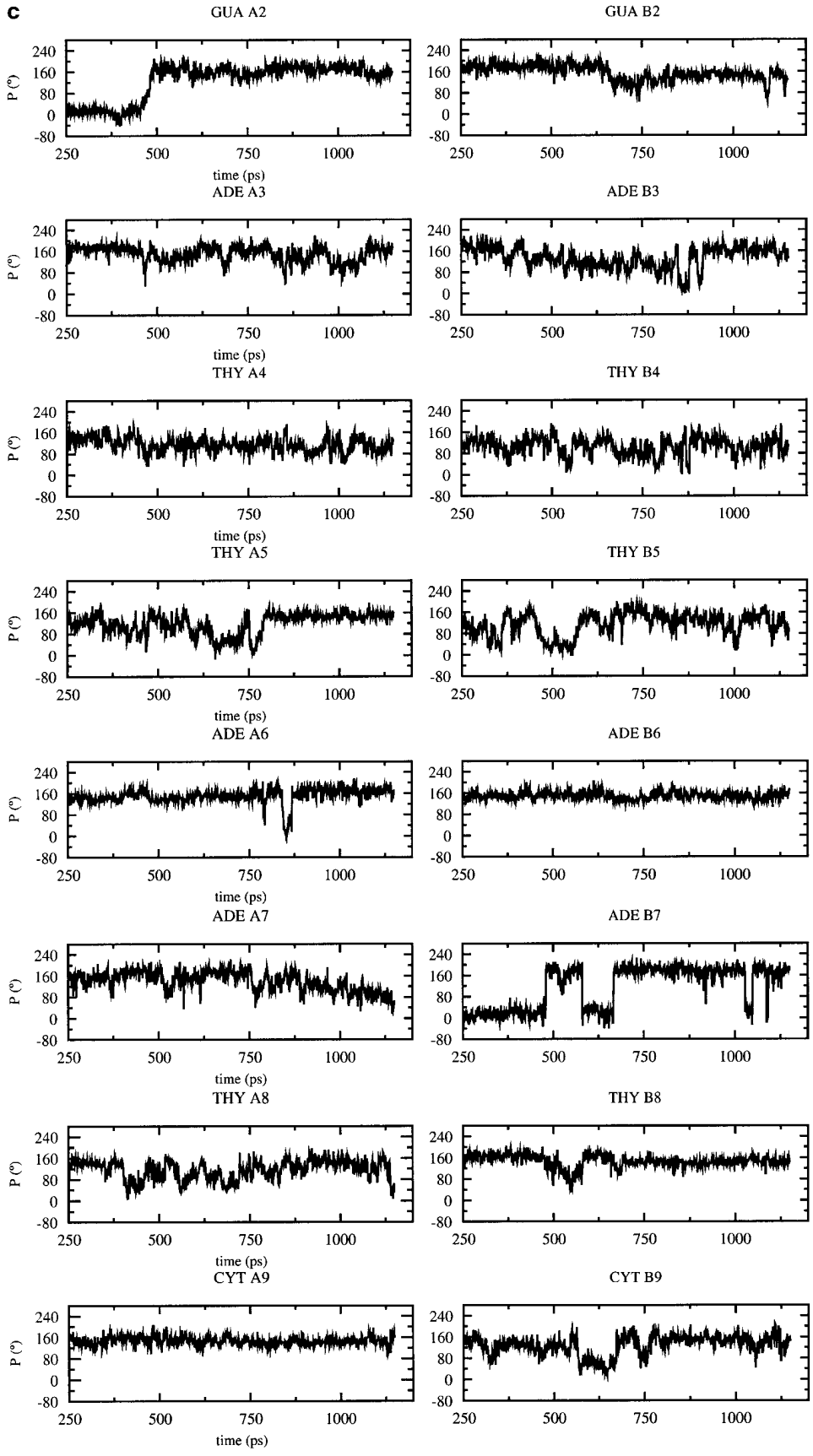


FIGURE 3 Continued

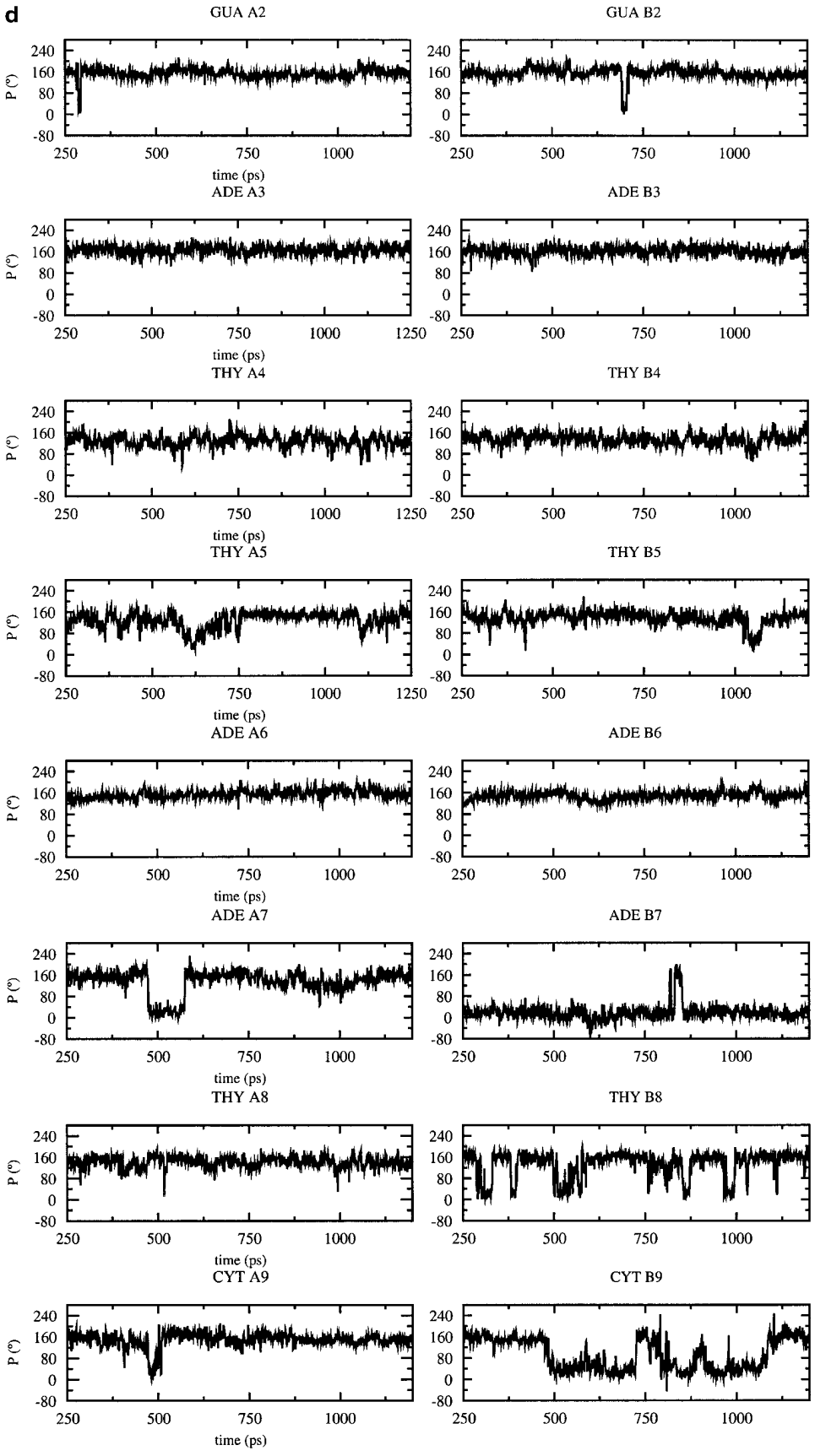


FIGURE 3 Continued

example. Similarly, the double helix, whose morphology can be evaluated using the base step parameters, can be locally distorted at some positions even if the average parameters are in agreement with a given canonical DNA. The rise between adjacent basepairs, for example, exhibits quite large fluctuations, which are equivalent in magnitude to the difference between the standard values for the A and B forms (see Table 1 and Fig. 4). Despite an A-like conformation, the structure obtained with CHARMM22 is characterized by an average rise closer to canonical B-DNA. Most of the basepair steps exhibit a bistable behavior with an initial value of rise generally close to the standard B value and a final value close to the standard A value. This is particularly evident at the ApT (AT/AT: A3A4/B7B8), TpT (TT/AA: A4A5/B6B7), and TpC (TC/GA: A8A9/B2B3) basepair steps (Fig. 4 *a*). The rise is closer to the standard B value with CHARMM27 (Fig. 4 *b*), but it oscillates between the standard A and B values except for the ApT (AT/AT: A3A4/B7B8) and the central TpA (TA/TA: A5A6/B5B6) basepair steps, which exhibit a rise closer to that of a standard B-DNA. Similar behavior is observed in the simulation with BMS (Fig. 4 *c*, Supplementary Material) whereas the fluctuations in rise observed with AMBER are generally smaller and centered around a standard rise value for the B form (Fig. 4 *d*, Supplementary Material). The most striking feature that is shared by the MD structures simulated with all the force fields is the high rise at the central TpA (A5A6) basepair step (Fig. 4 *b*). Although high rise is not specific to TpA steps, it is known as a flexible basepair step whereas ApT, ApA, and TpT are known as rigid basepair steps (El Hassan and Calladine, 1996).

The helical twist also shows fluctuations, which are larger than the difference between the standard A and B values. Although the helical twist is only approximately anticorrelated with the rise, the basepair steps with low rise tend to have high twist and vice versa for a given DNA form (Fig. 4, *a* and *b*). In the simulation with CHARMM22, most of the basepair steps have a high twist/low rise profile, then they tend to adopt a low twist/high rise profile (Fig. 4 *a*). For example, the basepair steps ApT (AT/AT: A3A4/B7B8 and A7A8/B3B4) and the TpT (TT/AA: A4A5/B6B7) switch from the high rise/low twist state to the low rise/high twist

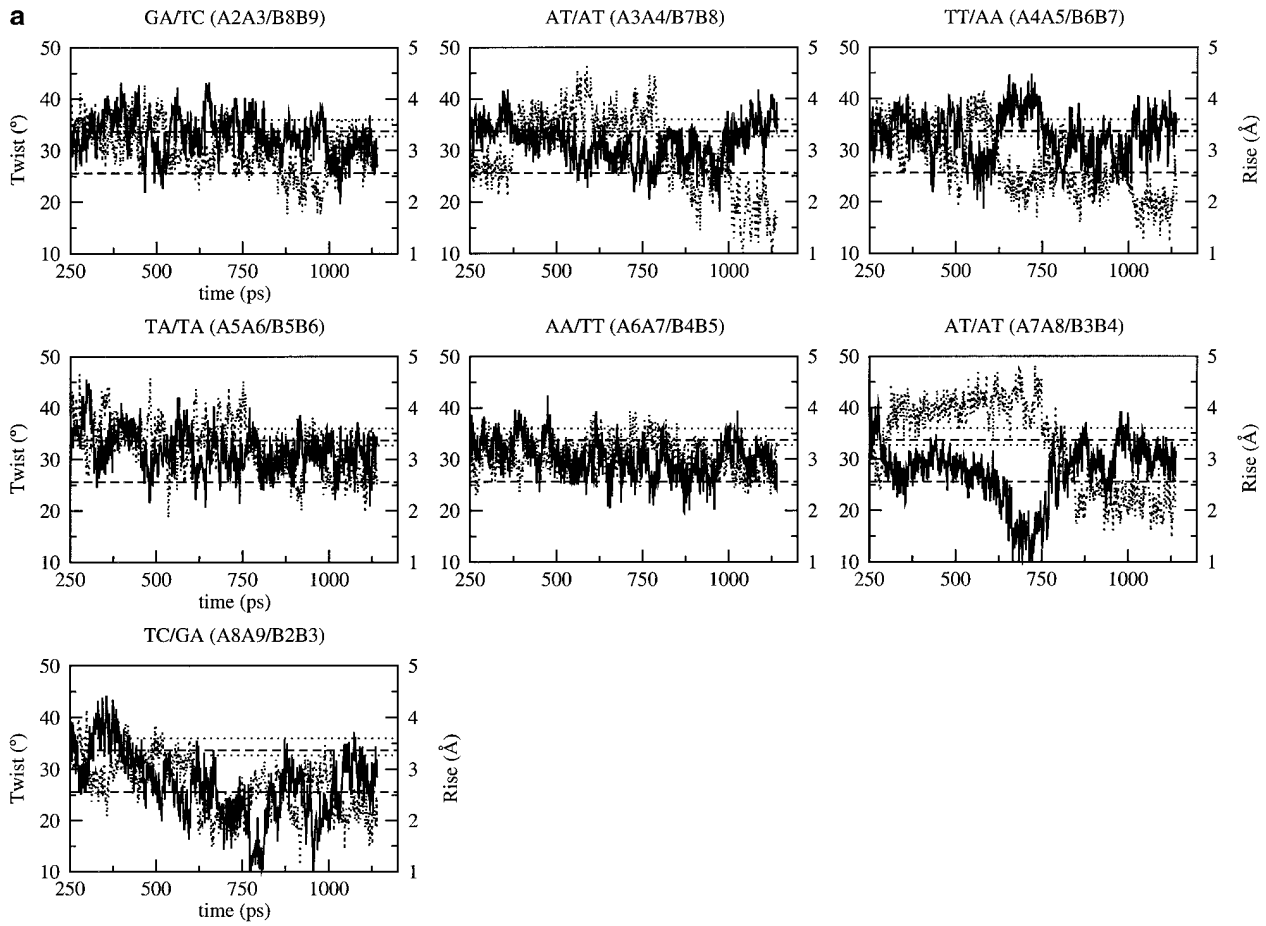


FIGURE 4 Time dependence of the twist (*solid line*) and rise (*dotted line*) at the basepair steps of the B-DNA decamer using the CHARMM force fields: (*a*) CHARMM22, (*b*) CHARMM27. The standard values for the canonical A- and B-DNA are represented by dotted lines for the twist and dashed line for the rise; in both cases, the upper one corresponds to the B value and the lower one to the A value. Fig. 4, *c* and *d* are in Supplementary Material.

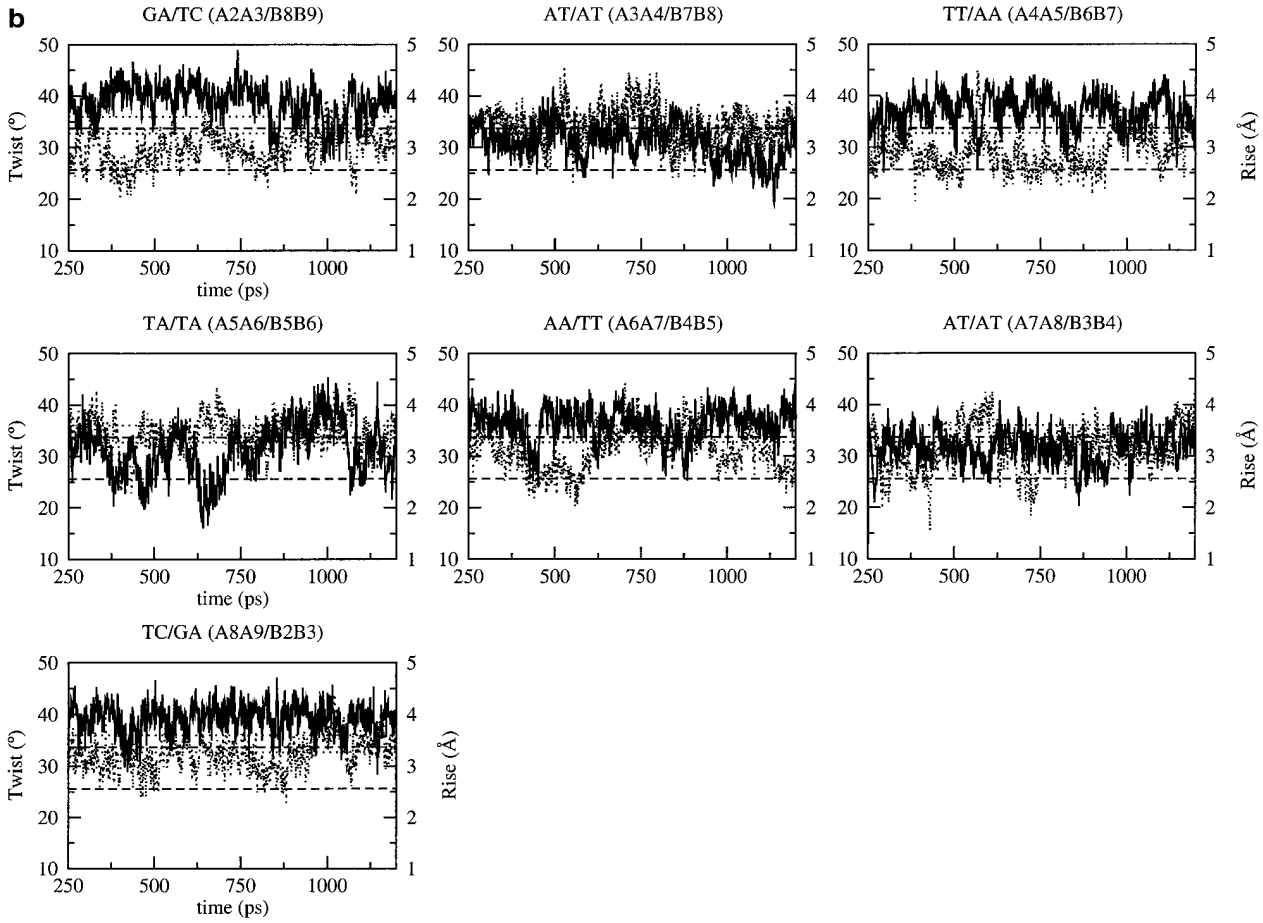


FIGURE 4 Continued

state after the B to A conversion in the simulation with CHARMM22 (Fig. 4 *a*). For the other basepair steps, the average twist value is generally closer to the A value. At the end of the simulation, all basepair steps present a twist/rise profile corresponding to A-like DNA (Fig. 4 *a*). For CHARMM27, the twist value is close to the standard B value with short transitions toward the A value (Fig. 4 *b*). Most of the basepair steps exhibit the high twist/low rise state; the twist value is slightly more than that of canonical B and the rise value is close to the canonical B value. It concerns more particularly the purine-purine (GA/TC and AA/TT) and pyrimidine-pyrimidine basepair steps (TT/AA and TC/GA). The exception concerns the two ApT basepair steps (AT/AT: A3A4/B7B8 and A7A8/B3B4), which exhibit a twist closer to the standard A value; this is true also for the central TpA basepair step (TA/TA: A5A6/B5B6). This central TpA basepair step shows a high rise/low twist state, a feature observed in the x-ray structure, as mentioned above for the twist. Structural analyses of DNA structures have shown that TpA steps can adopt the lowest twist (30.6° ± 6.7°, El Hassan and Calladine, 1996). The purine-purine and pyrimidine-pyrimidine basepair steps are the more prone to switch between high twist/low rise and low twist/high rise profiles (Nelson et al., 1987). It occurs at different times during the simulation for the GA/TC and AA/TT basepair steps (Fig. 4 *b*).

Similar features are observed with BMS and AMBER (Fig. 4, *c* and *d*, Supplementary Material). In the case of the GA/TC and TC/GA basepair steps, the deviations in twist and rise, when switching from the high twist/low rise to the low twist/high rise state, are more pronounced with AMBER (Fig. 4 *c*, Supplementary Material). The twist has a bistable behavior oscillating between the standard A and B values. A common feature in the simulations with CHARMM27, BMS, and AMBER are the very large fluctuations of twist at the central TpA basepair step. These fluctuations correspond to unstacking of the bases at the TpA step (high rise) and changes in the backbone conformation via B_I/B_{II} transitions (the B_{II} phosphate conformation requiring the two bases linked to the phosphate to be unstacked). This is related to the local widening of the minor groove in regions where B_{II} conformations are present for basepairs diagonally opposite each other across the groove.

The X-displacement, corresponding to the depth of the

major groove, changes in a concerted way at all basepair steps (Fig. 5). This behavior is more obvious in the case of the simulation with CHARMM22 because the wide major groove (B-DNA) becomes deep and narrow (A-DNA) after the B to A transition, taking place at 750 ps (Fig. 5 *a*). The transition is preceded by a sharp change at 600 ps corresponding to a more B-like conformation before the DNA continues a gradual conversion to the A form. CHARMM27 tends to overestimate the depth of the major groove, which oscillates between the standard A and B values in a concerted way at all the basepairs (Fig. 5 *b*). AMBER gives similar behavior, with an X-displacement that is only slightly overestimated and closer to the standard B value (Fig. 5 *c*, Supplementary Material). The X-displacement deviates very little with BMS and stays close to the standard B value (Fig. 5 *d*, Supplementary Material). The inclination of the basepairs also changes in a concerted way (Fig. 5). The inclination and X-displacement are strongly anticorrelated (the correlation coefficient is around −0.7 at most basepair steps) with CHARMM22 (Fig. 5 *a*). Despite their common behavior, the rise and X-displacement are not significantly correlated in the CHARMM27 simulation (the maximum correlation coefficient is −0.5), although high rise tends to be associated with low X-displacement and vice versa, in particular at the central basepair steps (Fig. 5 *b*). The inclination is overestimated with CHARMM27 at all the basepair steps. It is also overestimated, although slightly less, with AMBER but no significant anticorrelation is observed between inclination and X-displacement (the maximum correlation coefficient is –0.34 at the TT/AA basepair step and almost null at the other basepair steps; see Fig. 5 *c*, Supplementary Material) or with BMS which gives an inclination closer to its standard B value (the maximum correlation coefficient is −0.27 at the TC/GA basepair step and almost null at the other basepair steps, see Fig. 5 *d*, Supplementary Material). Feig and Pettitt (1998) suggested that the anticorrelation between the inclination and the X-displacement is related to a change in the accessibility of the major groove; that is, when the number of water molecules in the major groove decreases, going for example

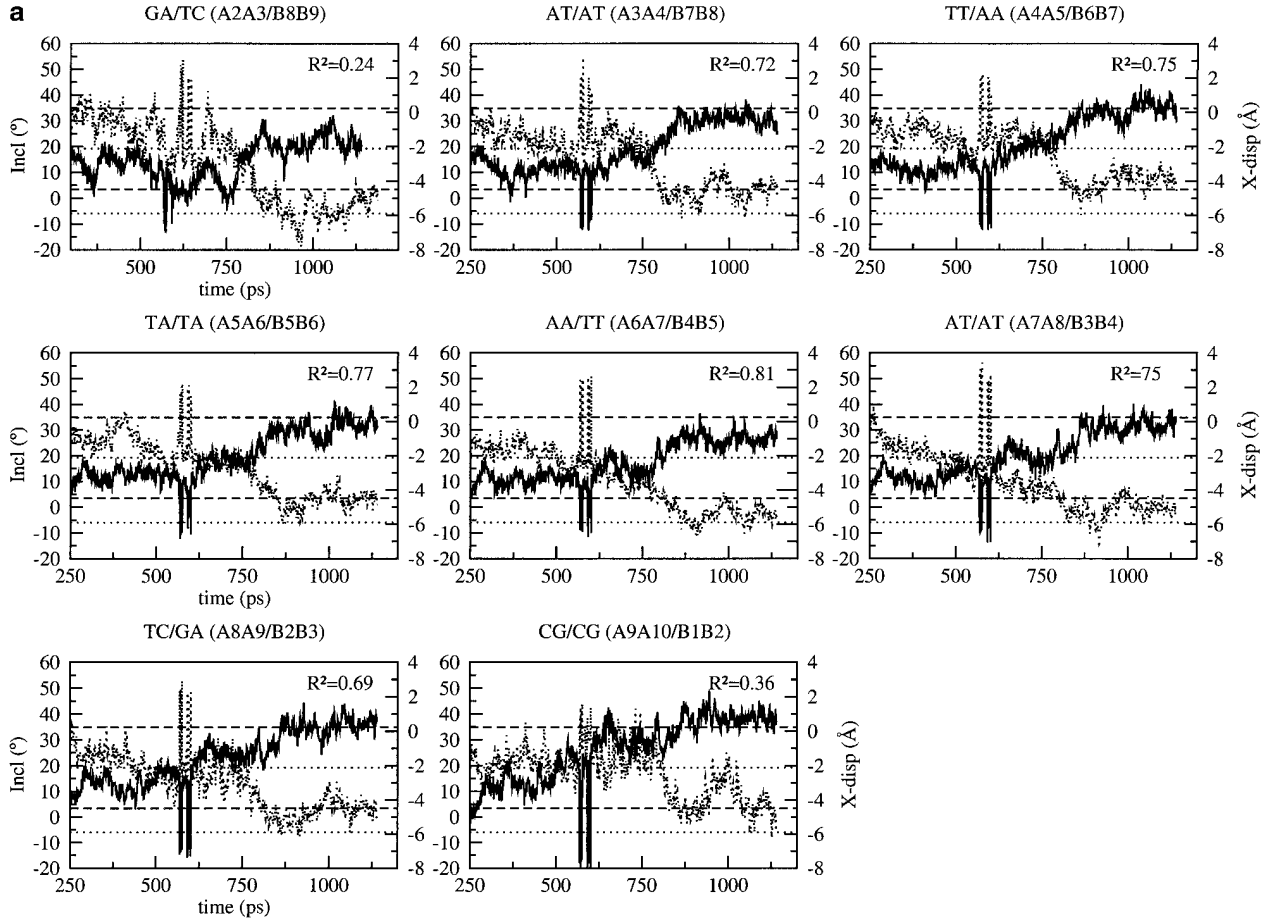


FIGURE 5 Time dependence of the inclination (*solid line*) and X-displacement (*dotted line*) at the basepairs of the B-DNA decamer using the CHARMM force fields: (*a*) CHARMM22, (*b*) CHARMM27. The standard values for the canonical A- and B-DNA are represented by dotted lines for the inclination (the upper one corresponds to the A value and the lower one to the B value). Dashed lines are used for the X-displacement (the upper one corresponds to the B value and the lower one to the A value). Fig. 5, *c* and *d* are in Supplementary Material.

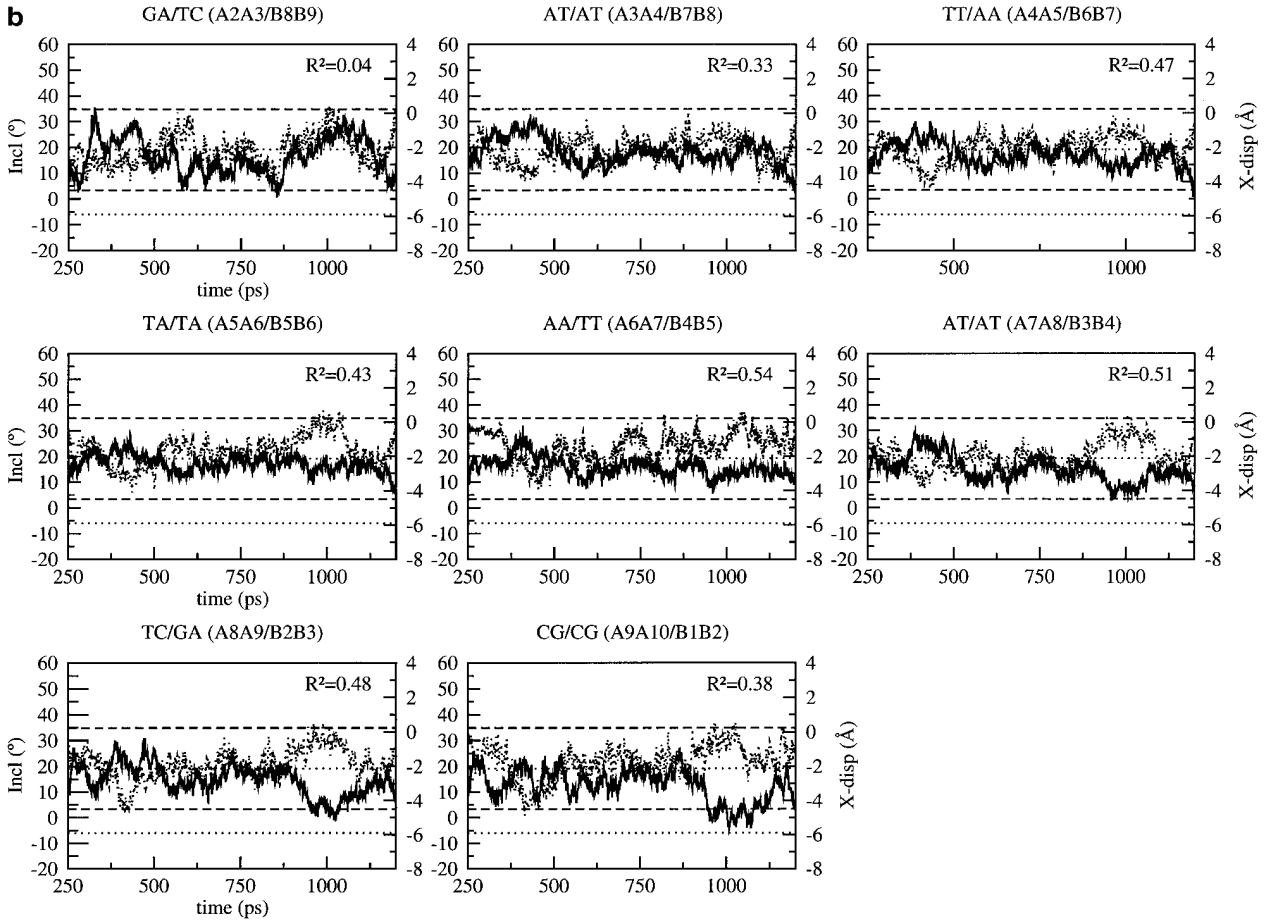


FIGURE 5 Continued

from a wide major groove (B-DNA) to a narrow major groove (A-DNA), the desolvation of some functional groups in the groove induces an increase of the base inclination angle. We observe a correlation between the fluctuations of the X-displacement and the solvent accessible surface of the major groove calculated by the Lee and Richards method (Lee and Richards, 1971). The stronger anticorrelation between X-displacement and inclination at T:A basepairs is consistent with the fact that adenines are more easily desolvated, in particular in A-DNA (Egli et al., 1998). This feature of DNA solvation is documented in the manuscript which describes the solvation changes associated with the A to B transition (F. Leclerc, S. Y. Reddy, and M. Karplus, unpublished).

Sequence dependence of the conformational flexibility

To determine the influence of the force field on the conformational flexibility and its sequence dependence, we have analyzed more precisely the backbone and base conformations and compared the simulation results with available experimental data and previous theoretical studies. The CHARMM22 force field calculations are not included because they deviate significantly from the experimental data on B-DNA, as described above. The dynamical behavior of the phosphate backbone can be described by two conformational substates, B_I and B_{II}, which are determined by the orientation of the phosphate group. The phosphate group can adopt two conformations: B_I with values of ε and ζ in the range 120–210° (*trans*, *t*) and 235–295° (*gauche*$^-$, g^-), respectively; and B_{II} with torsions ε and ζ values in the range 210–300° (*gauche*$^-$, g^-) and 150–210° (*trans*, *t*). The conformation of the phosphate group is a good criterion for the evaluation of the conformational flexibility because the interconversion between these two substates is dependent on base destacking and leads to changes in the width of the minor groove. To determine to what extent the force fields can reproduce sequence specific average structural features (comparison with the x-ray structure) and the local conformational flexibility (comparison with analyses of other DNA structures), we use the helicoidal parameters that were introduced in the previous section. For this analysis, we select the five representative conformers from the MD trajectory; they were identified by a cluster analysis, as described in the Methods section. Although we did omit from our

simulations the Mg^{2+} cation bound at the TpT base step in the x-ray structure, we do not expect any significant change in basepair or base step parameters. Because the Mg^{2+} cation is bound quite deep in the minor groove and not at the top of the minor groove, it has a negligible effect on roll, tilt, and bending (Chiu and Dickerson, 2000).

In the simulation with CHARMM27, substate B_I is found to be dominant for all positions, except B6 where the first half of the trajectory corresponds to B_{II} and the second half to B_I. There are B_I-B_{II} interconversions with a B_{II} life up to 100 ps at many but not all of the positions (the variations of the backbone torsions ε and ζ with the different force fields are given in Fig. 6). The nucleotides at the termini, A2:B9 and A9:B2, show B_I-B_{II} interconversions as expected because of the ease of base destacking at the ends the double helix. All the other nucleotides preserve a B_I conformation, except for those of the central TA base step (A5A6/B5B6). Generally, similar behavior is observed with the AMBER and BMS force fields, although the B_{II} conformation is more persistent and as populated as the B_I conformation at A5, A6, A9, and B8 for AMBER, at A5, A6, A9 for BMS; at B6, it is the more populated conformation for both AMBER and BMS. In the x-ray structure, apart from the nucleotides at the terminal basepairs A1 and B1, only A6 and B6 have a B_{II} conformation; B_I-B_{II} interconversions are observed at these two positions with all the force fields. The simulation results are in agreement with the x-ray data showing a preference for B_I (63–69%) over B_{II} (22–29%) in a survey of about 60 B-DNA structures (Winger et al., 1998). The three force fields can reproduce most of the sequence specific structural features of the DNA backbone related to the sugar pucker (C3′-endo conformation at B7) or the phosphate conformation (B_{II} conformation at A6 and B6).

The results can be compared with a structural analysis performed on eight DNA structures, which has shown that the basepair parameters have large deviations from those of

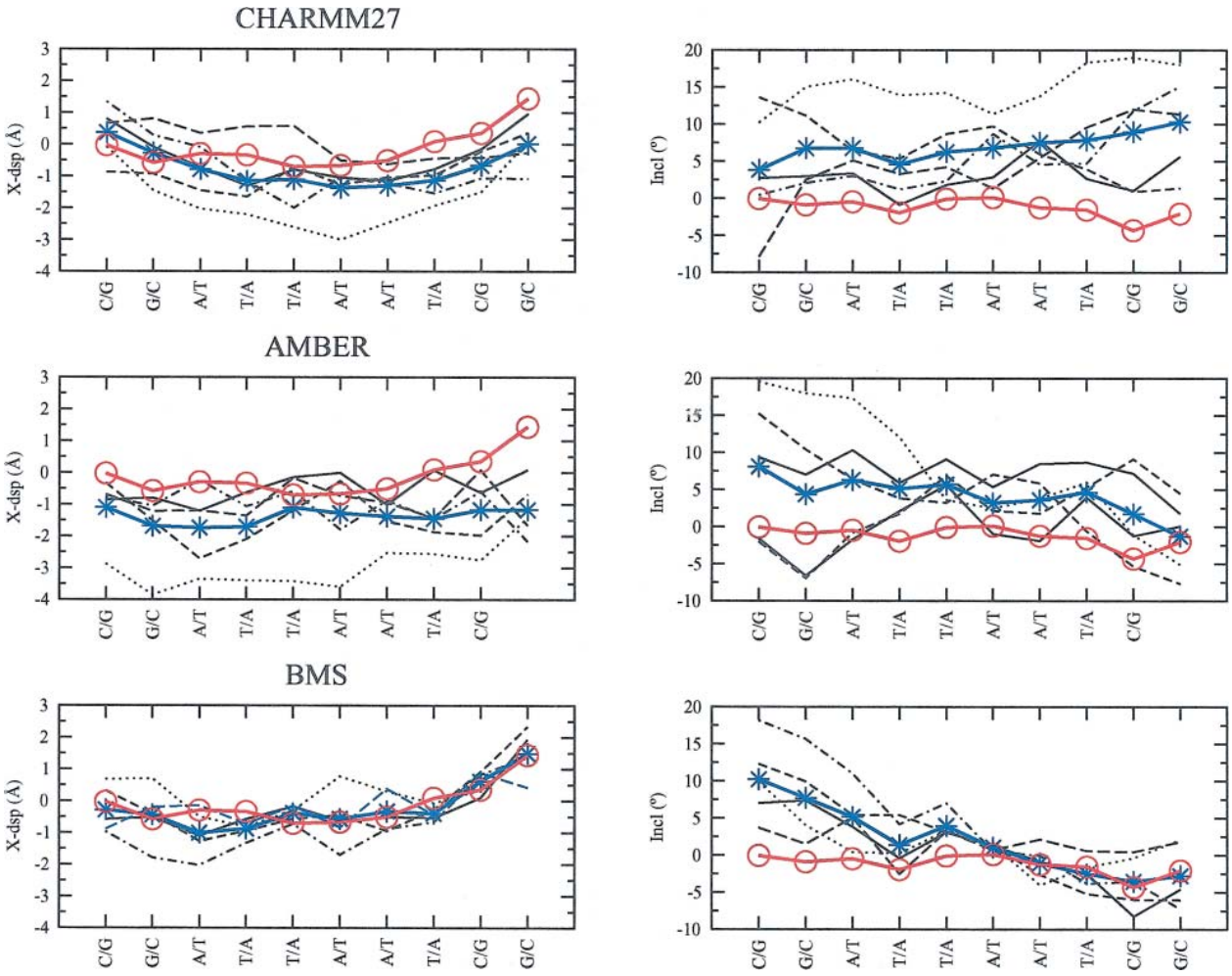


FIGURE 6 Basepair parameters (X-displacement and inclination) of the five representative conformations for: CHARMM27, AMBER, and BMS. The bold solid line marked by circles represents the basepair parameter of the x-ray structure; the solid line marked by stars represents the basepair parameter averaged over the five conformations; each of the other lines represents a different conformation.

ideal B-DNA, which depend on the nearest-neighbor bases. With CHARMM27, the three more sequence-dependent parameters, (tip, propeller twist, and buckle) show large deviations from the ideal values: up to 25° for the tip, 30° for the propeller twist, and 50° for the buckle at A:T basepairs; the maximum deviations observed in the analysis by Lam and Au-Yeung (1997) are 7°, 10°, and 9° respectively. In the x-ray structure of the $d(CGATTAATCG)_2$ decamer (Quintana et al., 1992), the maximum deviation in propeller twist between two A:T basepairs is 10.5° (between A6:B5 and B3:A8). However, despite these large variations in the simulations, the sequence-specific basepair variability is preserved on average. Two parameters show a slightly different behavior: the inclination and X-displacement. The plots of the X-displacement and inclination at the different basepair positions are shown in Fig. 6 for the simulations with the CHARMM27, AMBER, and BMS force fields. The basepair parameters are represented for each of the five representative conformers, the average structure, and the x-ray structure for comparison. With CHARMM27, the X-displacement is slightly underestimated on average (twice the maximum deviation of 0.5 Å observed by Lam and Au-Yeung) whereas the inclination is slightly overestimated with respect to the x-ray structure (less than the maximum deviation of 18° observed by Lam and Au-Yeung). This tendency to underestimate the X-displacement and overestimate the inclination was also identified in a previous study with CHARMM22 (Feig and Pettitt, 1998). Equivalent results are obtained with the AMBER force field although CHARMM27 tends to better preserve the sequence-specific variations (Fig. 6). On the contrary, the BMS force field reproduces very well the X-displacement and to a lesser extent the inclination of the x-ray structure: the X-displacement and inclination of the average structure are almost identical to those of the x-ray structure at all basepairs. The deviations in X-displacement and inclination between the five representative conformers and the x-ray structure are also very small (Fig. 6).

The variation in the basepair step parameters (tilt, slide, rise, roll, twist, cup) are rather small and close to those observed in the x-ray structure. For illustration, we have chosen to describe more in detail: on the one hand, two backbone independent parameters, the roll and tilt that contribute to the DNA bending (Fig. 7 *a*), on the other hand two sequence-context-dependent parameters, the twist and slide (Fig. 7 *b*).

With CHARMM27, there is high roll at TpA (TA/TA: A5A6/B5B6) basepair step and low roll at the TpT (TT/AA: A4A5/B6B7 and B4B5/A6A7) and ApT (AT/AT: A3A4/B7B8 and B3B4/B7B8) basepair steps; high tilt at the TpT (TT/AA: A4A5/B6B7) basepair step and low tilt at the ApT (AT/AT: A3A4/B7B8) and ApA (AA/TT: A6A7/B4B5) basepair steps. All these sequence-specific features are in agreement with the experimental data (Gorin et al, 1995). Another sequence-specific feature that is well described is the bistable character of TpA basepair steps: in the five conformers, when the slide is negative (Fig. 7 *b*), the roll is positive and when the slide is positive, the roll is low (Fig. 7 *a*), a rule established from the analysis of base stacking interactions (Hunter, 1993). All the parameters that are essentially backbone independent (roll, tilt, and rise) give variations with CHARMM27 that are compatible with the zsequence-specific variability observed in x-ray structure. On the other hand, the parameters which are backbone-dependent (twist) or strongly sequence-context-dependent (slide) deviate more at some basepair steps. For example, the twist tends to be significantly large, specifically at the central TA/TA basepair step (Fig. 7 *a*). The AMBER force field gives similar results: the roll and tilt at the ApA, TpT, ApT, and TpA basepair steps are close on average to what is observed in the x-ray structure (Fig. 7 *a*). The bistable character, in terms of roll and slide (Fig. 7 *b*) of the TpA basepair step is also reproduced by AMBER. The twist is significantly smaller (compared to that obtained with CHARMM27), except for the central TA/TA basepair step. This is also a feature observed with the BMS force field (Fig. 7 *b*). The latter force field again gives smaller deviations from the basepair parameters of the x-ray structure. The larger deviations in the basepair step parameters obtained with the CHARMM27 and AMBER force fields result from the larger size of the accessible conformational space, as compared with the more rigid BMS simulation. The larger deviations in roll and tilt between the MD simulated structures and the x-ray structure are observed at the central TpA step with CHARMM27 and AMBER. Because both roll and tilt contribute to the local DNA bending, the larger RMSD with respect to the x-ray structure observed with these two force fields in comparison with the BMS force field (Fig. 2) might be due to some local bending at this particular base step. The plots of the roll, tilt and bending versus time (data not shown) reveal that in the case of AMBER, the jump in RMSD around 500 ps (Fig. 2) is associated with an increase in DNA bending at the GA/TC (A2A3/B8B9) and TA/TA (A5A6/B5B6) basepair steps. In the case of CHARMM27, the first jump in RMSD at 250 ps is related to a general increase of DNA bending at almost all basepair steps. In contrast, no marked bending is observed with the BMS force field, including at the TpA basepair step.

Major and minor groove hydration and ion distribution

Experimental data (Eisenstein and Shakked, 1995; Tippin and Sundaralingam, 1997; Egli et al., 1998; Tereshko et al., 1999) and MD simulations (Young et al., 1997; Feig and Pettitt, 1999a,b) have revealed the existence of hydration patterns which are specific to the A and B forms of DNA, and related to specific sequences. The hydration of the major groove and minor groove show very distinct patterns depending on the DNA form. The exocyclic amino group of adenines in the major groove is generally not solvated in

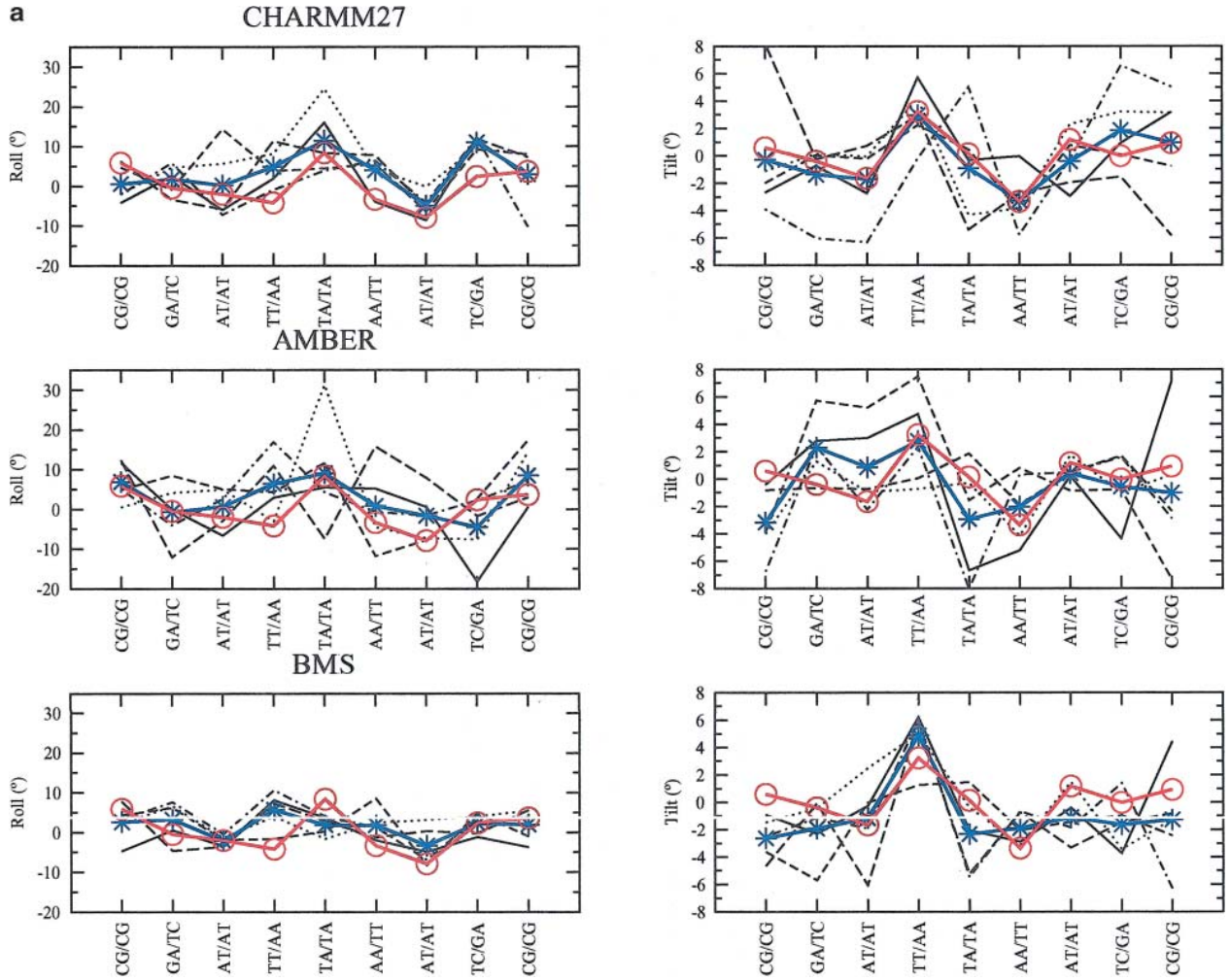


FIGURE 7 Basepair step parameters (Roll, Tilt, Twist, Slide) of the five representative conformations for CHARMM27, AMBER, and BMS. (*a*) Roll and Tilt parameters, (*b*) Twist and Slide parameters. The bold solid line marked by circles represents the basepair parameter of the x-ray structure; the solid line marked by stars represents the basepair parameter averaged over the five conformations; each of the other lines represents a different conformation.

A-DNA because of the formation of hydrogen bonds with stacked thymine bases. By contrast, a string of water molecules bridges adjacent adenines and thymines from opposite strands in the minor groove of B-DNA (Egli et al., 1998). Some sequence-specific patterns are also observed, such as the extended ''hydrat-ion'' spine in B-DNA minor groove at ApT basepair steps (Tereshko et al., 1999). This specific minor groove hydration is believed to be due to preferred monovalent metal ion coordination bridging the N3 and O2 atoms of adenine and thymine bases, respectively (Tereshko et al., 1999). Other hydration patterns are generally less extended, depending on the sequence, than the ''caterpillar-like'' structures typical of the minor groove hydration. They result from the presence of more local hydration sites at specific atomic positions, for example at the N4 atoms of cytosines (Feig and Pettitt, 1999a). More extended hydration patterns can also be observed in the major groove, at the four water oxygen atoms along the O6, N7, C8, and backbone phosphate atoms of the guanine bases (Feig and Pettitt, 1999a). Some of this behavior has been reproduced and analyzed with molecular dynamics simulations using the CHARMM22 (MacKerell, 1997) and AMBER1995 (Cheatham and Kollman, 1997) force fields or both force fields (Feig and Pettitt, 1997).

The local sequence-dependent hydration around the hydrogen-bond donors and acceptors of the DNA duplex is described in Fig. 8. The diagrams of local hydration around DNA sites show that the phosphodiester backbone (O1P and O2P) is more strongly hydrated. As found from the x-ray crystallographic analysis of the hydration of DNA at atomic resolution (Egli et al, 1998), the largest number of first-shell water molecules in the A (CHARMM22) or B form DNA duplexes (CHARMM27, AMBER, and BMS) are located around phosphate groups (oxygen atoms O1P, O5′, O3′, O2P, and O4′). This strong hydration is particularly concentrated around the nonbridging oxygens (O1P and

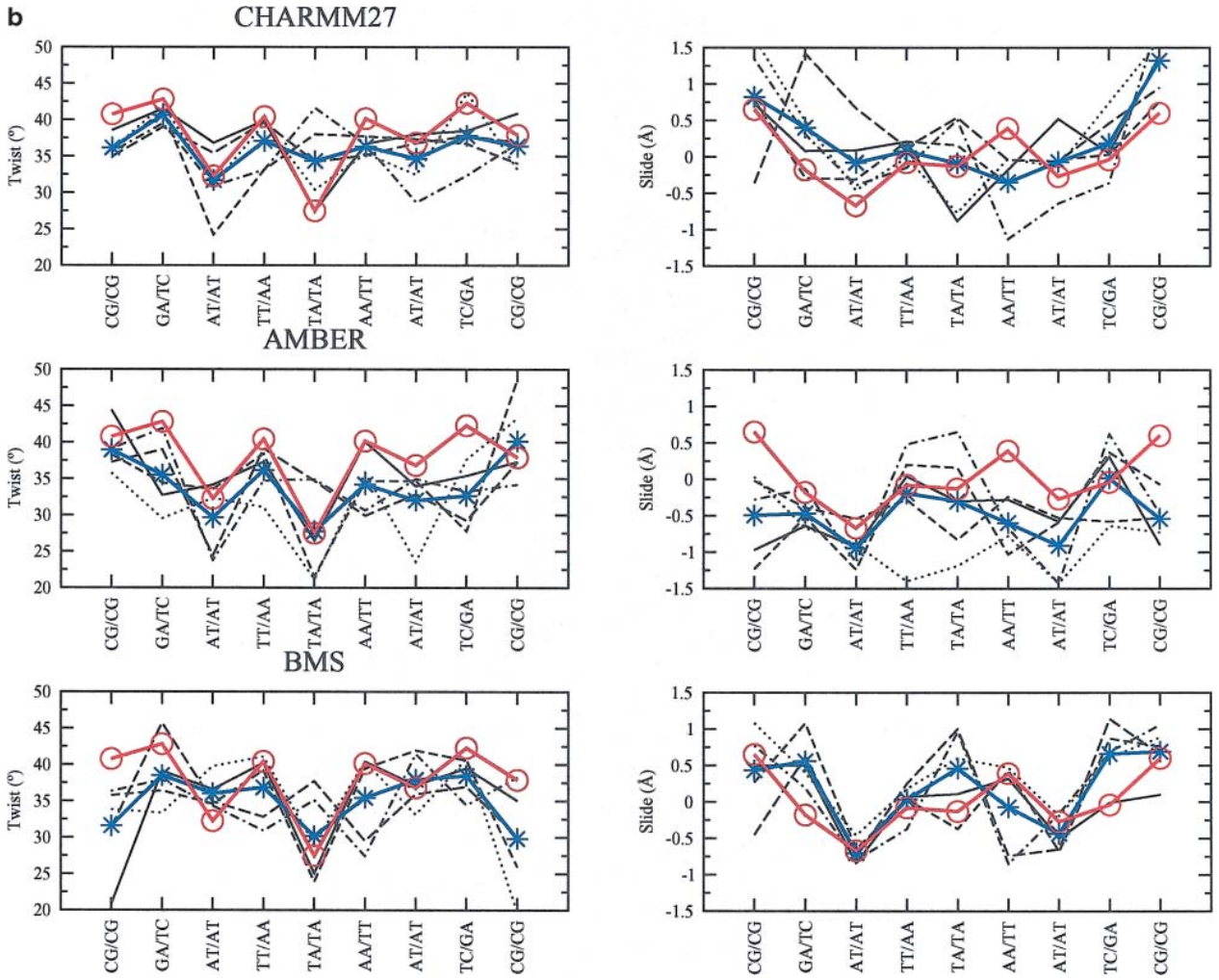


FIGURE 7 Continued

O2P) of phosphate groups, a feature observed with all the force fields (Fig. 8, *a–d*). The more exposed anionic oxygen of the phosphate group (O1P) is slightly more solvated (Fig. 8 *e*). The BMS force field and to a lesser extent the AMBER force field lead to a stronger solvation of the nonbridging oxygens whereas the CHARMM force fields (and more particularly the CHARMM27 force field) generate a stronger solvation of the sugar oxygen O4′, which contributes to the stabilization of the water spine in the B-DNA minor groove. The more favorable hydration sites of the nucleic acid bases are located around the O6/N6 atoms of purines, N2 of guanines, and N4 of cytosines (Fig. 8, *a* and *b*). A comparison between the different force fields reveals that the solvation of the nucleic bases is stronger with BMS and AMBER than with the CHARMM force fields (Fig. 8 *f*). At all the base sites, the strength of solvation decreases from BMS to AMBER, CHARMM27, and CHARMM22. This trend is slightly changed at O6/N6 sites where AMBER gives a stronger solvation. On the other hand, the trend is reversed at N7 sites. At the basepair level, the G:C (or C:G) basepairs are more solvated than A:T (or T:A) basepairs with any of the force fields (Fig. 8 *f*). Among the A:T (or T:A) basepairs (6 out of the 10 basepairs, Fig. 1), the symmetrical fourth and seventh basepairs, corresponding to ApT steps, are more strongly solvated, a feature common to the force fields leading to a stable B-DNA (AMBER gives a behavior slightly different where the eighth basepair is a bit more solvated than the seventh one). It is noteworthy that the residence time of water molecules around the nucleic acid bases in the grooves is not significantly longer than that around the phosphate groups; the proportion of short-lived and long-lived water molecules is similar (data not shown). Nevertheless, the presence of water-water contacts in the vicinity of the various and more or less strong hydration sites leads to clear differences in local water densities as discussed below.

To determine the ability of the different force fields to reproduce specific hydration patterns, we have compared the water and ion densities calculated from the MD trajectories (Figs. 9 and 10). For the CHARMM27 simulation, we first

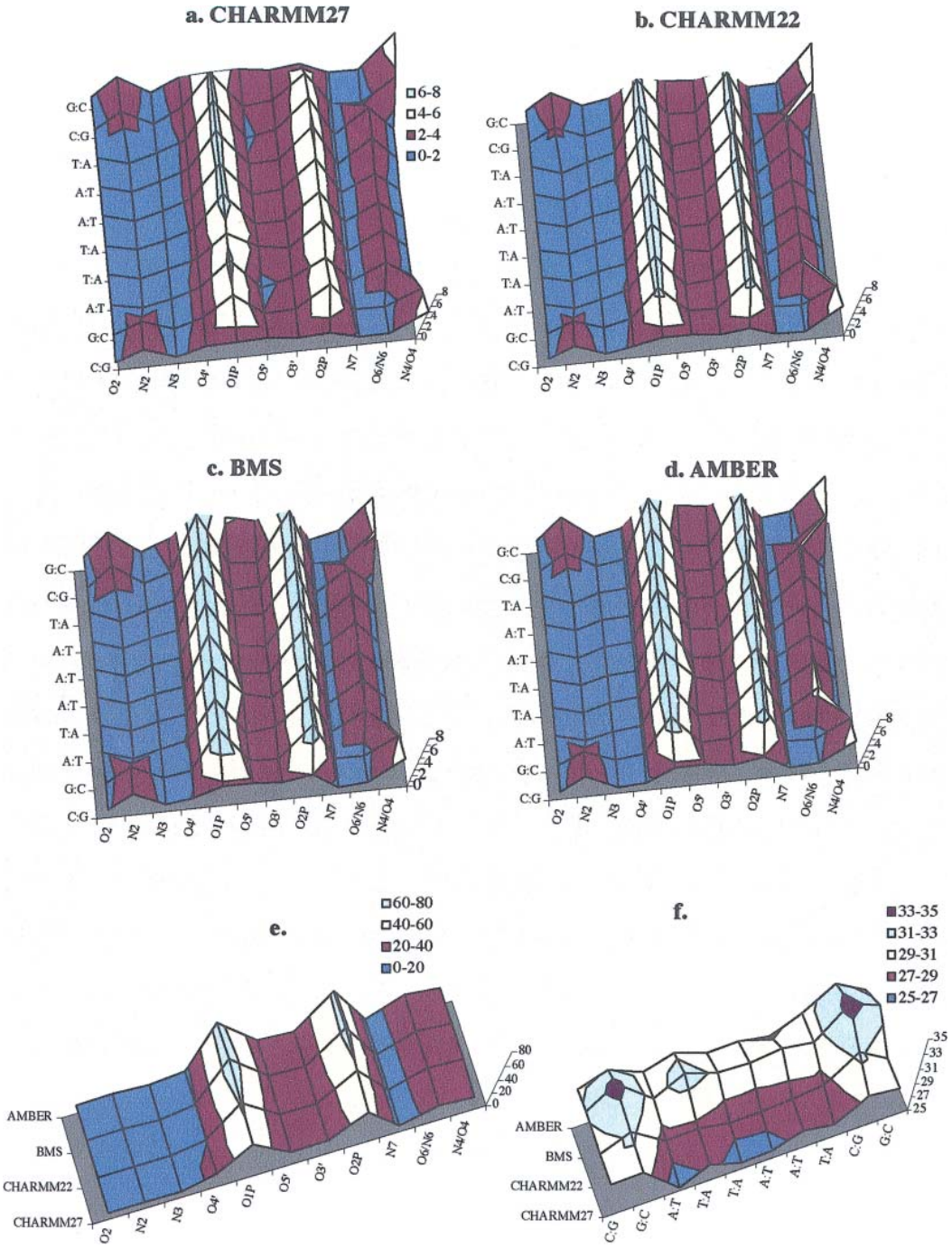


FIGURE 8 Diagram illustrating the number of water molecules in the hydration shells of individual oxygen and nitrogen atoms averaged over the DNA simulations for the different force fields: (*a*) CHARMM27, (*b*) CHARMM22, (*c*) BMS, (*d*) AMBER. The number of water molecules averaged over all atomic sites for all basepairs is shown in (*e*); the average number of water molecules at all atomic sites around each basepair is shown in (*f*). Water molecules are counted if they are less than 3.0 Å from the hydrogen-bonding partners in the DNA. Colors are used to make clearer the different basepair positions (from the first basepair at the bottom to the last one at the top).

compare the positions of high water density with those of the crystal water molecules. Sixty percents of the high-density spots in the map overlap positions of the crystal water molecules (data not shown). The phosphate groups tend to be more hydrated than the nucleic acid bases (Egli et al., 1998) because of the negative charge carried by the phosphate groups. In the calculated results, the high density spots are generally smaller around the phosphate groups than around the bases (Fig. 11). The organized waters around the phosphate (Fig. 11 *a*) correspond to the first hydration shell

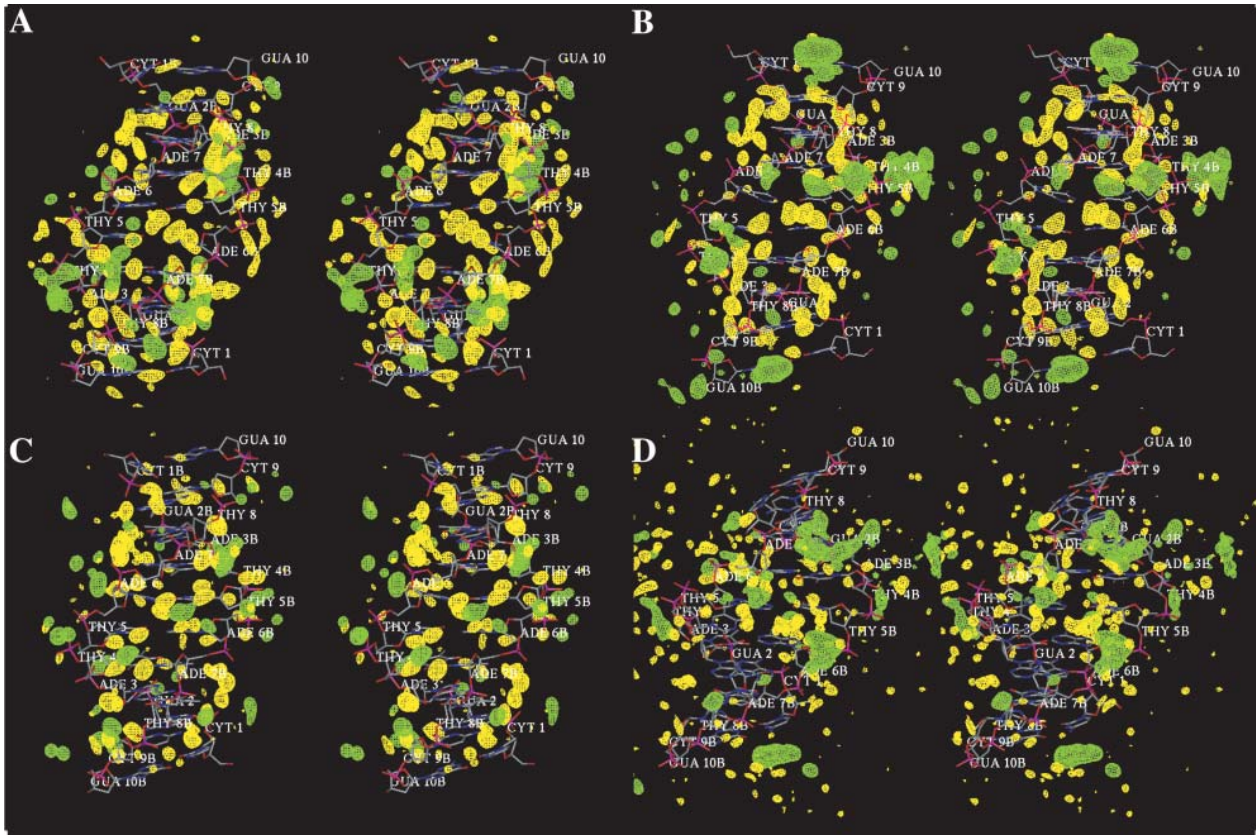


FIGURE 9 Wall-eyed stereo view of water and ion hydration patterns in the minor groove of B-DNA as obtained from the MD trajectories with (*a*) AMBER, (*b*) BMS, (*c*) CHARMM27, (*d*) CHARMM22. Density contours show water oxygen (*yellow*) and sodium ion (*green*) densities at level of 80 water molecules per nm^3 and 10 ions per nm^3 calculated from a grid with 0.25 Å resolution.

but are kinetically labile (Denisov et al., 1997); the second shell is unorganized. The water around the bases are well-organized and establish hydrogen bonded networks revealed by the presence of large and extended caterpillar-like density spots at specific sites in the grooves (Fig. 11 *b*). This is in agreement with observations on the hydration of phosphate groups (Schneider et al., 1998), the kinetics of DNA hydration (Denisov et al., 1997) and the x-ray data showing that the water is more ordered around the bases than around the phosphate groups (Egli et al., 1998). The difference of hydration between the phosphate backbone and the DNA grooves comes from the fact that the water molecules around the nucleic acid bases are more organized in the first and second hydration shells. The analysis of the time-dependent organization of water molecules around a high water density spot, in the major groove of ADE 7B, reveals the presence of various organized water molecules in the shells of hydration (Fig. 12). The first shell water molecules making contacts with either the O6 or N7 atoms of ADE 7B are stabilized by second shell water molecules that establish hydrogen bond contacts with those of the first shell and sometimes also with the neighboring phosphate groups.

The hydration patterns generated with the different force fields do not differ significantly, except for CHARMM22, which leads to small and very dispersed water density spots (Figs. 9 *d* and 10 *d*). This is due in part to the transition from B- to A-DNA that prevents the existence of organized and long-lived water molecules around the bases. For the other force fields, preferred hydration sites are localized at the O2 and N3 atoms of pyrimidines and purines, respectively, in the minor groove fields (Fig. 9, *a–c*), and at the O6 and N7 atoms of pyrimidines and purines, respectively, in the major groove (Fig. 10, *a –c*). In the minor groove, preferred hydration sites associated with high water density are observed at the N2 atoms of GUA A2 and B2, at the N3 atoms of ADE A3, B3, A6, and A7, at the O2 atoms of THY A4, B4, A8, B8, and at the O2 atoms of CYT A9 and B9. In the case of the AMBER and BMS force fields (Fig. 9, *a* and *b*), many extended high density spots are observed due to the proximity of hydrogen bond acceptors or donors between successive basepairs either on the same strand or on both strands. For example, the region defined by the N3 atoms of GUA B2, ADE B3, and the O2 atom of CYT A9 is associated with a unique and extended high density spot spread out across the minor groove. These density spots correspond to ordered water molecules in the minor groove of B-DNA identified in many x-ray structures (Drew and Dickerson, 1981; Kopka et al., 1983; Shui et al., 1998;

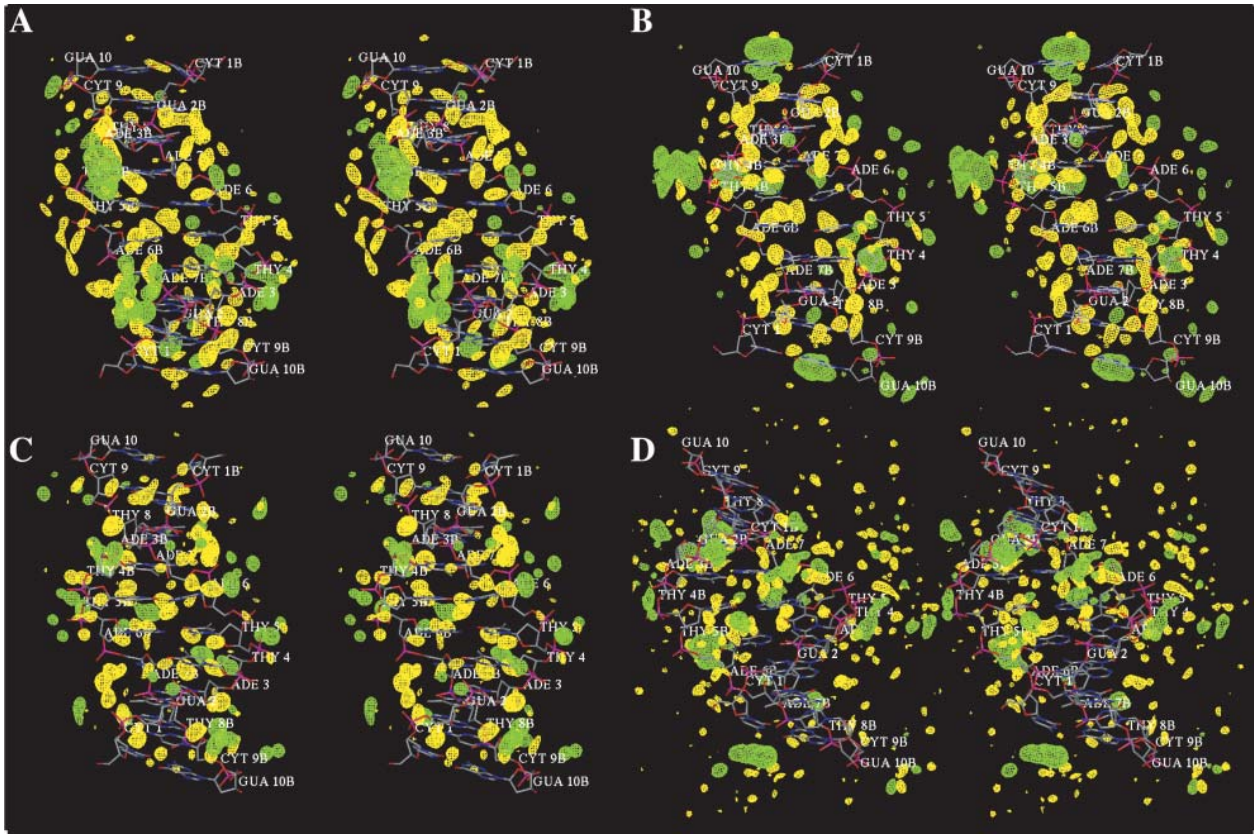


FIGURE 10 Wall-eyed stereo view of water and ion hydration patterns in the major groove of B-DNA as obtained from the MD trajectories with (*a*) AMBER, (*b*) BMS, (*c*) CHARMM27, (*d*) CHARMM22. Contours show water oxygen (*yellow*) and sodium ion (*green*) densities at level of 80 water molecules per nm^3 and 10 ions per nm^3 calculated from a grid with 0.25 Å resolution.

Tereshko et al., 1999; Egli et al., 1998). Such high density spots are also present around the O2 atoms of THY A4 and THY B8, which belong to two successive basepairs (THY A4/ADE B7 and ADE A3/THY B8), and similarly around the O2 atoms of THY B4 and THY A8. Another hydration site with a similar pattern is found with the AMBER and CHARMM27 force fields (Fig. 9, *a* and *c*): it involves the N2 atoms of GUA A2 and GUA B10. As mentioned above,

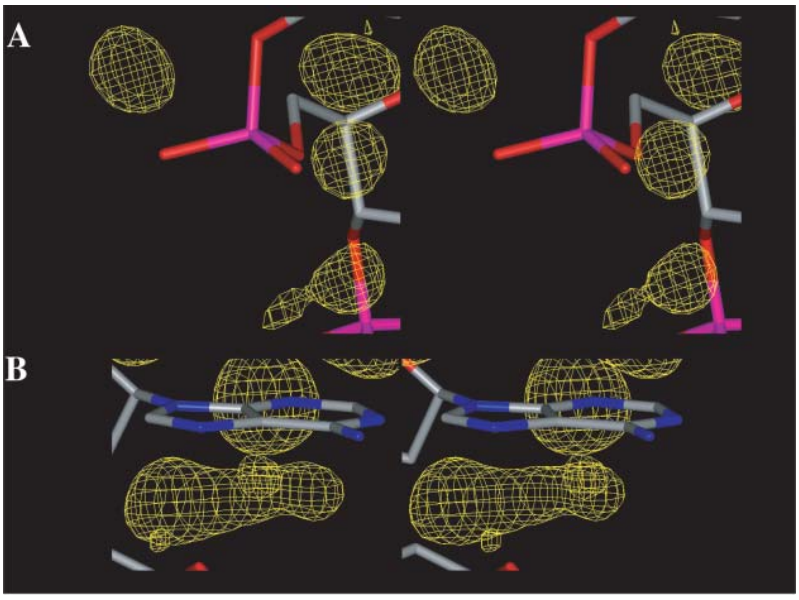


FIGURE 11 Details of water hydration patterns around the phosphate group of THY 5B (*A*) and around the major and minor groove edges of ADE 7B (*B*) obtained with CHARMM27.

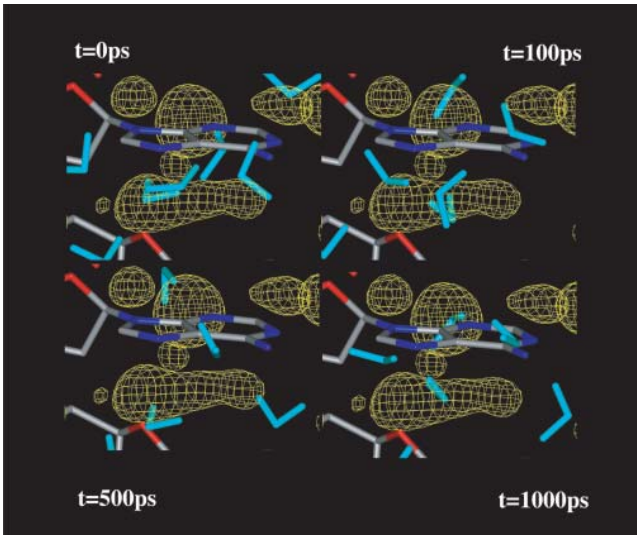


FIGURE 12 Snapshots at $t = 0,100,500,1000$ ps from the CHARMM27 simulation showing the water molecules (*cyan*) in the first and second hydration shells around the nucleotide ADE 7B corresponding to the high water density at this position in the major groove.

such hydration patterns are much less frequent in the CHARMM27 simulation. This is due to the presence of many counterions in the minor groove, revealed by the high ion density spots, that prevent the existence of extended water network. Thus, small high water density spots are juxtaposed with high ion density spots. In the major groove, more extended high water density spots are observed in the CHARMM27 simulation due to the lower population of counterions. High water density spots spread out across the major groove between successive basepairs are observed with the AMBER, BMS, and CHARMM 27 force fields (Fig. 10, *a–c*). They involve the O6 atoms of GUA B2 and THY A8 (and its symmetrical site with GUA A2 and THY B8), the O2 atom of THY B4 and the N6 atom of ADE A6, or the O2 atoms of THY A5 and THY B5, which are 2.62 Å and 2.92 Å far away, respectively, from the same crystal water molecule. Another common hydration pattern corresponds to the high water density spot around the N6 atom of CYT B9 that stretches to the phosphate group of ADE B7.

Sequence-specific binding sites for sodium ions in the minor groove, revealed by the presence of high ion density spots, are shown in details at the following basepair steps: CpG (Fig. 13), ApT (Fig. 14), TpA (Fig. 15), and TpT (Fig. 16). Some ion binding sites are common to DNA duplexes simulated with the different force fields. In other cases, high ion density spots can be substituted by high water density spots. For example, the high water density spot present around CYT A1, GUA A2, and CYT B9 observed with CHARMM27 (Fig. 13 *a*) is substituted with AMBER by a high water density spot in this region (Fig. 13 *b*). In a MD simulation on a B-DNA decamer using the AMBER force field, Young et al. (1997) have found high counterion density in a region of same sequence CpGpA (A1-A2-A3 and B1-B2-B3 in this DNA decamer).

The minor groove of AT basepairs is known to be a preferred location for monovalent cations (Sissoeff et al., 1976; Denisov et al., 1997; Halle and Denisov, 1998; Denisov and Halle, 2000). ApT steps are strong sites for monovalent cations localization because of a uniquely low electrostatic potential at these base steps (Young et al., 1997). The DNA structure contains two ApT basepair steps, A3A4/B7B8 and A7A8/B3B4, which are potential binding sites for monovalent ions as identified in x-ray structures determined

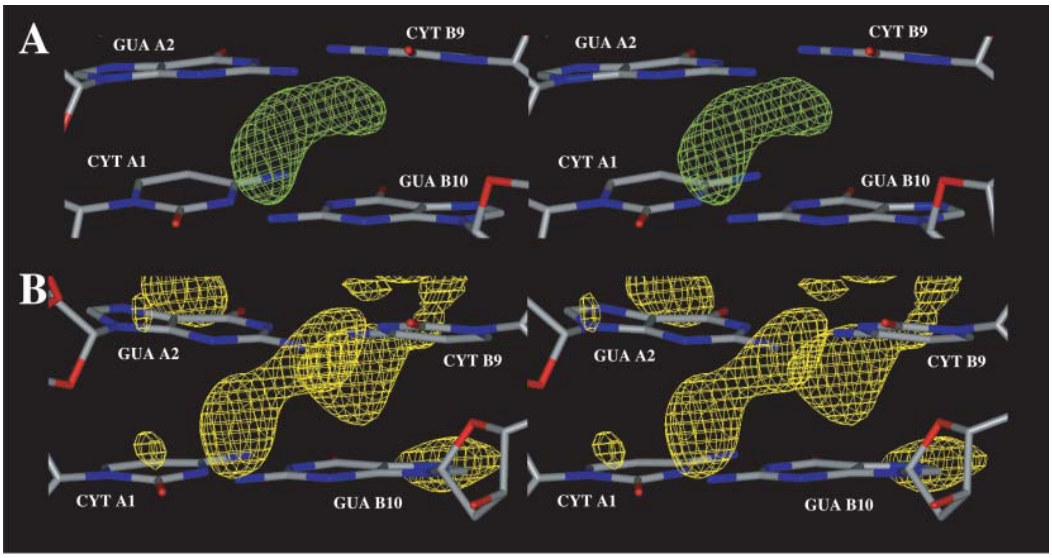


FIGURE 13 Zoom in view of sequence specific hydration and ion binding patterns (the contours for ions or water molecules are represented as described in Fig. 12) at the first CpG (A1A2/B9B10) basepair step obtained with: (*A*) CHARMM27 (counterions in *green*), (*B*) AMBER (water density spots in *yellow*).

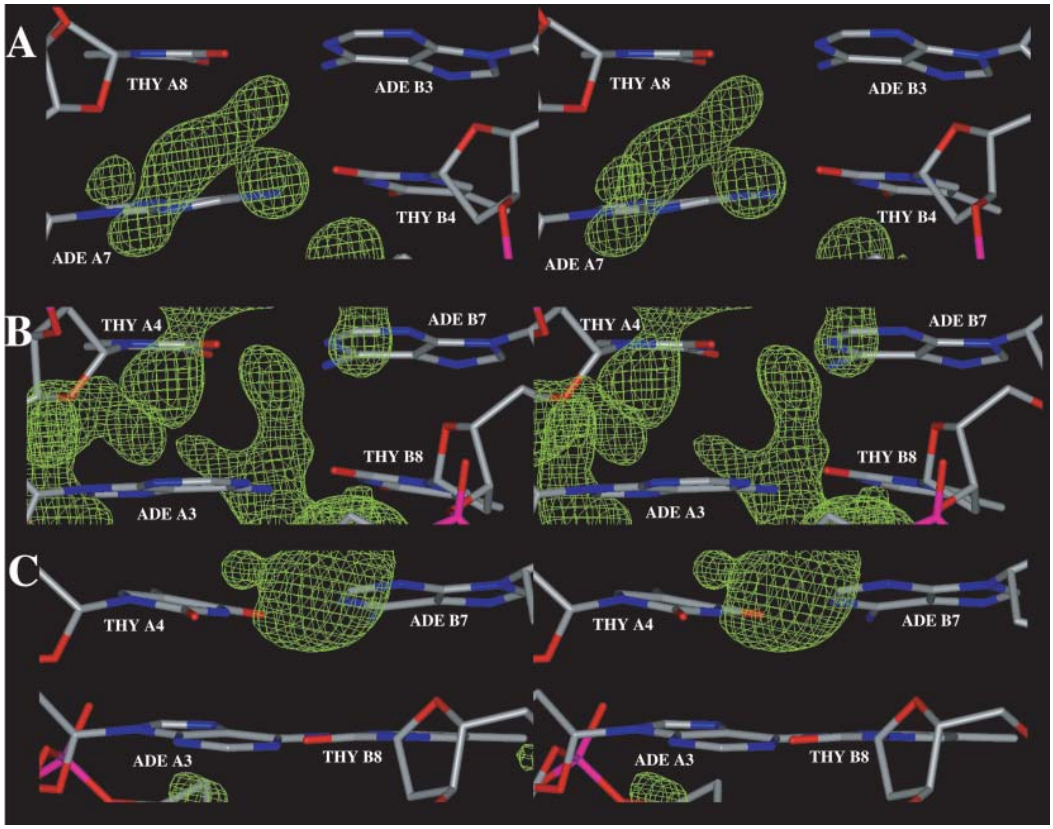


FIGURE 14 Zoom in view of sequence specific hydration and ion binding patterns at the ApT basepair steps observed with: (*A*) CHARMM27 (A7A8/B3B4), (*B*) AMBER (A3A4/B7B8), (*C*) BMS (A3A4/B7B8).

at very high resolution (Tereshko et al., 1999). One site, at the ApT basepair step A7A8, is identified by CHARMM27 (Fig. 14 *a*) and the alternate site, at the A3A4 basepair step, is identified by AMBER (Fig. 14 *b*); neither of the two is identified with BMS. An additional monovalent ion binding site is found in the minor groove with CHARMM27 at the central TpA basepair step (A5A6/B5B6, Fig. 15 *a*) although it does not correspond to a binding site as strong as those at

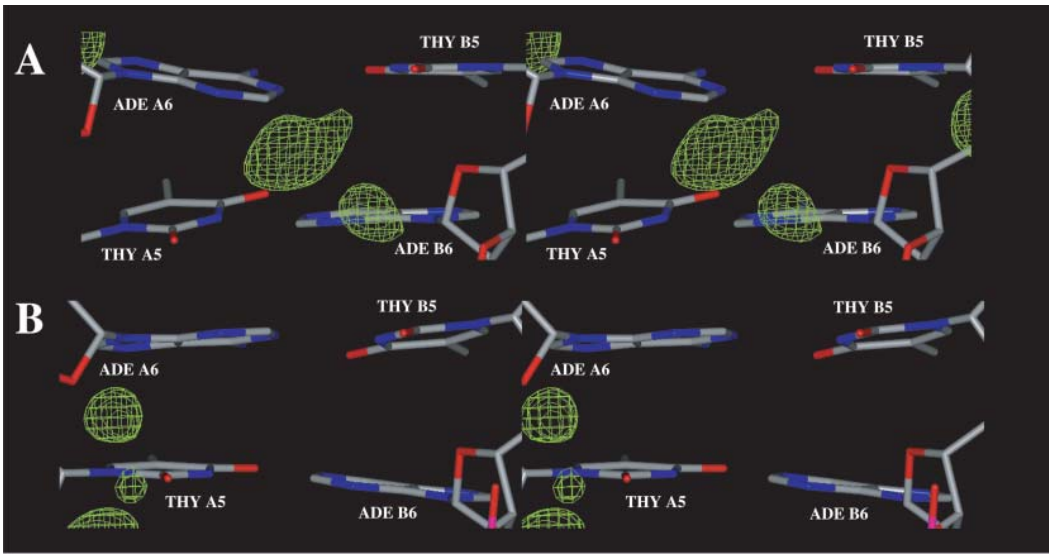


FIGURE 15 Zoom in view of sequence specific hydration and ion binding patterns at the TpA basepair step (A5A6/B5B6) observed with: (*A*) CHARMM27, (*B*) AMBER.

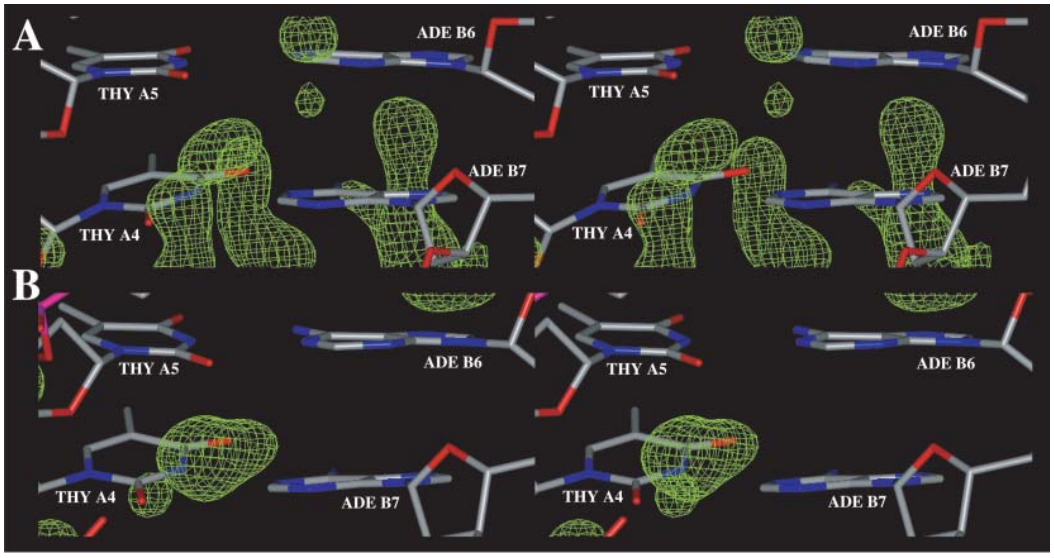


FIGURE 16 Zoom in view of sequence-specific hydration and ion binding patterns at the TpT basepair step (A4A5/B6B7) observed with: (*A*) AMBER, (*B*) CHARMM27.

the ApT steps (Denisov and Halle, 2000). No similar binding site is observed with BMS; a high counterion density spot is observed with AMBER around the residue A6 but at more than 8 Å away from the N3 of the adenine (Fig. 15 *b*). In the x-ray structure of the DNA decamer, there is one magnesium site centered at the TpT basepair step A4A5 (basepaired to B6B7). A high ion density spot, corresponding to a sodium ion binding site, is observed in this region with AMBER (Fig. 16 *a*) and CHARMM27 (Fig. 16 *b*). Such a sodium binding site has already been observed in MD simulations at TpT steps (Feig and Pettitt, 1999b). In the case of BMS, this region is occupied by high water density spots and the close high ion density spot is distant from the original magnesium binding site by more than 8.0 Å. For comparison, the positions of the different ion binding sites identified in the minor groove with the three force fields (AMBER, BMS, CHARMM27) are represented in Fig. 17. It appears that the presence of counterions in the minor groove (Fig. 17) leads to a local groove narrowing at the corresponding cross-strand phosphate groups or neighboring cross-strand phosphate groups (Table 1). A local narrowing is observed at the PA_4-PB_{10}, PA_5-PB_9, PA_6-PB_8 cross-strand positions with AMBER: two sodium binding sites are observed at the last two positions (Fig. 17 *a*). A local narrowing is also observed at other positions: PA_6-PB_8, PA_7-PB_7, PA_8-PB_6 with CHARMM27; two sodium binding sites are present at the two first positions and the third position is inserted between two sodium binding sites (Fig. 17 *c*). By contrast, the local narrowing observed with BMS seems to be more sequence-specific because no close sodium binding sites are observed in the minor groove.

The ions are mostly concentrated around the phosphate groups with CHARMM22 (Fig. 9 *d*). More counterions are present in the major groove, as we expected for A-DNA. They are also more organized than those in the minor groove with extended caterpillar-like structures (Fig. 10 *d*). In B-DNA duplexes, because of the larger solvent accessibility of the major groove, counterions can be distributed between multiple binding sites. The presence of only a few high density spots for sodium ions in the simulation structure with CHARMM27 (Fig. 10 *c*) is consistent with the idea of a reduced localization of counterions in the major groove of B-DNA compared to that of A-DNA (Feig and Pettitt, 1999a, b). Some rare and localized high ion density spots are observed in the major groove of the simulated structures; they can also form caterpillar-like structures and extend from the major groove edge to the phosphate groups, and they are generally located around the N7 of purines (Figs. 10 and 18). They are located at the GpA (A2A3/B8B9) and ApA (A6A7/B4B5) basepair steps with AMBER. A similar location is observed for the only ion density spot obtained with CHARMM27 at the ApA basepair step whereas they appear only at the terminal basepairs with BMS.

Because of the short simulation times, artifacts might appear in the structural and solvation properties of the DNA decamer. Thus, we extended the MD simulation with the CHARMM27 force field to 2 ns to determine the reliability of our results. The DNA structure remains stable and gets closer to both the x-ray structure and a canonical B-DNA after 1400 ps (Fig. 19). From the structural point of view, as a result, the sugar pucker tends to converge toward a C2′-endo pucker; C2′-endo to C3′-endo transitions are less frequent (THY B4) or absent from the second part of the simulation (ADE A3). The more stable C3′-endo conformation, in the first part of the simulation at ADE B7, switches to a C2′-endo conformation that is conserved until the end of

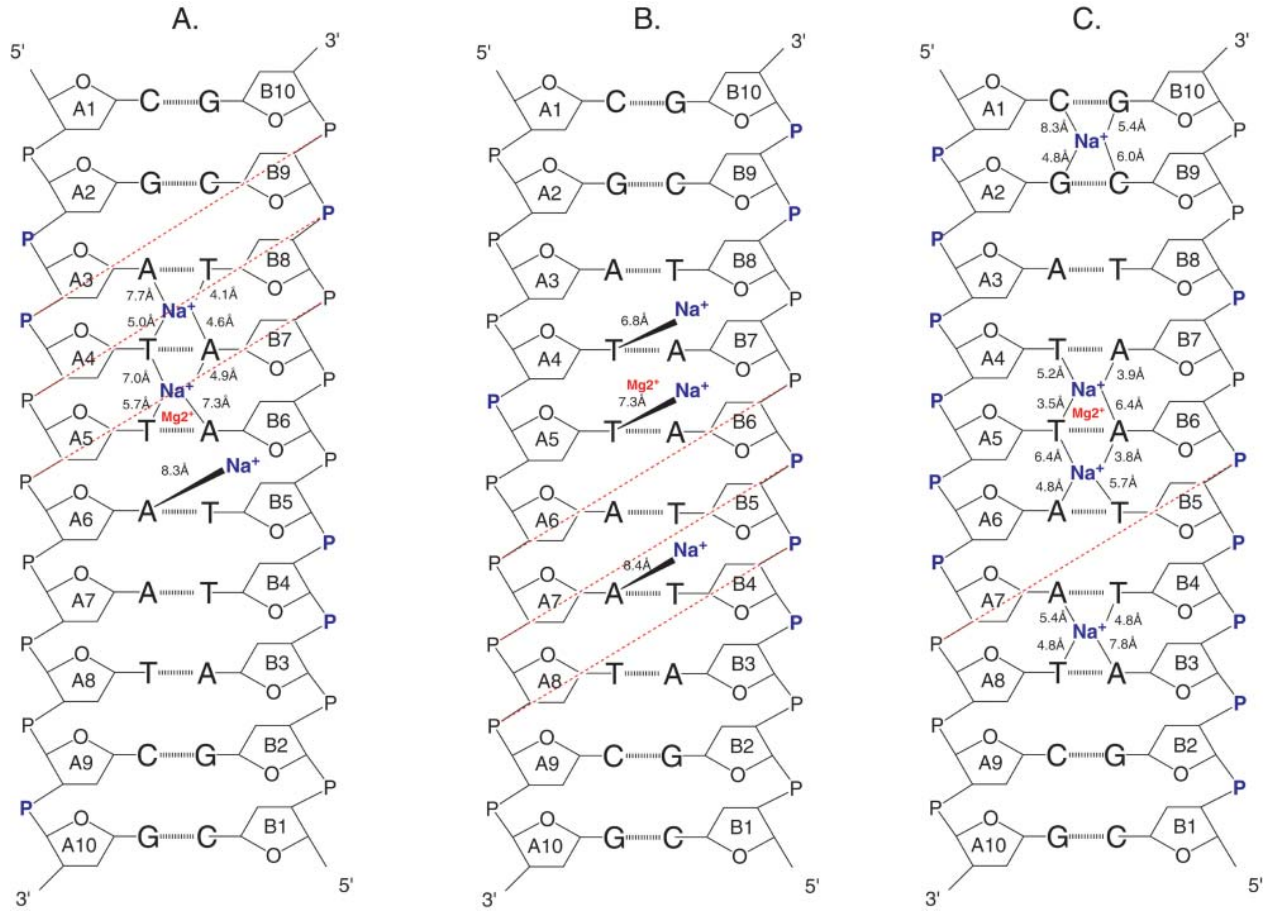


FIGURE 17 Schematic representation of the locations of counterions binding sites in the minor groove of simulated structures obtained with: (*A*) AMBER, (*B*) BMS, (*C*) CHARMM27. The positions of sodium ions (Na^+) are indicated by lines annotated with the coordination distance to N3 of purines and O2 of pyrimidines. The position of the magnesium binding (Mg^{2+}) site found in the x-ray structure is indicated for comparison. The presence of high counterion density spots around the phosphate groups are pointed out by a purple *P* atom. A red dashed line joins the cross-strand phosphate groups when their distance is lower than the standard value for a regular B-DNA helix (narrowing of the minor groove).

the simulation without any repuckering (Fig. 19). Similarly, the deviations of the basepair or base step parameters with respect to the x-ray structure and a canonical B-DNA are slightly reduced. In particular, the deviations in roll and tilt at the central TpA step are reduced and compare well with those observed at this base step with the BMS force field. From the point of view of solvation, some qualitative and quantitative changes are observed around the phosphate groups, which can be related to some changes in sugar pucker (downstream from the position B7 for example). However, no qualitative change is observed regarding the position of sodium binding sites in the minor groove.

CONCLUSIONS

The average simulated structures of the $d(CGATTAATCG)_2$ DNA double helix obtained with the CHARMM27, AMBER, and BMS force fields have a B-form geometry close to the x-ray structure; the starting B-DNA structure undergoes a transition to an A-like DNA structure with CHARMM22. Thus, all three of the newer force fields are clearly more useful for simulation studies of DNA than is CHARMM22. From a strict comparison of the average simulation structures with the x-ray data, the BMS force field gives the best agreement. The two other force fields, AMBER and CHARMM27, give similar results although each of them better describes certain structural features of B-DNA. A major difference between the BMS force field and the other two force fields is that the BMS simulation yields a significantly more rigid structure. The higher conformational flexibility obtained with AMBER and CHARMM27 may be more consistent with the available experimental data. In the simulation with the BMS force field, there are no ions in the primary solvation shell of the minor groove. The lower calculated conformational flexibility of B-DNA with BMS might be responsible for this, because ''breathing'' of the minor groove is necessary to accommodate some ions. Alternatively, the energetic penalty associated with the partial

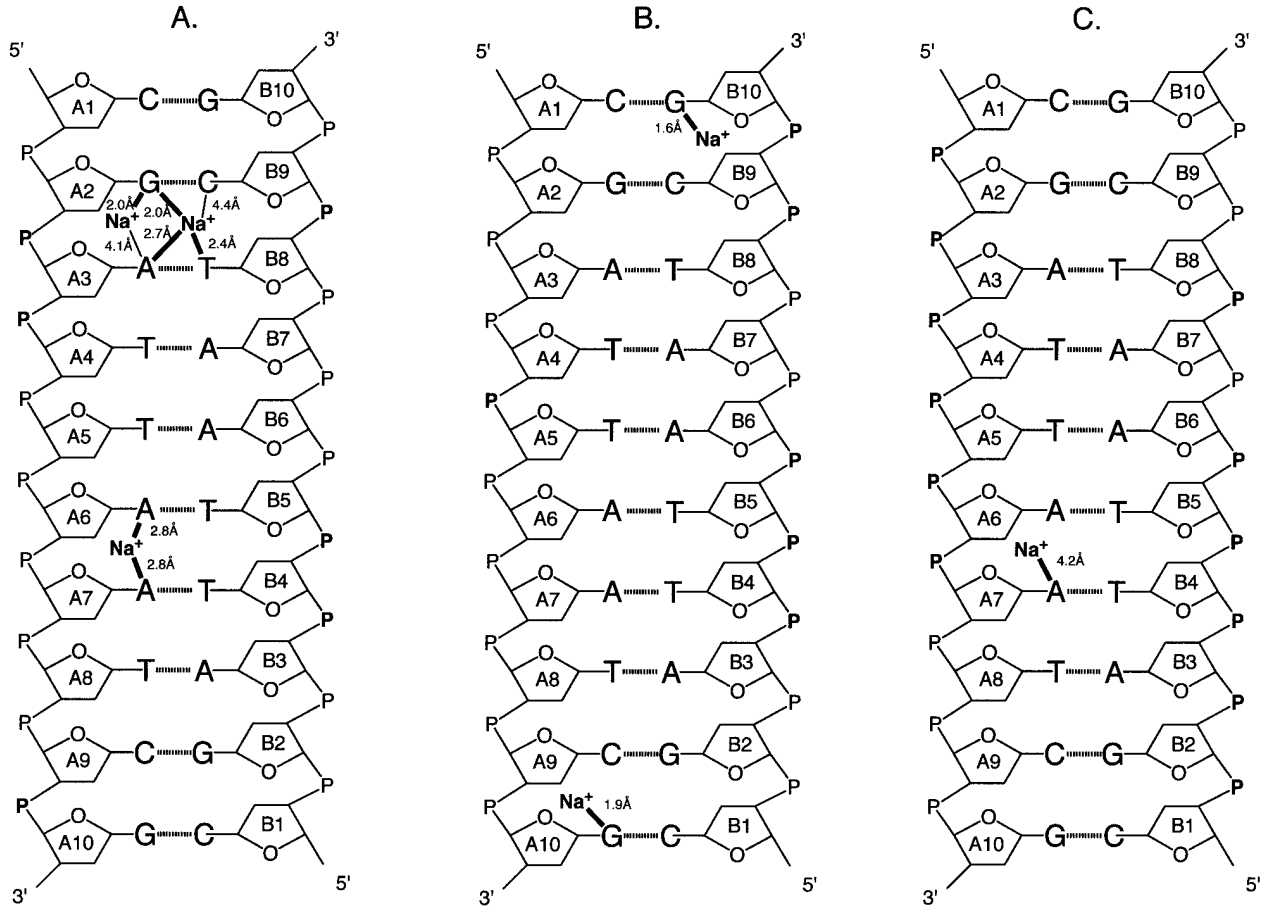


FIGURE 18 Schematic representation of the locations of counterions binding sites in the major groove of simulated structures obtained with: (*A*) AMBER, (*B*) BMS, (*C*) CHARMM27. The positions of sodium ions (Na^+) are indicated by lines annotated with the coordination distance to O6/N6 or N7 of purines and O4/N4 of pyrimidines.

dehydration of the counterions might be too high to allow the penetration of counterions into the minor groove. The BMS force field gives the strongest solvation of nucleic acid bases among all the force fields. The solvation of the minor groove in the AMBER and CHARMM27 simulations includes ions in the primary shell, in agreement with experimental data (Tereshko et al., 1999) and previous DNA simulations (Feig and Pettitt, 1999b). The minor groove of the average structures obtained with AMBER, BMS, or CHARMM27 tends to be nonspecifically wider than that of the x-ray structure (Table 1). However, some local narrowing of the minor groove is observed. The narrowing of the minor groove is generally associated with the presence of counterions in the minor groove (Feig and Pettitt, 1999b; Hamelberg et al., 2000). The presence of a divalent cation (Mg^{2+}) at the TpT basepair step (A4A5/B6B7) leads to a local widening of the minor groove in the x-ray structure (distance PA_6-PB_8 in Table 1). In the simulated structures with AMBER and CHARMM27, the substitution of the divalent cation by a monovalent cation induces instead a local narrowing of the minor groove. In the absence of any counterion at the position of the original magnesium binding site, the minor groove does not show any narrowing with the BMS force field.

The anticorrelation observed between basepair inclination and the depth of the major groove (measured by the X-displacement) is another example of the influence of solvation on the global DNA conformation in solution. Because these two features vary in an anticorrelated way, they can be used to monitor the conformational changes between A-like DNA conformations (high inclination and a deep less hydrated major groove) and B-like DNA conformations (low inclination and a more shallow more hydrated major groove).

At a more detailed atomic level, we observe that the phosphate groups are more strongly hydrated than the nucleic acid bases consistent with x-ray crystallographic data (Egli et al., 1998). The water molecules around the nucleic acid bases do not have calculated residence times significantly larger than those around the phosphate groups. However, because of the lack of reorientation of water molecules around the bases due to water contacts between the first and second solvent hydration shells, the solvent

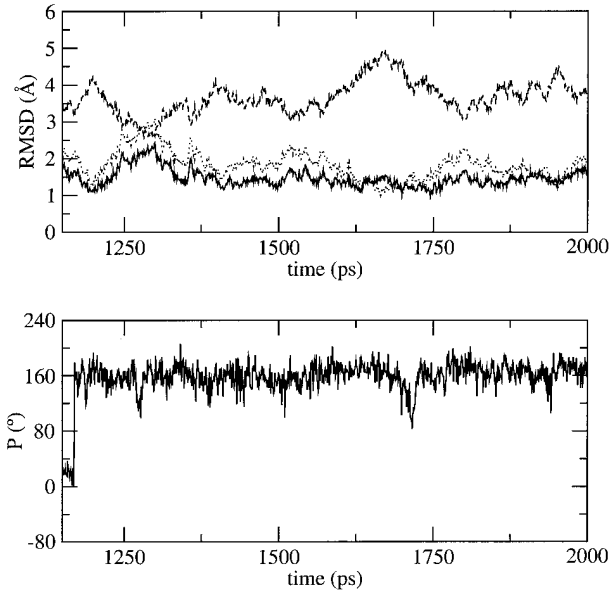


FIGURE 19 Selected structural changes observed during the MD simulation with CHARMM27 extended to 2 ns. (*top*) RMSD versus time from 1150 to 2000 ps. (*bottom*) Time dependence of the sugar pseudorotation angle at B7.

density is higher at specific base sites in the minor or major groove and leads to more extended caterpillar-like structures. Ion binding sites are generally more specific and localized in the minor groove for B-DNA duplexes. They can be partially dehydrated in the minor groove whereas they are generally fully hydrated in the major groove.

The results obtained with the CHARMM force fields show a significant improvement of the nucleic acid force field parameters in CHARMM27 relative to CHARMM22. The strong anticorrelation observed between the RMSD versus time with respect to the canonical A and B forms of DNA in the simulation with CHARMM27, together with the low anticorrelation between the base inclination and the X-displacement, indicates that the conformational space between the A and B forms of DNA is sampled. These results suggest that the CHARMM27 force field is appropriate for simulations of the influence of the environment on the form of DNA, including the relative stabilities and possible transitions between the A- and B-DNA forms.

We thank A. MacKerell for help with the CHARMM27 parameters and D. Langley for providing the BMS parameters before publication. We also thank R. Yelle, J. Ma, B. Matthias, and Y. Zhou for valuable discussions, and C. L. Brooks, M. Crowley, G. Ravishankar, and M. Feig for information on certain aspects of the PME method. The simulations were done on the T3E at Pittsburgh Supercomputing Center and at National Energy Research Scientific Computing Center under grants from National Institutes of Health and the Department of Energy, respectively. S.Y. was a recipient of a BOYSCAST fellowship from the Department of Science and Technology, (DST) INDIA. F.L. was a fellow of the Human Frontier Science Organization. The work was supported in part by a grant from National Institutes of Health.

REFERENCES

Ali, N., and R. Ali. 1997. High salt and solvent induced Z-conformation in native calf thymus DNA. *Biochem. Mol. Biol. Int.* 41:1227–1235.

Arnott, S., and D. W. Hukins. 1972. Optimised parameters for A-DNA and B-DNA. *Biochem. Biophys. Res. Commun.* 47:1504–1509.

Arnott, S., and D. W. Hukins. 1973. Refinement of the structure of B-DNA and implications for the analysis of x-ray diffraction data from fibers of biopolymers. *J. Mol. Biol.* 81:93–105.

Auffinger, P., and E. Westhof. 1997. RNA hydration: three nanoseconds of multiple molecular dynamics simulations of the solvated tRNA(Asp) anticodon hairpin. *J. Mol. Biol.* 269:326–341.

Auffinger, P., and E. Westhof. 1998. Simulations of the molecular dynamics of nucleic acids. *Curr. Opin. Struct. Biol.* 8:227–236.

Beveridge, D. L., and K. J. McConnell. 2000. Nucleic acids: theory and computer simulation, Y2K. *Curr. Opin. Struct. Biol.* 10:182–196.

Brooks, B. R., R. E. Bruccoleri, B. D. Olason, D. J. States, S. Swaminathan, and M. Karplus. 1983. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.* 4:187–217.

Brown, D., and J. H. R. Clarke. 1984. A comparison of constant energy, constant temperature and constant pressure ensembles in molecular dynamics simulations of atomic liquids. *Mol. Phys.* 51: 1243–1252.

Cheatham, T. E., 3rd., P. Cicplak, and P. A. Kollman. 1999. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol Struct. Dyn.* 16:845–862.

Cheatham, T. E., J. L. Miller, T. Fox, T. A. Darden, and P. A. Kollman. 1995. Molecular dynamics simulations highlight the structural differences among DNA:DNA, RNA:RNA, and DNA:RNA hybrid duplexes. *J. Am. Chem. Soc.* 117:4193–4205.

Cheatham, T. E., and P. E. Kollman. 1996. Observation of the A-DNA to B-DNA transition during unrestrained molecular dynamics in aqueous solution. *J. Mol. Biol.* 259:434–444.

Cheatham, T. E., and P. A. Kollman. 1997. Insight into the stabilization of A-DNA by specific ion association: spontaneous B-DNA to A-DNA transitions observed in molecular dynamics simulations of d[ACCCGCGGGT]2 in the presence of hexaamminecobalt(III). *Structure.* 5:1297–1311.

Cheatham, T. E., and P. E. Kollman. 2000. Molecular dynamics simulations of nucleic acids. *Annu. Rev. Phys. Chem.* 51:435–471.

Cheatham, T. E., and P. E. Kollman. 2001. Molecular dynamics simulation of nucleic acids: successes, limitations, and promise. *Biopolymers.* 56:232–256.

Cheatham, T. E., 3rd., and M. A. Young. 2000. Molecular dynamics simulation of nucleic acids: successes, limitations, and promise. *Biopolymers.* 56:232–256.

Chiu, T. K., and R. E. Dickerson. 2000. 1 Å crystal structures of B-DNA reveal sequence-specific binding and groove-specific bending of DNA by magnesium and calcium. *J. Mol. Biol.* 301:915–945.

Cornell, W. D., P. Cieplak, C. I. Baylay, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. 1995. A second generation force field for the simulation of proteins and nucleic acids. *J. Am. Chem. Soc.* 117:5179–5197.

Darden, T., D. York, and L. Pedersen. 1993. Particle mesh Ewald: an N.log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98:10089–10092.

Denisov, V. P., G. Carlström, K. Venu, and B. Halle. 1997. Kinetics of DNA hydration. *J. Mol. Biol.* 268:118–136.

Denisov, V. P., and B. Halle. 2000. Sequence-specific binding of counterions to B-DNA. *Proc. Natl. Acad. Sci. USA.* 97:629–633.

Dickerson, R. E. 1998. DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucleic Acids Res.* 26:1906–1926.

Drew, H. R., and R. E. Dickerson. 1981. Structure of a B-DNA dodecamer. III. Geometry of hydration. *J. Mol. Biol.* 151:535–556.

Egli, M., V. Tereshko, M. Teplova, G. Minasov, A. Joachimiak, R. Sanishvili, C. M. Weeks, R. Miller, M. A. Maier, H. An, P. Dan Cook, and M. Manoharan. 1998. X-ray crystallographic analysis of the hydration of A- and B-form DNA at atomic resolution. *Biopolymers*. 48:234–252.

Eisenstein, M., and Z. Shakked. 1995. Hydration patterns and intermolecular interactions in A-DNA crystal structures. Implications for DNA recognition. *J. Mol. Biol.* 248:662–678.

El Hassan, M. A., and C. R. Calladine. 1996. Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *J. Mol. Biol.* 259:95–103.

Ewald, P. P. 1921. Die Berechnung Optischer und elektrostatischer Gitterpotentiale (Calculation of optic and electrostatic lattice potential). *Ann. Phys.* 64:253–287.

Fang, Y., T. S. Spisz, and J. H. Hoh. 1999. Ethanol-induced structural transitions of DNA on mica. *Nucleic Acids Res.* 27:1943–1949.

Feig, M., and B. M. Pettitt. 1998. Structural equilibrium of DNA represented with different force fields. *Biophys. J.* 75:134–149.

Feig, M., and B. M. Pettitt. 1997. Experiment vs force fields: DNA conformation from molecular dynamics simulations. *J. Phys. Chem.* 101:7361–7363.

Feig, M., and B. M. Pettitt. 1999a. Modeling high-resolution hydration patterns in correlation with DNA sequence and conformation. *J. Mol. Biol.* 286:1075–1095.

Feig, M., and B. M. Pettitt. 1999b. Sodium and chlorine ions as part of the DNA solvation shell. *Biophys. J.* 77:1769–1781.

Field, M. J., and M. Karplus. 1992. CRYSTAL: Program for Crystal Calculations in CHARMM, Harvard University, Cambridge, MA.

Hamelberg, D., L. McFail-Isom, L. D. Williams, and W. D. Wilson. 2000. Flexible structure of DNA: ion dependence of minor-groove structure and dynamics. *J. Am. Chem. Soc.* 122:10513–10520.

Feller, S. E., R. W. Pastor, A. Rojnuckarin, S. Bogusz, and B. R. Brooks. 1996. Effect of electrostatic force truncation on interfacial and transport properties of water. *J. Phys. Chem.* 100:17011–17020.

Foloppe, N., and A. D. MacKerell. 2000. All-atom empirical force field for nucleic acids. I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comp. Chem.* 21:86–104.

Goodsell, D. S., M. Kaczor-Grzeskowiak, and R. E. Dickerson. 1994. The crystal structure of C–C-A-T-T-A-A-T-G-G. Implications for bending of B-DNA at T-A. *J. Mol. Biol.* 239:79–96.

Gorin, A. A., V. B. Zhurkin, and W. K. Olson. 1995. B-DNA twisting correlates with base-pair morphology. *J. Mol. Biol.* 247:34–48.

Gupta, G., M. Bansal, and V. Sasisekharan. 1980. Conformational flexibility of DNA: polymorphism and handedness. *Proc. Natl. Acad. Sci. USA*. 77:6486–6490.

Halle, B., and V. P. Denisov. 1998. Water and monovalent ions in the minor groove of B-DNA oligonucleotides as seen by NMR. *Biopolymers*. 48:210–233.

Hunter, C. A. 1993. Sequence-dependent DNA structure. The role of base stacking interactions. *J. Mol. Biol.* 230:1025–1054.

IUPAC-IUB joint commission on Biochemical Nomenclature. 1983. Abbreviations and symbols for the description of conformations of polynucleotide chains. Recommendations. *Eur. J. Biochem.* 131:9–15.

Jones, S., P. van Heyningen, H. M. Berman, and J. M. Thornton. 1999. Protein-DNA interactions: a structural analysis. *J. Mol. Biol.* 287: 877–896.

Jorgensen, W., J. Chandrasekhar, J. D. Madura, R. Impey, and M. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926–935.

Kielkopf, C. L., S. Ding, P. Kuhn, and D. C. Rees. 2000. Conformational flexibility of B-DNA at 0.74 Å resolution: d(CCAGTACTGG)(2). *J. Mol. Biol.* 296:787–801.

Kopka, M. L., A. V. Fratini, H. R. Drew, and R. E. Dickerson. 1983. Ordered water structure around a B-DNA dodecamer. A quantitative study. *J. Mol. Biol.* 163:129–146.

Kumar, G. S., and M. Maiti. 1994. DNA polymorphism under the influence of low pH and low temperature. *J. Biomol. Struct. Dyn.* 12: 183–201.

Lam, L., and S. C. F. Au-Yeung. 1997. Sequence-specific local structural variations in solution structures of d(CGXX'CG)2 and d(CAXX'TG)2 self-complementary deoxyribonucleic acids. *J. Mol. Biol.* 266:745–760.

Langley, D. R. 1998. Molecular dynamic simulations of environment and sequence dependent DNA conformations: the development of the BMS nucleic acid force field and comparison with experimental results. *J. Biomol. Struct. Dyn.* 16:487–509.

Laskowski, R. A. 1995. SURFNET: A program for visualizing molecular surfaces, cavities and intermolecular interactions. *J. Mol. Graph.* 13:323–330.

Lee, B., and F. M. Richards. 1971. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55:379–400.

Lee, H., T. Darden, and L. Pedersen. 1995. Accurate crystal molecular dynamics simulations using particle-mesh-Ewald: RNA dinucleotides-ApU and GpC. *Chem Phys Lett.* 243:229–235.

MacKerell, A. D., J. Wiorkiewicz-Juczera, and M. Karplus. 1995. An all-atom empirical energy function for the simulation of nucleic acids. *J. Am. Chem. Soc.* 117:11946–11975.

MacKerell, A. D., and N. Banavali. 2000. All-atom empirical force field for nucleic acids. II. Application to molecular dynamics simulations of DNA and RNA in solution. *J. Comp. Chem.* 21:105–120.

MacKerell, A. D. 1997. Influence of magnesium ions on duplex DNA structural, dynamic, and solvation properties. *J. Phys. Chem.* 101:646–650.

Minasov, G., V. Tereshko, and M. Egli. 1999. Atomic-resolution crystal structures of B-DNA reveal specific influences of divalent metal ions on conformation and packing. *J. Mol. Biol.* 291:83–99.

Nelson, H. C. M., J. T. Finch, B. F. Luisi, and A. Klug. 1987. The structure of an oligo(dA).oligo(dT) tract and its biological implications. *Nature*. 330:221–226.

Norberg, J., and L. Nilsson. 1996. Constant pressure molecular dynamics simulations of the dodecamers: d(GCGCGCGCGCGC)2 and r(GCGCGCGCGCGC)2. *J. Chem. Phys.* 104:6052–6057.

Petersen, H. J. 1995. Accuracy and efficiency of the particle mesh Ewald method. *J. Chem. Phys.* 103:3668–3679.

Quintana, J. R., K. Grzeskowiake, K. Yanagi, and R. E. Dickerson. 1992. Structure of a B-DNA decamer with a central T-A step: C-G-A-T-T-A-A-T-C-G. *J. Mol. Biol.* 225:379–395.

Reinert, K. E. 1999. DNA multimode interaction with berenil and pentamidine: double helix stiffening, unbending and bending. *J. Biomol. Struct. Dyn.* 17:311–331.

Ryckaert, J. P., G. Ciccotti, and H. J. C. Berendsen. 1977. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. 1977. *J. Comput. Phys.* 23:327–341.

Schneider, B., K. Patel, and H. M. Berman. 1998. Hydration of the phosphate group in double-helical DNA. *Biophys. J.* 75:2422–2434.

Shui, X., L. McFail-Isom, G. H. Hu, and L. D. Williams. 1998. The B-DNA dodecamer at high resolution reveals a spine of water on sodium. *Biochemistry*. 37:8341–8355.

Sissoeff, I., J. Grisvard, and E. Guille. 1976. Studies on metal ions-DNA interactions: specific behavior of reiterative DNA sequences. *Prog. Biophys. Mol. Biol.* 31:165–199.

Tereshko, V., G. Minasov, and M. Egli. 1999. A ''Hydrat-Ion'' spine in a B-DNA minor groove. *J. Am. Chem. Soc.* 121:3590–3595.

Tippin, D. B., and M. Sundaralingam. 1997. Comparison of major groove hydration in isomorphous A-DNA octamers and dependence on base sequence and local helix geometry. *Biochemistry*. 36:536–543.

Verlet, L. 1967. Computer ''experiments'' on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.* 159:98–103.

Winger, R. H., R. K. Liedl, S. Rudisser, A. Pichler, A. Hallbrucker, and E. Mayer. 1998. B-DNA's BI. . .BII conformer substate dynamics is coupled with water migration. *J. Phys. Chem.* 102:8934–8940.

Yang, L., and B. M. Pettitt. 1996. B to A transition of DNA on the nanosecond time scale. *J. Phys. Chem.* 100:2564–2566.

York, D. M., W. Yang, H. Lee, T. Darden, and L. G. Pedersen. 1995. Toward the accurate modeling of DNA: the importance of long-range electrostatics. *J. Am. Chem. Soc.* 117:5001–5002.

Young, M. A., G. Ravishanker, and D. L. Beveridge. 1997. A 5-nanosecond molecular dynamics trajectory for B-DNA: analysis of structure, motions, and solvation. *Biophys. J.* 73:2313–2336.

Two-Metal-Ion Mechanism for Hammerhead-Ribozyme Catalysis

Fabrice Leclerc*,†,‡,§ and Martin Karplus*,‡,§

Laboratoire de Maturation des ARN et Enzymologie Moléculaire, Université Henri Poincaré, Faculté des Sciences, B.P. 239, Bd. des Aiguillettes, 54506 Vandoeuvre-lès-Nancy, France, Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138, and Laboratoire de Chimie Biophysique, Université Louis Pasteur, Institut le Bel, 67000 Strasbourg, France

Received: July 12, 2005; In Final Form: November 2, 2005

The hammerhead ribozyme is one of the best studied ribozymes, but it still presents challenges for our understanding of RNA catalysis. It catalyzes a transesterification reaction that converts a 5′,3′ diester to a 2′,3′ cyclic phosphate diester via an S_N2 mechanism. Thus, the overall reaction corresponds to that catalyzed by bovine pancreatic ribonuclease. However, an essential distinguishing aspect is that metal ions are not involved in RNase catalysis but appear to be important in ribozymes. Although various techniques have been used to assign specific functions to metals in the hammerhead ribozyme, their number and roles in catalysis is not clear. Two recent theoretical studies on RNA catalysis examined the reaction mechanism of a single-metal-ion model. A two-metal-ion model, which is supported by experiment and based on ab initio and density functional theory calculations, is described here. The proposed mechanism of the reaction has four chemical steps with three intermediates and four transition states along the reaction pathway. Reaction profiles are calculated in the gas phase and in solution. The early steps of the reaction are found to be fast (with low activation barriers), and the last step, corresponding to the departure of the leaving group, is rate limiting. This two-metal-ion model differs from the models proposed previously in that the two metal ions function not only as Lewis acids but also as general acids/bases. Comparison with experiment shows good agreement with thermodynamic and kinetic data. A detailed analysis based on natural bond orbitals (NBOs) and natural energy decomposition (NEDA) provides insights into the role of metal ions and other factors important for catalysis.

1. Introduction

The discovery of catalytic RNA molecules (ribozymes) in the early 1980s,[1,2] at a time when proteins were thought to be the only enzymes, raised the fundamental question of how RNA enzymes work. Although ribozymes have been under intense study for the intervening years, no mechanism that provides a detailed description of the reaction is universally accepted for any ribozyme. Among the various known RNA enzymes, the best-characterized is the hammerhead ribozyme. It was the first ribozyme to be crystallized, and a series of X-ray structures corresponding to a biologically active ribozyme have been determined.[3−6] This ribozyme has also been the subject of numerous biochemical studies, yet questions remain regarding the reaction mechanism.[7−9] Like the RNA-catalyzed self-cleavage of other ribozymes,[10] the reaction catalyzed by the hammerhead ribozyme involves a transesterification step in the phosphate ester hydrolysis.[11] This step leads to isomerization from a 5′,3′ diester to a 2′,3′ cyclic phosphate diester. In a second step, the 2′,3′ cyclic phosphate is hydrolyzed to yield a 3′ phosphate and regenerate the 2′ OH group. The transesterification reaction has been shown to proceed via an S_N2(P) or "in-line" mechanism in which the attacking nucleophile (the 2′ oxygen) is aligned with the phosphorus atom and the 5′ oxygen atom of the phosphate group from the neighboring 3′ nucleotide.[12−14] Thus, the overall mechanism corresponds to that found in bovine pancreatic RNase,[15] although metal ions, which appear to play an essential role for the ribozymes, are not present in RNase. Models proposed for the reaction mechanism differ particularly with respect to the number of metal ions involved (single-metal-ion mechanisms[16−20] or two-metal-ion mechanisms[21,22]) and their specific role in the catalysis; that is, whether they act as a general acid/base, an electrophilic catalyst, or a Lewis acid[19] (Figure 1). When the metal is involved in the deprotonation of a nucleophile or in the protonation of a leaving group, it can function either as a Lewis acid or as a generalized acid/base. The metal acts as a Lewis acid if it stabilizes an anionic nucleophile or leaving group (by direct coordination of the metal to the oxygens of the phosphate group) but does not participate directly in the proton transfer, while it acts as a general acid or base when it is directly involved as a proton donor (from hydrated metal) or acceptor (by metal hydroxide). The metal functions as an electrophilic catalyst when it activates the electrophile (the phosphorus atom) by making it more susceptible to nucleophilic attack (by direct coordination of the metal to the nonbridging pro-R or pro-S oxygens). The single-metal-ion mechanisms are mostly based on a general acid/base model of catalysis (Figure 1B), while the two-metal-ion mechanisms are mostly based on a Lewis acid model of catalysis (Figure 1C and D). The experimental data, originally supporting a single-metal-ion mechanism (Figure 1A),[23−25] were shown subsequently to be more consistent with a two-metal-ion mechanism (Figure 1B and C).[26] Since then, additional experimental evidence has accumulated in favor of a two-metal-ion

* To whom correspondence should be addressed. E-mail: fabrice.leclerc@maem.uhp-nancy.fr (F.L.); marci@tammy.harvard.edu (M.K.).
† Université Henri Poincaré.
‡ Harvard University.
§ Université Louis Pasteur.

10.1021/jp053835a CCC: $33.50
Published on Web 02/01/2006

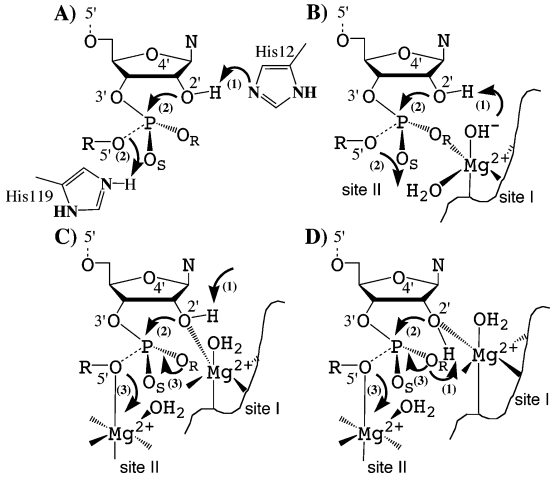


Figure 1. General acid/base catalysis vs metal-ion catalysis in the transesterification step of RNA hydrolysis. (A) General acid/base mechanism in RNase A. In this model, His12 acts as a general base to activate the 2′ oxygen as the nucleophile by abstraction of the proton from the 2′ OH (1), while His119 acts as a general acid to facilitate the departure of the leaving group by protonation of the 5′ oxygen (2). (B) General acid/base mechanism in the hammerhead ribozyme. In this single-metal-hydroxide-ion model,[16−20] the metal hydroxide (site I) activates the 2′ oxygen as the nucleophile (1). The activated 2′ oxygen attacks the phosphorus and induces the departure of the 5′ oxygen leaving group (2). The hydrated metal (site II) can be regenerated as a cofactor by giving away a proton from a coordinated water molecule to the 5′ oxygen (2). (C) Metal-ion catalysis in the hammerhead ribozyme. In this two-metal-ion model based on the dianionic mechanism,[24−28] an external Brönsted base (water molecule) activates the 2′ oxygen (1). The attack of the 2′ oxygen on the phosphorus (2) is followed by the departure of the leaving group (3). The metals at sites I and II act as Lewis acids by accepting the electrons from the 2′ and 5′ oxygens, respectively. (D) Alternative metal-ion catalysis in the hammerhead ribozyme. In this two-metal-ion model based on the triester-like mechanism (monoanionic mechanism),[25,37,52] the activation of the 2′ oxygen is accomplished by one of the nonbridging oxygens of the phosphate group that accepts the proton from the 2′ OH (1). The rest of the mechanism is similar to the mechanism in part B. The mechanisms in both parts B and C are shown as being sequential, though they could be concerted, in part.

mechanism.[27−30] The two-metal-ion mechanism is strongly supported by the differential metal-ion effects on the cleavage rate observed for the natural substrate (cleavage activation observed in the presence of Mg^{2+} and La^{3+})[28,30] and for a 5′ thio modified substrate.[27] The metal ions usually involved in chemical catalysis are divalent cations such as Mg^{2+} or Mn^{2+}. Recent studies have shown that monovalents cations, such as Na^+, Li^+, and NH_4^+, can also act as metal cofactors at extremely high concentrations (400-fold higher).[31−34] In fact, the hammerhead-catalyzed reaction is significantly more efficient in the presence of divalent ions that act as catalytic cofactors under physiological conditions. Even under more favorable but artificial conditions (hammerhead RNA with a small helix I domain and an extremely high concentration of monovalent ions), the rate enhancement in 4 M Li^+ is 10-fold less than that in 10 mM Mg^{2+}.[32,34] Studies where both divalent ions and monovalent ions are combined suggest that the divalent-metal-ion-catalyzed reaction may involve a Mg^{2+} ion with weak binding affinity in vivo;[35,36] this could explain why the difference in rate enhancement between the hammerhead cleavage reactions stimulated by monovalent ions and by divalent ions is not more significant in solution. In summary, since high concentrations of monovalent ions are inhibitory for the hammerhead ribozyme in cells,[35] all evidence points to the fact that the hammerhead cleavage reaction in vivo should follow a divalent-metal-ion-dependent channel,[35,36] as assumed here. Ribozymes can function by both base-catalyzed and acid-catalyzed mechanisms, depending on the pH of the solution. Both reaction mechanisms produce a 2′,3′ cyclic phosphate and a 5′ oxygen leaving group. The base-catalyzed reaction, which we consider here, is believed to involve a dianionic phosphorane species; the 2′ OH group is activated by an external base, which can be assisted by the direct coordination of a metal ion to the 2′ oxygen[29] to form a 2′ O^- oxyanion that attacks the adjacent phosphorus to generate the dianionic phosphorane.[25]

In the monoanionic or "triester-like" mechanism (based on a two-metal-ion model), the 2′ proton is transferred to the nonbridging phosphoryl oxygen (pro-Rp) to render the substrate triester-like. This latter mechanism was proposed to explain thio effects (loss of catalytic activity when substituting one of the nonbridging oxygens by sulfur) and rescue effects (restoration of catalytic activity by thiophilic metal ions) in hammerhead-catalyzed reactions.[37] SpS and RpS isomers of hammerhead-ribozyme substrates are much less reactive than the natural unmodified substrates (thio effect). The fact that the thio effect is much larger for the RpS isomer than the SpS isomer suggested the pro-Rp oxygen plays a more critical role. This would mean that the pro-Rp oxygen is the proton acceptor during the 2′ OH deprotonation. However, reinvestigation of the thio effect and rescue have shown that the data can be explained by the coordination of a divalent metal ion to the pro-Rp oxygen at the cleavage site.[38] The dianionic mechanism is also more consistent with other experimental data. Indeed, a pH-dependent conformational change of the hammerhead ribozyme associated with the chemical reaction suggests that a 2′ O^- oxyanion is formed by deprotonation of the 2′ OH group.[39] The deprotonation, taking place at basic pH (at or above pH 8.5), would drive the conformational change that initiates the reaction. Quantum mechanical studies of a small RNA model compound (a phosphorylated ribose with a 5′ O-methoxy group as the leaving group) and different phosphorothioate analogues also suggest there is no significant preference for the triester-like mechanism over the dianionic mechanism (Lopez et al., Leclerc et al., to be published). The calculated activation free energies in solution, with the unmodified analogue in the absence of metal ions, are 34.2 (Supporting Information Figure S1) and 22.9 kcal/mol (Supporting Information Figure S2) for the two mechanisms, respectively.

Although qualitative descriptions of the single- and two-metal-ion mechanisms have been available for some time,[23,24] a quantitative theoretical study of the reaction pathway for a single-ion mechanism was published only recently.[20] In the present paper, we propose a quantitative two-metal-ion model for the reaction mechanism of the transesterification step in the hammerhead catalysis. The study was guided by the insightful discussions of von Hippel and co-workers[26,28] of a two-metal-ion mechanism based on experimental data and a proposal of Steitz and Steitz[21] for a variety of protein and RNA enzymes that cleave phosphodiester bonds. The model was built from the small RNA model mentioned above by adding two solvated magnesium ions. Ab initio and density functional theory (DFT) methods were used to calculate the structures of the stationary points identified along the reaction path and the energetics (relative energy and free energy profiles) of the corresponding chemical processes using realistic quantum chemical models (including electron correlation effects and a large basis set) both in the gas phase and in solution. Reaction path calculations were performed for each transition state to ensure that it is connected to the corresponding starting and ending structures. The

TABLE 1: Geometries of Stationery Points on the Reaction Path for Transesterification of Methyl Ribose Phosphate[a]

	guess[b] 0	reactant I	TS1 II‡	intermediate 1 III	TS2 IV‡	intermediate 2 V	TS3 VI‡	intermediate 3 VII	TS4 VIII‡	product IX
P−O2′	2.845	2.805	2.817	2.766	2.167	1.916	1.829	1.783	1.708	1.607
H−O2′		1.035	1.131	2.487						
H(O2′)−O⁻(Mg$_I$)		1.538	1.321	0.971						
O2′−O⁻(Mg$_I$)		2.359	2.285	2.766						
P−O3′	1.555	1.559	1.556	1.560	1.589	1.614	1.609	1.614	1.599	1.604
P−O5′	1.730	1.606	1.607	1.620	1.663	1.692	1.697	1.709	2.027	2.955
P−O$_R$	1.533	1.548	1.549	1.549	1.566	1.581	1.592	1.590	1.554	1.531
P−O$_S$	1.476	1.478	1.479	1.478	1.489	1.499	1.519	1.530	1.497	1.475
Mg$_I$−O2′	1.946	2.075	2.031	1.832	1.886	1.939	1.954	1.992	2.096	2.763
Mg$_I$−O$_R$	2.050	1.968	1.963	2.019	1.999	1.978	1.948	2.013	1.999	2.027
Mg$_{II}$−O5′	2.146	2.310	2.317	2.215	2.215	2.105	2.067	1.998	2.204	2.101
Mg$_{II}$−O$_R$	2.057	1.972	1.965	1.994	1.980	1.978	1.948	2.000	1.999	2.027
Mg$_I$−Mg$_{II}$	3.910	3.700	3.702	3.866	3.873	3.857	3.768	3.850	3.768	3.685
Mg$_I$−OH$_2$[c]	2.273	4.782	4.762	4.688	4.524	4.450	3.649	2.079	2.038	2.578
O4′−OH[d]	1.763	1.578	1.567	1.545	1.583	1.617	1.660	1.631	1.605	1.616
O2′−P−O5′	140.1	161.6	161.3	157.2	162.1	162.9	160.7	165.6	163.7	157.5
O2′−H−O⁻(Mg$_I$)		131.8	137.4	96.23						

[a] Geometries optimized at the RHF/3-21+G* level. Distances are given in angstroms, and angles, in degrees. Numbers in italics correspond to significant variations in distance due to the change in protonation state of the 5′ oxygen associated with the departure of the leaving group. [b] Starting structure used as initial guess in the geometry optimizations (Figure 2). [c] Distance measured between the magnesium at the first metal site and the oxygen of the water molecule associated with the switch from a penta- to hexacoordinated magnesium. [d] Distance measured between the O4′ oxygen and the hydrogen of the water molecule associated with the switch from a penta- to hexacoordinated magnesium.

complete reaction path from the reactant, a 3′ phosphorylated ribose, to the product, a 2′,3′ cyclic phosphorylated ribose and a 5′ OH methyl leaving group, has been obtained. The results show the coordination of the metal ions along the transesterification pathway and demonstrate their role in catalysis, as well as the importance of solvent effects on the free energy profile of the reaction.

2. Model and Methods

The model used in this study is a phosphorylated ribose complexed with two hydrated magnesium cations. The model is based, in part, on the structures of two conformational intermediates of the hammerhead ribozyme obtained by crystallographic freeze-trapping: the "early" intermediate shows two Mg^{2+} ions in proximity to the cleavage site (one coordinated to the pro-Rp nonbridging phosphate oxygen and the second distant from the first one by 4.4 Å), in addition to three other Mg^{2+} ions that are further from the cleavage site;[3] the "late" intermediate shows a conformation compatible with an in-line attack mechanism.[6] The conformational changes associated with these two intermediates, relative to the "ground state" structure, are restricted to the catalytic pocket of the ribozyme. One of the metals is coordinated to the nonbridging pro-R oxygen of the phosphate group with a Mg^{2+}−O distance of 2.43 Å. In the proposed two-metal-ion model (see below), the stationary points exhibit geometries corresponding to an in-line mechanism involving a metal-to-metal distance between 3.69 and 3.87 Å and a metal-to-pro-R-oxygen distance between 1.95 and 2.0 Å (Table 1). Preliminary calculations on a model compound corresponding to a phosphorylated ribose allowed us to identify the stationary points and a unique transition state connecting the reactant and product of the transesterification step, in the base-catalyzed phosphate ester hydrolysis (Leclerc et al., to be published separately). In the absence of metal ions, the reaction follows a dianionic mechanism where the nucleophilic attack of the 2′ oxygen on the phosphorus is concerted with the departure of the 5′ oxygen leaving group. The geometry of the RNA part was taken from this transition state; this assumes that the metals stabilize the in-line conformation of the phosphorylated ribose, which corresponds to the active conformation in the catalytic pocket of the hammerhead ribozyme. Two solvated metal ions, hexa- or pentacoordinated, were added via inner-sphere coordinations with the 2′ oxygen and the 5′ oxygen, in accord with the proposed metal coordinations for the two-metal-ion model.[25,26,29] With these structural constraints, various penta- and hexacoordinated forms of hydrated magnesium ions[40] were constructed (a total of 10 geometries). After full optimization, some "guessed" geometries deviated significantly from their initial geometry while others preserved the in-line conformation ($\angle$(O2′−P−O5′) $\geq$ 140 °). Among those, several guessed geometries converged to the same or some equivalent conformation, which differs only by the number of solvating water molecules (for details, see the Supporting Information). In the case of equivalent conformations with the same metal inner-sphere coordinations, the more solvated ones were preferred. Finally, only conformations that exhibit conserved inner-sphere coordinations with the pro-Rp oxygen, as proposed in the two-metal-ion model,[25,26,29] were retained. Two nonequivalent in-line conformations with eight and nine water molecules in the metal coordination shells were selected after optimization. One of these, the conformation with eight water molecules (four water molecules in the coordination shell of each metal), inner-sphere coordination with the pro-Rp oxygen and 2′ oxygen at the first metal site and inner-sphere coordination with the pro-Rp oxygen and 5′ oxygen at the second metal site (Figure 2), was selected as the starting structure for the subsequent calculations. The model with nine water molecules attached to the metal ions was excluded because it has only a single inner-sphere coordination (the other is an outer-sphere coordination) with the pro-Rp oxygen, which is less consistent with experimental evidence.[28,41] Moreover, the distance between the two metal ions (more than 5.6 Å) in this model fit the X-ray data less well (distance between metal at sites 1 and 6 of 4.4 Å) than that with eight water molecules (3.9 Å, Table 1).

The geometry of the model structure, optimized at the HF/3-21+G* level, was used as the starting point for the reaction path calculations. The O2′−P and P−O5′ distances were used as the reaction coordinates; they correspond to the nucleophilic attack on the phosphorus and the departure of the leaving group, respectively. Nine geometries were generated with a distance range for the O2′−P and P−O5′ internal coordinates that include O2′−P bond formation (distance between 3.235 and 1.742 Å)

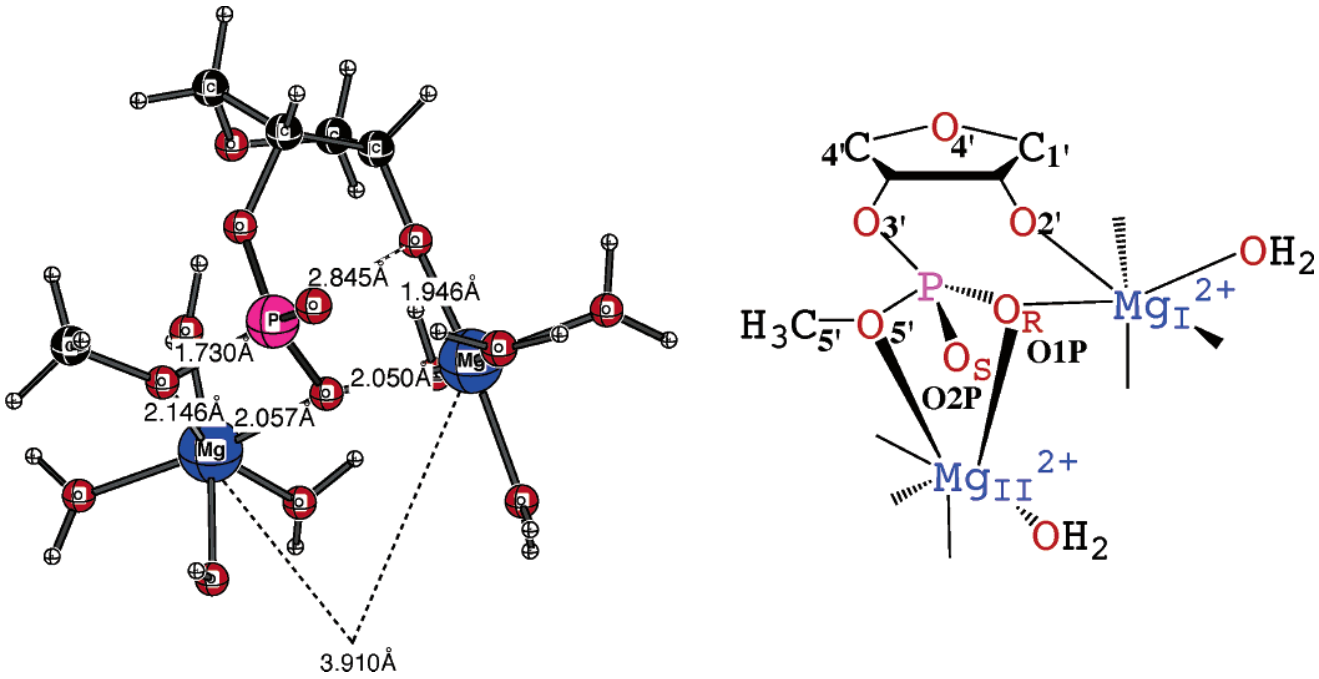


Figure 2. Model structure used to locate stationary points. The structure of the two-metal-ion model corresponding to the geometry optimized at the HF/3-21+G* level and its schematic representation with the stereochemistry of the nonbridging oxygens of the phosphate group are shown on the left and right sides, respectively. The atoms are colored according to the following code: carbon, black; hydrogen, white; oxygen, red; phosphorus, magenta; magnesium, blue.

and P−O5′ bond breaking (distance between 1.726 and 4.167 Å); the end points were chosen by following the reaction path (intrinsic reaction coordinate calculations) for the reaction of a phosphorylated ribose in the absence of metal ions (i.e., in the small RNA model mentioned above). Geometry optimizations and frequency calculations with O2′−P and P−O5′ frozen internal coordinates on those 10 geometries were then done to locate possible transition states corresponding to the nucleophilic attack, the departure of the leaving group, or both in the case of a concerted mechanism. Four out of the ten constrained geometries optimized at the HF/3-21+G* level had an imaginary frequency. After full relaxation (optimization and frequency calculations at the HF/3-21+G* level), only two geometries corresponded to transitions states: one to the nucleophilic attack and the other one to the departure of the leaving group. Reaction path following was performed from these transition states to determine the stationary points corresponding to possible intermediates. The stationary points corresponding to the 2′ OH activation were inferred from the intermediate (dianionic species) that just precedes the nucleophilic attack. In this way, five local minima corresponding to the reactant (R), product (P), and different intermediates along the reaction pathway (I1, I2, and I3) were obtained from the three saddle points (transition states) corresponding to the 2′ OH activation (TS1), the nucleophilic attack (TS2), and the departure of the leaving group (Table 1). The presence of more than (n + 1) local minima with respect to the number of saddle points indicated that a saddle point was missing along the reaction pathway. The synchronous transit-guided quasi-Newton (STQN) method was then used to locate this missing transition state, which connects the second and third intermediates (I2 and I3). The full reaction path involves nine stationary points: four transition states and three intermediates plus the reactant and the product. The free energies in the gas phase and solution of the stationary points relative to the reactant were calculated at the B3LYP/6-31+G**//HF/3-21+G* level (for details, see the Supporting Information).

All geometry optimizations were performed using Gaussian 98 (Gaussian, Inc., Pittsburgh, PA, 2001, revision A.10). The frequencies were scaled by an empirical factor of 0.9207 to correct, at the HF/3-21+G* level, for errors in the potential energy surface.[42] The vibrational contributions to the entropy and to the enthalpy, zero-point energy, and vibrational energy at 298 K were calculated from the frequencies. The other contributions (rotational and translational entropies and the work term (PV)) were calculated according to standard classical statistical mechanics (e.g., an ideal gas PV term was added to obtain the Gibbs free energy). Effective energies in solution were calculated for the geometries optimized in the gas phase using the solvation model (Poisson−Boltzmann solver) implemented in Jaguar (Jaguar, Schrödinger, Inc., 2002, version 4.2).[43] The eight explicit water molecules in the solvation shells of the two metals are treated as part of the solute in the solvation calculations. The use of a continuum model for the treatment of the solvation effects is based on the assumption that water molecules from the solvent do not modify the metal coordinations we have described. The assumption is supported by the results obtained with an explicit solvent model for each stationary point of the reaction path (Zdenek and Leclerc, data not shown). The latter used a combined quantum mechanical and molecular mechanical (QM/MM) method implemented in the CHARMM program interfaced with the ab initio quantum mechanical GAMESS program. Optimization of the geometry of each stationary point solvated in a 15 Å^3 waterbox resulted in geometries and coordinations very close to those obtained at the QM level in this paper (the maximum deviation is 0.07 Å for the bond length and 0.2 Å for the metal coordination).

3. Results

3.1. The Reaction Mechanism. The reaction is found to involve four chemical steps (Figure 3): (1) nucleophile activation (Figure 3A) to form a dianionic species (for comparison with the theoretical study published recently,[20] we have

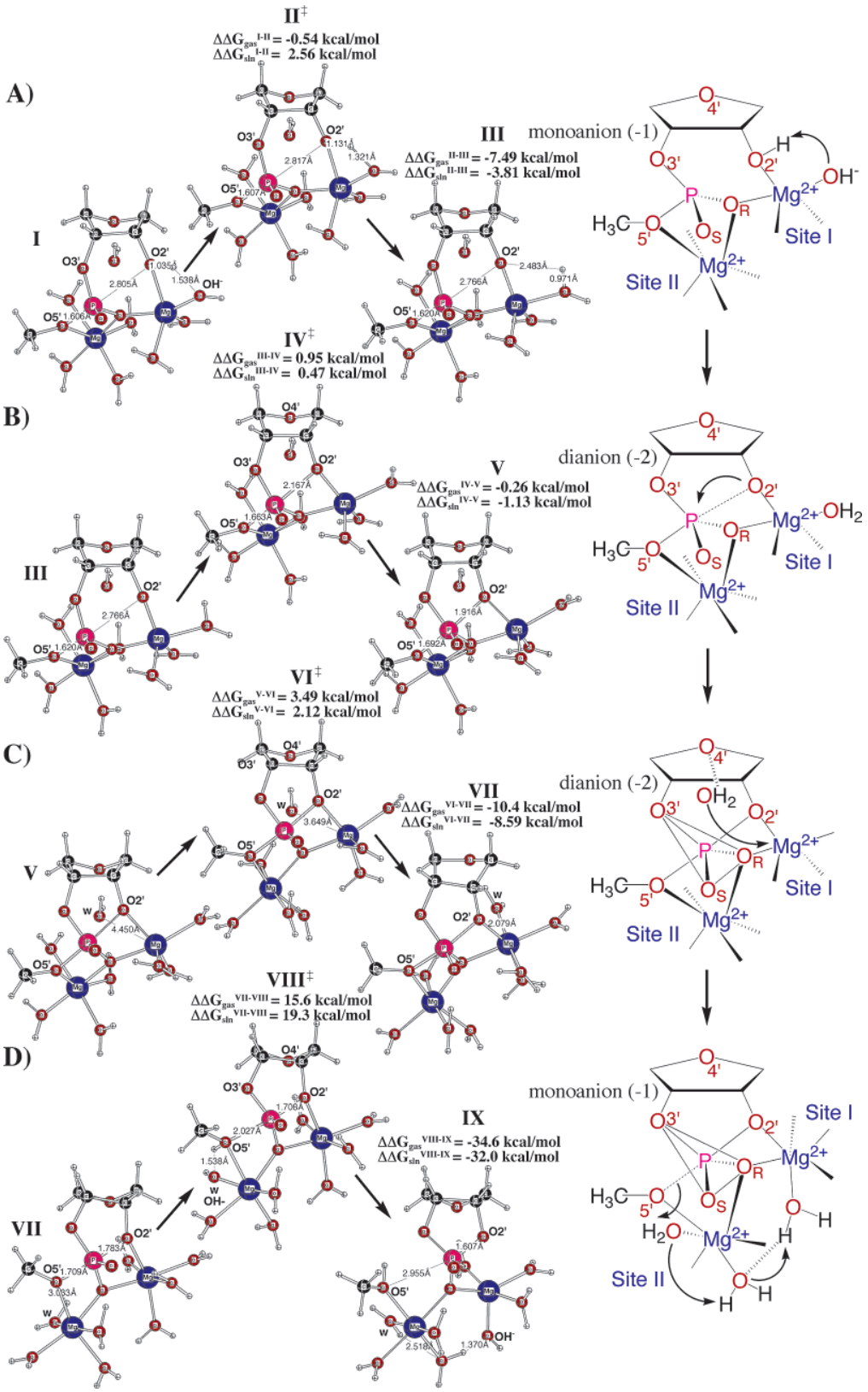


Figure 3. Proposed mechanism for hammerhead-catalyzed reactions. (A) First reaction step: nucleophile activation of the 2′ OH into a 2′ oxyanion. (B) Second reaction step: nucleophilic attack of the 2′ oxygen on the phosphorus. (C) Third reaction step: coordination change at the first metal site (site I). (D) Fourth reaction step: departure of the 5′ oxygen leaving group. The structures of the stationary points located along the reaction pathway are shown with indication of the forward and reverse energy barriers between each of them. The activation energies (backward and forward) are given at the B3LYP-6-31+G**//HF/3-21+G* level and at the HF/3-21+G* level in parentheses. Transition states are labeled by a double dagger. A set of distances relevant to each chemical step are indicated. The structures are represented with the atom color code used in the previous figure. A schematic representation (right side) shows the major features of each step: arrows indicate the bonding changes and dotted lines the bond formation or bond breaking.

TABLE 2: Relative Energies and Free Energies (at 298 K) for Stationery Points on the Transesterification Path of Methyl Ribose Phosphate with Respect to the Starting Molecule[a]

molecule[b]	ΔE	ΔZPE[c]	ΔE_{TRV}	$T\Delta S_{TRV}$	ΔG_{gas}[d]	ΔH[e]	ΔG_{sol}	ΔG_{sln}
I	0	0	0	0	0	0	0	0
II‡	0.544	−1.652	−2.084	−0.566	−0.542	−1.375	3.100	2.558
III	−6.773	−0.281	0.218	0.981	−8.035	−6.574	6.780	−1.255
IV‡	−6.667	−0.554	−0.487	0.084	−7.087	−7.116	6.308	−0.780
V	−6.773	−0.205	0.169	0.585	−7.346	−6.618	5.440	−1.906
VI‡	−3.521	−0.718	−0.703	−0.162	−3.861	−4.168	4.070	0.209
VII	−12.191	−0.879	−0.339	1.378	−14.230	−12.503	5.845	−8.385
VIII‡	1.790	−1.214	−1.385	−0.615	1.408	−1.377	9.520	10.928
IX	−27.484	−2.697	−1.661	3.231	−33.195	−29.013	12.100	−21.095

[a] B3LYP/6-31+G**//HF/3-21+G* values in kilocalories per mole. The HF/3-21+G* frequencies were scaled by an empirical factor of 0.9135 to correct for errors in the potential energy surface. The solvation energies were calculated from the geometry optimized in the gas phase using the SCRF model implemented in Jaguar (25): the van der Waals radii of the 2′ oxygen and nonbridging oxygens of the phosphate group were fitted to reproduce the solvation energies of H_3PO_4, $H_2PO_4^-$, and HPO_4^{2-} (Leclerc et al., unpublished). [b] States marked with a double dagger symbol are transition states, and the other states are minima on the potential energy surface. [c] Zero-point-energy contribution. [d] $\Delta G_{gas} = \Delta E + \Delta ZPE + \Delta E_{TRV} + \Delta(PV) - T\Delta S_{TRV}$, where ΔE_{TRV} and ΔS_{TRV} include the translational, rotational, and vibrational contributions. ΔG_{sol} is the solvation free energy, and ΔG_{sln} is the total free energy difference in solution. The entropy (S_{vib}), zero-point energy (ZPE), and vibrational energy (E_{vib}) were calculated from the frequencies and geometries according to standard statistical mechanical formulas (26). The rotational (E_{rot}) and translational (E_{trans}) energies and the work term (PV) were treated classically; an ideal gas PV term was added to obtain the Gibbs free energy of the reaction. [e] The enthalpies of reaction were calculated as the sum of the energy difference with respect to the reactant and the thermal correction to enthalpy: $\Delta H = \Delta E + \Delta(PV)$. The calculated values for I are $E = -1981.758\,766$ hartree, ZPE = 0.389 62 hartree, $\Delta(PV) = 0.420\,251$ hartree, $E_{TRV} = 263.712$ kcal/mol, $\Delta S_{TRV} = 189.433$ cal/mol, and $\Delta G_{sol} = -1.15$ kcal/mol.

considered that the 2′ OH activation proceeds via a solvated metal hydroxide as proposed for a single-ion mechanism, but we do not exclude other possible modes of activation; see Figure 1), (2) nucleophilic attack on the phosphorus (Figure 3B), (3) a coordination change at the first metal site from a pentacoordinated magnesium to a hexacoordinated magnesium (Figure 3C), and (4) the departure of the leaving group (Figure 3D). There are nine stationary points (numbered from I to IX) of which four are transition states (II‡, IV‡, VI‡, and VIII‡) and three are chemical intermediates (III, V, and VII), in addition to the reactant (I) and product (IX). The critical geometric parameters of the stationary points are listed in Table 1, and the calculated energies and free energies (see the Supporting Information) are given in Table 2. In the reactant, the first metal ion is pentacoordinated and not hexacoordinated, as it was in the starting structure used for the optimization; one water molecule has moved during the optimization to become hydrogen bonded to the O4′ sugar oxygen so that it no longer belongs to the first coordination shell of the metal. Interestingly, this change in coordination leads to better stabilization of the in-line conformation for an O2′−P distance more than 2.0 Å (see the Supporting Information for details); that is, the switch from a hexacoordinated metal to a pentacoordinated metal at site I is associated with an increase of the O2′−P−O5′ angle from 140 to 162° (compare stationary points 0 and I in Table 1). The second metal site is unchanged from its initial conformation and corresponds to a hexacoordinated magnesium. Overall, the resulting reactant structure has an in-line conformation stabilized by the positions and coordinations of the two metals (Figure 3A). The pentacoordinated magnesium ion at site I includes two inner-sphere ligands from the ribozyme (the oxygen O2′ and the nonbridging oxygen pro-R of the phosphate group); the other three ligands are two water molecules and one hydroxide ion. In this trigonal bipyramidal arrangement, the 2′ oxygen and the two coordinated water molecules (Figure 3A) occupy the equatorial positions; the hydroxide ion and the pro-R oxygen occupy the apical positions of the bipyramid (Figure 3A). The proton from the 2′ OH is in the plane of the bipyramid (defined by the pro-R and O2′ oxygens, the metal, and the hydroxide ions) and easily abstracted by the hydroxide ion; the distance between the O2′ and hydroxide-ion oxygens (less than 2.6 Å) corresponds to a relatively short hydrogen bond (present in the reactant and transition state, see Table 1). Thus, the deprotonation of the 2′ OH occurs without any major change in the geometry of the reactant: only the O2′−P distance is slightly shortened from 2.80 (Figure 3A, I) to 2.77 Å (Figure 3A, III). The hexacoordinated Mg^{2+} at site II is approximately octahedral with six ligands (the pro-R and O5′ oxygens from the RNA and four water molecules).

The nucleophilic attack also does not involve any major geometrical change. Only the O2′−P distance changes significantly (from 2.77 Å in III to 2.17 Å in IV), and the activation energy is very low; it equals 0.11 kcal/mol at the B3LYP/6-31+G**//HF/3-21+G* level (Figure 3B). The activation energy for the reverse reaction between the second intermediate (V) and the corresponding transition state (IV‡) is the same (0.11 kcal/mol); the second intermediate (V) has a further shortening of the O2′−P distance to 1.92 Å. The transition state geometry is essentially a trigonal bipyramid, which is maintained until product formation.

The third step corresponds to a structural change between the nucleophilic attack and the departure of the leaving group. Its essential feature is the change of the metal coordination at the first metal site from pentacoordinated to hexacoordinated, which is required for the reaction to proceed (Figure 3C); that is, the intermediate with two hexacoordinated metals (VII) corresponds to a symmetric state where the 2′ and 5′ bridging oxygens (after nucleophile activation) interact equally as strongly with both metals (the O2′−P bond starts to form with the O2′−P distance changing from 1.916 Å in V to 1.783 Å in VII, while the P−O5′ bond starts breaking with the P−O5′ distance changing from 1.692 Å in V to 1.709 Å in VII). This change occurs via the migration of the water molecule hydrogen bonded to the O4′ oxygen into the first coordination shell of the metal at site I, so that it again is hexacoordinated (Table 1). The activation barrier for this step is 3.3 kcal/mol. It leads to the formation of a third intermediate (VII), which is more stable (by 4.9 kcal/mol) than the second intermediate (V) and has a reverse activation energy of 8.7 kcal/mol.

The fourth (final) step involves the departure of the leaving group (5′ oxygen methyl), which is protonated to form CH_3OH by proton transfer from one of the water molecules coordinated to the second metal; this water becomes an OH^- ligand. As in the first step of the reaction that also involves a proton transfer,

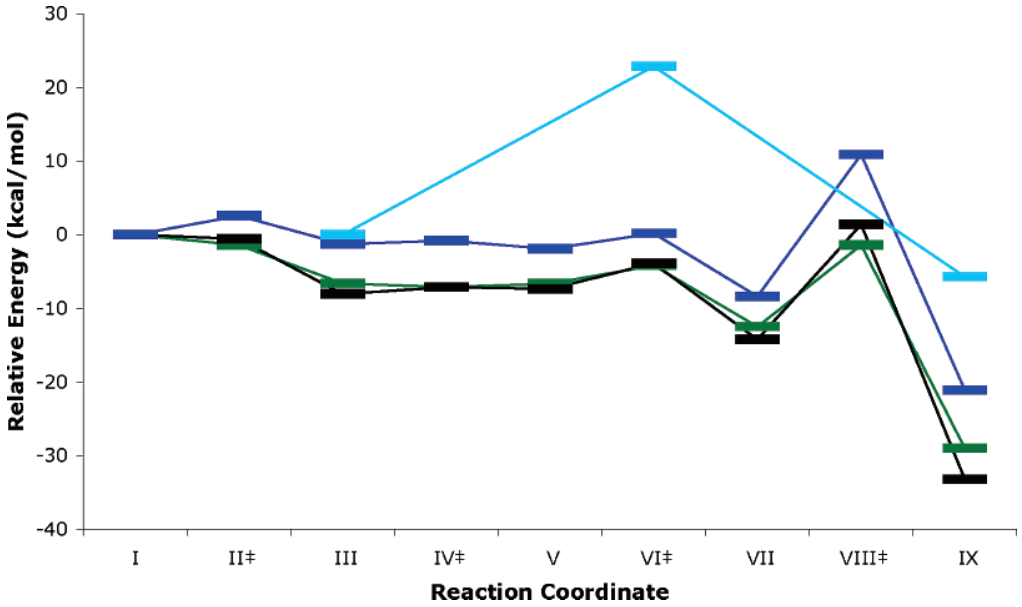


Figure 4. Reaction profiles of the proposed reaction mechanism. The energy profiles correspond to the total free energy difference (black line), the total free energy difference in solution (dark blue line), and the enthalpy difference (green line). For comparison, the energy profile corresponding to the total free energy difference in solution is also shown for a metal-free system (light blue line).

a short hydrogen bond is formed between the proton donor (one of the coordinated water molecules to the metal at site II) and the proton acceptor (the 5′ oxygen). However, the hydrogen bond is only formed in the transition state after the proton transfer has effectively occurred between the leaving group (CH_3O/Mg_{II}) and the water molecule (H_2O/Mg_{II}). The protonated leaving group (CH_3OH/Mg_{II}) is then dissociated from the 2′,3′ cyclic phosphate ribose but remains coordinated with the Mg^{2+} ion at site II (Figure 3D). This step has the highest forward activation energy (14.0 kcal/mol).

From the above description and the values in Table 2, the rate-limiting step for the overall reaction is the transition from the third intermediate (VII) to the product (IX) via the transition state (VIII‡). The calculated free energy barrier is $\Delta G^{\ddagger} = 19.3$ kcal/mol, which is close to the experimental value $\Delta G^{\ddagger} = 20.1$ kcal/mol. The measured value is for the overall reaction, but it has been suggested that the cleavage of the P−O5′ bond corresponds to the rate-limiting step for nonenzymatic and hammerhead-ribozyme-catalyzed reactions,[27,44] as found here. The calculated values for the activation energies ($\Delta H^{\ddagger} = 13.9$ kcal/mol, $E_a = 14.0$ kcal/mol) are in reasonable agreement with the experimental values ($\Delta H^{\ddagger} = 17.7$ kcal/mol, $E_a = 18.3$ kcal/mol). Since the activation free energies of the first three steps are all quite low (Table 2), the reaction could appear to be essentially concerted.

3.2. Energetics and Solvent Effects. The relative energies of all of the stationary points, with respect to the reactant as a reference, are listed in Table 2; the relative free energies both in the gas phase and in solution are also included. The calculations at the Hartree−Fock and Becke3LYP levels give similar energy profiles for the activation barriers; the rate-limiting step corresponds to the departure of the leaving group (see above and Figure 3D). In the calculations at the higher level of theory (B3LYP/6-31+G**//HF/3-21+G*), the energy barriers for the other steps are very small, the value is near zero for the nucleophilic attack ($\Delta\Delta G_{sln}^{III\rightarrow IV^{\ddagger}} = 0.47$ kcal/mol). The enthalpy makes the larger relative contribution to the free energy change in most of the reaction steps (Table 2 and Figure 4). The entropy represents between 19 and 45% of the free energy difference except for the stationary points V and VI‡, which correspond to the migration of the water molecule in the second reaction step (Figure 3C), for which the entropic component is more than 60% (97 and 64%, respectively). The free energy barriers of the reaction are larger than the activation barriers due to the entropic penalty (except for the formation of VII and IX). In solution, the free energy barriers are slightly lowered except for the departure of the leaving group where the barrier increases from 15.6 to 19.3 kcal/mol (Figure 4). The solvation free energy difference is large in the first step between the reactant and the first intermediate (6.78 kcal/mol between I and III, see Table 2) because the deprotonation of the 2′ OH required for the nucleophile activation involves a switch from a monoanion (−1) to a dianion (−2) of the RNA moiety (from I to III, Figure 3A), although the overall charge of the system does not change. By contrast, the differential solvent effect is small in the second and third steps of the reaction (the variation of solvation free energy is 2.7 kcal/mol or less between II‡ and VII‡) because the redistribution of charge is much smaller for the various stationary points with a trigonal bipyramidal geometry. In the last step, there is a second significant solvent effect that is related to two different events: the separation of charge between the partially dissociated products (the ribose 2′,3′ cyclic phosphate and the CH_3OH) and the delocalization of the hydroxide-ion charge formed after neutralization of the leaving group (CH_3OH). The redistribution of charge associated with the product formation is attenuated by the neutralization of the leaving group (from CH_3O^- to CH_3OH) via the proton transfer from a water molecule coordinated to the metal at site II (Figure 3D); the solvation free energy associated with this event is 3.7 kcal/mol (between VII and VIII‡, see Table 2). The delocalization of the hydroxide-ion charge onto three water molecules involves secondary proton transfers (two water molecules coordinated to the metal at site II and one water molecule coordinated to the metal at site I, see Figure 3D). The solvation free energy associated with this event is 2.6 kcal/mol. Nevertheless, the magnitude of the solvent effects is much smaller than that found, for example, in the nonenzymatic ionic ester hydrolysis. There, the solvation free energy change between the reactant and products is estimated to be 49.4 kcal/mol (Supporting Information Figure S2) versus 12.1 kcal/mol

TABLE 3: Interaction Energies between Donor and Acceptor Orbitals from the NBO Analysis[a]

vicinal interactions	reactant I	TS1 II‡	intermediate 1 III	TS2 IV‡	intermediate 2 V	TS3 VI‡	intermediate 3 VII	TS4 VIII‡	product IX
$n_{O^-(Mg(I))} \rightarrow \sigma^*_{O2'-H}$ [b]	47.4	105							
$n^*_{Mg(I)} \rightarrow \sigma_{O2'-H}$	4.80	6.40							
$n^*_{Mg(I)} \rightarrow \sigma_{O2'-H}$	1.96	4.60							
$n_{O2'} \rightarrow \sigma^*_{P-O5'}$	1.0	0.98	4.6	5.4	4.4	5.2	5.7		
$n_{O5'} \rightarrow \sigma^*_{P-O2'}$					5.2	5.3	5.9	32.4	0.48
$\sigma_{P-O2'} \rightarrow \sigma^*_{P-O5'}$					26.4	28.0	30.3		
$\sigma_{P-O5'} \rightarrow \sigma^*_{P-O2'}$					32.6	30.6	31.7		
$\sigma^*_{P-O2'} \rightarrow \sigma^*_{P-O5'}$					164	231	502		
$n_{O_{2'}} \rightarrow n^*_{Mg(I)}$	19.2	21.3	42.3	37.5	28.3	25.7	26.4	24.1	2.88
$n_{O_R} \rightarrow n^*_{Mg(I)}$	25.3	25.5	20.6	20.5	24.2	26.1	24.2	28.1	25.3
$E_{orbital} = \sum\{n_{O_{RNA}} \rightarrow n^*_{Mg(I)}\}$ [c]	44.5	46.8	62.9	58.0	52.5	51.8	50.6	52.2	28.2
$E_{orbital} = \sum\{n_{Oi} \rightarrow n^*_{Mg(I)}\}$ [d]	138	136	142	138	133	131	155	160	138
$n_{O_{5'}} \rightarrow n^*_{Mg(II)}$ [e]	13.5	13.3	16.4	23.4	22.5	25.8	26.7	13.0	24.8
$n_{O_R} \rightarrow n^*_{Mg(II)}$	27.6	27.8	26.7	26.6	28.7	31.2	28.3	27.9	23.2
$E_{orbital} = \sum\{n_{O_{RNA}} \rightarrow n^*_{Mg(II)}\}$	41.1	41.1	43.1	50.0	51.2	57.0	55.0	40.9	48.0
$E_{orbital} = \sum\{n_{Oi} \rightarrow n^*_{Mg(II)}\}$	145	145	148	154	155	156	154	154	153

[a] B3LYP/6-31+G(d,p)//HF/3-21+G* values in kilocalories per mole. [b] Mg(I) refers to metal at site I. [c] The energy value corresponds to the sum of the second-order perturbative estimates of the stabilization energies between donor−acceptor pairs where $n_{O_{RNA}}$ refers to the donor p lone pairs of the oxygen atoms belonging to the RNA moiety (pro-R and 2′ oxygen atoms at site I and pro-R and 5′ oxygen atoms at site II). [d] Same as footnote *c* except that n_{Oi} refers to the donor p lone pairs of the oxygen atoms from all ligands (RNA and water molecules). [e] Mg(II) refers to metal at site II.

for the present system. The total free energy change including the solvation contribution is then −21 kcal/mol (see Table 2). However, it should be noted that the geometries of the reactant and product are likely not fully relaxed in the catalytic pocket of the hammerhead ribozyme, as they are in this model system. Intramolecular interactions can make the reactant more stable or more likely in this case the product less stable and thus reduce the relative energy difference between the reactant and product. Such relative stabilization to equalize the reactant and product free energy has been found in many enzymes (e.g., triose phosphate isomerase) and could explain the observed reversibility of the reaction.[45]

3.3. Natural Orbital Analysis. One approach to analyze the electronic changes taking place during the reaction is to look at the behavior of natural bond orbitals (NBOs) which are commonly used to describe hybridization and covalency effects in molecules. They are localized on a small number of atoms and describe the Lewis-like molecular bonding pattern of electron pairs. The NBO analyses at the Becke3LYP/6-31+G**//HF-3-21+G* level show that the natural bond orbitals are different for the various steps of the reaction. We focus on three natural bond orbitals between (1) the 2′ oxygen and its proton before the O2′ H activation in the first step of the reaction ($\sigma_{O2'-H}$, I and II), (2) the phosphorus and the leaving group in the first three steps of the reaction ($\sigma_{P-O5'}$, I to VII), and (3) the phosphorus and the nucleophile in the last step of the reaction ($\sigma_{P-O2'}$, VII to IX). The $\sigma_{O2'-H}$ natural bond orbital is mostly localized on the 2′ oxygen with 83−86% (I to II) of the orbital consisting of basis functions on oxygen. The inner-sphere coordination to the hydroxide−metal complex strongly weakens the O2′−H bond ($n_{O^-(Mg(I))} \rightarrow \sigma^*_{O2'-H}$ from 47.4 kcal/mol for I to 105 kcal/mol for II, Table 3) and slightly increases the O2′−H polarization (in the triester-like mechanism, the contribution from the 2′ oxygen to $\sigma_{O2'-H}$ is less than 80% in the reactant that also involves a short hydrogen bond between the O2′ and the pro-Rp oxygen, Leclerc et al., to be published). The $\sigma_{P-O5'}$ natural bond orbital is mostly localized on the 5′ oxygen with 85−87% (I to VII) of the orbital consisting of basis functions on oxygen. In $\sigma_{P-O2'}$, the contributions from the phosphorus and oxygen to the $\sigma_{P-O2'}$ in V are 9 and 91%, respectively, and they end up at 15 and 85%, respectively, in the product (IX). The trigonal bipyramidal intermediates (V and VII) as well as the transition state that connects them (VI‡) both exhibit the natural bond orbitals $\sigma_{P-O2'}$ and $\sigma_{P-O5'}$. During the migration of the water molecule (associated with the coordination change of the metal at site I), the delocalization of the lone pair on the nucleophile into the P−O5′ antibonding orbital, which was present in the two first steps (nucleophile activation and nucleophilic attack), remains with an increasing interaction energy and contributes to weaken the P−O5′ bond ($n_{O2'} \rightarrow \sigma^*_{P-O5'}$ from 4.4 kcal/mol for V to 5.7 kcal/mol for VII). The three stationary points involved in this process (V to VII) are the only ones where the two natural bonds $\sigma_{P-O5'}$ and $\sigma_{P-O2'}$ are present at the same time; they also exhibit the presence of two specific non-Lewis NBOs ($\sigma_{P-O2'} \rightarrow \sigma^*_{P-O5'}$ and $\sigma^*_{P-O2'} \rightarrow \sigma^*_{P-O5'}$, see Table 3). The increasing σ-interactions into the P−O5′ antibonding orbital associated with the coordination change of the metal at site I contribute to weakening of the P−O5′ bond involved in the departure of the leaving group (P−O5′ bond going from 1.692 Å for V to 1.709 Å for VII, Table 1).

3.4. Reaction Path and Role of Metal Ions. The reaction path was followed by performing intrinsic reaction coordinate (IRC) calculations,[46] showing that the transition states connect the corresponding reactant, product, and intermediates (Supporting Information Figure S3). In the presence of metal ions, the reaction is sequential according to the model developed here, although, as pointed out above, the kinetics might appear concerted due to the small activation free energies for several of the early steps. There is no direct spectroscopic evidence for the existence of chemical intermediates, but the hammerhead-ribozyme kinetics have been interpreted as suggesting that an intermediate exists;[47] see also earlier ab initio molecular orbital calculations on phosphates and phosphoranes.[25] Interestingly, the model developed for the metal-free reaction suggests a concerted mechanism where the nucleophilic attack and the departure of the leaving group are coupled (Leclerc et al., to be published separately). A comparison of the energy profiles in solution between the metal-assisted reaction and the metal-free reaction shows that metals contribute significantly to lowering

TABLE 4: Natural Energy Decomposition Analyses (NEDA) of the RNA−Metal Complexes[a]

	total binding energy					
molecule	ΔE_{tot}[b]	ES	POL	CT	EX	DEF
I	−311	−382	−242	−256	−43.9	614
II‡	−324	−405	−253	−315	−47.5	696
III	−644	−667	−332	−230	−43.3	628
IV‡	−644	−674	−327	−232	−43.5	633
V	−646	−681	−322	−233	−43.6	634
VI‡	−652	−689	−340	−234	−45.3	656
VII	−646	−683	−337	−257	−46.5	676
VIII‡	−307	−375	−236	−259	−42.6	606
IX	−379	−436	−269	−235	−44.9	607

[a] RHF/6-31+G**//RHF/3-21+G* values in kilocalories per mole. The calculations are based on the decomposition into three molecular fragments corresponding to the RNA model and the two hydrated metal ions treated with their solvation shell. [b] The binding energy for bringing together the three molecular fragments is given by ΔE_{tot} = ES + POL + CT + EX + DEF, sum of the electrostatic (ES), polarization (POL), charge transfer (CT), exchange (EX), and deformation (DEF) contributions calculated by the NBO 5.0 program.[47] The metal ion at site I is pentacoordinated from I to IV and hexcoordinated from V to VII. The metal ion at site II is hexacoordinated and includes the water molecule in the second solvation shell (I to IV) which is then transferred to the metal at site I (V to VII).

of the energy barriers; the overall free energy barrier is reduced by 12.0 kcal/mol (Figure 4).

Natural Energy Decomposition. To obtain a more detailed understanding of the influence of the metal ions on the different steps of the reaction, natural energy decomposition analyses (NEDA)[48] were performed for the stationary points. On the basis of this analysis, the binding energy between the RNA model and the two hydrated metals and its variation along the reaction coordinate were determined (Table 4). The binding energy was decomposed into two-body interactions between pairs of fragments—RNA−Mg_I, RNA−Mg_{II}, and Mg_I−Mg_{II} (Table 5)—and a non-pairwise-additive three-body interaction (Supporting Information Table S1), which are further decomposed into different energetic contributions. The method and details of the results of the NEDA are described in the Supporting Information.

Although the binding energy corresponding to the process of bringing together the RNA and the two hydrated metals is always favorable (RNA−Mg_I and RNA−Mg_{II} are favorable, while Mg_I−Mg_{II} is unfavorable), it varies significantly in the two steps of the reaction corresponding to changes of charge on the RNA moiety, that is, during the nucleophile activation (first step: I to III) and during the departure of the leaving group (fourth step), which involve a switch from a monoanionic RNA species to a dianionic RNA species and vice versa (Figure 5). In the first step of the reaction (I to III, Figure 3A), there is a large increase (of more than 100%) in binding energy (>300 kcal/mol) associated with the monoanion-to-dianion conversion due to the 2′ OH deprotonation which occurs late along the reaction pathway (Supporting Information Figure S3A). As expected, the RNA−Mg_I interaction that involves a direct coordination between the 2′ oxygen and Mg_I represents the major contribution to the increased binding energy. However, the RNA−Mg_{II} interaction contributes about one-third of the increase (Table 5). In the reactant (I), the RNA−Mg_{II} interaction represents the major contribution to the total binding energy: 43% versus 32% for the RNA−Mg_I interaction. After formation of the first intermediate (III), the relative contributions of the RNA−Mg_{II} and RNA−Mg_I interactions are reversed (34% versus 43%) and the RNA−Mg_I interaction becomes predominant in the total binding energy (Figure 5A).

From the first intermediate to the product, the RNA−Mg_I interaction decreases while the RNA−Mg_{II} interaction increases. The two-body interaction term between the hydrated metals (Mg_I−Mg_{II}) remains essentially constant; it is destabilizing due to the metal−metal electrostatic repulsion (Table 5). The changes in the two RNA−metal terms (RNA−Mg_I and RNA−Mg_{II}) are compensatory and partly associated with the migration of the water molecule from the second solvation shell of Mg_{II} to the first solvation shell of Mg_I. The RNA−metal complex is destabilized by the three-body term which arises mainly from the polarization term that represents up to 74% of the unfavorable contribution (Supporting Information Table S1). The two dominant contributions are the electrostatic (ES) and deformation (DEF) terms; each term contributes about 35%. The ES term is favorable while the DEF term is unfavorable to the association of the RNA with the metals. The DEF term is predominantly associated with the Pauli repulsion that prevents the charge distribution of the RNA fragment from penetrating that of the hydrated metal fragments and vice versa. The polarization (POL) and charge transfer (CT) terms are the two other favorable terms that contribute significantly to the binding energy: POL contributes from 16 to 18% and CT from 12 to 17%. The exchange (EX) term represents only 2−3% of the binding energy. The decreasing ES and POL components for RNA−Mg_I and the increasing ES and POL components for RNA−Mg_{II} explain the compensation in the binding between these two pairs of fragments (Table 5). The stabilization energy comes in part from the delocalization of the lone pairs on the 2′ and 5′ oxygens into the metal orbitals (nonbonding lone pairs)

TABLE 5: Natural Energy Decomposition Analyses (NEDA) of the Two-Body Interactions[a]

	two-body term (RNA + Mg)[b]						two-body term (Mg + Mg)					
molecule	ΔE(I/II)	ES(I/II)	POL(I/II)	CT(I/II)	EX(I/II)	DEF(I/II)	ΔE	ES	POL	CT	EX	DEF
I	−202/−267	−247/−284	−119/−121	−128/−101	−23.5/−18.2	315/258	152	149	−20.8	−30.1	−3.92	57.8
II‡	−213/−267	−269/−286	−127/−122	−185/−103	−27.2/−18.5	395/262	153	150	−20.5	−28.6	−3.74	55.1
III	−520/−404	−510/−418	−216/−135	−108/−116	−22.9/−19.2	336/285	257	261	−19.4	−18.9	−2.13	37.0
IV‡	−504/−423	−500/−438	−196/−153	−109/−117	−21.6/−20.9	322/305	259	263	−19.6	−19.3	−2.12	37.4
V	−496/−436	−495/−450	−184/−163	−113/−116	−21.0/−22.0	316/316	261	265	−19.6	−18.4	−2.03	36.0
VI‡	−498/−441	−497/−461	−183/−173	−110/−123	−20.8/−24.0	313/340	263	269	−20.6	−16.4	−1.79	32.8
VII	−466/−469	−478/−476	−180/−171	−139/−121	−24.8/−21.3	355/320	266	271	−18.8	−13.6	−1.77	29.5
VIII‡	−287/−164	−296/−212	−145/−95.0	−110/−138	−20.4/−21.2	284/303	131	133	−16.1	−18.1	−2.10	34.5
IX	−137/−280	−159/−279	−69.7/−128	−77.3/−93.0	−14.0/−16.5	183/235	25.4	2.00	−84.3	−52.0	−14.3	174

[a] RHF/6-31+RHF/6-31+G*//RHF/3-21+G* values in kilocalories per mole. The binding energy calculated as previously (Table 3) is decomposed into two-body interactions between pairs of fragments (RNA + Mg_I, RNA + Mg_{II}, and Mg_I + Mg_{II}) and a non-pairwise-additive three-body term (Table 5, Supporting Information). [b] For comparison, the binding energies of the two (RNA + Mg) pairs, corresponding to the metal sites I and II, respectively, are given together separated by a slash.

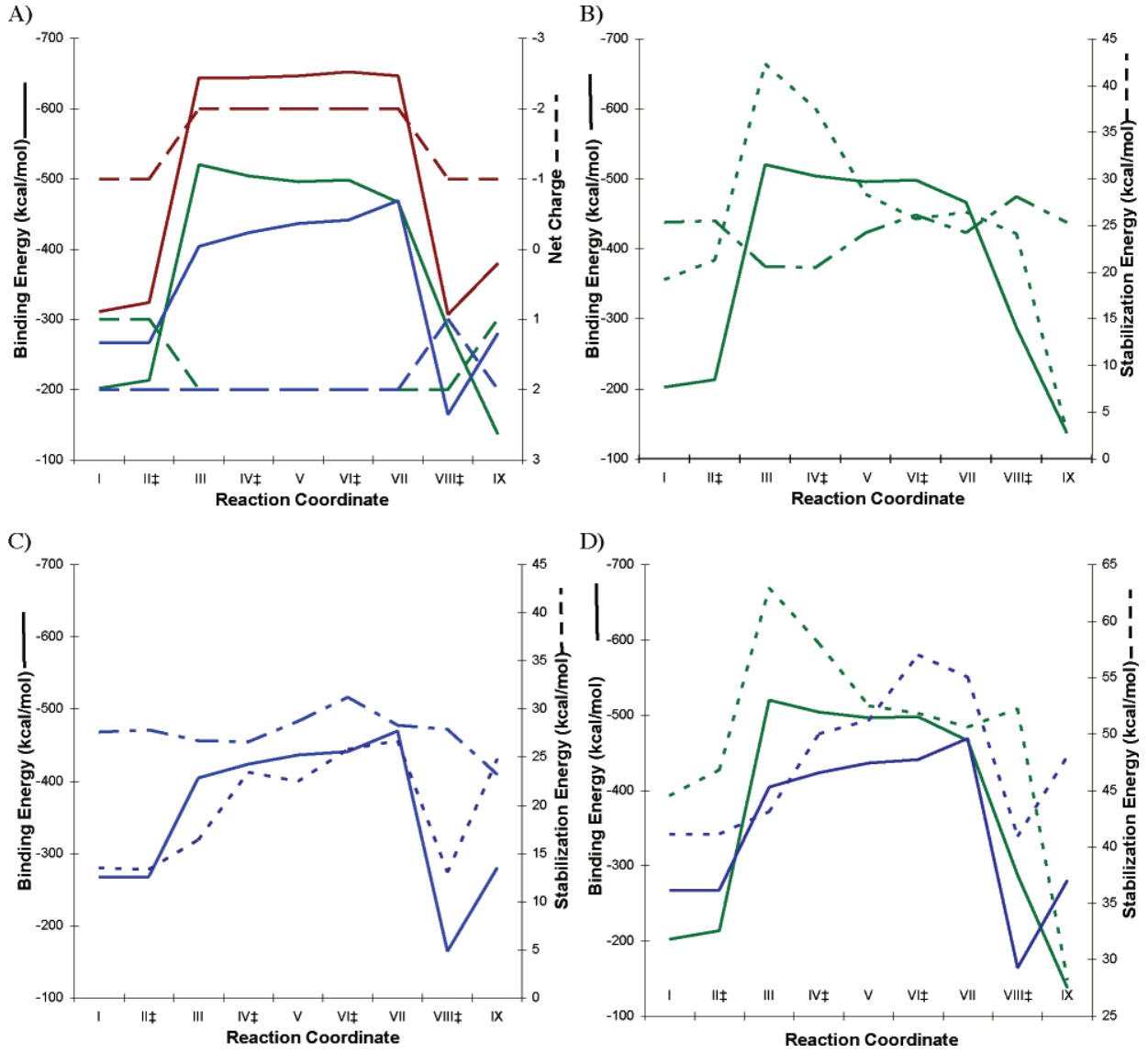


Figure 5. Binding energy along the reaction pathway of the RNA−metal complex and comparison with the stabilization energy from the delocalization of the lone pairs on the bridging oxygens into the metal orbitals. (A) The binding energy (left-hand vertical axis), calculated at the HF/6-31+G**//HF/3-21+G* level by NEDA, corresponds to the energy to bring together three fragments: the RNA moiety and each of the two hydrated metal ions (metal and solvation shells). The total binding energy between the three fragments (solid red line) and the relative contribution (two-body interactions) from the metal at site I (RNA−Mg_I, solid green line) and the metal at site II (RNA−Mg_{II}, solid blue line) interacting with the RNA moiety are shown. The dashed lines (right-hand vertical axis) indicate the net charge on the RNA moiety (dashed red line), on the hydrated metal at site I (dashed green line), and on the hydrated metal at site II (dashed blue line). (B) The contribution from the RNA−Mg_I interaction (solid green line) is compared with the stabilization energy (indicated as positive on the right-hand axis) from the delocalization of the lone pairs on the pro-R and 2′ oxygen into the metal I orbitals ($n_{O_R} \rightarrow n^*_{Mg_I}$, dashed−dotted line; $n_{O2'} \rightarrow n^*_{Mg_I}$, dotted green line). (C) The contribution from RNA−Mg_{II} interaction is compared with the stabilization energy from the delocalization of the lone pairs on the pro-R and 5′ oxygen into the metal II orbitals ($n_{O_R} \rightarrow n^*_{Mg_{II}}$, dashed−dotted blue line; $n_{O5'} \rightarrow n^*_{Mg_{II}}$, dotted blue line). (D) The contributions from the RNA−Mg_I and RNA−Mg_{II} interactions (solid green and blue lines) are compared with the total stabilization energy from the delocalization of both the lone pairs on the pro-R and 2′ oxygen into the metal I orbitals (dotted green line) or from the delocalization of both the lone pairs on the pro-R and 5′ oxygen into the metal II orbitals (dotted blue line). In parts B−D, the orbital delocalization contributions correspond to the right-hand vertical axis.

at Mg_I and Mg_{II} ($n_{O2'} \rightarrow n^*_{Mg_I}$ and $n_{O5'} \rightarrow n^*_{Mg_{II}}$) that represents covalency contributions to the RNA−metal interactions (Figure 5B and C).

Natural Orbital Analysis. As mentioned before, the coordination of the 2′ oxygen to Mg_I contributes to weaken the O2′−H bond in the first step of the reaction. The destabilization due to the delocalization of the antibonding orbitals of metal-ion lone pairs of Mg_I into the antibonding σ orbital of the O2′−H bond increases ($n^*_{Mg(I)} \rightarrow \sigma^*_{O2'-H}$ from 1.96 kcal/mol for I to 4.60 kcal/mol for II), while the corresponding stabilization due to the delocalization of the antibonding orbitals of metal-ion lone pairs of Mg_I into the bonding σ orbital of the O2′−H bond increases but to a lesser extent ($n^*_{Mg(I)} \rightarrow \sigma_{O2'-H}$ from 4.80 kcal/mol for I to 6.40 kcal/mol for II, Table 3). The net destabilization effect of the metal Mg_I on the O2′−H bond is thus 1.8 kcal/mol.

From the first intermediate to the product, the stabilization energy due to the delocalization of the lone pair on the 2′ oxygen, corresponding to the coordination to Mg_I, decreases ($n_{O2'} \rightarrow n^*_{Mg_I}$ from 42 to 2.9 kcal/mol). Simultaneously, the stabilization energy due to the delocalization of the lone pair on the 5′ oxygen corresponding to the coordination to Mg_{II}

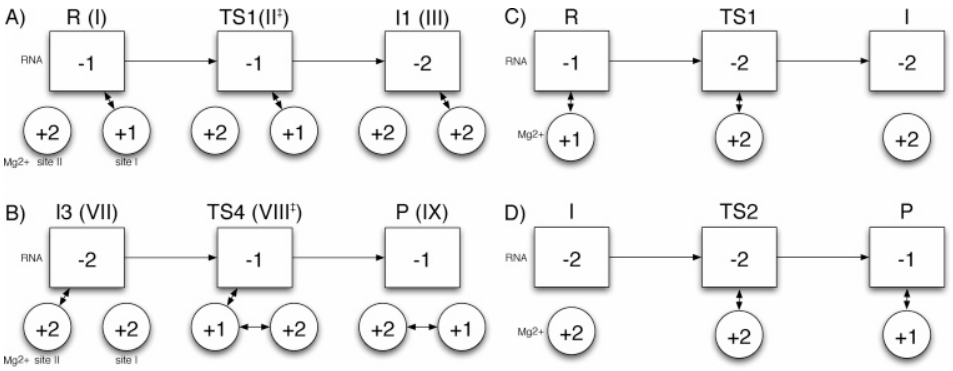


Figure 6. Changes of charge distribution between the RNA and hydrated metal moieties along the reaction pathway in the steps involving the monoanion/dianion interconversions for the current two-metal-ion model (A and B) and the single-metal-ion model (C and D, Torres et al., 2003). (A) Monoanion/dianion conversion in the first step of the reaction. The three corresponding stationary points I, II‡, and III are represented in a schematic way. (B) Dianion/monoanion conversion in the last and rate-determining step of the reaction. The three corresponding stationary points VII, VIII‡, and IX are represented similarly. (C) Monoanion/dianion conversion in the first step of the reaction. (D) Dianion/monoanion conversion in the last and rate-determining step of the reaction. The rectangle represents the RNA moiety, and the spheres represent the hydrated metal moieties (sites I and II are labeled as in Figure 3, and the fragment moieties are defined as in the NEDA calculations). The single arrows indicate the forward direction of the reaction. The double arrows indicate the moieties involved in the changes of charge distribution.

increases somewhat ($n_{O5'} \rightarrow n^*_{Mg_{II}}$ from 16 kcal/mol for III to 25 kcal/mol for IX). As the O2′−P bond forms, the delocalization of the lone pair on the 2′ oxygen into the antibonding orbitals of metal-ion lone pairs of Mg_I decreases (Figure 5B). Inversely, as the P−O5′ bond breaks, the delocalization of the lone pair on the 5′ oxygen into the antibonding orbitals of metal-ion lone pairs of Mg_{II} increases (Figure 5C). We can define a covalency contribution to RNA−metal binding as the sum of the stabilization energies associated with the delocalization of the lone pairs on all of the RNA oxygens into the antibonding orbitals of metal-ion lone pairs of Mg_I and Mg_{II}. It includes the contribution from the 2′ and pro-R oxygens interacting with Mg_I ($n_{O2'} \rightarrow n^*_{Mg_I}$ and $n_{O_R} \rightarrow n^*_{Mg_I}$) and that from the 5′ and pro-R oxygens interacting with Mg_{II} ($n_{O5'} \rightarrow n^*_{Mg_{II}}$ and $n_{O_R} \rightarrow n^*_{Mg_{II}}$). Thus, there is a correlation between the RNA−metal binding energy (calculated with the NEDA) and the covalency contribution to RNA−metal binding; the major covalency contributions come from the delocalization of the lone pair on the 2′ and 5′ oxygens into the antibonding orbitals of metal-ion lone pairs of Mg_I and Mg_{II}, respectively (Figure 5B−D). The contributions from the delocalization of the lone pair on the pro-R oxygen into the antibonding orbitals of both metals have small variations along the reaction path (Figure 5B and C). The total stabilization energy for σ-interactions that transfer charge from occupied lone-pair natural bond orbitals on the oxygen atoms, belonging to both RNA and water molecules, into empty non-Lewis orbitals on the metal ions ($E_{orbital} = \sum\{n_{Oi} \rightarrow n^*_{Mg(I)}\}$ and $E_{orbital} = \sum\{n_{Oi} \rightarrow n^*_{Mg(II)}\}$, Table 3) varies to a lesser extent than the stabilization energy corresponding to the covalency contribution to RNA−metal binding. Indeed, the decreasing stabilization energy from III to IX for the σ-interactions in Mg_I−O2′ and the increasing stabilization energy from I to VII for the σ-interactions in Mg_{II}−O5′ are compensated by increasing and decreasing the stabilization energies, respectively, arising from water ligands (Table 3). The total stabilization energy for the σ-interactions in metal−oxygen bonds is greater in the case of the hexacoordinated metal Mg_{II} (I to VI), but the difference with the pentacoordinated metal Mg_I can be small (142 kcal/mol versus 148 kcal/mol in III, Table 3) because of the stronger covalency contribution to RNA−Mg_I binding in comparison with RNA−Mg_{II} binding (62.9 kcal/mol versus 43.1 kcal/mol in III, Table 3). Nevertheless, the noncovalency contributions to RNA−metal binding, especially the electrostatic contribution, are more important than the covalency contributions.

During the last step of the reaction, the total binding energy drops by more than 40% in going from the 3rd intermediate (VII) to the product (IX) when the 5′ oxygen is protonated and the proton of a water molecule coordinated to Mg_{II} is transferred to the leaving group (Table 4). This large change in binding energy is again associated with a modification of the net charge on the RNA moiety (a dianion-to-monoanion conversion in this case) that arises from the partial separation of the leaving group from the ribose 2′,3′ cyclic phosphate (Figure 5A). The destabilization of the RNA−metal complex is mainly electrostatic (the ES and POL contributions decrease), and it is only partly compensated by the decrease in the unfavorable two-body term Mg_I−Mg_{II}, which is reduced by 90% (Table 4) and by the three-body term which also becomes less destabilizing (Supporting Information Table S1). This destabilization is even more pronounced in the last transition state (VIII‡) than in the product (IX) and is associated with a double change in the charge distribution from the third intermediate (VII) to the fourth transition state (VIII‡) to the product (IX). The first change in the charge distribution is due to the first proton transfer from a coordinated water molecule to the 5′ oxygen which occurs early in the reaction pathway, while the second one is due to the second proton transfer between two water molecules each coordinated to one of the metal ions (Supporting Information Figure S3D).

Monoanion/Dianion Interconversions. The two steps of the reaction that involve a change in the net charge on the RNA moiety, that is, the nucleophile activation of the 2′ oxygen (first step) and the departure of the leaving group (last step), correspond to a monoanion/dianion conversion (first step) and a reverse dianion/monoanion conversion (last step); that is, the net charge on the RNA moiety switches from −1 (I) to −2 (III) in the first step and back from −2 (VII) to −1 (IX) in the last steps (Figure 6). As the calculations have shown, the free energy barrier is low in the first step ($\Delta\Delta G_{sln}^{I \rightarrow II‡} = 2.56$ kcal/mol) while it is high in the last step ($\Delta\Delta G_{sln}^{VII \rightarrow VIII‡} = 19.3$ kcal/mol). Both reactions lead to a lowering of the energy ($\Delta\Delta E$) and free energy ($\Delta\Delta G_{sln}$) of the system, as expected. Whether there is a low activation energy (as in the first reaction) or a high activation energy (as in the last reaction) depends on

whether the stabilizing interactions appear as the system goes from the reactant to the transition state or from the transition state to the product, respectively.

The two steps both involve proton transfers between the RNA moiety and a hydrated metal: in the first step, the proton transfer is the only chemical process and it occurs late along the reaction pathway (i.e., after the transition state), while the proton transfer occurs early in the last step (i.e., before the transition state) and is concurrent with the P−O5′ bond breaking. The chemical process which corresponds to the higher energy barrier during the two first steps of the reaction is the proton transfer ($\Delta\Delta G_{sln}^{I\rightarrow II^{\ddagger}} = 2.56$ kcal/mol) and not the nucleophilic attack ($\Delta\Delta G_{sln}^{III\rightarrow IV^{\ddagger}} = 0.48$ kcal/mol). In the first step, the existence of a low barrier transition state for the proton transfer ($\Delta E = 0.544$ kcal/mol, $\Delta G_{gas} = -0.542$ kcal/mol, $\Delta G_{sol} = 3.1$ kcal/mol, see Table 2) is consistent with the results from a quantum chemical study on the proton transfer in RNase A catalysis involving the formation of short, strong hydrogen bonds ($\Delta E = 1.07$ kcal/mol, $\Delta G_{gas} = -1.24$ kcal/mol, 2.85 kcal/mol $\leq \Delta G_{sol} \leq$ 3.89 kcal/mol[49]). This low barrier for proton transfer is also in agreement with the explanation proposed for rapid enzyme-catalyzed proton abstraction associated with late transition states.[50] In the particular case of phosphodiesters, the low barrier for proton transfer in the reaction catalyzed by RNase A was also explained by the stabilization of the developing negative charge in the transition state that reduces the structural reorganization between the reactant and the transition state.[51] In this two-metal-ion model, the difference in the magnitude of the energy barriers for the two monoanion/dianion interconversions lies in the synchronization (or lack thereof) between the actual proton transfers and the associated molecular processes. In general, it is expected that a product stabilizing factor that lags behind bonding changes or the loss of a reactant stabilizing factor that is ahead of bonding changes enhances the barrier. By contrast, the late loss of a reactant stabilizing factor or the early development of a product stabilizing factor would lower the barrier.

In the first step, there is only one chemical process (cleavage of the 2′ O−H bond) that corresponds to the proton transfer associated with the nucleophile activation and a single change of charge distribution between the three fragments of the complex: the RNA moiety, the hydrated metal at site I, and the hydrated metal at site II (between $II^{\ddagger}$ and III, Figure 6A). The proton transfer occurs late along the reaction pathway (Supporting Information Figure S3A), so that the reactant I and the transition state $II^{\ddagger}$ are very close in structure (Table 1 and Figure 3A) and in energy ($\Delta\Delta G_{gas}^{I\rightarrow II^{\ddagger}} = -0.54$ kcal/mol, Table 2) and conserve the same net charge of −1 on the RNA moiety (monoanion). This similarity between the reactant I and the first transition state $II^{\ddagger}$ is typical for a late loss of reactant stabilizing factors that leads to a low intrinsic barrier (the cleavage of the 2′ O−H bond occurs after the transition state).

The last step of the reaction is considerably more complex than the first, since there are several concurrent chemical processes. The proton transfer to the 5′ oxygen of the leaving group is associated with the first change of the charge distribution, and there is the breaking of the P−O5′ bond and a second proton transfer between the two hydrated metals, where the negative charge (OH^-) formed on the metal coordinated to the leaving group is transferred from Mg_{II} in $VIII^{\ddagger}$ to Mg_I in IX (Figure 6B). The main process, corresponding to the P−O5′ bond breaking, occurs rather late along the reaction pathway, while the first proton transfer occurs prior to the transition state (Supporting Information Figure S3D). The imbalance of the transition state is reinforced by the poor synchronization between each of the two chemical processes and the concurrent molecular processes corresponding to the charge delocalization and solvation. The first proton transfer corresponding to the change of net charge on the RNA moiety between the reactant (dianion VII) and the transition state (monoanion $VIII^{\ddagger}$) is characterized by an unfavorable electrostatic contribution to the RNA−metal interaction ($\Delta\Delta E_{ES}^{VII\rightarrow VIII^{\ddagger}} = 308$ kcal/mol, Table 4). On the other hand, the late second proton transfer induces a charge delocalization between the two hydrated metals that corresponds to an electrostatic stabilizing factor ($\Delta\Delta E_{ES}^{VIII^{\ddagger}\rightarrow IX} = -61$ kcal/mol, Table 4). This product stabilization is developed late along the reaction pathway, after the bond breaking at the TS (in the IRC profile: shoulder observed after the TS, Figure 5D). It can be assigned as an electrostatic contribution (noncovalency contribution) to the RNA−metal binding, which is less unfavorable for the product than for the transition state ($\Delta\Delta E_{ES}^{VII\rightarrow IX} =$ 247 kcal/mol vs $\Delta\Delta E_{ES}^{VII\rightarrow VIII^{\ddagger}} = 308$ kcal/mol). In summary, the high energy barrier of the rate-limiting step arises from the early destabilization of the third intermediate and the late stabilization of the product, associated with the P−O5′ bond breaking and the secondary proton transfers.

In the single-metal-ion model developed by Torres et al.,[20] there are two equivalent steps corresponding to monoanion/dianion interconversions (Figure 6C and D). Interestingly, the first transition state that corresponds to a monoanion-to-dianion conversion exhibits a high energy barrier (activation barrier of 18.6 kcal/mol). On the other hand, in the last step of the reaction, the second transition state corresponding to a reverse dianion-to-monoanion conversion exhibits a relative low energy barrier (relative activation barrier of 2.2 kcal/mol from the intermediate).[20] In the first step (Figure 6C), the proton transfer is early (and concurrent with the nucleophilic attack) and ahead of the transition state, while it is late in the last step of the reaction (Figure 6D) and lags behind the transition state. Although the energetic trend is opposite with respect to that of the two-metal-ion model, we can find common features between the two models (Figure 6). The comparison of the two models suggests that the monoanion/dianion interconversions induced by proton transfer exhibit (1) a high energy barrier when the charge redistribution is ahead of the transition state and (2) a low energy barrier when the charge redistribution lags behind the transition state. In the two-metal-ion model, the particularly low energy barrier involved in the monoanion-to-dianion conversion (first step) is associated with a proton transfer between two common metal ligands (O2′H and OH^- at site I) and thus only involves a very localized charge redistribution between the reactant and the first transition state ($II^{\ddagger}$).

3.5. Novelty of the Two-Metal-Ion Model. The two-metal-ion model described here is similar in spirit to the dianionic mechanism proposed by von Hippel et al.[26] However, the present work provides quantitative calculations of the mechanism which make possible a detailed understanding not available from the earlier, more qualitative description. The data supporting a two-metal-ion model[24−28] suggest that the metal ions act as Lewis acids (i.e., stabilizing the negative charge on the bridging 2′ and 5′ oxygens, Figure 1C) and not as general acids/bases, as was proposed for the single-metal-ion model (Figure 1B); the latter was inspired by the RNase A mechanism for transphosphorylation (Figure 1A). In the two-metal-ion model described here, the metal Mg_I which is coordinated to the 2′ oxygen acts as a Lewis acid by polarizing the 2′ O−H bond and thus facilitating the deprotonation. The calculated activation free energy barrier for the 2′ OH deprotonation and for the

nucleophilic attack is 2.6 kcal/mol. This contrasts with single-metal-ion-model estimates of 12 kcal/mol[52] and 18.6 kcal/mol.[20] The large energy difference is explained by the fact that, in the two-metal-ion model, the pK_a of the 2′ OH is lowered by the direct (inner-sphere) coordination of Mg_I to the 2′ oxygen, which polarizes the hydroxide bond, and that the dianion formed after the nucleophile activation is stabilized by both metals. These two features are absent from the single-metal-ion model. The NEDA results (see section 3.4) suggest that the stabilization of the RNA−metal complex, in which the RNA moiety is a dianion, is due both to Mg_I (2/3) and Mg_{II} (1/3). The presence of two metal ions not only makes the nucleophile activation more favorable, but it also stabilizes the RNA−metal complex in a conformation for in-line attack. As result, the free energy barrier in the solvated system for the second step of the reaction (the nucleophilic attack) is significantly lowered; it is 22.8 kcal/mol in the absence of the metal ions and 0.47 kcal/mol in their presence. The calculations performed with different geometries indicate that the pentacoordinated (rather than hexacoordinated) metal coordinated to the 2′ oxygen plays an important role in stabilizing the geometries which facilitate the two first steps of the reaction and contributes to lower their energy barriers (data not shown). However, the calculations indicate that the departure of the leaving group prefers a hexacoordinated state for both metals, so that a switch from Mg_I(V) to Mg_I(VI) takes place in the formation of the last intermediate. The NBO analyses show that the process of coordination change, associated with the natural bond orbitals $\sigma_{P-O2'}$ and $\sigma_{P-O5'}$, prepares the system for the departure of the leaving group. The increasing delocalization of the lone pair on the 2′ oxygen into the P−O5′ antibonding orbital, during the coordination change (V to VII), contributes to slightly weaken the P−O5′ bond before the departure of the leaving group. Product formation is facilitated by the delocalization of the hydroxide ion through secondary proton transfers involving two water molecules; the first one belongs to the solvation shell of Mg_{II} (in the axial position opposite the 5′ oxygen), and the second one belongs to the solvation shell of Mg_I (in the axial position opposite the 2′ oxygen). The proton transfer that occurs in the protonation of the 5′ oxygen in the rate-limiting step of the reaction is reminiscent of the single-metal-ion model (Figure 1B).

A theoretical model based on a single-metal-ion model and involving a similar proton transfer was proposed recently.[20] However, since the metal is not directly coordinated to the 5′ oxygen, it mostly behaves as a general acid/base. In the two-metal-ion model developed here, Mg_{II} acts both as a general acid and as a Lewis acid by giving a proton from one of its coordinated water molecules to the 5′ oxygen and accepting its electrons, a feature that has not been proposed in previous discussions of two-metal-ion models.[24−28,37,52] A two-metal-ion model was proposed by Boero et al. after this paper was completed, using a Car−Parrinello molecular dynamics (MD) simulation method for a model system.[53] In this model, the base generated by the spontaneous deprotonation of a water molecule located close to the 5′ leaving group is involved in the neutralization of the 5′ oxyanion. A similar neutralization occurs in the current model, but the water molecule involved is in the inner-sphere coordination of the metal at site II and the proton transfer precedes the P−O5′ bond breaking (Supporting Information Figure S3D). Both models follow a reaction path in which the nucleophile activation proceeds via a metal hydroxide as the general base, the subsequent nucleophilic attack leads to the formation of a trigonal bipyramidal structure, and the rate-limiting step is the departure of the leaving group. The major differences between the two models concern the interactions with the metal ions; that is, the metal solvation shells (metal/water or metal/OH−) and the metal coordinations with the RNA moiety differ at all steps of the reaction (in particular the coordinations with the nonbridging oxygens, Boero et al.,[53] Tateno, personal communication). As pointed out by Boero et al.,[53] the activation free energies calculated from first-principles MD simulations are higher than those obtained for the single-metal-ion model.[20] The lower activation barrier obtained in the present model agrees well with the experimental data and is likely to come from an optimal reaction pathway which was not sampled by Boero et al. It is noteworthy that the relative free energy difference between the noncatalyzed (no metal) and metal-catalyzed (two-metal-ion models) reactions is similar in the two models: it is 16.3 kcal/mol for Boero's model and 13.7 kcal/mol for our model.

3.6. Solvent Isotope Effect. The occurrence of a proton transfer during the rate-limiting step of the reaction is controversial. A large solvent deuterium isotope effect ($k_{cleav}^{(H_2O)}/k_{cleav}^{(D_2O)} = 4.3$) has been observed for the hammerhead ribozyme.[24] In the case of the participation of a metal hydroxide in the catalysis (Figure 1B), the observed isotope effect was imputed to a change in the equilibrium concentration of active species (Mg^{2+}-bound 2′ alkoxide).[24,54] Consequently, the isotope effect was not interpreted as a proton transfer occurring in the transition state (departure of the leaving group) but instead as an apparent isotope effect; that is, the isotope effect simply reflects a difference in the concentration of the activated 2′ oxyanion in D_2O (which is severalfold lower than that in H_2O) at a given pH.[17,24] Thus, the intrinsic isotope effect associated with the 2′ OH deprotonation would be equal to 1. However, experimental evidence supports the proton transfer when Mg^{2+} is substituted by NH_4^+ as a cofactor in the reaction.[35] To determine whether the two-metal-ion model proposed here is consistent with the observed isotope effect, we have used a simple model to estimate the kinetic solvent isotope effect (KIE) for the rate-determining (fourth) reaction step (Figure 3D).

Neglecting tunneling, the kinetic isotope effect can be approximated in transition state theory (TST) by the change in activation free energy of H_2O versus D_2O.[55] This gives

$$KIE_{cl} \simeq e^{(\Delta G_D^{\ddagger} - \Delta G_H^{\ddagger})/k_B T} \quad (1)$$

where $\Delta G^{\ddagger}$ is the barrier height corrected for zero-point-energy (ZPE) changes between the reactant (V, Figure 3D) and the transition state (VI, Figure 3D) in H_2O ($\Delta G_H^{\ddagger}$) and in D_2O ($\Delta G_D^{\ddagger}$) calculated at the HF/3-21+G** level (Supporting Information Table S2). The calculations give a KIE that is close to unity ($KIE_{cl} = 1.2$) and is consistent with the reaction asymmetry for this proton transfer.[55] This suggests that the KIE observed experimentally is likely to be due to a change in the equilibrium concentration of charged species in the presence of Mg^{2+}, as proposed earlier,[24,35] but that is not incompatible with a proton transfer to the leaving 5′ oxygen in the rate-limiting step of the reaction. In the case of a nucleophile activation model, based on a metal hydroxide at site I, an intrinsic KIE corresponding to a proton transfer on the 5′ oxygen, combined with a change in the equilibrium concentration of charged species, would lead to a much larger apparent KIE, as is observed for the NH_4^+-mediated reaction ($k_{cleav}^{(H_2O)}/k_{cleav}^{(D_2O)} = 7.68$.[35]

4. Concluding Discussion

A two-metal-ion model for hammerhead-ribozyme catalysis, based on density functional quantum mechanical calculations,

is described and analyzed. It is found that the reaction involves a series of steps with three intermediates and four transition states. The calculated free energy barriers for the solvated system (Table 2), confirmed by reaction path following (Supporting Information Figure S3), indicate that the intermediates involved in the first three steps of the reaction (nucleophile activation, nucleophilic attack, and formation of an intermediate involved in the departure of the leaving group) would have lifetimes too short to be kinetically significant; that is, the free energy barriers of the steps from the reactant to the final intermediate are all small, with the largest arising in the nucleophile activation (2.6 kcal/mol). The high free energy barrier which corresponds to the rate-determining step, involves the departure of the leaving group. The calculated free energy barrier (19.3 kcal/mol) is in good agreement with the measured value 20.1 kcal/mol. The high barrier is suggested to result from an "imbalanced" transition state; that is, there is an early destabilization of the last intermediate VII (due to the proton transfer from a coordinated water molecule at site II on the 5′ oxygen, Figure 3D), which occurs before the P−O5′ bond breaking, and a late stabilization of the product (due to the delocalization of the OH^- formed at site II and transferred to site I) that lags behind the breaking of the P−O5′ bond.

Comparison of a model for RNA catalysis corresponding to the base-catalyzed reaction in solution in the absence of metal ions (Leclerc et al., to be published separately) and the two-metal-catalyzed reaction in the hammerhead ribozyme, as obtained in the present paper, suggests that the metal ions contribute to the catalysis in several essential ways: (1) by lowering the pK_a of the 2′ OH, (2) by stabilization of the in-line conformation of the reactant, (3) by stabilization of the trigonal bipyramidal structure of the transition states and intermediates, and (4) by stabilization of the leaving group. One of the metal ions, identified as Mg_{II}, acts as a Lewis acid and a general acid, while the other metal ion, identified as Mg_I, functions as a general base and Lewis acid, so there is a clear distinction between the roles played by the two metals. Their roles as Lewis acids evolve along the reaction path; that is, a strong Lewis acid is needed at site I in the first steps (roles 1−3 described above) and at site II in the last steps of the reaction (roles 3 and 4 described above). The role of both metals as a Lewis acid is reinforced at site I by a pentacoordinated state of Mg_I in the two first steps of the reaction and at site II by a hexacoordinated state of Mg_I that contributes to delocalize the negative charge developed around the leaving group in the two last steps of the reaction. Given the above, we expect that any factor that minimizes the imbalance of the transition state in the rate-limiting step by making the proton transfer more synchronized with the P−O5′ bond breaking would accelerate the reaction. This can be accomplished, for example, by a later destabilization of the third intermediate (VII) or an earlier stabilization of the product (IX) with respect to the transition state ($VIII^{\ddagger}$). As proposed previously,[28] the presence of a stronger Lewis acid at site II would allow a better stabilization of the negative charge on the leaving group. In light of the present two-metal-ion model, the replacement of Mg^{2+} by La^{3+} at site II, which leads to an enhanced apparent rate constant for the hammerhead-ribozyme cleavage reaction,[28] lowers the energy barrier by a slightly earlier product stabilization (i.e., the P−O5′ bond would break earlier, making the departure of the leaving group more synchronized with the proton transfer) along the reaction pathway (likely by sequestrating the OH^- at site II).

In summary, the calculations presented here have provided a detailed reaction path for the two-metal-ion ribozyme catalysis and identified the metal-ion contribution to the reaction. Although the results are in agreement with experiments, additional measurements (KIE, for example) are necessary to confirm the analysis. We hope that having a specific proposal will stimulate new studies of this reaction, which is of fundamental importance in living systems.

Acknowledgment. The authors are grateful to Qiang Cui for many profitable discussions. We thank Eric Gledening and Baudilio Tejerina for technical support with the NBO program and Michael E. Harris for personal comments. F.L. was a fellow of the Human Frontier Science Program (1998−2000). This work was partially supported by grants from the Department of Energy (DOE, U.S.A.) and from the National Institute of Health (NIH) and by CNRS funding for young investigators (ATIP, France). The calculations were performed, in part, at the National Energy Research Scientific Computing Center (NERSC), at the Institut du Développement et des Ressources en Informatique Scientifique (IDRIS, France), and at the Centre Informatique National de l′Enseignement Supérieur (CINES, France).

Supporting Information Available: Detailed procedures for construction of guess geometries, geometry optimizations and energetics, location of transition state structures, and NBO and NEDA calculations. Data on the triester-like and dianionic mechanisms in solution and data on IRC calculations. Complementary data on the NEDA calculations of the three-body interactions and on the calculated barrier heights used to estimate the kinetic solvent isotope effect. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Cech, T.; Zaug, A.; Grabowski, P. *Cell* **1981**, *27*, 487−496.
(2) Zaug, A.; Cech, T. *Nucleic Acids Res.* **1982**, *10*, 2823−2838.
(3) Scott, W.; Murray, J.; Arnold, J.; Stoddard, B.; Klug, A. *Science* **1996**, *274*, 2065−2069.
(4) Scott, W.; Finch, J.; Klug, A. *Cell* **1995**, *81*, 991−1002.
(5) Murray, J.; Szoke, H.; Szoke, A.; Scott, W. *Mol. Cell* **2000**, *5*, 279−287.
(6) Murray, J.; Terwey, D.; Maloney, L.; Karpeisky, A.; Usman, N.; Beigelman, L.; Scott, W. *Cell* **1998**, *92*, 665−673.
(7) McKay, D. *RNA* **1996**, *2*, 395−403.
(8) Hammann, C.; Lilley, D. *ChemBioChem* **2002**, *3*, 690−700.
(9) Blount, K.; Grover, N.; Mokler, V.; Beigelman, L.; Uhlenbeck, O. *Chem. Biol.* **2002**, *9*, 1009−1016.
(10) Lilley, D. *Curr. Opin. Struct. Biol.* **1999**, *9*, 330−338.
(11) Fedor, M.; Uhlenbeck, O. *Biochemistry* **1992**, *31*, 12042−12054.
(12) van, T. H.; Buzayan, J.; Feldstein, P.; Eckstein, F.; Bruening, G. *Nucleic Acids Res.* **1990**, *18*, 1971−1975.
(13) Koizumi, M.; Ohtsuka, E. *Biochemistry* **1991**, *30*, 5145−5150.
(14) Slim, G.; Gait, M. *Nucleic Acids Res.* **1991**, *19*, 1183−1188.
(15) Raines, R. *Chem. Rev.* **1998**, *98*, 1045−1066.
(16) Wang, S.; Karbstein, K.; Peracchi, A.; Beigelman, L.; Herschlag, D. *Biochemistry* **1999**, *38*, 14363−14378.
(17) Kuimelis, R.; McLaughlin, L. *Biochemistry* **1996**, *35*, 5308−5317.
(18) Hermann, T.; Auffinger, P.; Scott, W.; Westhof, E. *Nucleic Acids Res.* **1997**, *25*, 3421−3427.
(19) Kuimelis, R.; McLaughlin, L. *Chem. Rev.* **1998**, *98*, 1027−1044.
(20) Torres, R.; Himo, F.; Bruice, T.; Noodleman, L.; Lovell, T. *J. Am. Chem. Soc.* **2003**, *125*, 9861−9867.
(21) Steitz, T.; Steitz, J. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 6498−6502.
(22) Uebayasi, M.; Uchimaru, T.; Koguma, T.; Sawata, S.; Shimayama, T.; Taira, K. *J. Org. Chem.* **1994**, *59*, 7414−7420.
(23) Dahm, S.; Derrick, W.; Uhlenbeck, O. *Biochemistry* **1993**, *32*, 13040−13045.
(24) Sawata, S.; Komiyama, M.; Taira, K. *J. Am. Chem. Soc.* **1995**, *117*, 2357−2358.
(25) Zhou, D.; Taira, K. *Chem. Rev.* **1998**, *98*, 991−1026.

(26) Pontius, B.; Lott, W.; von Hippel, P. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 2290−2294.
(27) Zhou, D.; Zhang, L.; Taira, K. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 14343−14348.
(28) Lott, W.; Pontius, B.; von Hippel, P. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 542−547.
(29) Lyne, P.; Karplus, M. *J. Am. Chem. Soc.* **2000**, *122*, 166−167.
(30) Nakamatsu, Y.; Warashina, M.; Kuwabara, T.; Tanaka, Y.; Yoshinari, K.; Taira, K. *Genes Cells* **2000**, *5*, 603−612.
(31) Zhou, J.; Zhou, D.; Takagi, Y.; Kasai, Y.; Inoue, A.; Baba, T.; Taira, K. *Nucleic Acids Res.* **2002**, *30*, 2374−2382.
(32) Curtis, E.; Bartel, D. *RNA* **2001**, *7*, 546−552.
(33) Murray, J.; Seyhan, A.; Walter, N.; Burke, J.; Scott, W. *Chem. Biol.* **1998**, *5*, 587−595.
(34) O'Rear, J.; Wang, S.; Feig, A.; Beigelman, L.; Uhlenbeck, O.; Herschlag, D. *RNA* **2001**, *7*, 537−545.
(35) Takagi, Y.; Inoue, A.; Taira, K. *J. Am. Chem. Soc.* **2004**, *126*, 12856−12864.
(36) Inoue, A.; Takagi, Y.; Taira, K. *Nucleic Acids Res.* **2004**, *32*, 4217−4223.
(37) Zhou, D.; He, Q.; Zhou, J.; Taira, K. *FEBS Lett.* **1998**, *431*, 154−160.
(38) Scott, E.; Uhlenbeck, O. *Nucleic Acids Res.* **1999**, *27*, 479−484.
(39) Murray, J.; Dunham, C.; Scott, W. *J. Mol. Biol.* **2002**, *315*, 121−130.
(40) Bock, C.; Kaufman, A.; Glusker, J. *Inorg. Chem.* **1994**, *33*, 419−427.
(41) Cunningham, L.; Li, J.; Lu, Y. *J. Am. Chem. Soc.* **1998**, *120*, 4518−4519.
(42) Scott, A.; Radom, L. *J. Phys. Chem.* **1996**, 100, 16502−16513.
(43) Marten, B.; Kim, K.; Cortis, C.; Friesner, R.; Murphy, R.; Ringnalda, M.; Sitkoff, D.; Honig, B. *J. Phys. Chem.* **1996**, *100*, 11775−11788.
(44) Zhou, D.; Usman, N.; Wincott, F.; Matulic-Adamic, J.; Orita, M.; Zhang, L.; Komiyama, M.; Kumar, P. K. R.; Taira, K. *J. Am. Chem. Soc.* **1996**, *118*, 5862−5866.
(45) Bash, P.; Field, M.; Davenport, R.; Petsko, G.; Ringe, D.; Karplus, M. *Biochemistry* **1991**, *30*, 5826−5832.
(46) Gonzalez, C.; Schlegel, H. *J. Phys. Chem.* **1990**, *94*, 5523−5527.
(47) Scott, W. *Q. Rev. Biophys.* **1999**, *32*, 241−284.
(48) Glendening, E.; Badenhoop, J.; Reed, A.; Carpenter, J.; Bohmann, J.; Morales, C.; Weinhold, F. Theoretical Chemistry Institute, University of Wisconsin, Madison, 2001.
(49) Vishveshwara, S.; Madhusudhan, M.; Maizel, J. J. *Biophys. Chem.* **2001**, 89, 105−117.
(50) Gerlt, J.; Gassman, P. *J. Am. Chem. Soc.* **1993**, *115*, 11552−11568.
(51) Gerlt, J.; Gassman, P. *Biochemistry* **1993**, *32*, 11943−11952.
(52) Boero, M.; Terakura, K.; Tateno, M. *J. Am. Chem. Soc.* **2002**, *124*, 8949−8957.
(53) Boero, M.; Tateno, M.; Terakura, K.; A, O. *J. Chem. Theory Comput.* **2005**, *1*, 925−934.
(54) Takagi, Y.; Taira, K. *J. Am. Chem. Soc.* **2002**, *124*, 3850−3852.
(55) Kiefer, P.; Hynes, J. *J. Phys. Chem.* **2003**, *107*, 9022−9039.

CHAPTER FIFTEEN

A Dedicated Computational Approach for the Identification of Archaeal H/ACA sRNAs

Sébastien Muller, Bruno Charpentier, Christiane Branlant, *and* Fabrice Leclerc

Contents

Abstract

Whereas dedicated computational approaches have been developed for the search of C/D sRNAs and snoRNAs, as yet no dedicated computational approach has been developed for the search of archaeal H/ACA sRNAs. Here we describe a computational approach allowing a fast and selective identification of H/ACA sRNAs in archaeal genomes. It is easy to use, even for biologists having no special expertise in computational biology. This approach is a stepwise knowledge-based approach, combining the search for common structural features of H/ACA motifs and the search for their putative target sequences. The first step is based on the ERPIN software. It depends on the establishment of a secondary structure-based "profile." We explain how this profile is built and how to use ERPIN to optimize the search for H/ACA motifs. Several examples of applications are given to illustrate how powerful the method is, its limits, and how the results can be evaluated. Then, the possible target rRNA sequences corresponding to the identified H/ACA motifs are searched by use of a descriptor-based method (RNAMOT). The principles and the practical aspects of this method are also explained, and several examples are given here as well to help users in the interpretation of the results.

Laboratoire de Maturation des ARN et Enzymologie Moléculaire, Nancy Université, Faculté des Sciences et Techniques, Vandoeuvre-les-Nancy, France

Methods in Enzymology, Volume 425
ISSN 0076-6879, DOI: 10.1016/S0076-6879(07)25015-3

1. Introduction

Pseudouridine (Ψ) is one of the most abundant posttranscriptionally modified nucleotides found in tRNAs, rRNAs, and UsnRNAs (Rozenski *et al.*, 1999). Uridine to pseudouridine conversion can be catalyzed by a single protein with RNA recognition capacity and RNA/Ψ-synthase activity. It can also be catalyzed by H/ACA RiboNucleoProtein complexes (H/ACA sRNPs) containing a guide RNA with H/ACA boxes, called snoRNAs and scaRNAs in Eukarya and sRNAs in Archaea. These RNAs are associated with a set of proteins, Nhp2p, Nop10p, Cbf5, and Gar1p in yeast, NHP2, NOP10, Dyskerin (or NAP57), and GAR1 in human (Henras *et al.*, 1998; Khanna *et al.*, 2006; Lafontaine *et al.*, 1998; Meier and Blobel, 1994; Wang and Meier, 2004; Watkins *et al.*, 1998), and their homologs L7Ae, aNOP10, aCBF5, and aGAR1 in Archaea (Baker *et al.*, 2005; Charpentier *et al.*, 2005; Rozhdestvensky *et al.*, 2003). Proteins aCBF5 and Cbf5/Dyskerin belong to the TruB family of RNA/Ψ-synthases (Hamma *et al.*, 2005; Lafontaine *et al.*, 1998; Manival *et al.*, 2006; Meier and Blobel, 1994; Rashid *et al.*, 2006; Wang and Meier, 2004).

The guiding properties of H/ACA RNAs are based on the formation of a complex structure: two sequences from the RNA substrate, that are separated by a 5′-UN-3′ dinucleotide, base pair with the two strands of an internal loop of the guide RNA (pseudouridylation pocket). The U residue in the 5′-UN-3′ dinucleotide is the target site of the reaction (Balakin *et al.*, 1996; Ganot *et al.*, 1997a,b; Ni *et al.*, 1997).

Several studies combining computational and experimental approaches have been dedicated to the identification of eukaryal H/ACA snoRNAs in human, mouse, *D. melanogaster, S. cerevisiae*, and *S. pombe* (Huang *et al.*, 2004, 2005; Kiss *et al.*, 2004; Li *et al.*, 2005; Schattner *et al.*, 2004, 2006; Torchet *et al.*, 2005; Yuan *et al.*, 2003). Most of these RNAs were found to have a two stem-loop structure. Most generally, each of the two stem loops contains a pseudouridylation pocket. These stem loops are linked by a single-stranded sequence containing the ANANNA H box and the 3′ stem loop is flanked by a single-stranded element containing an ANA trinucleotide. Whereas the global architecture of eukaryal snoRNAs is highly conserved, the sizes and base compositions of their stem-loop structures are highly variable.

Knowledge regarding archaeal H/ACA sRNAs is more limited, because they have only been characterized in *Archaeoglobus fulgidus* (Tang *et al.*, 2002a), *Pyrococcus* species (Baker *et al.*, 2005; Charpentier *et al.*, 2005; Rozhdestvensky *et al.*, 2003), and *Methanocaldococcus jannaschii* (Thebault *et al.*, 2006). Identified Ψ residues in archaeal rRNAs are also very limited: 4, 6, and 3 Ψ residues were detected in the 23S rRNA of *Halobacterium*

halobium, Sulfolobus acidocaldarius, and *Haloarcula marismortui*, respectively (Del Campo *et al.*, 2005; Massenet *et al.*, 1999; Ofengand and Bakin, 1997). However, the available data reveal some specific features of archaeal H/ACA sRNAs compared with their eukaryal counterparts. Indeed, although the target U residue is identified by the same base-pairing rules between the guide RNA and the targeted sequence, the global architecture of H/ACA sRNAs is more variable. One, two, or three contiguous stem-loop structures containing a pseudouridylation pocket can be present. They may even contain an additional stem-loop structure without a pseudouridylation pocket (see Pf6, Rozhdestvensky *et al.*, 2003). In addition, compared with their eukaryal counterparts, they are structurally more constrained, with a strong conservation of the relative positions, sizes, and base compositions of their structural elements. In each stem-loop structure, the stems delineating the pseudouridylation pocket are highly enriched in G-C and C-G pairs. Each stem loop also displays a K-turn motif or a K-loop motif (Charpentier *et al.*, 2005; Charron *et al.*, 2004; Hamma *et al.*, 2005; Klein *et al.*, 2001; Li and Ye, 2006; Rozhdestvensky *et al.*, 2003; Vidovic *et al.*, 2000) in its apical part. K-turn and K-loop motifs are characterized by the presence of an A•G and G•A sheared pair flanked by a U residue. Each stem-loop structure is tailed by an ANA box that is more frequently an ACA triplet. By inspection of the secondary structure of the identified H/ACA sRNAs, we observed that the K-turn or K-loop motif on the one hand and the ACA box on the other hand are located at a well-defined distance from the targeted U residue in the pseudouridylation pocket.

Therefore, although not unique, the characteristic features of the H/ACA sRNA stem-loop structures, flanked by their associated ANA box, are specific enough to represent a signature that can be used for H/ACA sRNA gene identification by computational search in sequence databases. For simplification, we will designate these modular elements as H/ACA motifs. On the basis of the peculiar structure formed by the H/ACA sRNA and the RNA substrate, the finding by computational analysis of putative target sequences can also be used as a complementary source of information to discriminate true H/ACA sRNA genes from false-positive DNA sequences.

Taking advantage of the specificity of archaeal H/ACA sRNAs, we developed a knowledge-based approach that combines the consecutive searches for sequences coding for H/ACA motifs in complete sequences of archaeal genomes and their complementary RNA targets in rRNA sequences. The identification of H/ACA motifs is performed by use of a profile-based approach, taking advantage of the present knowledge on H/ACA sRNAs. The knowledge used includes data from comparative sRNA sequence analysis (Baker *et al.*, 2005; Charpentier *et al.*, 2005; Rozhdestvensky *et al.*, 2003; Tang *et al.*, 2002a; Thebault *et al.*, 2006), *in vitro* reconstitution of active H/ACA sRNPs by use of various WT and mutated H/ACA sRNAs and various RNA substrates (Charpentier *et al.*, 2005; S. Muller *et al.*, unpublished

results) and the recent determination of the 3D structure of a complete H/ACA sRNP particle (Li and Ye, 2006). Then, the identification of possible RNA targets associated with the identified H/ACA motifs is performed by use of a descriptor-based approach, taking into account the rules of complementarities that were also inferred from the data from *in vitro* reconstitution (S. Muller *et al.*, unpublished data). Because it is a knowledge-based approach, each unknown H/ACA motif associated with an identified target can, after an appropriate validation, contribute to increase and refine our knowledge on the H/ACA structural features and their target recognition. Thus, the general performance of the approach may be improved.

This chapter is dedicated to biologists who have no special expertise in computational methods. Our goal is to teach how to use the proposed strategy to get an H/ACA sRNA gene identification that will be as extensive as possible with a minimized number of false-positive results. The use of a computational approach for the prediction of H/ACA sRNA genes in Archaea and the prediction of their target sites are as important as *in vitro* reconstitution assays that are available to validate the results (see Chapter 16 by Charpentier *et al.*). The combined use of computational predictions and *in vitro* tests should increase considerably our knowledge on archaeal sRNAs in the near future.

2. Method

Our strategy for the detection of H/ACA sRNAs follows a stepwise and iterative procedure in which the first step is the search for H/ACA-like motifs through archaeal genomes, and the second step is the determination of the associated RNA target(s) in rRNAs (Fig. 15.1). In the first step (Fig. 15.1, step 1), H/ACA-like motifs are detected by use of the profile-based ERPIN program (Gautheret and Lambert, 2001). This program has been applied to the search of a wide range of RNA motifs (Lambert *et al.*, 2002, 2004; Legendre *et al.*, 2005). Once H/ACA-like motifs are identified, their putative target(s) in rRNAs are searched (Fig. 15.1, step 2) by use of the descriptor-based RNAMOT program. This program has also been extensively applied to the identification of several kinds of RNA motifs, and the data obtained have been experimentally validated (Bourdeau *et al.*, 1999; Laferriere *et al.*, 1994; Lescure *et al.*, 1999, 2002). In the present strategy, the use of RNAMOT is atypical, because this program is applied to the search of RNA motifs formed by partial base pairing of the internal loops of the H/ACA motifs with ribosomal RNAs. Because of the peculiar application of both ERPIN and RNAMOT in the present strategy, we had to create an appropriate H/ACA profile for ERPIN and several RNA target descriptors for RNAMOT. We will describe the rationale used for their generation

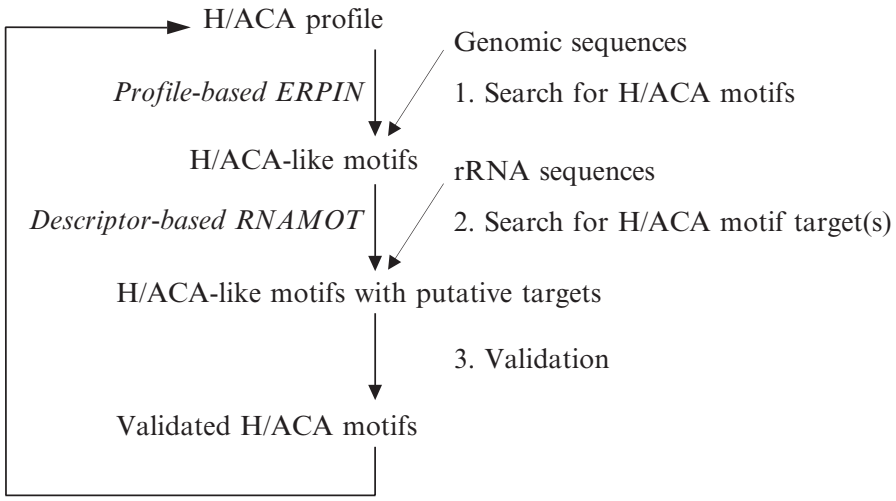


Figure 15.1 The multistep strategy proposed for the identification of H/ACA motifs. A description of the three steps leading to the identification of new H/ACA sRNA candidates: first an ERPIN-based step is used for the search of H/ACA-like motifs, then, a descriptor-based step is used for the search of their targets. Each of them includes tests of the validity of the results and a final global validation step is performed, taking into account the present knowledge on H/ACA sRNAs and rRNA pseudouridylation. The sequences of the new validated H/ACA motifs can be integrated in the H/ACA profile, and the search can be repeated on the same genome or run on other genomes.

and how to use them to achieve maximal efficiency and specificity of the delivered data.

By this procedure, H/ACA-like motifs that exhibit at least one putative RNA target are identified. They are then subjected to an evaluation step (Fig. 15.1, step 3). This evaluation step is based on: (1) current knowledge on Ψ positions in archaeal rRNA, (2) comparison with already identified archaeal H/ACA sRNAs and, (3) experimental data obtained from structure–function analysis of H/ACA sRNA by use of the *in vitro* reconstituted system. Obviously, the final proof of the validity of the prediction requires experimental tests of the proposed guiding property by use of the *in vitro* assembly procedure (see Chapter 16 by Charpentier *et al.*) and the identification of a Ψ residue at the targeted position in the rRNA by the CMCT approach (see Chapter 2). When the demonstration is completed, the newly identified H/ACA motifs can be used to enrich the ERPIN profile.

2.1. Search for H/ACA-like motifs

The search for H/ACA-like motifs requires the establishment of a profile, as implemented in the ERPIN program. This H/ACA profile is based on our present understanding of the common structural features of archaeal H/ACA motifs. Once the profile is built, the search can be started. The basic ERPIN command, parameters, and options and the present stage of their optimization

are explained by use of various examples of searches of H/ACA motifs in archaeal genomes. Through the proposed interpretations of the data obtained and by giving some tips, we intend to help users speed up their search, analyze and evaluate the results, and improve the initial profile.

2.1.1. Requirements

2.1.1.1. Hardware and software The ERPIN program used in this approach runs and has been tested under most UNIX platforms (for more details about compatible operating systems, see references). The ERPIN program is available as a stand-alone version that can be obtained from D. Gautheret (Gautheret and Lambert, 2001) and as a web server (http://tagc.univ-mrs.fr/erpin/) (Lambert *et al.*, 2004). The results presented here were obtained with the most recent release of ERPIN (version 5.5).

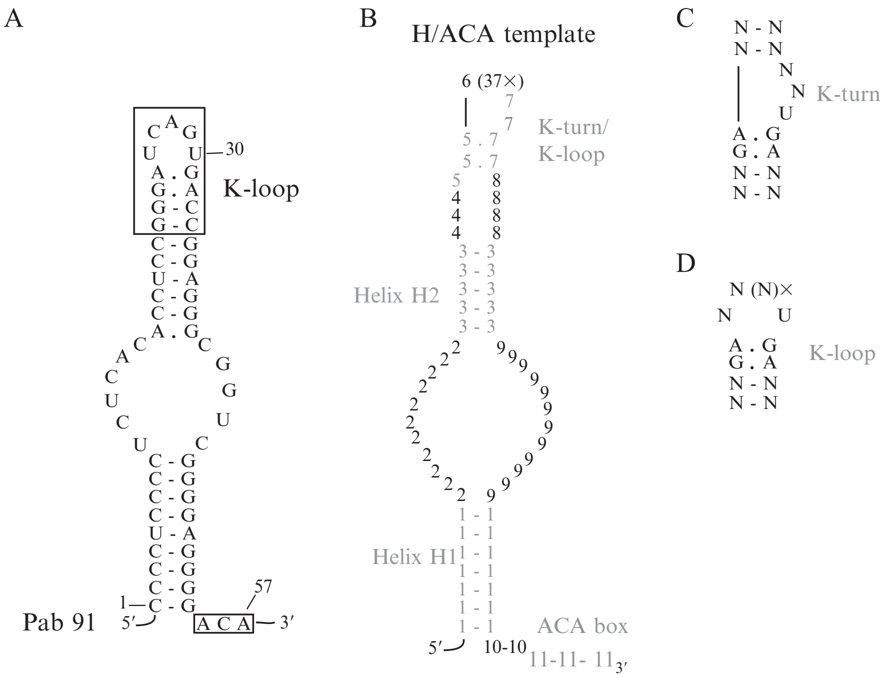
2.1.1.2. Sequence data The sequence format used for ERPIN applications is FASTA. Already published archaeal genomic sequences can be downloaded from the NCBI ftp site: ftp://ftp.ncbi.nih.gov/genomes/Bacteria. The sequences of the 26 already identified H/ACA sRNAs sequences were used to establish the H/ACA profile (Baker *et al.*, 2005; Charpentier *et al.*, 2005; Rozhdestvensky *et al.*, 2003; Tang *et al.*, 2002a). This profile is available at http://tagc.univ-mrs.fr/asterix/erpin/.

2.1.2. Establishment of the ERPIN profile

The H/ACA profile is established on the basis of a sequence alignment of all the H/ACA motifs of the already identified H/ACA sRNAs (Baker *et al.*, 2005; Charpentier *et al.*, 2005; Rozhdestvensky *et al.*, 2003; Tang *et al.*, 2002a). It is based on and includes secondary structure information, and its correct establishment is critical for the sensitivity and specificity of the ERPIN search.

To build this profile, the sequences aligned are considered as a succession of single-stranded (sx) and double-stranded (Hx) elements, x is a number defined according to the position of the element in the H/ACA sequence (5′ to 3′). Each nucleotide in a defined structural element carries the number attributed to this element (Fig. 15.2B). These numbers correspond to the first lines of the profile (Fig. 15.2E). Segments known to be single-stranded, like the ACA triplet or the two guide sequences, or segments that are single-stranded or double-stranded, depending on the considered H/ACA motif, are each defined by a unique stretch of a given number. In contrast, obligatory helices are identified by the presence of two stretches of residues that are all assigned to the same number and have identical lengths. When two stretches with an identical numbering are defined in the profile, the RNA sequences that do not contain the corresponding base-pair interaction are not selected in the search.

Constitution of the profile is complicated by the fact that, despite a strong conservation, H/ACA motifs show some structural variations. For



E

H1 s2 H2 s5 s7 H2 s9 H1 s11

```
>HACAsRNAs-List
0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000011111
1111111222222222223333344455566666666666666666666666666666666666666777788883333399999999999911111110011
>Afu46
AGCTCCGCC--CCTCACGCCCGGG--TGA-------GAAGC-------------------------ATGA--TCCCGGGTCGGTT----GGCGGAGCT--ACA
>Pab35-2
GGGCCGGTG----CATCCGCCCGCG-GGA---------------------TCA-------------ATGACCGCCGGGTCTCTGT----TGCCGGCCT--ACA
>Pho35-2
GGGCCGGTG----CATCCGCCCGCG-GGA------------TTA----------------------ATGACCGCTGGGTCTCTGT----TGCCGGCCC--ACA
>Pfu35-2
GGGCCGGTG----CATCCGCCCGCG-GGA---------------------TCA-------------ATGACCGCTGGGTCTCTGT----TGCCGGCCC--ACA
>Pab91
CCCCTCCCC--TCTCACACCTCCG--GGA--------TCA--------------------------GTGAC-CGGAGGGCGGTC-----GGGGAGGGG--ACA
>P-ho91
CCCCTCCCC--TCTCACACCCCCG--TGA--------TCG--------------------------GTGA-TCGGGGGGCGGTC-----GGGGAGGGG--ACA
>P-fu91
CCCCTCCCC--TCTCACACCCCCG--TGA---------------GAA-------------------GTGA-GCGGGGGGCGGTC-----GGGGAGGGG--ACA
>Pab40-3
GGCCCGTCTG---GGTTAGCCCGCC-TGATCATGCCGTTGGCTTAG--------------------ATGA-AGGCGGGTGTTA-----CGGGCGGGCT--ACA
>Pho40-3
GGCCCGTCTG---GGTTAGCCCGCC-CAATTTTGCCGAGGGCTTAG--------------------ATGA-GGGCGGGTGTTA-----CGGGCGGGCC--ACA
>Pfu40-3
GGCCCACCTG---GGTTAGCCCGCC-TGA------GAATGCATACATGCTACG-------------ATGA-GGGCGGGTGTTA-----CGGGTGGGCC--ACA
>Afu190
GCCCGCCAG---CCATGCGTTCCC--TGA--------TTG--------------------------GTGA--GGGGAATTCCAC-----TTGGCGGGC-TACA
```

Figure 15.2 The structural elements of H/ACA sRNAs taken into consideration in the ERPIN H/ACA profile and an example of H/ACA profile. (A) Secondary structure proposed for the *P. abyssi* Pab91 sRNA (Charpentier *et al.*, 2005). The ACA triplet at the 3′ end of the motif, the G/C rich helices 1 and 2, and the K-loop structure on top of helix 2 are shown. (B) The structural elements taken into consideration to build the H/ACA profile are shown, with their respective numbers (1–11), which are defined by reference to the 5′ to 3′ orientation of the sequence. (C) and (D) The K-turn and K-loop motifs used in the search are shown, respectively. (E) A portion of the sequence alignment of the H/ACA profile is given. The two stretches of identical numbers correspond to the helical elements H1 and H2. As the total number of defined structural elements exceeds 9, the digits are presented in two lines. The annotations on the top of the alignment indicates the various elements taken into consideration to build the profile, with the associated names: the basal helix (H1), the apical helix (H2), the K-turn motif 5′ and 3′ elements (s5 and s7), the ACA box (s11) and the 5′ and 3′ guide sequences (s2 and s9, respectively). The buffer elements are not indicated. For each of the two priority orders used in the search, only the H1, H2, s5, s7 and s11 elements were unmasked step by step.

instance, the sizes of helices 1 and 2 slightly vary from one motif to the other, and a limited number of bulge residues or mismatches can be present, in particular in helix 2. An extreme case of variation is exemplified in the second motif of the *P. furiosus* Pf7 sRNA, where the 5′ strand of the pseudouridylation pocket is separated from helix 1 by an additional stem-loop structure (see Pf7, Rozhdestvensky *et al.*, 2003). Finally, the presence of either an apical K-loop or an apical K-turn motif with a terminal stem loop increases the diversity of the H/ACA motifs. For an optimized prediction of H/ACA motifs, we had to take this variability into consideration. This was of high importance, because ERPIN does not select sequences that contain bulge residues and internal loops in the double-stranded regions identified by the H/ACA profile. To overcome this program limitation, two base-paired elements, H1 and H2, were defined, corresponding to the estimated minimal regular succession of base pairs present in helices 1 and 2 (7 bps for H1 and 5 bps for H2, respectively). Furthermore, to accommodate the size variations of helices 1 and 2, segments that can be either single stranded or double stranded were included in the profile (''buffer'' segments). A maximal size was defined for each ''buffer'' segment, by taking into account constant distances in H/ACA motifs (distances between the ACA trinucleotide and the pseudouridylation pocket and between the K-turn/K-loop motif and this pocket, respectively), as well as the overall lengths of known H/ACA motifs.

To be able to select RNA with either a K-loop or a K-turn motif, we defined these elements in the H/ACA profile by their common structural features: namely, two A•G and G•A sheared pairs with sequence constraints at the position 5′ to the GA sequence in the 5′ strand and at the two successive positions 5′ to the GA sequence in its 3′ strand (Fig. 15.2). Therefore, the two possible motifs are defined by the presence of a BGA (where B is any nucleotide except A) and RUGA elements that are separated by a loop. The maximal size of this loop was fixed to 37nts.

The 41 H/ACA motifs of the 26 identified H/ACA sRNAs were manually aligned, taking into account the following structural elements (Fig. 15.2B and E): a basal helical element H1 (No. 1), the 5′ and 3′ strands of the pseudouridylation pocket (Nos. 2 and 9, maximum lengths 11 and 12nts, respectively), an upper helical element H2 (No. 3), one K-turn or K-loop motif defined by a BGA element in the 5′ strand, a RUGA element in the 3′ strand and an apical loop (Nos. 5, 7, and 6, respectively), the ANA triplet (No. 11) and 3 ''buffer'' elements (Nos. 4, 8, and 10) that can be single or double stranded and are used for flexibility. Only three of the conserved structural elements have a sequence imposed by the presence of conserved residues in all the aligned RNA sequences of the profile, namely, the 5′ and 3′ elements of the K-turn/K-loop motif and the ANA box (Nos. 5, 7, and 11, respectively). The ''buffer'' elements numbered 4, 8, and 10 can be considered as gaps, which can be filled or not. The maximal size of elements 4, 8, and 10 were fixed to 3, 4, and 2nts, respectively.

The preceding description of the rationale for generation of the ERPIN profile will help the user of this approach to introduce new H/ACA sequences in the profile and will also facilitate the interpretation of the data obtained by use of this profile.

2.1.3. Practical procedure

The ERPIN searches based on the defined profile are performed by use of a local version of the program or the ERPIN web server version accessible at http://tagc.univ-mrs.fr/asterix/erpin/. Note that the current version of the ERPIN server does not allow the modifications of the profile and optional parameters. However, this server access provides a 2D structure representation of the selected putative H/ACA motif.

First, one needs to specify the genome sequences and the H/ACA profile to be used. To this end, the ''compulsory'' commands <profile> and <genome> allow the user to enter the names and locations of the ERPIN profile and sequence database, respectively. By default, the search is performed on both strands of the genome. Then, various options are available to finely tune the ERPIN search.

When activating the ''no mask'' option, the genomic screening will involve the simultaneous search of all the structural elements defined in the profile. We do not recommend this possibility; because of the great number of elements defined in the H/ACA profile, the computer search will be very slow. To increase the performance of the search (both speed and efficiency) screening for the presence of some selected structural elements can be done with a priority order. This order is defined by the use of successive masks. The first mask restricts the search to one or more selected element(s) (the unmasked elements, command ''umask''). The masked elements will influence the search by their delimited sizes. The sequences of the unmasked elements of the H/ACA motifs, which are aligned in the profile, are compared with the inspected genomic sequences. Then, the remaining selected masked elements are unmasked step by step (''-add'' option). The first step is fast, because only a few elements are searched through the entire genome. In the second step, the added elements are only searched within the sequence portions of the genome identified in the first step and so on in the following steps.

In this priority order strategy, a cutoff value can be specified to determine how stringent the search is at each of the steps defined by the successive masks (i.e., how similar the elements have to be in comparison with those included in the profile to be identified as hits). Different cutoff values can be defined for the different steps of the selection, depending on the relative importance of the elements specified in the mask. The definition of the masks, their order of use, and the cutoff values used at each step have an impact on the results.

To explain how these cutoff values are defined, we have to introduce the notion of lod-scores and scores. By comparison of one residue (single-stranded

element) or a pair of residues (double-stranded elements) in a given element of one aligned H/ACA motif in the profile with the corresponding residues, or pairs of residues, in the other aligned motifs, ERPIN can establish a score of similarity (lod-score) for individual residues or pairs of residues. On the basis of the sum of the lod-scores of all the residues or pairs of residues in a given element, ERPIN defines a score for this element. By extension, when a mask is used, a score can be established for the overall unmasked regions of each of the H/ACA motifs aligned in the profile. The cutoff value for the step of the search where this mask is used will be defined by reference to the score values established for all the motifs aligned in the profile.

The cutoff values can be defined as absolute values or as percentages. We will use percentages. They indicate the minimal percentage of sequences in the profile that are captured as hits on the basis of their calculated scores. When a 100% or higher percentage is used as cutoff (100% is the default cutoff value if no specification is given), all the sequences aligned in the profile will be captured as hits. Lower percentages indicate that not all the sequences in the profile are selected, thus, the search in the analyzed genome will be highly stringent. On the other hand, a higher cutoff value (>100%) indicates that sequences with some divergence relative to the aligned motifs of the profile can be selected. Note that the possibility of defining cutoff values >100% is available only for releases of ERPIN >5.3.

The two highly discriminating structural elements in the H/ACA motifs are the helical elements H1 and H2 and the K-turn/K-loop structure, respectively. Searches can be done by giving prevalence to one or the other of these two elements. In the following applications, we illustrate the relative efficiencies of two ordered series of three successive masks. In order 1, preference is given to the selection of the helical elements H1 and H2. Indeed, the presence of these two elements H1 (No. 1) and H2 (No. 3) is first tested, then the 5′ and 3′ strands of the K-turn/K-loop (Nos. 5 and 7) are searched, and finally, the presence of a putative ACA motif (No. 11) is investigated. When the second series of masks is used, designated as order 2, the first two steps are inverted, so that priority is given to the two strands of the K-turn/K-loop in the search. In these two procedures, the elements 2, 4, 6, 8, 9, and 10, namely the buffer elements, the apical loop and the two strands of the internal loop, are always masked.

2.1.4. Illustrative tests performed on archaeal genomes whose H/ACA sRNAs have been characterized

We will first present various tests run on genomes for which H/ACA sRNAs were characterized (Baker *et al.*, 2005; Charpentier *et al.*, 2005; Rozhdestvensky *et al.*, 2003; Tang *et al.*, 2002a; Thebault *et al.*, 2006). These tests illustrate the influence of the priority order and the cutoff values on the detection of H/ACA-like motifs. They also show how the ERPIN

parameters are optimized, taking into consideration the present state of the profile. In the blind tests presented in Table 15.1 (columns denoted blind tests), no prior knowledge on the H/ACA sRNAs identified in the analyzed species or in the genus, in the case of the *Pyrococcus* species, was included. To this end, the *A. fulgidus* genome was subjected to an ERPIN search by use of a profile in which the five identified *A. fulgidus* H/ACA motifs were eliminated. Similarly, the *M. jannaschii* genome was subjected to an ERPIN search by use of a profile in which the six identified *M. jannaschii* H/ACA motifs were eliminated. Finally, for the searches performed on each of the *Pyrococcus* genomes, the 30 motifs identified in *P. abyssi, P. furiosus*, and *P. horikoshii* were eliminated from the profile. In parallel, we evaluated the effect of the presence in the H/ACA profile of motifs belonging to three species of the same genus (*Pyrococcus*) on searches performed on the *A. fulgidus* and *M. jannaschii* genomes. To this end, we performed blind tests on the *A. fulgidus* and *M. jannaschii* genomes by use of an H/ACA profile containing motifs from only one of the *Pyrococcus* species (*P. abyssi*) (Table 15.1B). For each of these blind tests, the two priority orders (1 and 2) were applied, as well as three different cutoff values (95%, 100%, and 110%). The putative H/ACA motifs selected in the assay are next compared with the already identified H/ACA motifs. "Positive results" in Table 15.1 correspond to motifs already identified, whereas the other ones are denoted "likely false-positive results."

As illustrated in Table 15.1A, priority order 1 always allows the identification of the larger number of validated H/ACA motifs. The detection of 60–90% of the known H/ACA motifs with a cutoff value of 110% demonstrates the performance of the approach. The number of likely false-positive results increases with this cutoff value. However, as will be explained later, the inspection of the motifs and their location in the genome allows an efficient discrimination of likely false-positive results.

Some interesting observations can be made by inspection of Table 15.A and B: the presence of motifs from three species of the *Pyrococcus* genus in the profile did not bias the search. On the contrary, it has a marked positive effect on the selection of true H/ACA motifs in both *A. fulgidus* and *M. jannaschii*. The presence of all the H/ACA motifs from *Pyrococcus* is even required for an efficient selection of the H/ACA motifs from *A. fulgidus*.

For training in data interpretation and to demonstrate that secondary structure analysis of the candidate motifs helps discriminate the false-positive results, we provide the potential secondary structures of the candidate motifs selected for *M. jannaschii* in the blind tests presented in Table 15.1A (Fig. 15.3), and we comment on these results. As evidenced in Fig. 15.3 (panel B1), all of the four selected motifs corresponding to validated H/ACA motifs (denoted FW1, FW3, RC1, and RC3 in the search compilation) have an ACA triplet and a canonical K-turn structure, whereas the two likely false-positive sequences (FW2 and RC2 motifs, respectively) contain a CCA and an ACG motif instead of the ACA triplet (panel B2). Neither of them

Table 15.1A Tests of the effects of priority order, cutoff values and the use of a second step of selection in ERPIN searches

	Number of H/ACA sRNA motifs already identified	Cutoff	Blind Tests				Second Step Tests			
			Order (1)		Order (2)		Order (1)		Order (2)	
			Number of positive results	Number of likely false-positive results	Number of positive results	Number of likely false-positive results	Number of positive results	Number of likely false-positive results	Number of positive results	Number of likely false-positive results
A. fulgidus	5 [1]	95%	1	0	1	0	1	0	1	0
		100%	2	1	2	1	2	0	2	0
		110%	3	3	3	3	4	3	3	2
P. abyssi	10 [2, 3, 4]	95%	4	0	2	0	9	1	5	0
		100%	4	0	2	0	9	1	5	0
		110%	9	1	4	1	9	2	5	2
P. furiosus	10 [2, 3, 4]	95%	5	0	3	0	9	0	7	0
		100%	5	0	3	0	9	0	7	0
		110%	7	1	4	0	9	1	7	0
P. horikoshii	10 [2, 3, 4]	95%	3	0	2	0	9	0	6	0
		100%	3	0	2	0	9	0	6	0
		110%	6	2	4	0	9	1	6	0
M. jannaschii	6 [5]	95%	3	1	3	0	3	1	3	0
		100%	3	1	3	0	3	1	3	0
		110%	4	2	4	1	4	1	5	0

[1] Tang *et al.*, 2002.
[2] Rozhdestvensky *et al.*, 2003.
[3] Charpentier *et al.*, 2005.
[4] Baker *et al.*, 2005.
[5] Thébault *et al.*, 2006.

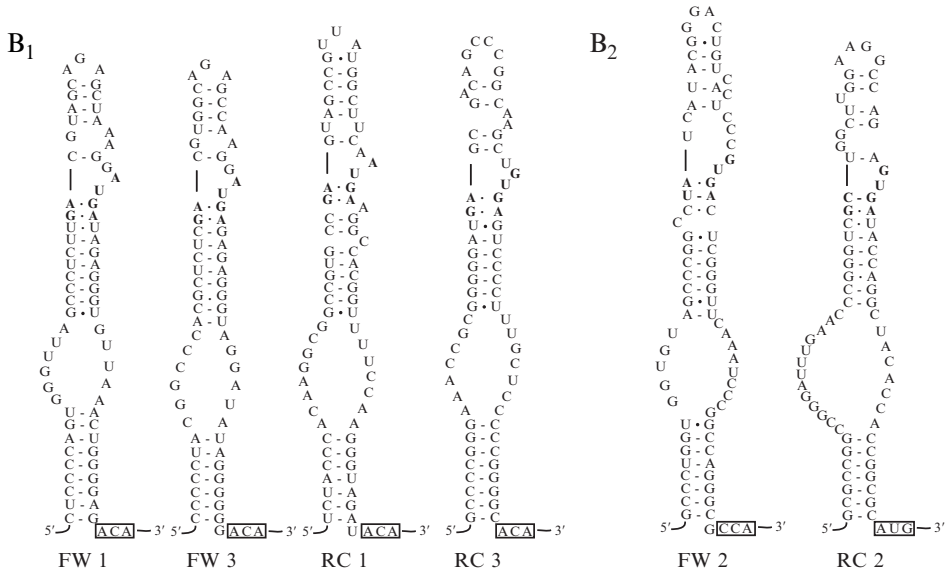
Table 15.1B Tests of the effects of priority order, cutoff values and the use of a second step of selection in ERPIN searches

			Blind Tests			
			Order (1)		Order (2)	
	Number of H/ACA sRNA motifs already identified	Cutoff	Number of positive results	Number of likely false-positive results	Number of positive results	Number of likely false-positive results
A. fulgidus	5 (1)	95%	1	0	1	0
		100%	1	0	1	0
		110%	1	1	1	1
M. jannaschii	6 (5)	95%	3	1	3	0
		100%	3	1	3	0
		110%	3	1	3	1

(A) Blind tests performed on each of the archaeal genomes whose H/ACA sRNAs have been studied (*A. fulgidus, P. abyssi, P. furiosus, P. horikoshii and M. jannaschii*). In the first step of these blind tests, for each of the studied genomes, the known H/ACA motifs of this species or of species of the same genus were removed for the H/ACA profile used for the ERPIN search. Searches were run with the two orders of priority and with three different cutoff values (95, 100 and 110%). In a given test, identical cutoff values were used for each of the masks.The name of the studied genome is given in the first column. The number of H/ACA motifs already identified in this species is indicated in the second column. the numbers of true H/ACA motifs (positives results) and likely false positive H/ACA motifs found in the ERPIN search are given for each assay. Results obtained with each of the priority orders are shown in parallel. In the second step of these blind tests, the true H/ACA motifs that were found in the first step were included in the H/ACA profiles. The tests use the same parameters as in the first step.
(B) Similar blind tests were performed on the *A. fulgidus* and *M. jannaschii* genomes with a profile lacking the *P.furiosus* and *P. horikoshii* H/ACA motifs.

A

```
>Methanocaldococcus jannaschii DSM2661 complete genome
FW 1 216280..216350 35.25 4.17e-06
CTCCCCAgtgggttag----CCCTCt--TGA.cgtagcagagctaaagg----ATGAta--GAGGGtgttaaac----TGGGGAG--ACA
FW 2 864064..864141 30.10 1.20e-04
GCCCTGGtggtgta------GCCCGgc-CTA.tcatacgggactgtcactcccGTGAct--CGGGTtcaaatcccgg-CCAGGGCg-CCA
FW 3 986084..986151 32.88 2.12e-05
CCCCCTAcggccca------CGCTCt--CGA.cgtggcagagccaagg-----ATGAga--GAGGGtaggata-----TAGGGGGg-ACA
>Methanocaldococcus jannaschii DSM2661 complete genome
RC 1 118060..118133 18.34 3.75e-02
CTACCCAcaaggcgg-----CCGTGc--CGA.gtagccgttatggcttca---ATGAaggcCACGGttttcca-----AGGGTAGatACA
RC 2 1150252..1150326 18.71 3.21e-02
GCGCCGGccgggatttgaacCCGGGt--CGC.tggcttggaaggccaga----GTGAta--CCAGGctacacca----CCGGCGC--ATG
RC 3 1659450..1659520 42.10 1.34e-08
GCCCCGGgaaaccgc-----GGGGGa--TGA.gcgacagcccggcaagct---GTGAgt--CCCCTttgctccc----CCGGGGC--ACA
```



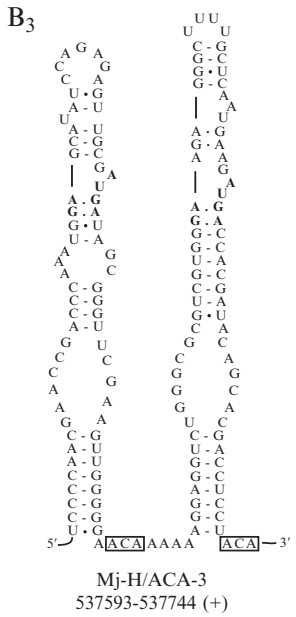


Figure 15.3 Results of an ERPIN search run on the *M. jannaschii* genomic sequence and the secondary structures proposed for the identified candidates. (A) The results of the ERPIN program, as displayed in output, are shown. The ERPIN program was run on the *M. jannaschii* genome sequence, using as a profile an alignment of the known H/ACA motifs, except for the ones identified in *M. jannaschii* (Thebault *et al.*, 2006) (see Table 15.1). The names of the candidates are defined by reference to the DNA strand screened and the order of appearance. ERPIN gives the positions relative to the forward

contain the canonical A•G and G•A sheared pairs tandem. In addition, they correspond to tRNA sequences, $tRNA^{Asp}$(GUC) and $tRNA^{Gly}$(UCC), respectively. Therefore, these likely false-positive H/ACA-like motifs can be unambiguously discarded from the ERPIN results. Note that most tRNA sequences can be folded into a stem-loop structure, explaining their detection with the profile that we used. The undetected motifs correspond to two motifs present in a unique sRNA (Mj-H/ACA-3, Thebault *et al.*, 2006). The 5′ motif has probably been discarded because of the internal loop present in helix 2. The second one probably differs by the sequences of helices 1 and 2 compared with the H/ACA motifs of the alignment.

Interestingly, when the four validated H/ACA motifs found in the blind test (FW1, FW3, RC1, and RC3) are included in the profile, the 3′-terminal motif of RNA Mj-H/ACA-3 is found in the search by use of the priority order 2 and a cutoff value of 110% (Table 15.1A). Hence, the use of both priority orders may be interesting in some specific cases.

More generally, a stepwise investigation including a second run after the inclusion in the profile of the validated H/ACA motifs selected in a first run, strongly improves the number of selected motifs and decreases the number of likely false-positive results (Table 15.1A). When the priority order 1 is used, only one of the previously identified H/ACA motifs is not found in the three *Pyrococcus* species, the one proposed to guide U to Ψ conversion at position 2575 in the *P. furiosus* LSU rRNA (see Pf3, stem I, Rozhdestvensky *et al.*, 2003).

2.1.5. Example of the search of H/ACA-like motifs in a yet unexplored genome

The genome from *Thermococcus kodakarensis*, a species belonging to the same order as species of the *Pyrococcus* genus, was used in this teaching example. The profile used contained the 41 known archaeal H/ACA motifs. The search was first performed with priority order 1, and the cutoff value defined for each step of the selection was 110%. Hence, the following ERPIN command was used:

erpin<profile><genome>−1,11−umask 1 3−add 5 7−add 11−cutoff 110%110%110%

"1,11" indicates the beginning and the end of the chain of structural elements considered in the profile. In this case, all the elements of the profile are considered.

strand for each selected H/ACA-like motif. (B) A secondary structure is proposed for each of the candidates. The motifs in Panel B1 were retained after inspection of their structures, whereas those in panel B2 were discarded. Panel B3 represents the proposed secondary structures of the two known H/ACA motifs that were not detected in the first step of the blind search.

The search time was short: less than 3 min of CPU time on a 3.2-Ghz P IV. Sixteen candidate motifs were identified. Six of them were encoded by the plus strand and 10 by the minus strand. At the last step of the selection, ERPIN computes an e-value reflecting the statistical significance of the selected H/ACA-like motifs. This e-value represents the probability of encountering this motif at random (Lambert *et al.*, 2005). Hence, it depends on the scores calculated for all the structural elements and on the size of the database. The output file displays the candidate sequences that satisfy the cutoff values defined at each selection step and the e-values of these sequences (Fig. 15.4). Only candidates with a negative e-value deserve further analysis.

In the search performed on the *Thermococcus kodakarensis* genome, nine candidates had negative e-values between 1e-06 and 4e-02, which was an indication for a good fit with the H/ACA motifs in the profile. In contrast, seven candidates had e-values between 1e-01 and 4, suggesting that they may correspond to false-positive results. The possible secondary structure of each candidate is shown in Fig. 15.4 (panels B1 and B2). The nine candidates with satisfying e-values displayed all the expected structural elements (panel B1). Only the FW3 candidate has an AUA motif instead of the ACA motif. Despite its high e-value, the FW1 motif has a canonical K-loop, an ACA motif, and helices 1 and 2 with correct lengths. The high e-value observed may be due to the presence of two G•U pairs in helix 1. Furthermore, motifs FW1, FW2 and FW3 may belong to the same H/ACA sRNA,

A

```
>Thermococcus-kodakarensis
FW   1  462059..462114   21.90  7.56e-01
TGCCCCTgcgcgagga----CCCCG---GGGaga----------------------ATGAac--CGGGGggcgatgc----GGGGGCA--ACA
FW   2  462115..462195   33.44  1.32e-04
GCCCGGCctcagcgaggtccCCGCG---GGAgggccttccgcgtcccggagcacg--ATGAc---CGCGGgaaacccag---GCCGGGC--ACA
FW   3  462201..462272   27.43  3.49e-02
GGCCCGCcagggttag----CCCGCc--CAAgggtggcgtacgccttcg--------ATGAgg--GCGGGagttaccg----GCGGGCC--ATA
FW   4  899083..899151   32.54  3.74e-04
GCCCCCGggaaaccgc----GGGGGa--TGAgcgcctcggcgagcc-----------GTGAtt--CCCCTtcgctccc----CGGGGGC--ACA
FW   5 1545183..1545239  35.55  7.44e-06
GCCCTCCcctctcacac---CCCCG---GGAgcg----------------------GTGAc---CGGGGggcggtcgg---GGAGGGC--ACA
FW   6 1769611..1769691  19.95  1.64e+00
CGCCCGGgtggtgta-----GCCCGgc-CCAtcatacgggactgtcactccc-----GTGAcc--CGGGTtcaaatcccggcCCGGGCGccATA
>Thermococcus-kodakarensis
RC   1   47849..47907    35.19  1.28e-05
GGCCTGGcgtcccgccct--CCCCGg--GGAaac----------------------GTGAac--CGGGGcttcctg-----CCAGGCCt-ACA
RC   2  465056..465126   34.46  3.57e-05
GGGCCCGgtctccgga----CCGCT---GGAgggggtctctggccgtca--------ATGAg---AGCGGggggatcagc---CGGGCCC--ACA
RC   3  465127..465197   33.20  1.77e-04
CCGCCCGgtggcccgt----GTCCC---GGAgggggccgaggcccaca---------GTGAa---GGGATgagacgatgc--CGGGCGG--ACA
RC   4  865615..865690   19.53  1.91e+00
GGGAGCAtcatgac------TCCGT---CAActccagccctgacgaccttcctaatgATGAtcc-TCGGGttaacg------TGCTCCC--AGA
RC   5  986951..987023   19.50  1.93e+00
AGCCACGctggagttcatg-CCCAGctcGGAaaagagttc-----------------GTGAagacCGGGGaaatgcccg---CGAGGCTttACA
RC   6 1107274..1107331  34.45  3.64e-05
GGGCCGGgagcatccac---CCGCG---GGAgca----------------------GTGAc---CGCGGgcctctgtac--CCGGCCC--ACA
RC   7 1107332..1107400  30.26  3.64e-03
GGGCTCGgtacaaccgc---CTCCG---CAAggtatcgggttcc-------------GTGAg---CGGAGcgtgctcacgc-CGAGCCC--ACA
RC   8 1371594..1371659  18.26  2.94e+00
GTCCTCTacctgaaatg---GCGCT---CAAggtgtccattc---------------GTGAg---AAAGTtagaagcct---GGAGGATg-AAA
RC   9 1404046..1404129  21.71  8.22e-01
CGGCGCCtattcctgcgc--TGGCGa--TGAtaccctgaggtatagctgagagaccgGTGAggc-CAACGgtaagacc----GGCGCCGa-ATA
RC  10 1460837..1460911  17.72  3.52e+00
TCGAGGGtctgaaagt----CCAGAa--TGAaacggaagacagt-------------ATGAcagtCTTGGttaagtttaaaaCCCTCGGcgAGA
```

Figure 15.4 (*continued*)

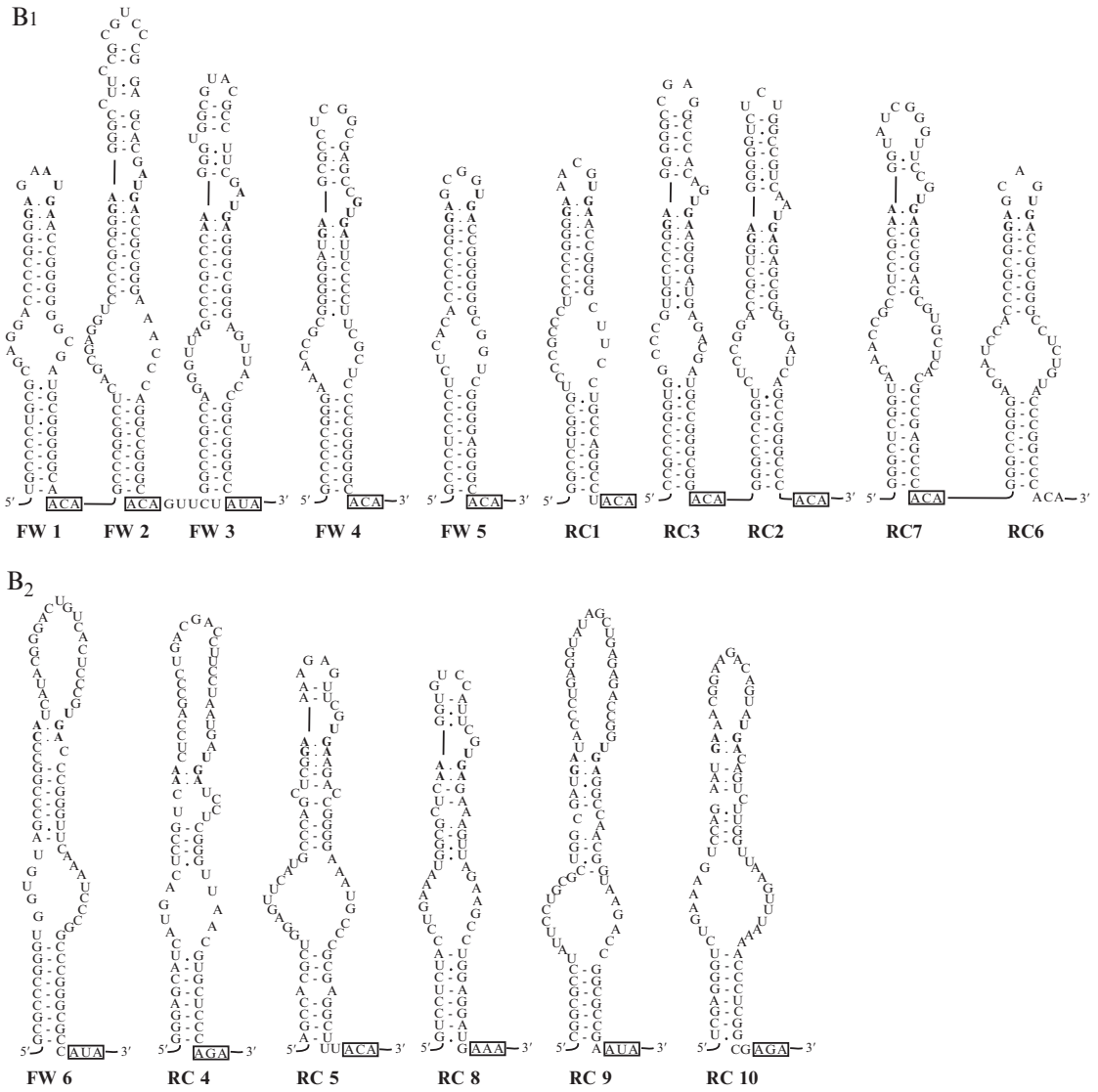


Figure 15.4 Results of an ERPIN search run on a yet-unstudied genome, *Thermococcus kodakarensis*, and proposed secondary structures of the identified candidates. Same legend as in Fig. 15.3. Panel (A) also indicates the absolute score and the e-value of the candidate. Panel (B1) displays the candidates retained after inspection of their structures, and panel (B2) displays the discarded candidates.

because their coding sequences are adjacent in the genome. Two pairs of motifs with a negative e-value are also encoded by adjacent sequences (RC3 and RC2 on the one hand, and RC7 and RC6 on the other hand). Therefore, they likely belong to two H/ACA sRNAs, each containing two H/ACA motifs. In addition, the coding sequences of the 10 motifs represented in panel B1 are all located between ORFs.

In contrast, the six candidate motifs with an e-value between 1e-01 and 4 (panel B2) can be discarded for several reasons: (1) except for motif RC5, they do not contain an ACA triplet; (2) most of them contain bulge residues or an internal loop in helix 1 or in helix 2; (3) the large terminal loops of motifs FW6, RC4, RC9, and RC10 are not structured. Finally, the

candidate motif FW 6 corresponds to a $tRNA^{Asp}(GUC)$, and the coding sequences of the five other motifs are located within ORFs. On the basis of our present knowledge, sRNA genes may partially overlap ORFs but are not included in ORFs. A strong argument for the validation of the 10 H/ACA motifs in *T. kodakarensis* (Fig. 15.4, panel B1) is their homology with the 10 H/ACA sRNAs found in the *Pyrococcus* species. Hence, we predict that the 10 motifs detected in *T. kodakarensis* are the homologs of the 10 experimentally demonstrated *Pyrococcus* H/ACA motifs.

Interestingly, no new putative motifs with a satisfying negative e-value were found when a second screening was performed after insertion of the 10 validated motifs in the profile by use of the same parameters as in the first screening step (order 1 and cutoff values of 110%). Only the number of motifs with a high e-value increased, suggesting that all of the true H/ACA motifs were detected in the first search. Therefore, the 41 known H/ACA motifs aligned in the profile seem to provide enough information for an exhaustive selection of the *T. kodakarensis* motifs in a single run.

We next tested whether introduction of the probable *T. kodakarensis* H/ACA motifs in the profile would increase the selection of the *Pyrococcus* species known H/ACA motifs in blind tests (Table 15.2). This is, indeed, the case. When a blind ERPIN search is performed on the *P. abyssi* genome with a profile including the H/ACA motifs from *A. fulgidus, M. jannaschii, P. furiosus*, and *P. horikoshii*, nine of the known H/ACA motifs are detected with the priority order 1 and a 95% cutoff value (Table 15.2). No false-positive result is found with a 95% cutoff value. When the *T. kodakarensis* motifs are also added in the alignment, the 10 *P. abyssi* motifs are found when a cutoff value of 95% is used. However, the number of false-positive results increases in this case (Table 15.2).

2.1.6. Recommendations

On the basis of the data presented in Tables 15.1 and 15.2, the cutoff values used should be defined, taking into account the number and origin of the H/ACA motifs aligned in the profile. If they belong to species that are phylogenetically closely related to the studied species, small cutoff values can be used. They are expected to allow a selection of the true motifs with a minimum of noise. When a phylogenetically distant species is studied, we recommend the use of a cutoff value of 110% for each step to select true motifs that may exhibit divergences relative to the canonical structure. Hence, the closer the species, the smaller the cutoff values can be.

As in most cases, use of priority order 1 gives better results; we recommend the use of this order and only exceptionally, when needed, order 2.

As illustrated by the blind tests in Table 15.1, to complete the identification of H/ACA motifs in a new species, it may be worthwhile performing a second screening, after inclusion in the profile of the H/ACA-like motifs that were identified in the first step and contain all the needed characteristic structural features.

Table 15.2 Tests of the effect of the integration of the *T. kodakarensis* H/ACA motifs identified in this work, on blind tests performed on the *P. abyssi* genome

			Blind Tests performed with the profile A[a]				Bind Tests performed with the profile B[a]			
			Order (1)		Order (2)		Order (1)		Order (2)	
	Number of H/ACA sRNA motifs already identified	Cutoff	Number of positive results	Number of likely false-positive results	Number of positive results	Number of likely false-positive results	Number of positive results	Number of likely false-positive results	Number of positive results	Number of likely false-positive results
P. abyssi	10 [2,3,4]	95%	9	0	7	0	10	1	7	1
		100%	10	1	8	0	10	2	8	2
		110%	10	1	8	1	10	4	8	5

[a] The profile A contains the H/ACA motifs of *A. fulgidus, M. jannaschii, P. horikoshii* and *P. furiosus*. The profile B corresponds to the profile A with, in addition, the *T. kodakarensis* H/ACA motifs.

The H/ACA motifs detected in *T. kodakarensis* are included in the H/ACA profile available at http://tagc.univ-mrs.fr/asterix/erpin/. We are currently analyzing the H/ACA motifs of all completely sequenced archaeal genomes, and an updated H/ACA profile including not experimentally verified motifs will soon be available at the same URL.

Concerning the interpretation of the data, we recommend the selection of H/ACA-like motifs with negative e-values for the subsequent analysis (screening for the target sequence). Motifs with e-values between 1 and 1e-02 can be maintained if they contain almost all the needed structural elements; a default in one of the structural elements may be accepted at this stage of the selection, because screening for target sequences will also help in discriminating true candidates.

Note that one strong discriminatory feature of ERPIN results is the overlap of candidate coding sequences with sequences coding for tRNAs, rRNAs, or proteins. Note also that, as observed in *T. kodakarensis*, when the apical loop closing the K-turn motif is long enough, an important criterion for the validation of the H/ACA motif is the possibility of forming a stem-loop structure in this apical region.

Finally, the newly identified sRNAs have to be experimentally confirmed with the approaches described in the third section of this chapter.

2.2. Search for targets of the H/ACA-like motifs

The identification of a putative RNA target for a given H/ACA motif is initially based on the search for a dinucleotide containing a uridine residue at the 5′ position. This dinucleotide should be flanked by two sequences complementary to the two strands of the expected pseudouridylation pocket. To this end, one has to first establish the most likely secondary structure of the H/ACA motif, and we will comment in the following on the difficulties that may be encountered at this step. Once the 2D structure is established, a series of RNA descriptors, describing the various possible base-paired structures that can be formed by the target RNA and the two guiding sequences, have to be settled (Fig. 15.5). The different steps in target identification are explained in the following.

2.2.1. Requirements

2.2.1.1. Hardware and software The RNAMOT program (Gautheret *et al.*, 1993) version 2.1 can be run on most UNIX platforms. The source and a tutorial are available at http://pages-perso.esil.univ-mrs.fr/~dgaut/download/. Any equivalent simple descriptor-based software such as RNAMotif (Macke *et al.*, 2001) can also be used. We selected the RNAMOT program because it is simple and easy to use.

2.2.1.2. Sequence data The sequence format used for RNAMOT is FASTA. The ribosomal RNA sequences of each of the studied archaeal species were extracted from their genomic sequence subsets, including all the noncoding RNAs (file with extension .frn obtained at ftp://ftp.ncbi.nih.gov/genomes/Bacteria).

2.2.2. Procedure to build the RNAMOT descriptors and run the program

2.2.2.1. Construction of the descriptors On the basis of our present knowledge of the archaeal H/ACA system, there is no requirement concerning the identity of the residue at the 3′ position in the single-stranded dinucleotide (Fig. 15.5). Altogether, the two base-paired regions most generally contain at least nine base pairs, and one wobble pair can be included in these nine pairs.

In Fig. 15.5, we illustrate the various descriptors that have to be built to look for the possible targets of the H/ACA motif Pab91 of *P. abyssi*. A roughly equivalent number of base pairs in the two helical regions is expected to be the most frequent situation (4 and 5 base pairs or 5 and 4 base pairs). However, one cannot exclude the possibility of having highly asymmetrical base-pair interactions (1 + 8, 2 + 7, 3 + 6, 6 + 3, 7 + 2, and 8 + 1).

We will explain how one of the descriptors needed to look for one of the base-paired configurations is built; all of the other descriptors are built in the same way. The descriptor used as an example in Fig. 15.5 (panel C) is devoted to the search of a target sequence forming base-pair interactions including 4 and 5 residues of the 3′ and 5′ strands of the pseudouridylation pocket, respectively (descriptor designated as 4 + 5). The descriptor file includes two lines: the first one identifies the searched sequence, s1 in this example. The second line defines the sequence requirements.

```
s1
s1 11 : 11 YUYYUNRGGYU
```

The two numbers separated by a colon correspond to the minimum and maximum nucleotide lengths of the searched target sequence; Y and R correspond to pyrimidine and purine residues, respectively.

If no positive result is obtained when the first series of descriptors corresponding to a total of nine base pairs is used, it is recommended to look for possible targets able to form only eight or seven base pairs. The use of this low number of base pairs may allow the selection of interactions that include mismatches. Indeed, such interactions cannot be selected directly by the RNAMOT descriptors described previously.

In targets that include mismatches, an extension of the possible base-pair interactions should be tested by inspection of the guide and target

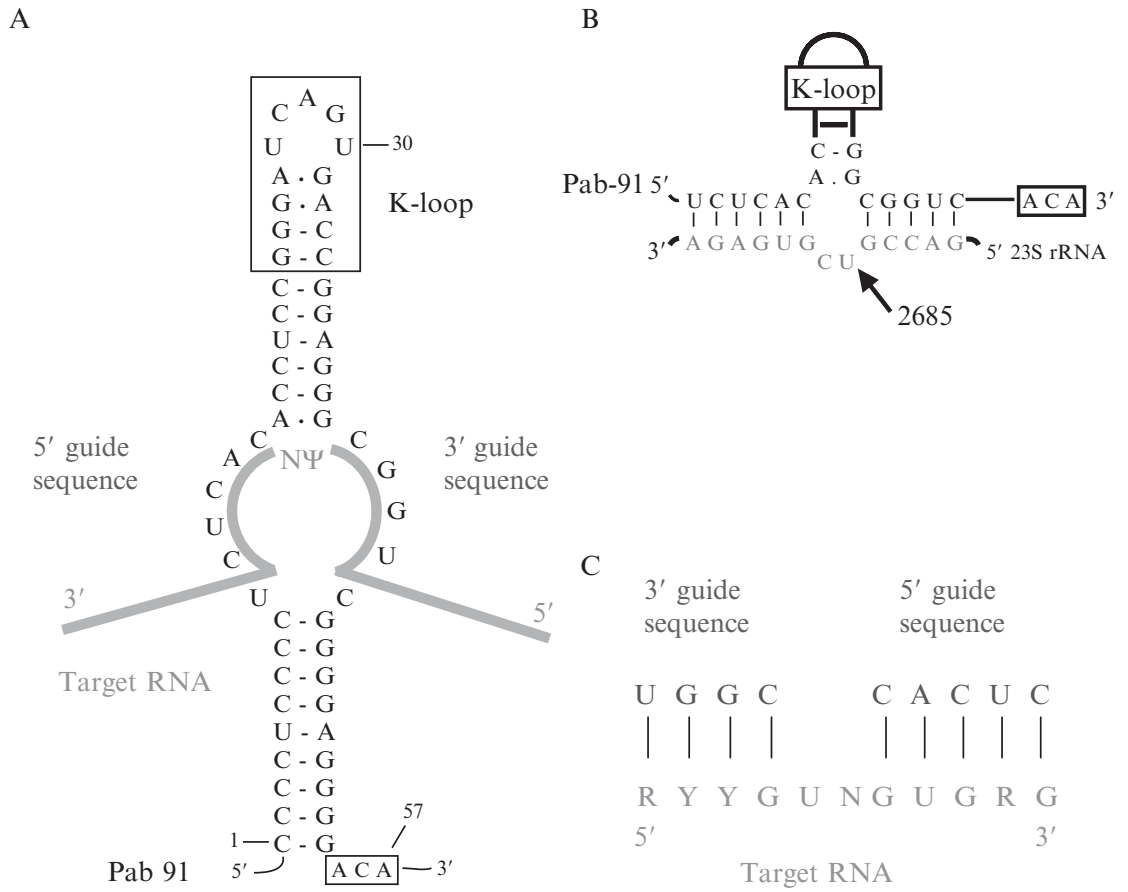


Figure 15.5 The strategy used for the identification of the targets of the H/ACA-like motifs. (A) The secondary structure of the *P. abyssi* Pab91 sRNA is shown with a schematic representation of the target RNA base pair interactions formed with the 5′ and 3′ guide sequences of the H/ACA motif (Charpentier *et al.*, 2005). The ΨN dinucleotide obtained after U to Ψ conversion is shown. (B) The known interaction formed between the *P. abyssi* Pab91 sRNA and its known target sequence in *P. abyssi* 23S rRNA. The targeted U2685 residue is indicated (Charpentier *et al.*, 2005). (C) As an example, we show the structure that would be used to define a 5 + 4 descriptor for the selection of target sequences of RNA Pab91. Once the 5′ and 3′ guide sequences are specified, the possible complementary sequences are written, taking into account possible wobble G·U base-pair interactions (Y = pyrimidines, R = purines) and the chosen numbers of base pairs to be formed with the 5′ and the 3′ guide sequence, respectively. The rRNA sequences are then screened with a descriptor of these possible complementary sequences by use of RNAMOT.

sequences. Even when a putative interaction including nine base pairs is found, we recommend looking for possible extensions of the base-pair interactions by inspection of the flanking sequences.

2.2.2.2. How to start the search? The program RNAMOT is started using the following command line:

rnamot -s -s <rRNA sequence > -d<descriptor > -o<output > -t

The name of the executable is ''rnamot.'' The first ''-s'' specifies the mode of use (-s for search, by opposition to -a for alignment); the second ''-s'' is used to provide the name and location of the file containing the rRNA sequences that will be screened, and ''-d'' the name and location of the descriptor file. Finally, the ''-o'' option specifies the file in which the results are written. The ''-t'' option activates the display of the search stage and the number of positive results obtained. A search has to be executed for each of the descriptors. However, whatever descriptor is used, the search time does not exceed a few seconds.

2.2.3. Example of application: search for the target sequences of the *M. jannaschii* FW1 motif

In the *M. jannaschii* motif FW1, helices 1 and 2 contain nine and six successive Watson–Crick base pairs, respectively (Fig. 15.6, panel A). In addition, one wobble G·U pair can be present at the extremity of helix 2. The formation or the absence of this wobble pair modifies the pseudouridylation pocket and, therefore, the putative target sequences. Hence, two series of descriptors were built as described earlier. In the first series, the G·U pair at the extremity of helix 2 was considered to be formed. In the second series, it was considered to be opened. No putative target sequence was detected for this second series of descriptors. In contrast, two putative target sequences were detected with the first series, suggesting the formation of the G·U pair in the guide sRNA. One of the target sequences (Fig. 15.6, panel B, hit sequence 1) was obtained by use of a descriptor designed for the search of sequences forming three and six base pairs with the 3′ and 5′ guiding sequences, respectively (3 + 6 descriptor). The second target sequence (panel B, hit sequence 2) was selected with a 4 + 5 descriptor. Hit sequence 1 belongs to 16S rRNA, and residue U1261 would be the target. Hit sequence 2 belongs to 23S rRNA, and residue U2015 would be the target.

Inspection of both rRNA flanking sequences reveals the possible extension by 2 (G-C and G·U) and 1 (G-C) base pairs of the interactions that can be formed with the 5′ guide sequences for hit sequences 1 and 2, respectively. Therefore, the base-pair interaction formed for hit sequence 1 involves two stretches of three and eight base pairs. However, each of them contains two wobble pairs. Thus far, such a high number of G·U pairs has never been described for the previously studied H/ACA sRNAs and snoRNAs. Therefore, hit sequence 1 may correspond to a false-positive result. In contrast, hit sequence 2 forms two stretches of four and six uninterrupted Watson–Crick base pairs. A strong argument in favor of the validity of hit sequence 2 is the observation of H/ACA motifs proposed to target similar positions in the 23S rRNAs of *P. furiosus* (H/ACA motif Pf7) and *A. fulgidus* (H/ACA motif Afu4) (Rozhdestvensky *et al.*, 2003) (Fig. 15.6, panels C and D).

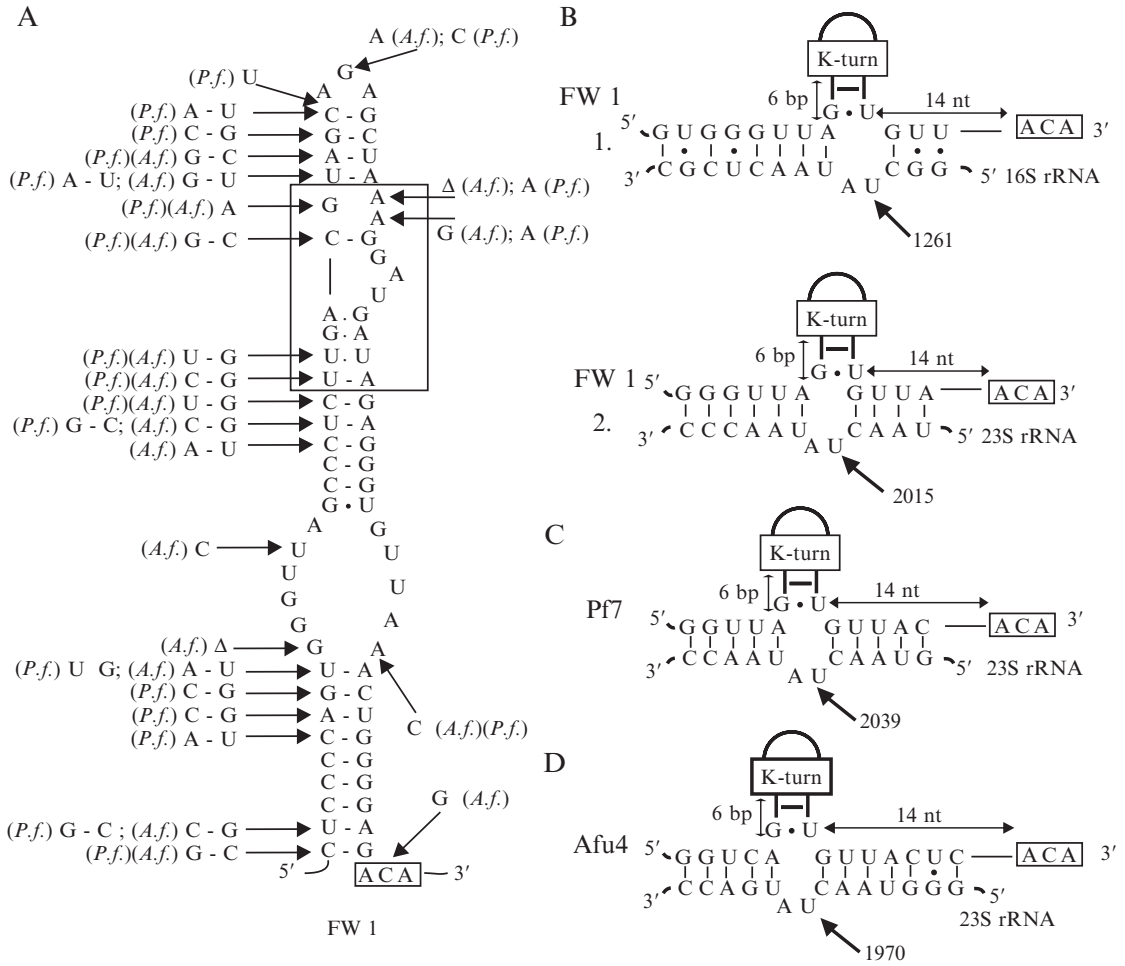


Figure 15.6 Target sequences found for the *M. jannaschii* FW1 motif by an RNAMOT search. (A) Secondary structure proposed for the *M. jannaschii* FW 1 candidate. The compensatory mutations and single base substitutions, insertions, or deletions found in the *A. fulgidus* Afu4 (*A.f.*) and *P. furiosus* Pf7 (*P.f.*) motifs compared with the *M. jannaschii* FW1 motif are indicated. Δ indicates one missing nucleotide. (B) The possible base-pair interactions between the *M. jannaschii* FW1 motif and the two putative targeted sequences identified for this motif by the RNAMOT search (hit sequences 1 and 2) are shown. The distances between the basal extremity of helix 2 and either the K-loop or the ACA box are indicated. The target sequences proposed for the third H/ACA motif of the *P. furiosus* Pf7 sRNA (C) and the third H/ACA motif of the *A. fulgidus* Afu4 sRNA (Rozhdestvensky *et al.*, 2003) (D), and the interactions that they can form with the guide RNA are represented.

2.2.4. Application to the search of target sequences for the 10 H/ACA-like motifs found in *T. kodakarensis*

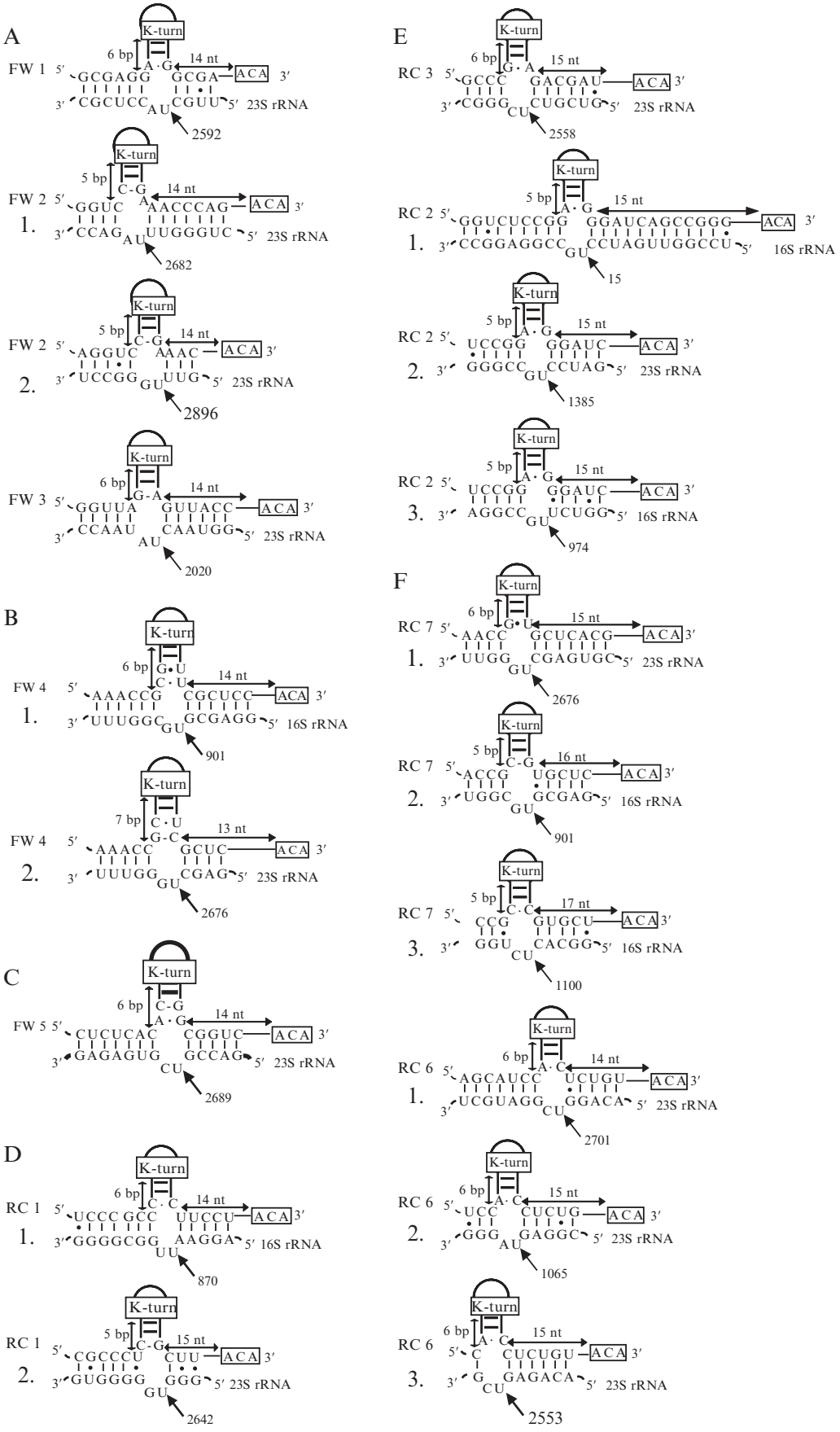
As described previously, we screened the *T. kodakarensis* genome for H/ACA-like motifs. We will describe how we looked for their possible target sequences with RNAMOT. For each of the 16 hit H/ACA motifs obtained by the ERPIN search, different series of descriptors were built to take into consideration the slight possible length variations of helix 2. For example, in the FW1 motif, a noncanonical G•A pair may be included at the basal extremity of helix 2 or

may be opened so that its two nucleotides can be included in the guide sequences (Fig. 15.7, panel A). At least two alternative lengths of helix 2 are possible for almost all the selected motifs of *T. kodakarensis* (Fig. 15.7).

When the various series of descriptors built for each of the motifs are used, no putative target sequence was found for the six motifs that we discarded after inspection of their primary and secondary structures. In contrast, up to 3 putative target sequences were found for each of the 10 motifs that we retained after this inspection. In Fig. 15.7, we present the corresponding base-pair interactions by order of stability for each of the motifs. The order of presentation also takes into account the fact that some of the H/ACA motifs are expected to belong to the same sRNA (FW1, FW2, and FW3; RC2 and RC3; RC6 and RC7).

Most generally, when two or three possible target sequences were detected for a given motif, they differ by their stabilities and/or by the distance between helix 2 and the ACA trinucleotide. For instance, among the two hit sequences found for each of the FW2 and RC1 motifs, the hit sequences numbered 2 form shorter base-pair interactions with the guide sequences compared with the hit sequences numbered 1. In cases in which three putative target sequences were found (motifs RC2, RC6 and RC7), most frequently, one of them forms significantly more stable interactions than the two other ones. Therefore, in each case, we prefer the hit sequence forming the more stable interaction (hits numbered 1). However, peculiar results were obtained for the RC2 motif: both hits 1 and 2 form stable interactions (9 Watson–Crick base pairs and a G·U wobble pair for hit sequence 2 and 19 Watson–Crick pairs and two G·U pairs for hit sequence 1). By use of hit sequence 2, the K-turn structure and the ACA triplet are located at correct distances from the basal extremity of helix 2 (Fig. 15.7E). Thus, despite the lower stability of the interaction formed by hit sequence 2 compared with hit sequence 1, we have no criterion to eliminate this hit sequence. Furthermore, it should be noted that the interactions that can be formed with hit sequence 1 are unusually long. Formation of these interactions would result in the disruption of a large part of helix 1. Because helix 1 is expected to be important for the folding and the activity of the H/ACA sRNP complex, the possible interaction between the hit sequence 1 and motif RC2 may have another function than RNA pseudouridylation. Interestingly, this hit sequence is located 3 nts downstream from the 16S rRNA 5′ extremity. We may imagine an RNA chaperone role of motif RC2, in addition to its possible role in pseudouridylation.

Note that no Watson–Crick base pair in helix 1 is opened in any of the interactions formed with the other H/ACA motifs and hit sequences. Only the G·U pair at the extremity of helix 1 in motif RC1 may be formed or opened, depending on whether a G·U pair is included or not in the interaction established between hit sequence 1 and the 5′ strand of the guide sequence (Fig. 15.7D).


A
FW 1
K-turn
6 bp
A·G
14 nt
ACA
5′ GCGAGG GCGA 3′
3′ CGCUCC CGUU 5′ 23S rRNA
AU
2592
FW 2
1.
K-turn
5 bp
C-G
14 nt
ACA
5′ GGUC AACCCAG 3′
3′ CCAG UUGGGUC 5′ 23S rRNA
AU
2682
FW 2
2.
K-turn
5 bp
C-G
14 nt
ACA
5′ AGGUC AAAC 3′
3′ UCCGG UUG 5′ 23S rRNA
GU
2896
FW 3
K-turn
6 bp
G-A
14 nt
ACA
5′ GGUUA GUUACC 3′
3′ CCAAU CAAUGG 5′ 23S rRNA
AU
2020
B
FW 4
1.
K-turn
6 bp
G•U
C·U
14 nt
ACA
5′ AAACCG CGCUCC 3′
3′ UUUGGC GCGAGG 5′ 16S rRNA
GU
901
FW 4
2.
K-turn
7 bp
C·U
G-C
13 nt
ACA
5′ AAACC GCUC 3′
3′ UUUGG CGAG 5′ 23S rRNA
GU
2676
C
FW 5
K-turn
6 bp
C-G
A·G
14 nt
ACA
5′ CUCUCAC CGGUC 3′
3′ GAGAGUG GCCAG 5′ 23S rRNA
CU
2689
D
RC 1
1.
K-turn
6 bp
C·C
14 nt
ACA
5′ UCCC GCC UUCCU 3′
3′ GGGGCGG AAGGA 5′ 16S rRNA
UU
870
RC 1
2.
K-turn
5 bp
C-G
15 nt
ACA
5′ CGCCCU CUU 3′
3′ GUGGGG GGG 5′ 23S rRNA
GU
2642
E
RC 3
K-turn
6 bp
G·A
15 nt
ACA
5′ GCCC GACGAU 3′
3′ CGGG CUGCUG 5′ 23S rRNA
CU
2558
RC 2
1.
K-turn
5 bp
A·G
15 nt
ACA
5′ GGUCUCCGG GGAUCAGCCGGG 3′
3′ CCGGAGGCC CCUAGUUGGCCU 5′ 16S rRNA
GU
15
RC 2
2.
K-turn
5 bp
A·G
15 nt
ACA
5′ UCCGG GGAUC 3′
3′ GGGCC CCUAG 5′ 23S rRNA
GU
1385
RC 2
3.
K-turn
5 bp
A·G
15 nt
ACA
5′ UCCGG GGAUC 3′
3′ AGGCC UCUGG 5′ 16S rRNA
GU
974
F
RC 7
1.
K-turn
6 bp
G•U
15 nt
ACA
5′ AACC GCUCACG 3′
3′ UUGG CGAGUGC 5′ 23S rRNA
GU
2676
RC 7
2.
K-turn
5 bp
C-G
16 nt
ACA
5′ ACCG UGCUC 3′
3′ UGGC GCGAG 5′ 16S rRNA
GU
901
RC 7
3.
K-turn
5 bp
C·C
17 nt
ACA
5′ CCG GUGCU 3′
3′ GGU CACGG 5′ 16S rRNA
CU
1100
RC 6
1.
K-turn
6 bp
A·C
14 nt
ACA
5′ AGCAUCC UCUGU 3′
3′ UCGUAGG GGACA 5′ 23S rRNA
CU
2701
RC 6
2.
K-turn
6 bp
A·C
15 nt
ACA
5′ UCC CUCUG 3′
3′ GGG GAGGC 5′ 23S rRNA
AU
1065
RC 6
3.
K-turn
6 bp
A·C
15 nt
ACA
5′ C CUCUGU 3′
3′ G GAGACA 5′ 23S rRNA
CU
2553

On the basis of the data presented in Fig. 15.7, the selected hit sequences can form from 4 up to 7 base pairs with the 5′ and 3′ guide sequences. Interestingly, for most of the more favorable hit sequences, a noncanonical pair is present at the basal extremity of helix 2, which may be required to ensure enough flexibility in the cruciform structure.

2.2.5. Tips for searching and analyzing H/ACA targets

As evidenced by the preceding examples, a major difficulty in the prediction of target sequences comes from the necessity of establishing the correct secondary structure of the putative H/ACA-like motif and to define the basal extremity of helix 2. Because of the irregularity and size variations of helix 2, several possible conformations can sometimes be proposed for a given motif. In contrast, helix 1, which consists in a regular stretch of Watson–Crick pairs, is generally easier to predict.

In addition, even when the correct secondary structure has been established, the size variability of the base-pair interactions formed with the 5′ and 3′ guide sequences increases the number of RNAMOT searches that have to be run for a given motif.

As illustrated by the *T. kodakarensis* example given previously, we recommend the initial use of a series of descriptors corresponding to interactions including 9 base pairs. A reduction of this number of base pairs in the first round of the selection would unnecessarily increase the number of false-positive results and limit the chance of obtaining positive results. Hits with a lower level of complementarity can be searched in a second step, if no hit is obtained with the 9 base pair descriptors. Restriction of the starting descriptors to interactions only including Watson–Crick pairs would avoid the selection of interactions containing one wobble G•U pair such as the one found for the *T. kodakarensis* motif FW1.

For interpretation of the results note that: (1) a non-Watson–Crick base pair is often present at the basal extremity of helix 2, (2) the distances between the basal extremity of helix 2 and the K-turn/K-loop sequence on the one hand, and the ACA triplet on the other hand, are most often 5–6 bps and 14–15nts, respectively, (3) the base-pair interactions with both guide sequences include most frequently 4–7 residues; however, more asymmetrical base-pair interactions cannot be excluded.

Figure 15.7 Putative target sequences identified for the 10 H/ACA motifs of *T. kodakarensis* by an RNAMOT search. Panels A–F display the base-pair interactions that can be formed between the putative target sequences identified by use of RNAMOT and each of the 10 H/ACA motifs identified for *T. kodakarensis*. Each panel corresponds to a given H/ACA sRNA. For each motif the possible base-pair interactions are represented according to their stability.

As illustrated by the *T. kodakarensis* example used in this chapter, the search of putative target sequences in rRNAs by RNAMOT can confirm the discrimination of true H/ACA motifs made by inspection of the H/ACA-like motifs obtained in the ERPIN search. However, if an H/ACA motif bears all the characteristic features of archaeal H/ACA motifs and if no putative target sequence is found in rRNAs, its target sequence may be contained in another kind of RNA. In this case, the search for possible target sequences in other RNAs will be useful.

2.3. Phylogenetic and experimental validation of the results

2.3.1. Phylogenetic validation

When new H/ACA motifs and their putative targets have been selected by the proposed approach, several comparisons with known data on archaeal H/ACA motifs may be helpful for their validation. This was illustrated in the interpretation of the RNAMOT hit sequences found for the FW1 motif of *M. jannaschii*, where three homologous H/ACA motifs from *P. furiosus, A. fulgidus*, and *M. jannaschii*, that differ by a limited number of base substitutions, insertions or deletions (Fig. 15.6), guide modifications at similar positions in 23S rRNA. Therefore, as a first step in the validation of the data, we recommend the comparison of the putative H/ACA motifs obtained for a new archaeal genome with the already identified H/ACA motifs. However, note that phylogenetic comparisons are limited by the rapid evolution of sRNAs.

Although pseudouridylation sites in archaeal rRNAs have only been identified in a limited number of species and are not strongly conserved from one species to the other (Del Campo *et al.*, 2005; Ofengand, 2002), a comparison of the putative target sites detected by RNAMOT with known pseudouridylation sites in archaeal rRNAs may also be useful, especially when Ψ residues have been experimentally identified in rRNAs from phylogenetically related species.

We will again use the *M. jannaschii* FW1 motif to illustrate this point. One of the two putative target sites found by the RNAMOT (U2015 in 23S rRNA) was experimentally found to be pseudouridylated in the 23S rRNA of archaeal species (Ofengand and Bakin, 1997; and for review, Ofengand, 2002; corrected in Del Campo *et al.*, 2005). The corresponding positions in the yeast and human large subunit rRNAs are also pseudouridylated, and the H/ACA snoRNAs snR191 in yeast and hU19 in human are, respectively, proposed to guide the modifications (Badis *et al.*, 2003; Bortolin and Kiss, 1998).

On the contrary, no counterpart of the second putative target residue (U1261) of motif FW1 was found to be pseudouridylated, neither in archaea nor in eukarya. Therefore, taken together, the FW1 motif of *M. jannaschii* can reasonably be considered as guiding U to Ψ conversion at position U2015 in 23S rRNA.

2.3.2. Experimental validations

The situation described previously for the FW1 motif of *M. jannaschii* is a highly favorable one in terms of validation. It is not always possible to find homologous H/ACA motifs in other species and/or the presence of a Ψ residue at the targeted position in another archaeal species. When this is not the case, direct experimental proof is particularly important to validate the data.

A first simple and important experiment to do is a Northern blot or primer extension analysis on total RNA extracted from the studied archaea. This will allow verification of the presence of the proposed H/ACA sRNA. Radiolabeled primers complementary to the H/ACA sRNAs candidates can be used for this purpose. In addition, these kinds of experiments will give information on the length of the sRNAs because, as explained before, archaeal sRNAs have variable lengths. A protocol for primer extension analysis is proposed in Chapter 2 in this volume by Motorine *et al.*

Another validation approach involves localizing the pseudouridylation positions in rRNAs by use of the CMCT-RT approach as described by Motorine *et al.* in Chapter 2. This is not an easy technique to handle. However, a detailed protocol and tips are given by Motorine *et al.*

Finally, the complete verification of the proposed target sequences can be obtained by *in vitro* reconstitution of an active H/ACA sRNP by use of *in vitro* transcribed guide RNA and target sequence and recombinant aCBF5, aNOP10, L7Ae, and aGAR1 proteins, as described in Chapter 16 by Charpentier *et al.*, in this volume.

3. Conclusions

The knowledge-based approach described herein, which combines the search for H/ACA motifs and their respective target(s), is an efficient approach as illustrated in the numerous examples given in this chapter. Its efficiency will be further improved, on enrichment of the H/ACA profile by inclusion of new validated motifs. One does not need much expertise in computing to use the ERPIN and RNAMOT softwares, which are easy to use. Moreover, a web server version of ERPIN is available for the user to become familiar with the method. No data other than the structural or functional features of H/ACA sRNAs, which are described in this chapter, is necessary.

Compared with the results recently obtained by application of the MilPat tool to the identification of H/ACA-like motifs in the genomes from *M. jannaschii* and three *Pyrococcus* species (Thebault *et al.*, 2006), the approach proposed in this chapter is more directed and, therefore, gives a very limited number of false-positive results. By the use of the H/ACA profile

enriched by the H/ACA motifs of *T. kodakarensis* identified in this work, we detected, in the blind test presented in Table 15.2, the 10 H/ACA motifs present in *P. abyssi, P. furiosus*, and *P. horikoshii*. When the MilPat approach is used, among the 89 to 148 candidates H/ACA motifs found for the different *Pyrococcus* species studied, only 6 to 7 of the known motifs were detected, depending on the species. We think that after inclusion of the H/ACA motifs from a limited number of archaeal species belonging to different archaeal orders, which is currently being done by our team, we will be able to propose soon (at site http://tagc.univ-mrs.fr/asterix/erpin/), a highly powerful tool for the search of H/ACA sRNAs in any archaeal species. The strength of our approach is the coupling of the search of H/ACA-like motifs to the search of their target sequences. As evidenced in the given examples, both steps in this strategy participate in the selection of the true H/ACA motifs among the identified hit motifs.

We would like to point out that the strategy proposed in this chapter is well suited to the search of archaeal H/ACA motifs because of their structural specificities (presence of a K-turn or a K loop and of G-C rich helices). This strategy would be much less efficient if applied to the search of snoRNA coding sequences. However, a specific computational approach dedicated to eukaryal H/ACA snoRNAs has been developed (Schattner *et al.*, 2006).

REFERENCES

Badis, G., Fromont-Racine, M., and Jacquier, A. (2003). A snoRNA that guides the two most conserved pseudouridine modifications within rRNA confers a growth advantage in yeast. *RNA* **9,** 771–779.

Baker, D. L., Youssef, O. A., Chastkofsky, M. I., Dy, D. A., Terns, R. M., and Terns, M. P. (2005). RNA-guided RNA modification: Functional organization of the archaeal H/ACA RNP. *Genes Dev.* **19,** 1238–1248.

Balakin, A. G., Smith, L., and Fournier, M. J. (1996). The RNA world of the nucleolus: Two major families of small RNAs defined by different box elements with related functions. *Cell* **86,** 823–834.

Bortolin, M. L., and Kiss, T. (1998). Human U19 intron-encoded snoRNA is processed from a long primary transcript that possesses little potential for protein coding. *RNA* **4,** 445–454.

Bourdeau, V., Ferbeyre, G., Pageau, M., Paquin, B., and Cedergren, R. (1999). The distribution of RNA motifs in natural sequences. *Nucleic Acids Res.* **27,** 4457–4467.

Charpentier, B., Muller, S., and Branlant, C. (2005). Reconstitution of archaeal H/ACA small ribonucleoprotein complexes active in pseudouridylation. *Nucleic Acids Res.* **33,** 3133–3144.

Charron, C., Manival, X., Clery, A., Senty-Segault, V., Charpentier, B., Marmier-Gourrier, N., Branlant, C., and Aubry, A. (2004). The archaeal sRNA binding protein L7Ae has a 3D structure very similar to that of its eukaryal counterpart while having a broader RNA-binding specificity. *J. Mol. Biol.* **342,** 757–773.

Del Campo, M., Recinos, C., Yanez, G., Pomerantz, S. C., Guymon, R., Crain, P. F., McCloskey, J. A., and Ofengand, J. (2005). Number, position, and significance of the pseudouridines in the large subunit ribosomal RNA of Haloarcula marismortui and Deinococcus radiodurans. *RNA* **11,** 210–219.

Ganot, P., Bortolin, M. L., and Kiss, T. (1997a). Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. *Cell* **89,** 799–809.

Ganot, P., Caizergues-Ferrer, M., and Kiss, T. (1997b). The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes Dev.* **11,** 941–956.

Gautheret, D., and Lambert, A. (2001). Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J. Mol. Biol.* **313,** 1003–1011.

Gautheret, D., Major, F., and Cedergren, R. (1993). Modeling the three-dimensional structure of RNA using discrete nucleotide conformational sets. *J. Mol. Biol.* **229,** 1049–1064.

Hamma, T., Reichow, S. L., Varani, G., and Ferre-D'Amare, A. R. (2005). The Cbf5-Nop10 complex is a molecular bracket that organizes box H/ACA RNPs. *Nat. Struct. Mol. Biol.* **12,** 1101–1107.

Henras, A., Henry, Y., Bousquet-Antonelli, C., Noaillac-Depeyre, J., Gelugne, J. P., and Caizergues-Ferrer, M. (1998). Nhp2p and Nop10p are essential for the function of H/ACA snoRNPs. *EMBO J.* **17,** 7078–7090.

Huang, Z. P., Zhou, H., He, H. L., Chen, C. L., Liang, D., and Qu, L. H. (2005). Genome-wide analyses of two families of snoRNA genes from *Drosophila melanogaster*, demonstrating the extensive utilization of introns for coding of snoRNAs. *RNA* **11,** 1303–1316.

Huang, Z. P., Zhou, H., Liang, D., and Qu, L. H. (2004). Different expression strategy: Multiple intronic gene clusters of box H/ACA snoRNA in *Drosophila melanogaster*. *J. Mol. Biol.* **341,** 669–683.

Khanna, M., Wu, H., Johansson, C., Caizergues-Ferrer, M., and Feigon, J. (2006). Structural study of the H/ACA snoRNP components Nop10p and the 3′ hairpin of U65 snoRNA. *RNA* **12,** 40–52.

Kiss, A. M., Jady, B. E., Bertrand, E., and Kiss, T. (2004). Human box H/ACA pseudouridylation guide RNA machinery. *Mol. Cell Biol.* **24,** 5797–5807.

Klein, D. J., Schmeing, T. M., Moore, P. B., and Steitz, T. A. (2001). The kink-turn: A new RNA secondary structure motif. *EMBO J.* **20,** 4214–4221.

Laferriere, A., Gautheret, D., and Cedergren, R. (1994). An RNA pattern matching program with enhanced performance and portability. *Comput. Appl. Biosci.* **10,** 211–212.

Lafontaine, D. L., Bousquet-Antonelli, C., Henry, Y., Caizergues-Ferrer, M., and Tollervey, D. (1998). The box H + ACA snoRNAs carry Cbf5p, the putative rRNA pseudouridine synthase. *Genes Dev.* **12,** 527–537.

Lambert, A., Legendre, M., Fontaine, J. F., and Gautheret, D. (2005). Computing expectation values for RNA motifs using discrete convolutions. *BMC Bioinformatics* **6,** 118.

Lambert, A., Fontaine, J. F., Legendre, M., Leclerc, F., Permal, E., Major, F., Putzer, H., Delfour, O., Michot, B., and Gautheret, D. (2004). The ERPIN server: An interface to profile-based RNA motif identification. *Nucleic Acids Res.* **32,** W160–W165.

Lambert, A., Lescure, A., and Gautheret, D. (2002). A survey of metazoan selenocysteine insertion sequences. *Biochimie* **84,** 953–959.

Legendre, M., Lambert, A., and Gautheret, D. (2005). Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics* **21,** 841–845.

Lescure, A., Gautheret, D., Carbon, P., and Krol, A. (1999). Novel selenoproteins identified *in silico* and *in vivo* by using a conserved RNA structural motif. *J. Biol. Chem.* **274,** 38147–38154.

Lescure, A., Gautheret, D., and Krol, A. (2002). Novel selenoproteins identified from genomic sequence data. *Methods Enzymol.* **347,** 57–70.

Li, L., and Ye, K. (2006). Crystal structure of an H/ACA box ribonucleoprotein particle. *Nature* **443,** 302–307.

Li, S. G., Zhou, H., Luo, Y. P., Zhang, P., and Qu, L. H. (2005). Identification and functional analysis of 20 Box H/ACA small nucleolar RNAs (snoRNAs) from *Schizosaccharomyces pombe*. *J. Biol. Chem.* **280,** 16446–16455.

Macke, T. J., Ecker, D. J., Gutell, R. R., Gautheret, D., Case, D. A., and Sampath, R. (2001). RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.* **29,** 4724–4735.

Manival, X., Charron, C., Fourmann, J. B., Godard, F., Charpentier, B., and Branlant, C. (2006). Crystal structure determination and site-directed mutagenesis of the *Pyrococcus abyssi* aCBF5-aNOP10 complex reveal crucial roles of the C-terminal domains of both proteins in H/ACA sRNP activity. *Nucleic Acids Res.* **34,** 826–839.

Massenet, S., Motorin, Y., Lafontaine, D. L., Hurt, E. C., Grosjean, H., and Branlant, C. (1999). Pseudouridine mapping in the *Saccharomyces cerevisiae* spliceosomal U small nuclear RNAs (snRNAs) reveals that pseudouridine synthase pus1p exhibits a dual substrate specificity for U2 snRNA and tRNA. *Mol. Cell Biol.* **19,** 2142–2154.

Meier, U. T., and Blobel, G. (1994). NAP57, a mammalian nucleolar protein with a putative homolog in yeast and bacteria. *J. Cell Biol.* **127,** 1505–1514.

Ni, J., Tien, A. L., and Fournier, M. J. (1997). Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA. *Cell* **89,** 565–573.

Ofengand, J. (2002). Ribosomal RNA pseudouridines and pseudouridine synthases. *FEBS Lett.* **514,** 17–25.

Ofengand, J., and Bakin, A. (1997). Mapping to nucleotide resolution of pseudouridine residues in large subunit ribosomal RNAs from representative eukaryotes, prokaryotes, archaebacteria, mitochondria and chloroplasts. *J. Mol. Biol.* **266,** 246–268.

Rashid, R., Liang, B., Baker, D. L., Youssef, O. A., He, Y., Phipps, K., Terns, R. M., Terns, M. P., and Li, H. (2006). Crystal structure of a Cbf5-Nop10-Gar1 complex and implications in RNA-guided pseudouridylation and dyskeratosis congenita. *Mol. Cell* **21,** 249–260.

Rozenski, J., Crain, P. F., and McCloskey, J. A. (1999). The RNA Modification Database: 1999 update. *Nucleic Acids Res.* **27,** 196–197.

Rozhdestvensky, T. S., Tang, T. H., Tchirkova, I. V., Brosius, J., Bachellerie, J. P., and Huttenhofer, A. (2003). Binding of L7Ae protein to the K-turn of archaeal snoRNAs: A shared RNA binding motif for C/D and H/ACA box snoRNAs in Archaea. *Nucleic Acids Res.* **31,** 869–877.

Schattner, P., Barberan-Soler, S., and Lowe, T. M. (2006). A computational screen for mammalian pseudouridylation guide H/ACA RNAs. *RNA* **12,** 15–25.

Schattner, P., Decatur, W. A., Davis, C. A., Ares, M., Jr., Fournier, M. J., and Lowe, T. M. (2004). Genome-wide searching for pseudouridylation guide snoRNAs: Analysis of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res.* **32,** 4281–4296.

Tang, T. H., Bachellerie, J. P., Rozhdestvensky, T., Bortolin, M. L., Huber, H., Drungowski, M., Elge, T., Brosius, J., and Huttenhofer, A. (2002a). Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. *Proc. Natl. Acad. Sci. USA* **99,** 7536–7541.

Thebault, P., de Givry, S., Schiex, T., and Gaspin, C. (2006). Searching RNA motifs and their intermolecular contacts with constraint networks. *Bioinformatics* **22,** 2074–2080.

Torchet, C., Badis, G., Devaux, F., Costanzo, G., Werner, M., and Jacquier, A. (2005). The complete set of H/ACA snoRNAs that guide rRNA pseudouridylations in *Saccharomyces cerevisiae*. *RNA* **11,** 928–938.

Vidovic, I., Nottrott, S., Hartmuth, K., Luhrmann, R., and Ficner, R. (2000). Crystal structure of the spliceosomal 15.5kD protein bound to a U4 snRNA fragment. *Mol. Cell* **6,** 1331–1342.

Wang, C., and Meier, U. T. (2004). Architecture and assembly of mammalian H/ACA small nucleolar and telomerase ribonucleoproteins. *EMBO J.* **23,** 1857–1867.

Watkins, N. J., Gottschalk, A., Neubauer, G., Kastner, B., Fabrizio, P., Mann, M., and Luhrmann, R. (1998). Cbf5p, a potential pseudouridine synthase, and Nhp2p, a putative RNA-binding protein, are present together with Gar1p in all H BOX/ACA-motif snoRNPs and constitute a common bipartite structure. *RNA* **4,** 1549–1568.

Yuan, G., Klambt, C., Bachellerie, J. P., Brosius, J., and Huttenhofer, A. (2003). RNomics in *Drosophila melanogaster*: Identification of 66 candidates for novel non-messenger RNAs. *Nucleic Acids Res.* **31,** 2495–2507.

Nucleic Acids Research Advance Access published February 27, 2008

Nucleic Acids Research, 2008, 1–17
doi:10.1093/nar/gkn077

Combined *in silico* and experimental identification of the *Pyrococcus abyssi* H/ACA sRNAs and their target sites in ribosomal RNAs

Sébastien Muller, Fabrice Leclerc, Isabelle Behm-Ansmant, Jean-Baptiste Fourmann, Bruno Charpentier and Christiane Branlant*

Laboratoire de Maturation des ARN et Enzymologie Moléculaire, UMR 7567 CNRS-UHP, Nancy Université, Faculté des Sciences et Techniques, 54506 Vandoeuvre-lès-Nancy, France

Received January 15, 2008; Revised February 7, 2008; Accepted February 8, 2008

ABSTRACT

How far do H/ACA sRNPs contribute to rRNA pseudouridylation in Archaea was still an open question. Hence here, by computational search in three *Pyrococcus* genomes, we identified seven H/ACA sRNAs and predicted their target sites in rRNAs. In parallel, we experimentally identified 17 Ψ residues in *P. abyssi* rRNAs. By *in vitro* reconstitution of H/ACA sRNPs, we assigned 15 out of the 17 Ψ residues to the 7 identified H/ACA sRNAs: one H/ACA motif can guide up to three distinct pseudouridylations. Interestingly, by using a 23S rRNA fragment as the substrate, one of the two remaining Ψ residues could be formed *in vitro* by the aCBF5/aNOP10/aGAR1 complex without guide sRNA. Our results shed light on structural constraints in archaeal H/ACA sRNPs: the length of helix H2 is of 5 or 6 bps, the distance between the ANA motif and the targeted U residue is of 14 or 15 nts, and the stability of the interaction formed by the substrate rRNA and the 3′-guide sequence is more important than that formed with the 5′-guide sequence. Surprisingly, we showed that a sRNA–rRNA interaction with the targeted uridine in a single-stranded 5′-UNN-3′ trinucleotide instead of the canonical 5′-UN-3′ dinucleotide is functional.

INTRODUCTION

Conversion of uridine into pseudouridine (Ψ) residues is one of the most abundant post-transcriptional modifications of tRNAs, rRNAs and UsnRNAs (1). Compared to U residues, Ψ residues can form an additional hydrogen bond at the N1–H position. Furthermore, the carbon–carbon link between the ribose and the base limits the flexibility of the ribose backbone of Ψ residues, which favours and stabilizes base-pair interactions (2). A role of Ψ residues in stabilization of the tRNA 3D structure has also been well documented (3), and the functional importance of Ψ residues in U2 snRNA for the activity of splicing complexes was demonstrated (4–7). Eukaryal rRNAs contain a large number of Ψ residues compared to bacterial rRNAs and they are concentrated in rRNA regions expected to play important functional roles, in particular in domains IV and V, which are directly involved in the peptidyl transferase activity (8–10). Taken individually, pseudouridylations in rRNAs are not essential for cell growth. However, the complete abolition of Ψ formation in rRNAs is deleterious for ribosome assembly and activity (10–12). Recent data suggest their possible involvement in: (i) subunit interaction (12–14), (ii) the translocation step (14,15), (iii) translation termination (12,16) and (iv) folding of 23S rRNA in an active form at the peptidyl transferase centre (PTC) (11,12,17). The large number of pseudouridylation sites in eukaryal rRNAs compared to bacterial rRNAs is explained by the use of different catalysts: stand-alone enzymes carrying both the RNA recognition capability and the catalytic activity are used in bacteria (10,18), whereas U to Ψ conversions are catalyzed by H/ACA snoRNPs in eukarya (19,20). The H/ACA snoRNPs contain four proteins and an H/ACA snoRNA that defines the targeted U residue by base-pair interaction with the rRNA. Recent data revealed a similar RNA-guided system in archaea (21–24). Most of the eukaryal H/ACA snoRNAs contain two characteristic stem-loop structures, with an internal loop (pseudouridylation pocket), that is complementary to two nucleotide stretches bordering the targeted U residue. Each of the stem-loops is flanked by a conserved motif (H and

*To whom correspondence should be addressed. Tel: +33 3 83 68 43 03; Fax: +33 3 83 68 43 07; Email: christiane.branlant@maem.uhp-nancy.fr

ACA, respectively). In the following part of this manuscript, one stem-loop structure containing a pseudouridylation pocket flanked by one of the conserved 3′ sequence will be denoted an H/ACA motif. In the RNA duplex formed between the H/ACA motif and the substrate rRNA sequence, the targeted U residue and its 3′ nucleotide are both single-stranded. The archaeal sRNA counterparts of the snoRNAs have more diverse architectures. They are composed of one, two or three H/ACA motifs (21–26). In addition, each archaeal H/ACA stem-loop structure contains a K-turn or a K-loop which binds protein L7Ae (22). The conserved 3′-flanking sequence is an ANA trinucleotide (most frequently an ACA trinucleotide).

As successful reconstitutions of active H/ACA sRNPs were achieved using an *in vitro* transcribed H/ACA sRNA and the recombinant archaeal H/ACA sRNP proteins (23,24), strong progresses were recently made in the understanding of the H/ACA sRNP structure and function. Like the eukaryal H/ACA snoRNPs, the archaeal H/ACA sRNPs contain four proteins. Protein aCBF5 is the RNA: Ψ-synthase, aNOP10 is required for aCBF5 activity, L7Ae binds the K-turn or K-loop motif, and aGAR1 may facilitate the H/ACA sRNP turnover (22–24). Both L7Ae and aGAR1 strongly reinforce the sRNP activity (23). The crystal structures of H/ACA sRNP protein complexes and of an entire H/ACA sRNP were recently solved at high resolution (27–31). The ANA sequence is needed for binding of aCBF5 to the guide RNA (28,29,31). In the crystal structure, aCBF5 also interacts with the pseudouridylation pocket, helix H1 and helix H2 of the H/ACA motif (31). Recently a 3D structure obtained for an H/ACA sRNP devoid of the L7Ae protein and bound to its RNA substrate, revealed contacts between aCBF5 and the target rRNA–H/ACA sRNA duplex (30). However, little is known on the structural constraints required for formation of an active rRNA target–H/ACA sRNA interaction. The two NMR structures established for a complex formed between an H/ACA stem-loop structure and a small target RNA revealed the capability of the two RNAs to interact together, in the absence of protein, at the high concentration used for NMR analysis (32,33). In these structures, the heterologous helix formed by interaction of the RNA target with the 3′-guide element of the sRNA is stacked on helix H1, while the heterologous helix formed by interaction of the RNA target with the 5′-guide element of the sRNA is stacked on helix H2.

At present, in contrast to this extensive knowledge on H/ACA sRNPs, little is known on the number and location of Ψ residues in archaeal rRNAs. Their presence has only been investigated in *Halobacterium halobium* (34), *Haloarcula marismortui* (35,36), *Sulfolobus acidocaldarius* (37) and *Archaeoglobus fulgidus* (21). Unfortunately, the search for putative H/ACA sRNA genes by computational analysis was made for other archaeal species: *Pyrococcus furiosus* (22,24,38), *Methanococcus jannaschii* (26) and *Thermococcus kodakarensis* (25). The utilization of RNomics approaches for the search of H/ACA sRNAs in archaea is even more scarce, it was only applied to the *A. fulgidus* (21) and *Sulfolobus solfataricus* (39) species: three H/ACA sRNAs and one single sRNA were identified, respectively. For a better definition of the rules that govern the H/ACA sRNP specificities and efficiencies, it was of high importance to identify all the putative H/ACA sRNAs of a given species as well as Ψ residues in its rRNAs and then, to try to assign the detected Ψ residues to the detected H/ACA sRNAs.

To this end, here, we used different experimental approaches in order to define the target sites of the putative H/ACA sRNAs that we identified by a computational analysis of the genomic sequences of three *Pyrococcus* species. These three species, *Pyrococcus abyssi*, *Pyrococcus furiosus* and *Pyrococcus horikoshii*, are hyperthermophiles (optimal growth temperature between 95 and 100°C). As a first step, we developed and used various computational approaches to identify all the putative H/ACA sRNAs of these species and their putative target sites in rRNAs. Then, to test for the presence of Ψ residues at the predicted sites in 16S and 23S rRNAs, we applied the RT-CMCT approach to large segments of the *P. abyssi* 16S and 23S rRNAs that were including the predicted pseudouridylation sites. Finally, to confirm the role of the identified H/ACA motifs in formation of the identified Ψ residues, we tested the activities of reconstituted H/ACA sRNPs on small rRNA fragments containing the expected target U residues. Finally, on the basis of the data obtained, we drew conclusions on structural constraints to which H/ACA sRNAs and the H/ACA sRNA–rRNA interactions are subjected.

MATERIALS AND METHODS

Extraction of inter-coding-regions (ICR) from *Pyrococcus* genomes

The sequences and annotations of the archaeal genomes *Pyrococcus abyssi GE5*, *Pyrococcus horikoshii OT3* and *Pyrococcus furiosus DSM3638* were downloaded in Fasta and Genbank formats from NCBI ftp site, ftp://ftp.ncbi.nih.gov/genomes/Bacteria. In each genome, the positions of DNA sequences corresponding to ORFs or template sequences of known stable RNAs (rRNA, tRNA, RNaseP and 7S RNA) were listed in a table denoted 'position table'. Based on this table, the remaining segments of the genomes were extracted automatically from the genomic sequences, by using a script written in awk and perl languages (ExtractICR). Only sequences exceeding 15 nts were collected. ExtractICR flanks each of the linking sequences by their two 15-nt long bordering sequences. These exported sequences, denoted ICRs, were formatted and assembled in data bases, using Readseq (ftp://iubio.bio.indiana.edu/molbio/readseq) and Formatdb (NCBI toolkit).

Selection of conserved ICRs and search for H/ACA sRNA genes

A Blast version adapted for multi-aligment was used to compare the ICR sequences. First, repeated elements (more than 50-nt long segments repeated several times in the genome) were removed by comparison of all the

ICRs extracted from one given genome. Then, the ICRs from one given pyrococcal species were compared to those of the two other species. Based on a statistical analysis, the criteria retained to consider that an ICR shows a significant degree of conservation in the three species was the presence of at least a stretch of 18 nts with an identical sequence in these species or a 21-nt long sequence with only one base-pair substitution in one or two of the three species. The selected ICRs were then aligned using Clustal-W (40). GeneMarks (41) was used for elimination of the conserved ICRs corresponding to mis-annotated ORFs. An RNAMOT (42) descriptor, that was designed for C/D box sRNA gene detection, was used to identify these genes in the conserved ICRs. Finally, another RNAMOT descriptor was used to screen for the presence of H/ACA sRNA genes. It was based on some of the known structural features of the archaeal H/ACA motifs: the presence of two helices H1 (at least 7 bps) and H2 (at least 5 bps) flanking an internal loop (each strand including 5 to 11 nts). The length of the apical loop was allowed to vary between 8 and 35 nts. The descriptor included an ACA trinucleotide flanking helix H1. Only H/ACA motifs having at least one putative target site in rRNAs or tRNAs that fitted to the rules defined for H/ACA sRNA–rRNA interactions were retained. To this end, RNAMOT descriptors were built for each putative H/ACA motif as described in (25).

Search for H/ACA sRNA genes in entire genomic sequences

Then, based on the results obtained with RNAMOT, we used the ERPIN (43) software for searches with higher constraints in the entire *Pyrococcus* genomes. ERPIN builds helix and single-strand lod score profiles from sequence alignments and screens genome sequences for occurrences of these profiles. The H/ACA sRNAs that we identified by screening the conserved ICRs, together with the H/ACA sRNAs identified experimentally in *A. fulgidus* (21), were used to build the ERPIN profile as described in (25). As above, for each new candidate detected by this approach, RNAMOT was used to predict target sites in rRNAs (25).

The nucleotide sequences of all the H/ACA sRNAs detected in this study and the positions of their template sequences in the archaeal genomes are accessible at http://tagc.univ-mrs.fr/erpin/.

P. abyssi cultures

P. abyssi strain GE5 cells were grown as described (44) at 95°C in Vent Sulfothermophiles Medium (20 g/l NaCl, 0.25 g/l KCl, 0.05 g/l NaBr, 0.5 g/l $SrCl_2$-$6H_2O$, 0.08 g/l boric acid, 3 g/l PIPES, 1 g/l yeast extracts, 4 g/l peptone, 1 g/l Resazurine, 200 g/l $MgSO_4$, 50 g/l $CaCl_2$, 50 g/l KH_2PO_4 pH 6.8). *P. abyssi* cell growth was stopped at the end of the exponential phase.

Total RNA isolation

The *P. abyssi* cells were collected by centrifugation. After washing in 1 M sorbitol, 25 mM Hepes, pH 7.0, they were frozen and stored at −80°C. The method described by Chomczynski and Sacchi (1986) (45) was used for extraction of total RNA. About 10^{10} cells were dissolved in 4 ml of solution D (4 M guanidinium thiocyanate, 25 mM sodium citrate pH 7.0, 0.5% sarcosyl, 0.1 M β-mercaptoethanol). The extracted RNAs were recovered by phenol/chloroform extraction, followed by ethanol precipitation using 0.3 M sodium acetate. They were dissolved in bi-distilled water and quantified.

Northern blot analysis

About 10 µg of *P. abyssi* total RNA and 5′-end labelled DNA size markers (100 bp DNA ladder, MBI Fermentas) were fractionated in parallel on 6% denaturing polyacrylamide gel (8 M urea, 0.5× TBE buffer). After electrotransfer on a Hybond-N+ membrane (Amersham) and by UV-crosslinking, a pre-hybridization was carried out for 1 h at 58°C in SSPE buffer (0.9 M NaCl, 47 mM Na_2HPO_4-$2H_2O$, 6 mM EDTA pH 7.4, containing 1 g/l Ficoll, 1 g/l polyvinylpyrolidone, 1 g/l BSA, 0.5% SDS). Oligonucleotide probes complementary to the predicted H/ACA sRNAs (Table S1 in Supplementary data) were 5′-end labelled with [γ-^{32}P]ATP and T4 polynucleotide kinase for 1 h at 37°C. Hybridization was carried out at 58°C for 16 h in the presence of 100 ng of the labelled oligonucleotide. The membranes were washed four times in SSPE buffer at 42°C for 5 min and the hybridization bands were visualized on a Typhoon 9410 (Amersham Biosciences).

Mapping of pseudouridine (Ψ) residues in the *P. abyssi* rRNAs

The *N*-cyclohexyl-*N*′-(2-morpholinoethyl)-carbodiimide metho-*p*-toluolsulfonate (CMCT) RNA treatment was adapted from Bakin and Ofengand (1993) (46), as previously described (47). CMCT modifications were performed on 5 µg of *P. abyssi* total RNA. Positions of CMCT modifications were identified by primer extension analysis, using the AMV RT (QBiogene, USA). The sequences of the 5′-end labelled primers that we used are given in Table S1 (Supplementary data). RNA sequencing was done on 20 µg of *P. abyssi* total RNA.

Recombinant protein production

The recombinant GST-L7Ae, GST-aNOP10, GST-aGAR1, GST-aCBF5 proteins were produced in *E. coli* BL21 CodonPlus cells (Novagen). The GST-tag was cleaved by the precision protease in the course of purification performed on Glutathione-Sepharose 4B (Pharmacia), as previously described (23).

In vitro transcription of H/ACA sRNAs and their rRNA substrates

The Pab19, Pab21, Pab35, Pab40, Pab91, Pab105 and Pab160 sRNA sequences were PCR amplified from the *P. abyssi* GE5 genomic DNA using a forward primer containing the T7 RNA polymerase promoter and a second primer complementary to the expected 3′ end of the sRNA (Table S1 in Supplementary data). The amplified DNA fragments were cloned in the pTAdv vector (Clontech) and sequenced. *In vitro* transcriptions with the T7 RNA polymerase were performed as

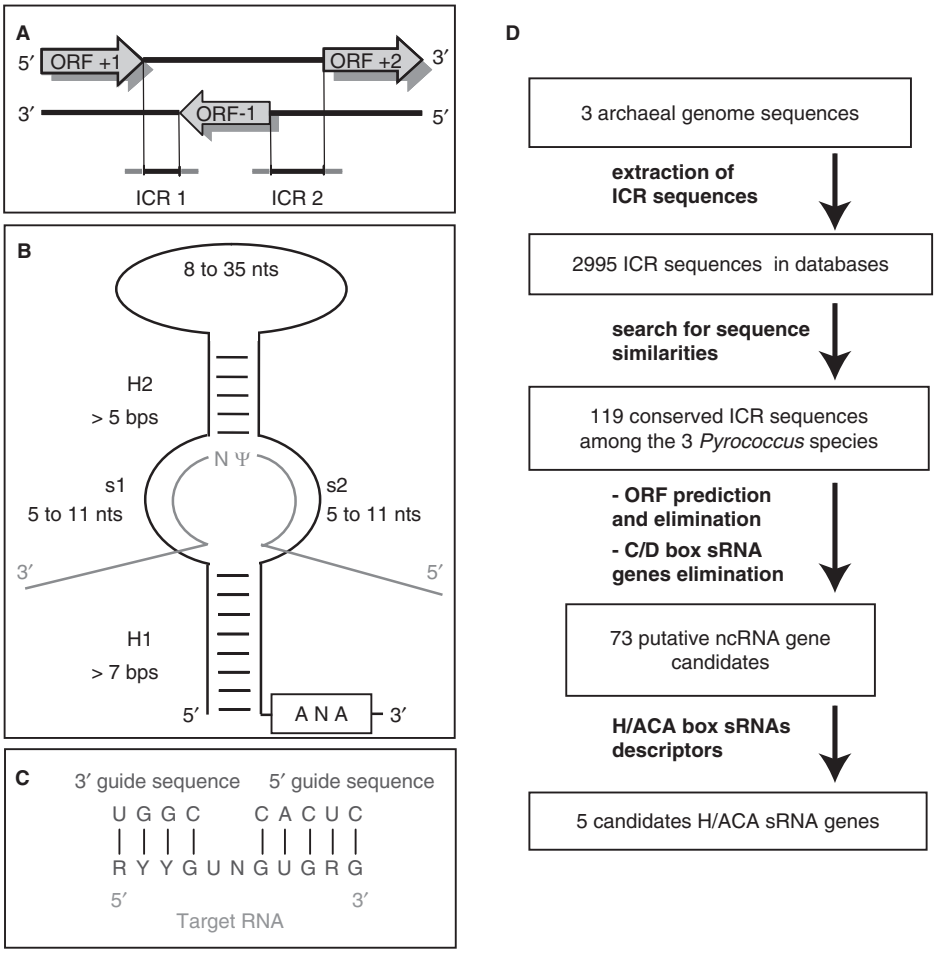


Figure 1. The computational process used for the search of H/ACA genes in ICRs. Panel A: ICR definition. ICRs include 15 bps upstream and downstream of the segments linking the ORFs or template sequences for known RNAs. Panel B: the helices H1 and H2, and loop size criteria used to design the RNAMOT descriptor devoted to the search for H/ACA sRNA genes are shown. Panel C: Example of base-pair criteria used to search for possible rRNA targets of the candidate H/ACA motifs. Panel D: the 5 computer based-steps used for the search of H/ACA genes: (1) sequences and annotations were downloaded, (2) sequences of ICRs were extracted from the genomic sequences and assembled in databases, (3) after elimination of ICRs containing repeated sequences, the ICR sequences from one species were compared to those of other species by using Blast, 119 conserved ICRs were selected, (4) protein-coding sequences and known C/D box sRNA genes were filtered out of the conserved sequences and (5) RNAMOT descriptors were used for the search of H/ACA genes in ICRs, five putative genes were detected.

previously described (23). The H/ACA sRNA transcripts were purified by gel electrophoresis.

Template DNA encoding the RNA targets derived from *P. abyssi* rRNAs were obtained by annealing two complementary oligonucleotides. *In vitro* transcription was performed as previously described (23), in the presence of 20 μCi [α-^{32}P] NTP (800 Ci/mmol), 0.13 mM of the same NTP and 4 mM of each of the three other NTPs. When using RNase P1 for the digestion, [α-^{32}P]UTP was used for labelling. When RNase T2 was used, the identity of the [α-^{32}P]NTP used was defined by the residue located 3′ to the targeted U residue. The sequences of the *in vitro* transcribed substrates are given in Table S2 in Supplementary data.

H/ACA sRNP reconstitution and test of their activity

As previously described (48), the unlabelled sRNA (4 pmol) and [α-^{32}P] NTP-labelled targets (150 fmol) were mixed with the four L7Ae, aCBF5, aNOP10 and aGAR1 recombinant proteins at a 200 nM concentration, at room temperature in buffer D (150 mM KCl; 1.5 mM $MgCl_2$; 0.2 mM EDTA; 20 mM HEPES, pH 7.9). The mixture was incubated at 65°C for 80 min. Then, formation of Ψ residue was detected after either T2 or P1 RNase digestion. The 3′-mono phosphate nucleotides produced by RNase T2 digestion or the 5′-monophosphate residues obtained after P1 RNase digestion were fractionated by thin-layer chromatography (TLC) (mono or 2D) as previously described, using the N1 buffer for 1D TLC and N1-N2 or N1-R2 buffers for 2D TLC (48). The radioactivity in the spots or the bands was quantified with a phosphorimager, by using the ImageQuant software. When digestion was achieved with RNase T2, the yield of Ψ formation (expressed in mol of Ψ residue per mol of target RNA) was estimated taking into account the total number of residues located 5′ to the incorporated

labelled nucleotide. After P1 RNase digestion, we only took into consideration the total number of U residues in the target RNA.

RESULTS

Analysis of conserved ICRs in *Pyrococcus* identifies five conserved putative H/ACA sRNAs

In order to make a link between H/ACA sRNAs and pseudouridylation in rRNAs, in the *P. abyssi* species, we first made a complete identification of H/ACA–sRNA motifs in *P. abyssi*. This was done in two steps because, when we started this study, only a limited number of archaeal H/ACA sRNAs had been identified. As a first step, we applied a phylogenetic approach to the *Pyrococcus* genus. The idea was that DNA segments, which link ORFs and/or template sequences for stable known RNAs (rRNAs, tRNAs, RNase P, 7S RNA) (Inter-Coding-Regions, ICRs) and that bear long stretches of conserved sequences in three *Pyrococcus* species *P. abyssi*, *P. furiosus* and *P. horikoshii*, may correspond to genes for functional non-coding RNAs. Based on the few H/ACA sRNAs known at that time (three from *A. fulgidus*) (21) some characteristic features of H/ACA sRNA motifs could be delineated. They were used to build an RNAMOT descriptor for the search of H/ACA sRNA genes. Then, RNAMOT profiles were built for the search of the possible rRNA target sites of each candidate H/ACA motif. By using this approach (see the details in Materials and Methods), we detected five putative H/ACA sRNA genes, that were common to the three species. Four of them have been recently characterized in *P. furiosus* by the use of a computational approach based on G/C content analysis, namely, the Pf1, Pf3, Pf6 and Pf7 sRNAs (22,49). The counterparts that we identified in *P. abyssi* and *P. horikoshii* are designated as Pab21, Pab105, Pab35 and Pab40 sRNAs (*P. abyssi*) and Pho21, Pho 105, Pho35 and Pho40 sRNAs (*P. horikoshii*), respectively (Figure 2). The fifth common H/ACA sRNA, which had not been characterized by other teams, is denoted Pab91, Pfu91 and Pho91 in *P. abyssi*, *P. furiosus* and *P. horikoshii*, respectively. We previously used it to settle conditions for *in vitro* reconstitution of active H/ACA sRNPs (23). Taking into account the numerous compensatory base-pair mutations in the three species studied and in *T. kodakarensis*, we could propose relevant secondary structures for each of the five sRNAs (Figure 2). Only for sRNA Pab21 and for motif 2 in sRNA Pab40, it was difficult to make a choice between two possible 2D structures that were both containing a K-loop motif (Figure 2). Production of the H/ACA sRNAs Pab21, Pab35, Pab40, Pab91 and Pab105 in *P. abyssi* was verified by Northern blot analysis (Figure 3), and two forms of Pab21 sRNA with or without the C/D box motif were found to be present *in vivo* (Figure 3).

Screening of complete genomes with ERPIN identifies two additional common putative H/ACA sRNAs in Pyrococci

Some H/ACA motifs may have escaped detection when using the above comparative approach. Therefore, we took advantage of the increased knowledge on H/ACA sRNAs, which was brought by the 15 pyrococcal H/ACA sRNAs that we identified, to develop another computational approach based on the ERPIN software (25,43). To this end, the sequences of the 3 known H/ACA sRNAs from *A. fulgidus* and the 15 pyrococcal H/ACA sRNAs were aligned by ERPIN, on the basis of their proposed 2D structures and conserved sequence elements. As recently described (25), by comparison of such a profile with complete archaeal genomic sequences, ERPIN can predict H/ACA sRNA genes with a high degree of efficiency. By using different sets of constraints for the lengths of the 3′- and 5′-guide strands of the pseudouridylation pocket, we detected two other putative H/ACA sRNAs common to the three *Pyrococcus* species. One of them was also recently detected in *P. furiosus* (sRNA Pf9) by another approach (24). Its counterparts in *P. abyssi* and *P. horikoshii* were denoted Pab160 and Pho160, respectively (Figure 2F). The second H/ACA sRNA identified (Pab19, Pfu19 and Pho19 in *P. abyssi*, *P. furiosus* and *P. horikoshii*, respectively) had not been detected previously, probably because of its extended 5′-guide sequence (Figure 2). Nucleotide sequence conservation in the three species and secondary structure conservation by compensatory base changes strongly suggest the presence of a small 5′ stem-loop structure upstream from the H/ACA motif in sRNA Pab19. Note that the Pab19 pseudouridylation pocket is larger than usual due to the length of its 5′-guide sequence (15 nts) (Figure 2F). By applying the ERPIN search approach with other kinds of relaxed constraints to each of the 3 *Pyrococcus* genomes, we did not find any other putative H/ACA sRNA gene. Therefore, we concluded that *Pyrococcus* species probably contain only seven H/ACA sRNAs, which are highly conserved in this genus. In addition, they are also conserved in *Thermococcus kodakarensis* that belongs to the same order as the *Pyrococcus* genus (25).

The seven pyrococcal H/ACA sRNAs are predicted to target 21 sites in rRNAs

We used the RNAMOT software (25) to predict possible target sites in rRNAs for each of the seven identified putative H/ACA sRNAs, and this in the three species studied. One up to four distinct target sites were predicted for each putative pseudouridylation pocket (Figure 4). Taken together, 14 and 7 target sites were predicted in 23S and 16S rRNAs, respectively. Note that motif 1 in sRNA Pab35 and sRNA Pab160 are both predicted to guide modifications at position 922 in 16S rRNA and at position 2672 in 23S rRNA (Table 1). Only 8 of the 17 sites that we predicted for the five previously identified *P. furiosus* H/ACA sRNAs had been proposed (22). Strong conservation of the possibility to form the rRNA–sRNA interactions are observed in the three species (Figure 4). Note that in the non-canonical interaction proposed for sRNA Pab19, two single-stranded residues are found 3′ to the targeted U residue (5′-UGC-3′) (Figure 4G). We verified that none of the seven identified H/ACA sRNAs can act on tRNAs. In addition, no other putative H/ACA

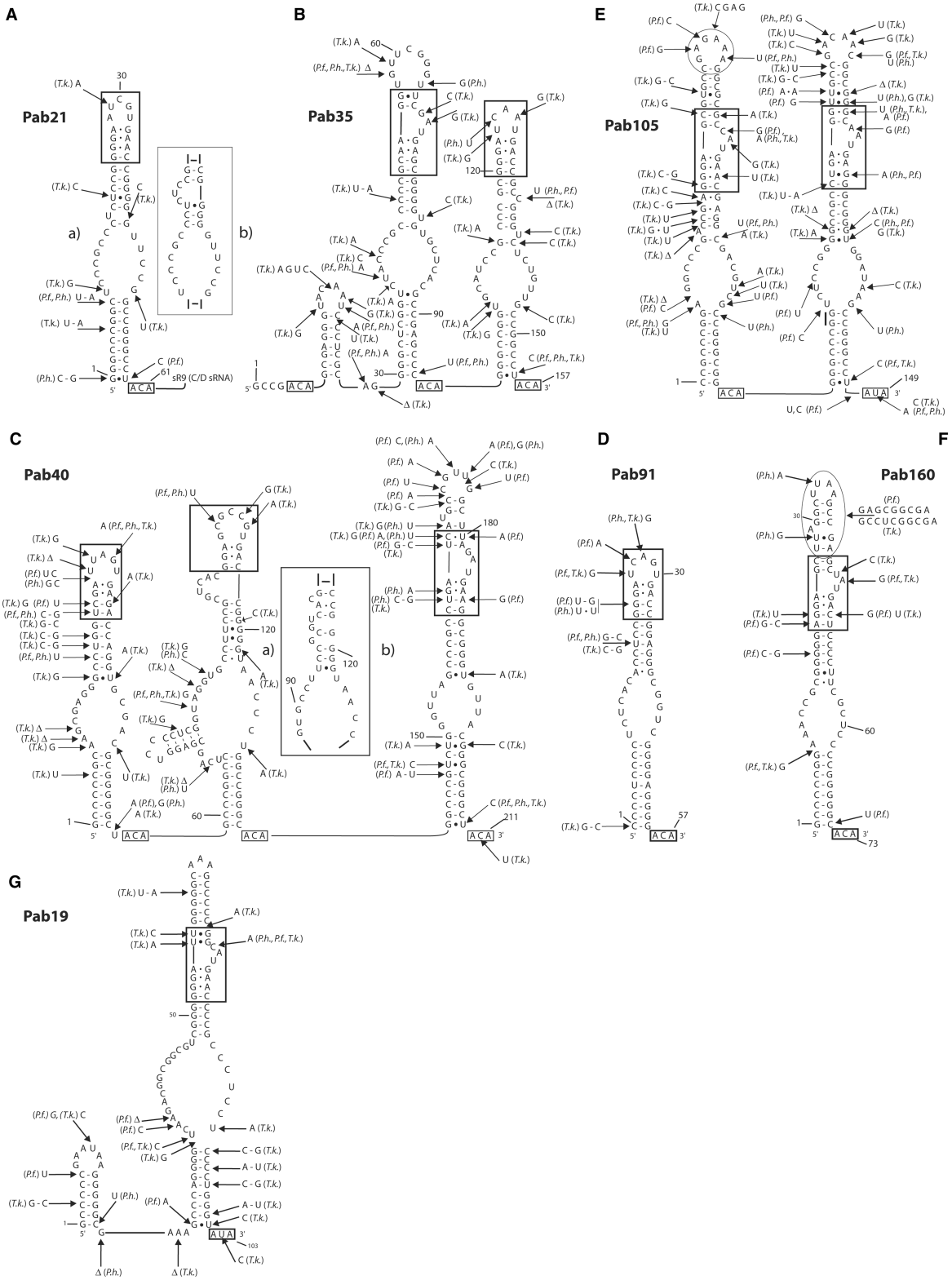


Figure 2. Sequences and proposed secondary structures of the seven pyrococcal H/ACA sRNA candidates. The sequences and proposed secondary structures for the seven candidate *P. abyssi* sRNAs, Pab21 (**A**), Pab35 (**B**), Pab40 (**C**), Pab91 (**D**), Pab105 (**E**), Pab160 (**F**) and Pab19 (**G**) sRNAs, are shown. The ANA sequence at the 3′ end of the RNA and the K-turn or K-loop motif in the apical part of the H/ACA motif are boxed. Base substitutions in the *P. furiosus (P.f.)*, *P. horikoshii* (*P.h.*) and *T. kodakarensis* (*T.k.*) (25) sRNAs are shown. The two putative foldings, which can be proposed for sRNA Pab21 and for motif 2 in sRNA Pab40, are shown (insets in panels A and C). Only the conformation shown in the entire molecule turned to be functional.

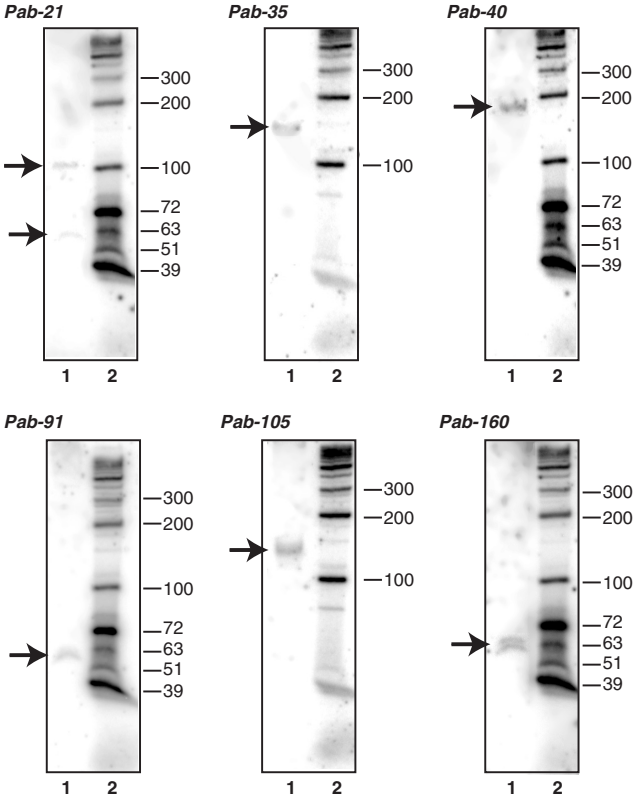


Figure 3. Northern blot analysis of the candidate *P. abyssi* H/ACA sRNAs. The expression of each gene encoding a putative H/ACA sRNA was tested by Northern blot analysis of *P. abyssi* total RNA extracts, as described in Materials and Methods (Lane 1 in each panel). The nucleotide sequences of the specific 5′-end radio-labelled probes used for each sRNA are given in Table S1 in Supplementary data. A 5′-end labelled DNA ladder was loaded in Lane 2. The detected transcripts are shown with an arrow on the left of each autoradiogram.

sRNA that may guide modification of a tRNA was detected in any of the three species.

Experimental search for Ψ residues at the predicted target sites in rRNAs

We used the CMCT-RT approach (47) to test for the presence of Ψ residues at the rRNA positions predicted to be targeted by the seven putative H/ACA sRNAs. To this end, total RNA from *P. abyssi* was treated with CMCT with or without further alkaline incubation. Positions of the alkaline-resistant Ψ residues were detected by primer-extension analyses using a large series of oligonucleotide primers (Table S1 in Supplementary data). Altogether, the 16S and 23S rRNA segments that were probed represented 20 and 27% of the entire molecules, respectively. In particular, large parts of the 23S rRNA domains located at the peptidyl-transferase centre (domains IV and V) were analysed. We identified 3 and 11 Ψ residues in *P. abyssi* 16S and 23S rRNAs, respectively, 8 of them are located in domain V of 23S rRNA.

The experimental analysis confirmed Ψ formation at 12 of the 21 predicted sites (positions 27, 891, 922 in 16S rRNA, and positions 1932, 2549, 2554, 2588, 2672, 2685, 2697, 2794 and 2930 in 23S rRNA) (Table 1 and Figure 5). In addition, the absence of Ψ formation was clearly demonstrated at six of the predicted positions (positions 892, 995 and 1122 in 16S rRNA and 278, 2250, and 2552) (Table 1 and Figure 5). Note that three of these unmodified positions were previously proposed to be modified in *P. furiosus* (22). Based on the present data, they are unlikely to be modified in this species. For three of the positions predicted to be modified in *P. abyssi*, the experimental analysis was obscured by the presence of an RT pause at the level of the stop expected after CMCT modification (positions 1017 in 16S rRNA, 1377 and 2016 in 23S rRNA). Therefore, we could not determine whether U to Ψ conversion occurred at these positions. However, for one of them (2016 in 23S rRNA), formation of a Ψ residue was detected at the corresponding position in two other archaeal species: *H. halobium* (34) and *H. marismortui* (35,36), and in nearly all bacterial and eukaryal organisms which were studied up to now (10).

Interestingly, in the course of this analysis, Ψ residues were found at two positions that were not predicted to be targeted by any of the identified H/ACA sRNAs (positions 2585 and 2603 in 23S rRNA). Even by relaxing the constraints in the ERPIN search, we could not detect putative H/ACA sRNAs for these positions. Therefore, their formation may be sRNA independent.

The seven H/ACA sRNPs are active *in vitro*

To get experimental supports for the relationships that we established between Ψ residues in *P. abyssi* rRNAs and the identified H/ACA sRNAs, we used the H/ACA sRNP *in vitro* reconstitution method which was developed in the laboratory (23). To this end, each of the seven identified H/ACA sRNAs was transcribed, and the four recombinant aCBF5, aNOP10, aGAR1 and L7Ae proteins were produced. Assembly of individual proteins and different combinations of them on the H/ACA sRNAs was tested by electrophoresis-mobility shift assay (EMSA). Each of the identified H/ACA sRNA could be assembled into a sRNP (data not shown), which was a clear demonstration that they were true H/ACA sRNAs. Then, we measured the activities of the reconstituted particles on all their predicted target sites. To this end, we used small *P. abyssi* rRNA fragments containing from 18 up to 31 nts, most of them had a single target U residue located in the middle of the molecule, except 3 substrates that were containing 2 or 3 target U residues, because of the close vicinity of these U residues in 23S rRNA (see Table S2 in Supplementary data). Therefore, the activity of one up to three distinct sRNPs was tested on a given RNA substrate. This activity was measured by the nearest-neighbour method as previously described (23,48). T2 RNase digestion was performed on RNA substrates labelled with [α-^{32}P]ATP, CTP, GTP or UTP. P1 RNase hydrolysis was done on [α-^{32}P]UTP labelled RNA substrates. Each RNA substrate was incubated with the four core proteins (LCNG) at 65°C as

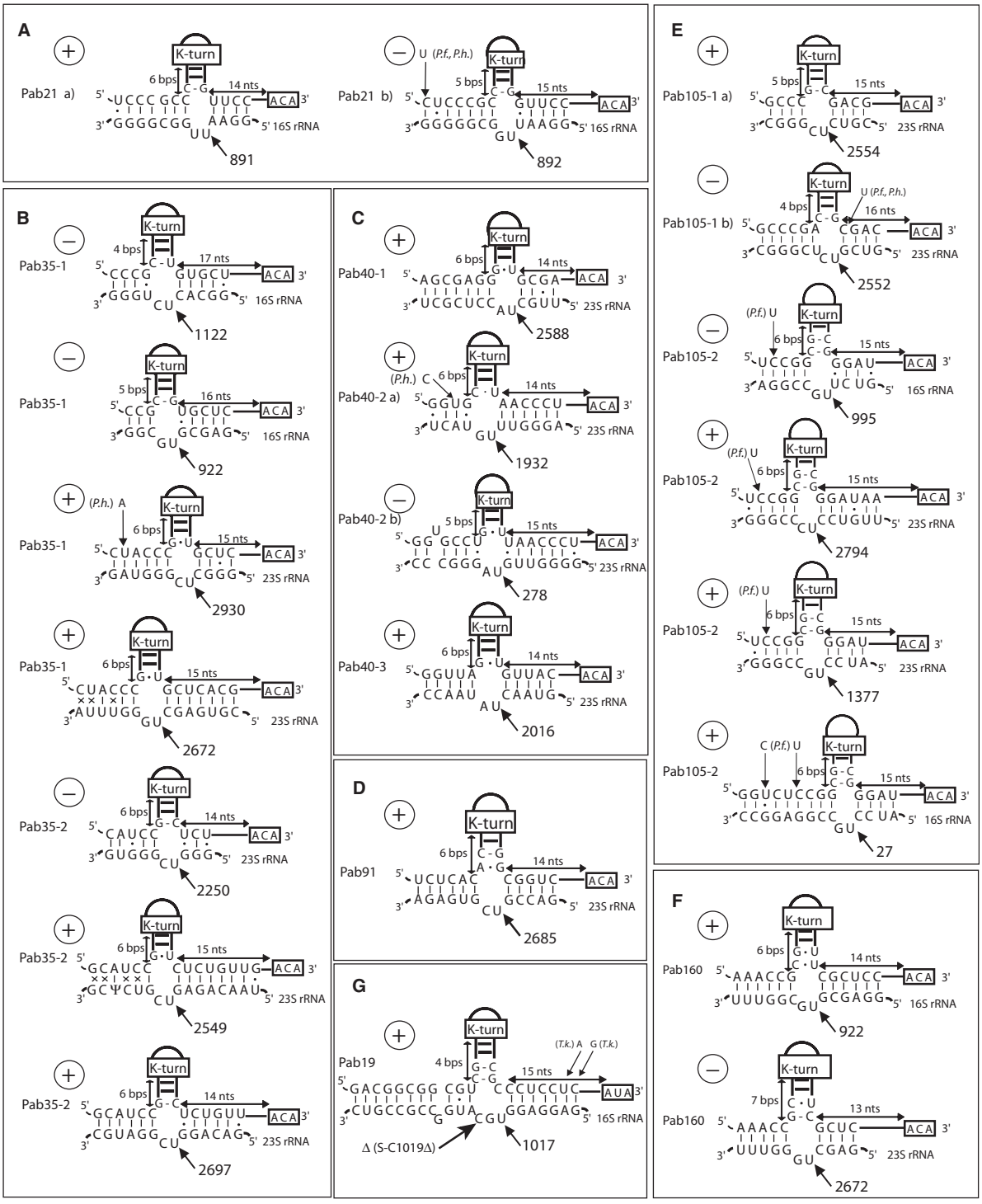


Figure 4. Predicted rRNAs target sites of the 11 H/ACA motifs. Interactions between each H/ACA motif and its putative rRNA target sequences are shown for the seven candidates sRNAs Pab21 (**A**), Pab35 (**B**), Pab40 (**C**), Pab91 (**D**), Pab105 (**E**), Pab160 (**F**) and Pab19 (**G**). The expected targeted U residues are indicated by arrows and numbers giving their positions in the 16S or 23S rRNAs. The distance between the K-turn motif and the pseudouridylation pocket is given in bps, the length between the ANA sequence and the pseudouridylation pocket is given in nts. The two possible intermolecular interactions proposed for sRNA Pab21 and motif 2 in sRNA Pab40 result from the two possible alternative conformations of these H/ACA motifs (Figure 2 Panels A and C). The interactions found to be functional in *in vitro* assays are indicated by (+), the inactive ones are shown by (−).

Table 1. Predicted and experimentally identified target sites of the identified *P. abyssi* H/ACA sRNAs and comparison with *P. furiosus*

P. abyssi H/ACA	Predicted targets	RT-CMCT analysis	*In vitro* activity	*P. furiosus* counterparts	Predicted targets
Pab21	16S rRNA 891	+	++	Pf1	16S rRNA 879
	16S rRNA 892*	–	–		
Pab35	motif 1: 16S rRNA 1122	–	–	Pf6	
	motif 1: 16S rRNA 922	+	–		
	motif 1: 23S rRNA 2930	+	++		motif 1: 23S rRNA 2953
	motif 1: 23S rRNA 2672	+	++		motif 1: 23S rRNA 2695
	motif 2: 23S rRNA 2250	–	–		
	motif 2: 23S rRNA 2549	+	++		motif 2: 23S rRNA 2572
	motif 2: 23S rRNA 2697*	+	+++		motif 2: 23S rRNA 2720*
Pab40	motif 1: 23S rRNA 2588*	+	+++	Pf7	motif 1: 23S rRNA 2611*
	motif 2: 23S rRNA 278*	–	–		
	motif 2: 23S rRNA 1932	+	+		
					motif 2: 23S rRNA 2701
	motif 3: 23S rRNA 2016*	?	+++		motif 3: 23S rRNA 2039*
Pab91	23S rRNA 2685	+	+++	Pfu91	23S rRNA2708
Pab105	motif 1: 23S rRNA 2552*	–	–	Pf3	
	motif 1: 23S rRNA 2554	+	++		motif 1: 23S rRNA 2577
	motif 2: 16S rRNA 995	–	–		
	motif 2: 16S rRNA 27*	+	++		motif 2: 16S rRNA 15*
	motif 2: 23S rRNA 1377	?	+++		motif 2: 23S rRNA 1400
	motif 2: 23S rRNA 2794	+	+++		motif 2: 23S rRNA 2817
Pab160	23S rRNA 2672	+	–	Pf9	
	23S rRNA 922*	+	+++		16S rRNA 910*
Pab19	16S rRNA1017	?	+++	Pfu19	16S rRNA 1005

The predicted target positions in 16S and 23S rRNAs are indicated for each of the candidates H/ACA motifs. Detection by CMCT-RT analysis of a Ψ residue at the predicted position in *P. abyssi* rRNAs is indicated by '+' in the second lane. '?' indicates the presence of an RT pause that obscured the analysis. Detection of an *in vitro* activity of the reconstituted H/ACA sRNP at the predicted site is indicated by '+','++' or '+++' in the third lane. The number of '+' is proportional to the rate of modification detected after a 80 min incubation (25–50%, 50–80% and 80–100%, respectively). Column 5 gives the name of the *P. furiosus* counterpart sRNAs (22,24). The target positions that were previously predicted for these *P. furiosus* Pf1, Pf3, Pf6, Pf7 and Pf9 sRNAs (22,24) are indicated by an asterisk in column 2. Based on the rules that we established from our *P. abyssi* experimental data, we predicted modification positions for both the previously identified *P. furiosus* sRNAs and the two additional sRNAs detected in this study (column 6). The validated previously proposed modified positions are marked by an asterisk.

previously described, in the presence or the absence of the H/ACA sRNA (23). Then, the RNAs were extracted, digested, and fractionated by thin layer chromatography (as described in Materials and Methods) (Figure 6).

Only 15 of the 21 predicted sites were found to be modified by the reconstituted sRNPs and the rates of U to Ψ conversions were ranging from 30 to 100% (Figure 6 and Table 1). The rRNA segments predicted to be modified by two distinct H/ACA motifs were in fact modified by only one of them (motif 1 in sRNA Pab35 modifies position 2672 in 23S rRNA and sRNA Pab160 acts at position 922 in 16S rRNA). The data obtained confirmed that each of the 11 H/ACA motifs is capable to guide pseudouridylation. Even the Pab19 H/ACA motif, which has a large internal loop and forms a non-canonical interaction with its substrate, was active (Figure 6D). Deletion of the additional single-stranded C residue, which is located 3′ to the 5′-UN-3′dinucleotide (mutant S-C1019Δ) (Figures 4G and 6D), showed that the Pab19 sRNP modifies the WT and mutated RNA substrates at very similar rates. Among the two possible conformations of motif 2 in sRNA Pab40, only conformation denoted a in Figure 2C is active *in vitro*. Interestingly, this structure includes an additional stem-loop in the 5′-guide strand and a poorly stable helix 2 (Figure 2C). This may explain the low yield of U to Ψ conversion found for this H/ACA motif as compared to the other *P. abyssi* H/ACA motifs (Figure 6C).

Some of the H/ACA motifs can guide modification at more than one position in rRNAs

Some of the H/ACA motifs can guide modifications at two or even three distinct positions. For instance, altogether, motifs 1 and 2 in sRNA Pab35 can guide U to Ψ conversion at positions 2672, 2930, 2549 and 2697 in 23S rRNA. Motif 1 in sRNA Pab105 only guides modification at one position (2554 in 23S rRNA), while motif 2 in this sRNA can act at positions 1377 and 2794 in 23S rRNA and 27 in 16S rRNA. Therefore, both sRNAs Pab35 and Pab105 can guide modifications at four positions in rRNAs. Interestingly, we found that all the H/ACA sRNAs that contain a single H/ACA motif (Pab21, Pab91, Pab160 and Pab19) guide modification at a unique position in rRNAs. Altogether, the 11 H/ACA motifs of the 7 *P. abyssi* sRNAs can guide U to Ψ conversion at 15 sites in the *P. abyssi* rRNAs (Table 1). Presence of Ψ residue was detected at 12 of these sites (Figure 5 and Table 1). As mentioned above, for technical reasons, modification could not be tested experimentally at the three other positions. However, the strong activity measured *in vitro* at these three rRNA positions is a strong argument for the occurrence of these modifications *in vivo*.

Constraints on the H/ACA sRNA structure and H/ACA sRNA–target RNA interaction

Based on the above data, we tried to define rules for H/ACA sRNA structure and interactions. In several of

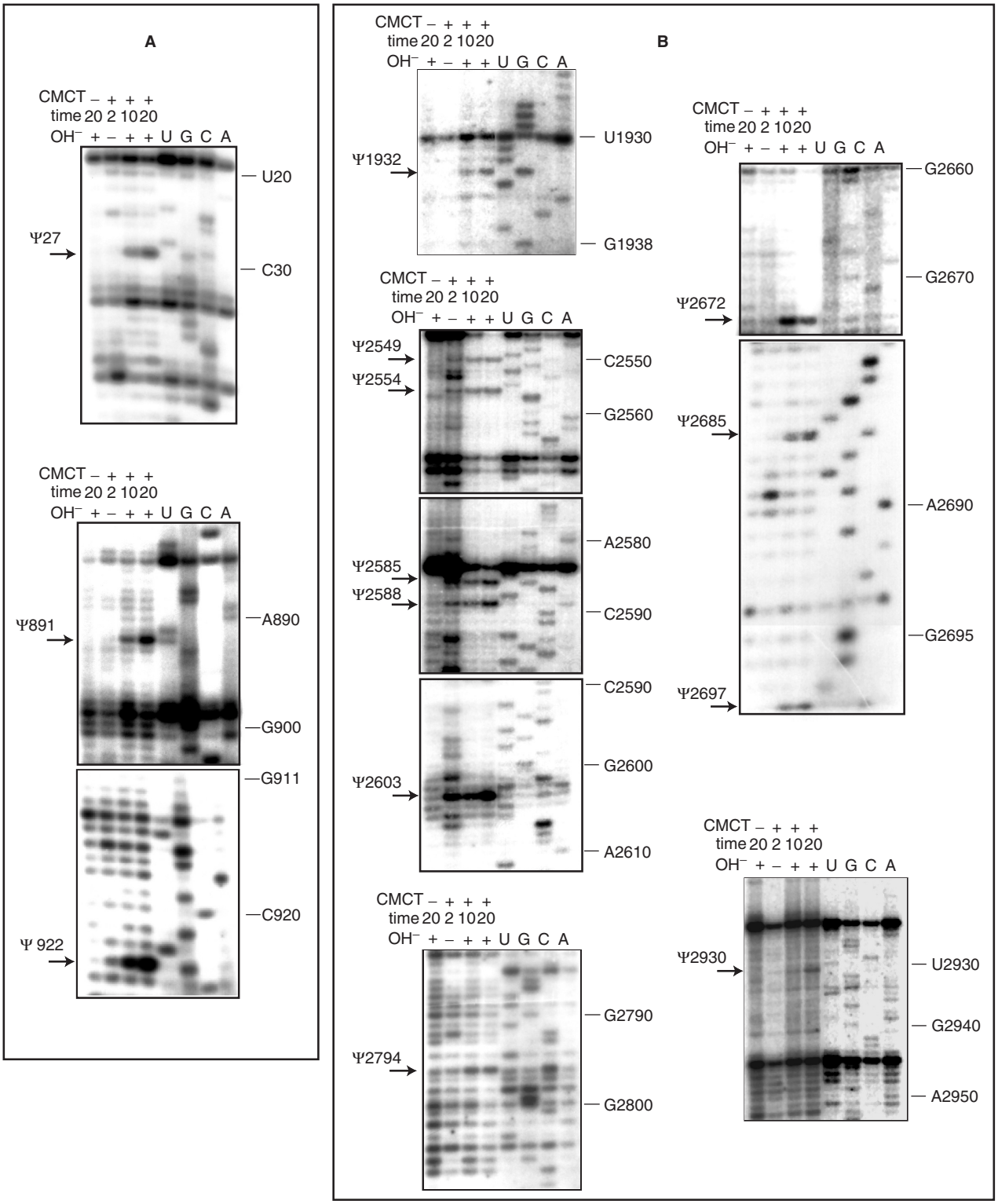


Figure 5. Identification of Ψ residues in the *P. abyssi* 16S (Panel A) and 23S (Panel B) rRNAs by the RT-CMCT method. The *P. abyssi* total RNA was treated in the absence (–) or the presence of CMCT (+), for 2, 10, or 20 min as indicated on the top of the lanes. The CMCT treatment was (+) or was not (–) followed by an alkaline treatment at pH 10.4 as indicated above the lanes. The appearance of an RT stop of increasing intensity after this treatment reveals the presence of a Ψ residue. They are indicated by horizontal arrows. Lanes U, G, C and A correspond to the RNA sequencing ladder. Positions of residues in 16S or 23S rRNAs are given on the right side of the autoradiograms.

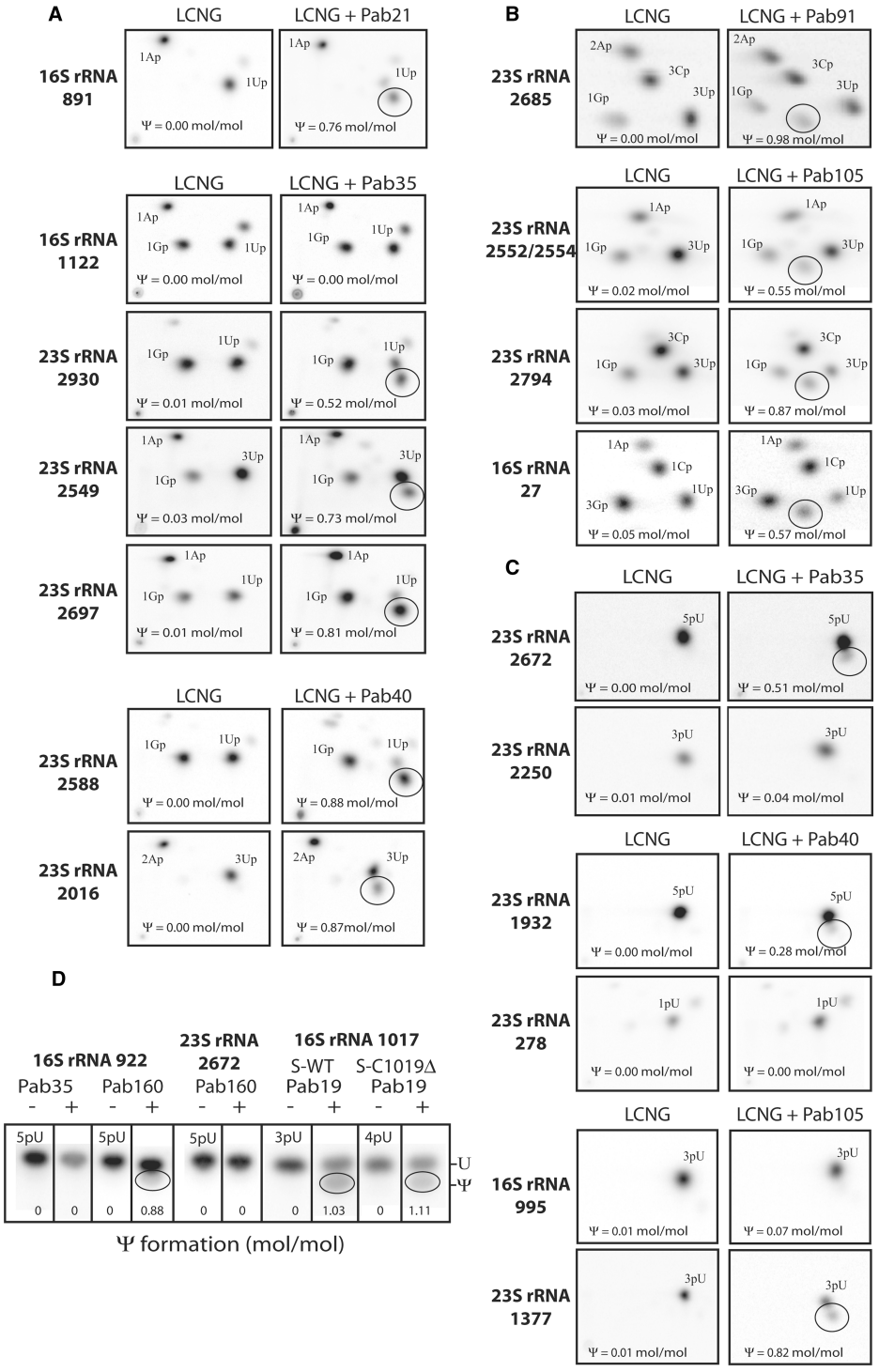


Figure 6. *In vitro* activity tests of the seven reconstituted H/ACA sRNPs. H/ACA sRNP particles were assembled by using 4 pmol of *in vitro* transcribed H/ACA sRNA, the four recombinant proteins L7Ae (L), aCBF5 (C), aNOP10 (N) and aGAR1 (G) (LCNG) (200 nM each) and 150 fmol of radio-labelled *in vitro* transcribed RNA substrate as previously described (23). The putative target position in the RNA substrate is given on the left side of the chromatograms. After a 80-min incubation at 65°C, the RNAs were extracted and digested with T2 (Panels A and B) or P1 nuclease (Panels C and D). The nucleotide 3′ or 5′ monophosphate, that are respectively released, were fractionated by 2D (Panel A, B and C) or 1D (Panel D) thin layer chromatography. When RNase T2 was used for the digestion, the RNA template was labelled by incorporation of [α-^{32}P] ATP, CTP, GTP or UTP depending on the identity of the residue located 3′ to the targeted U residue. When P1 nuclease was used for digestion, labelling was achieved by [α-^{32}P] UTP incorporation. 2D TLC on cellulose plates were performed using either the N1-N2 (Panels A and C), or the N1-R2 (Panel B) buffers and the N1 buffer was used for 1D TLC. Positions of Ap, Cp, Gp, Up and Ψp or pU and pΨ in the chromatograms are indicated as well as their numbers in the substrate RNAs. The Ψ spots are circled. The indicated numbers of Ψ moles formed per mole of RNA substrates were calculated as explained in Materials and Methods.

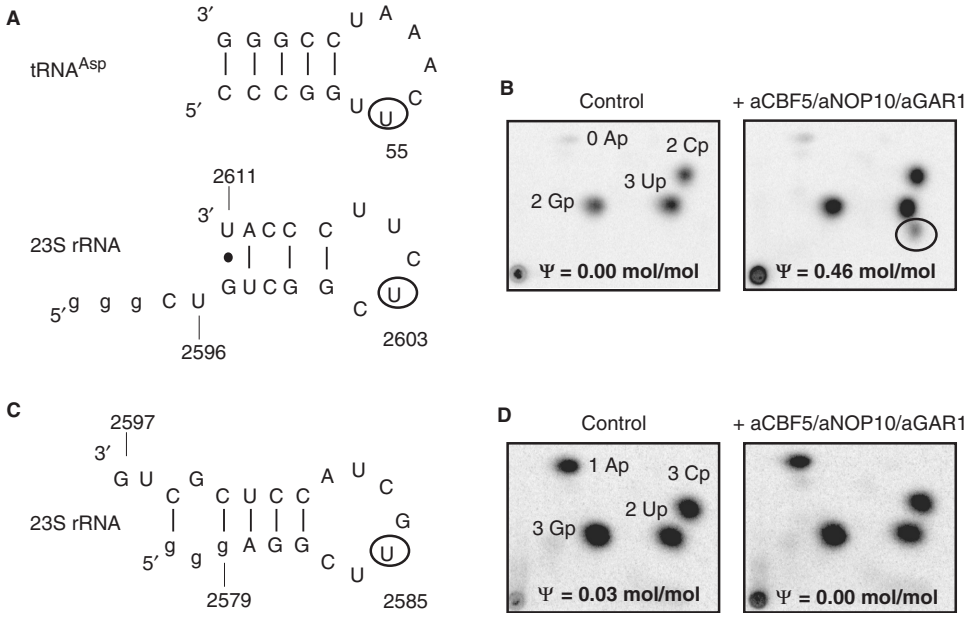


Figure 7. The aCBF5/aNOP10/aGAR1 complex catalyses U to Ψ *in vitro* conversion at position 2603 in a 23S rRNA fragment. Panel A: Comparison of the 2D structure of the TΨC arm of the *P. abyssi* $tRNA^{Asp}$ with the 2D structure that can be formed by the *P. abyssi* 23S rRNA fragment used to test the pseudouridylation activity of the aCBF5/aNOP10/aGAR1 at position 2603 in this rRNA. The three G residues located at the 5′ extremity of the rRNA substrate arose from the T7 promoter. The U residues targeted by the aCBF5/aNOP10/aGAR1 complex are circled (Panel B). Panel B: The 20-nt long RNA substrate (150 fmol) was incubated with the aCBF5/aNOP10/aGAR1 protein complex (200 nM each). The experiment was performed in conditions described in Figure 6, except the absence of guide RNA. A control experiment was performed in the absence of the protein complex. Positions of the Ap, Cp, Gp, Up and Ψp spots are indicated on the chromatograms, as well as their specific ratios in the target RNA. The Ψ spot is circled. The yield of Ψ formation was assessed and is indicated. Panel C: Proposed secondary structure for the 23S rRNA fragment used to test the *in vitro* activity of the aCBF5/aNOP10/aGAR1 protein complex at position 2585 in the *P. abyssi* 23S rRNA. Panel D: Absence of *in vitro* activity of the aCBF5/aNOP10/aGAR1 complex at position 2585 in the *P. abyssi* 23S rRNA.

the predicted sRNA–target RNA interactions that turned to be non-functional, helix H2 was too short (4 bps) or too long (7 bps). The most frequent length in functional interactions (14 out of 15) is of 6 bps. Some of the non-functional interactions also showed too long or too short distances between the ACA trinucleotide and the targeted U residue (17 or 13 nts). The distance in the active interactions is of 15 or 14 nts. Interestingly also, the base-pair interaction established with the sRNA 3′-guide sequence is more important than that formed with the 5′-guide sequence. This is exemplified by the efficient modifications at positions 2549 and 2672 in 23S rRNA, which are guided by the H/ACA motifs 1 and 2 of sRNA Pab35, respectively (Figure 4). In both cases, the 5′-guide sequence forms a weak base-pair interaction with the rRNA substrate. On the contrary, in spite of a canonical length of helix H2 (6 bps) and a canonical distance between the ACA trinucleotide and the target U residue (15 nts), motif 2 in sRNA Pab105 did not guide modification at position 995 in 16S rRNA, probably because of the low stability of the interaction formed with the sRNA 3′-guide sequence (4 bps including two G.U pairs). Therefore, the length of helix H2, the distance between the ACA triplet and the targeted residue as well as the stability of the base-pair interaction established by the 3′-guide sequence and the rRNA substrate are essential criteria for activity. In contrast, unexpectedly, the presence of the 5′-UN-3′ single-stranded dinucleotide between the two intermolecular interactions is not a strict rule.

One of the two orphan Ψ residues in 23S rRNA can be formed *in vitro* without guide sRNA

As we detected no guide H/ACA sRNA for residues Ψ2585 and Ψ2603 in *P. abyssi* 23S rRNA, and as the aCBF5/aNOP10/aGAR1 complex can modify position 55 in tRNAs in the absence of guide sRNA (50–52), we tested the *in vitro* activity of this complex at these two 23S rRNA positions. The assays were performed using two 20-nt long rRNA fragments (Table S1) containing, respectively, residue U2585 or U2603 (Figure 7). The aCBF5/aNOP10/aGAR1 complex was active on the rRNA fragment containing residue U2603 (46% yield) (Figure 7), but not on that containing residue U2585. Interestingly, we observed that the sequence containing the residue 2603 can be folded into a stem-loop structure showing some similarity with the TΨC stem-loop of tRNA (Figure 7). We concluded that formation of residue Ψ2603 may be catalyzed by the free aCBF5-aNOP10-aGAR1 complex, while an unidentified catalyst may act at position U2585.

DISCUSSION

Application of the various computational approaches that we developed for the search of H/ACA sRNAs in archaeal genomes turned to be highly efficient, since two of the seven sRNAs detected in *P. furiosus* were not found by other approaches previously applied to one of this species (22,24). Our blind detection of C/D box

sRNAs in conserved ICRs allowed the detection of 45 out of the 49 known C/D box sRNAs which are conserved in the three pyrococcal species. This finding illustrates the efficiency of our ICR-based approach for the search of non-coding RNAs of unknown structural characteristics. Up to now, computational predictions of the H/ACA sRNA–rRNA targets were not reliable. This is illustrated by the numerous corrections that we made in previous predictions. Our experimental search for Ψ residues in *P. abyssi* rRNAs, together with the use of the H/ACA sRNP reconstitution and activity assays, turned to be essential to identify the sites which are really targeted by the identified sRNAs. However, by using the structural rules that we established for functional sRNA–rRNA interactions, computational predictions will be more reliable.

A high number of Ψ residues in *P. abyssi* rRNAs as compared to other archaea

Based on experimental analysis of Ψ residues in rRNAs and reconstitution of H/ACA sRNPs, the *P. abyssi* rRNA regions that we studied (20 and 27% of the 16S and 23S rRNAs, respectively) probably contain 17 Ψ residues. These rRNA regions were selected because they contain the highest number of post-transcriptionally modified residues in all living organisms (10). However, we cannot exclude the possibility that some additional Ψ residues are formed without the use of guide RNA, in *P. abyssi* rRNA segments located outside of these regions. Up to now, 17 Ψ residues is the highest number of Ψ residues found in archaeal rRNAs. Only three and four Ψ residues were found upon complete analysis of the *H. marismortui*, *H. halobium* 16S and 23S rRNAs, respectively (34–36). Six Ψ residues were detected in domains II, IV and V of the *S. acidocaldarius* 23S rRNA and five Ψ residues were found in *A. fulgidus* rRNAs when looking for H/ACA sRNA target sites (21,37). Interestingly, the two halophile species, *H. marismortui* and *H. halobium*, which have the smallest number of Ψ residues, grow at 50 and 42°C, respectively. *S. acidocaldarius* and *A. fulgidus* grow at 80 and 83°C, respectively, whereas *P. abyssi* optimally grows at 98°C. Therefore, as already proposed for archaeal 2′-*O*-methylations (53), there may be a correlation between the number of Ψ residues in archaeal rRNAs and the growth temperature of the organism. Confirmation of this statement requires further analysis on a larger number of archaeal species growing at different temperatures.

Ψ residues in *P. abyssi* 23S rRNA are concentrated at the PTC

Interestingly, 2 and 8 of the 13 Ψ residues expected to be present in *P. abyssi* 23S rRNA are located in the functional domains IV and V, respectively (Figure 8). Noticeably, residue 2016 in domain IV corresponds to one of two highly conserved Ψ residues in stem-loop structure 69 (SLS69). Its conservation is probably explained by the high functional importance of SLS69: (i) it contacts both the 16S rRNA and the tRNA bound at the acceptor site (A site) (12,14), (ii) it is located at the subunit interface and, (iii) it was proposed to play a role in tRNA translocation (9,54,55) and in translation termination (12,16). Formation of two Ψ residues in the terminal loop of SLS69, including the *P. abyssi* Ψ2016 counterpart, was found to confer a growth advantage in *S. cerevisiae* (12,17).

Eight Ψ residues are concentrated in domain V of the *P. abyssi* 23S rRNA. This domain is expected to be directly involved in the catalytic activity of the PTC. Interestingly, these 8 Ψ residues are all located in helical segments (Figure 8). They may stabilize the conformation of this 23S rRNA region at the high growth temperature of *P. abyssi*. A cooperative effect of Ψ residues in stabilizing RNA conformation, was previously proposed (10). A need for this cooperative effect may explain the presence of 4, 2 and 2 Ψ residues in close vicinity in helices 89, 90 and 93, respectively. Only three of these eight Ψ residues were detected in several eukarya, including *S. cerevisiae* (Figure 8) (34,56).

Altogether, four pseudouridylation sites in *P. abyssi* 23S rRNAs are highly conserved in the large eukaryal rRNAs (positions 2016 in domain IV, 2549, 2588 and 2603 in domain V) (Figure 8). In addition, residues Ψ2930 detected in domain VI is also present in mouse and human rRNAs. Position 2698 in *P. abyssi* 23S rRNA corresponds to a frequently pseudouridylated position in eukarya. Instead of U2698, U2697 is converted into a Ψ residue in *P. abyssi*. Altogether, this comparison of *P. abyssi* and eukaryal pseudouridylation in rRNAs suggests that conserved pseudouridylation sites in the large rRNA may have appeared early in evolution. Interestingly, one of the conserved Ψ residues in domain V (Ψ2603) is not guided by an H/ACA sRNA in *P. abyssi*.

A base substitution in sRNA Pab40 may generate a different specificity as compared to sRNAs Afu4, Pf7 and Pho40

Like the *P. abyssi* Pab40 sRNA, the Afu4, Pf7, and Pho40 sRNAs contain three H/ACA motifs. The target sites proposed for motifs 1 and 3 in Afu4 and Pf7 sRNAs (21,22) are identical to the ones determined experimentally for motifs 1 and 3 in Pab40 sRNA, respectively. In contrast, whereas motif 2 in the Afu4 and Pf7 sRNAs were proposed to guide modification at positions 2601 and 302 in the *A. fulgidus* and *P. furiosus* 23S rRNAs, respectively (corresponding to positions 2647 and 278 in *P. abyssi* 23S rRNA, respectively), we found that motif 2 in sRNA Pab40 is active at position 1932 in *P. abyssi* 23S rRNA. A difference of specificity between motif 2 in Pab40 and motifs 2 in Afu4, PFf7, and Pho40 sRNAs can be explained by a point mutation in the K-turn motif (a G to U substitution at position 85 as referred to the Pab40 numbering, Figure S1). Indeed, this base substitution modifies the secondary structure of motif 2 and its guiding specificity (Figure S1). Nevertheless, the target sites previously proposed for motifs 2 in Afu4 and Pf7 sRNAs are not in agreement with the structural rules that we established for active H/ACA sRNA–rRNA interactions. According to these rules, positions 2632 in the *A. fulgidus* 23S rRNA and 2701 in the *P. furiosus* 23S rRNA, that both correspond to position 2678 in *P. abyssi* 23S rRNA, are expected to be the true target

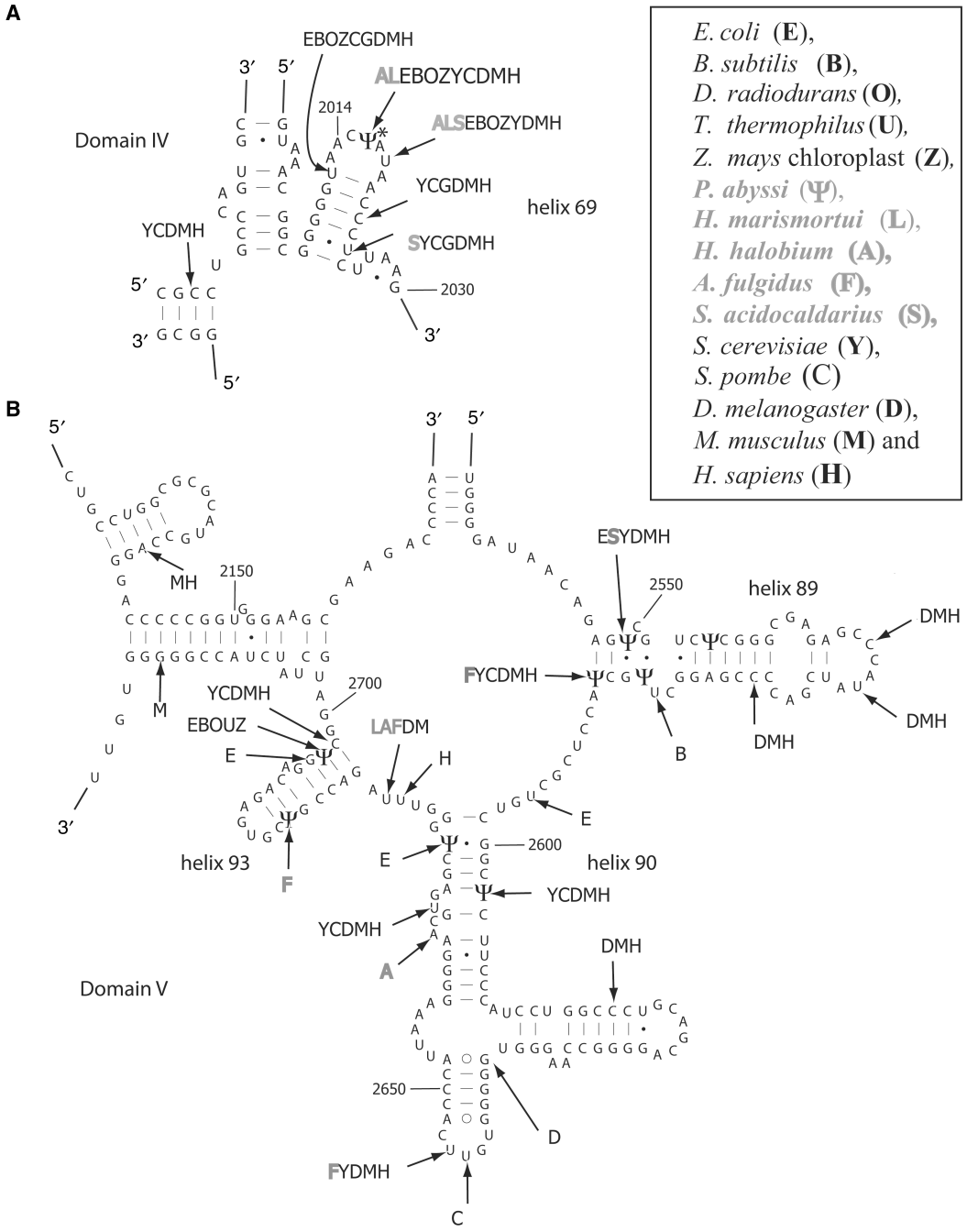


Figure 8. Location of Ψ residues in domains IV (A) and V (B) of the *P. abyssi* 23S rRNA. The secondary structure of the *P. abyssi* 23S rRNA is adapted from one of the *Thermococcus celer* 23S rRNA (M67497) (Gutell website, www.rna.icmb.utewas.edu). Numbering of residues is that of *P. abyssi* 23S rRNA. The Ψ residues detected in this work are indicated by Ψ symbol. The one in domain IV, which could not be detected in 23S rRNA because of an RT pause but was formed *in vitro*, is marked by an asterisk. Positions of pseudouridylation in *E. coli*, *B. subtilis*, *D. radiodurans*, *T. thermophilus*, *Z. mays* chloroplasts, *H. marismortui*, *H. halobium*, *A. fulgidus*, *S. acidocaldarius*, *S. cerevisiae*, *D. melanogaster*, *M. musculus* and *H. sapiens* are indicated by arrows marked by E, B, O, U, Z, L, A, F, S, Y, D M and H, respectively. Archaeal species are shown in grey.

sites (Figure S1). U 2678 is not converted into a Ψ residue in *P. abyssi*, however, this is the case in *H. marismortui* (35,36), *H. halobium* (34), *Drosophila melanogaster* and *Mouse musculus* (10) (Figure 8). The absence of modification at position 2678 in the *P. abyssi* 23S rRNA shows how a single point mutation can dramatically modify the specificity of an H/ACA motif.

Note that two other proposed target sites in *P. furiosus* rRNAs (corresponding to positions 892 in 16S rRNA and 2552 in *P. abyssi* 23S rRNA and which were expected to be guided by sRNAs Pf1 and Pf3 respectively, Table 1), are invalidated by our experimental rules. In addition, some *bona fide* target sites were not previously predicted in *P. furiosus* rRNAs (Table 1). Altogether, these data

strengthen the importance to verify experimentally the proposed H/ACA target sites.

Poor conservation of pseudouridylation sites between archaeal orders

Except for motif 2 in sRNA Pab40, H/ACA sRNAs and their target sites are conserved in the three *Pyrococcus* species studied as well as in the phylogenetically related *T. kodakarensis* species (25). In contrast, conservation of Ψ positions is poor between species of different orders. Only one of the three Ψ residues found in the *H. halobium* and *H. marismortui* 23S rRNAs and only one of the six Ψ residues found in the *S. acidocaldarius* 23S rRNA are conserved in *P. abyssi* (Figure 8). *A. fulgidus* is more closely related to *Pyrococcus* species than *Sulfolobus*, *Halobacterium*, and *Haloarcula* species. Accordingly, our experimental data show that three of the six Ψ residues detected in the *A. fulgidus* 23S rRNA are conserved in *P. abyssi* as well as one modified position in 16S rRNA.

The four Ψ residues detected in 16S rRNA are located in functional areas of the 30S subunit

The location of Ψ residues in the bacterial and eukaryal SSU rRNAs is rather variable from one species to the other. Interestingly, two of the four Ψ residues that we detected in the *P. abyssi* 16S rRNA (positions 27 and 891), are located within or very close to the essential central pseudoknot of the 16S rRNA (57). They may play a role in its formation or stability. This central pseudoknot is in close vicinity of the P site (58). Residue Ψ922 also belongs to a 16S rRNA segment located at the P site, and residue Ψ1017 belongs to a segment involved in A site formation. Hence, the four Ψ residues present in the *P. abyssi* 16S rRNA are located at or very near the A and P sites. Noticeably, no Ψ residue was detected in the 16S rRNA from *H. volcanii* while *S. solfataricus* 16S rRNA was estimated to contain five Ψ residues based on mass spectrometry analysis (59).

The aCBF5/aNOP10/aGAR1 complex may act on rRNA without guide sRNA

Among the small *P. abyssi* rRNA substrates that we used to test the activity of reconstituted sRNPs (Table S2), only that containing residue U2603 was modified in the absence of H/ACA sRNA (Figure 7). Therefore, the aCBF5/aNOP10/aGAR1 complex may be a specific catalyst for this position in 23S rRNA. The stem-loop structure formed by the small RNA substrate used in the assay is not the one expected to be formed in the 50S subunit (Figures 7 and 8). However, we cannot exclude the possibility that it is formed at some stage during 23S rRNA synthesis or 50S subunit assembly. Although a stable stem-loop structure could also be formed by the small substrate containing residue U2585 (Figure 7), the aCBF5/aNOP10/aGAR1 complex did not modify it. No activity was detected with the recombinant Pus10 enzyme (data not shown). Two other RNA: Ψ synthases are expected to be present in archaea, the *E. coli* TruA and TruD (Pus7) homologues (60). The implication of Pus7 is unlikely, since residue 2585 is not located in a sequence that fits the consensus sequence recognized by this enzyme (RSUNΨAR (R = purine, S = G/C, N = any nucleotide) (61 and our unpublished data). Therefore, it would be interesting to test the activity of the TruA homologue at position 2585 in *P. abyssi* 23S rRNA.

In agreement with the absence of *E. coli* RluE, RluB and RluC homologues in *P. abyssi*, the pseudouridylations at the three positions modified by these enzymes in *E. coli* are catalyzed by one H/ACA sRNA, Pab35.

Structural determinants for H/ACA sRNA specificity

Our data shed light on structural determinants of the sRNA-rRNA interaction that will be useful for the identification of new H/ACA sRNAs and their target sites. They can be explained taking structural data into account: (i) NMR analysis of the H/ACA snoRNA–rRNA interaction showed that the P1S and P2S intermolecular helices formed with the 3′- and 5′-guide sequences, respectively, can be coaxially stacked on helices H1 and H2 of the sRNA (32,33), and in this structure, the substrate is folded into a U-shape, with the targeted U residue protruding in the middle (32), (ii) based on the sRNP crystal structure, the H2-P2S pseudo-helix likely interacts with protein L7Ae bound to the K-turn or K-loop (22–24) and to a lesser extent with aNOP10 and aCBF5 through helix H2 (31), (ii) the various X-ray structures which have been established show that several interactions are formed between the aCBF5 and aNOP10 proteins (27–31) and the L7Ae and aNOP10 proteins (31) and (iv) finally, the H1-P1S pseudo-helix is expected to interact with protein aCBF5 bound to the ACA triplet (23,24,30,31). Therefore, it is highly conceivable that variations of the distances that separate the ACA box and the targeted uridine residue on the one hand (14 or 15-nt long in archaea, *versus* 16-nt long in eukarya) (62), and the K-turn and targeted U residue on the other hand (5 or 6 bps), disturb the positioning of aCBF5 relative to the targeted U residue and/or prevent some protein–protein interactions in the core protein structure (23,51). Importantly, we noticed that helix H2 never contains more than one bulged nucleotide.

Noticeably, for the first time, we show a greater importance of the rRNA interaction formed with the 3′-guide sequence (P1S), compared to that formed with the 5′-guide sequence (PS2). This difference can be explained by the close contact of aCBF5 with the 3′-guide sequence, compare to its limited contact with the 5′-guide sequence in the sRNP 3D structure (30). Remarkably, also the minimal length of the overall rRNA–sRNA base-pair interaction seems to be shorter in archaea compared to eukaryal system. This decreased stringency in archaea may explain why, in archaea but not in eukarya, a given H/ACA motif can guide Ψ formation at up to three different positions in rRNAs.

SUPPLEMENTARY DATA

Supplementary Data is available at NAR Online.

ACKNOWLEDGEMENTS

S. Muller and J-B Fourmann were doctoral fellows supported by the French Ministère de la Recherche et des Nouvelles Technologies (MRNT). The work was supported by the Centre National de la Recherche Scientifique (CNRS), the MRNT and the Pôle de Recherche Scientifique et Technologique (PRST) « Bioingénierie » of Région Lorraine. Fujihiko Matsunaga and Patrick Forterre (UMR CNRS 8621, Orsay, France) are warmly thanked for their generous gift of *P. abyssi* cells. J. Ugolini is acknowledged for her efficient technical assistance. Funding to pay the Open Access publication charge was provided by CNRS.

Conflict of interest statement. None declared.

REFERENCES

1. Rozenski,J., Crain,P.F. and McCloskey,J.A. (1999) The RNA Modification Database: 1999 update. *Nucleic Acids Res.*, **27**, 196–197.
2. Arnez,J.G. and Steitz,T.A. (1994) Crystal structure of unmodified $tRNA^{Gln}$ complexed with glutaminyl- tRNA synthetase and ATP suggests a possible role for pseudo-uridines in stabilization of RNA structure. *Biochemistry*, **33**, 7560–7567.
3. Auffinger,P. and Westhof,E. (2001) An extended structural signature for the tRNA anticodon loop. *RNA*, **7**, 334–341.
4. Yu,Y.T., Shu,M.D. and Steitz,J.A. (1998) Modifications of U2 snRNA are required for snRNP assembly and pre-mRNA splicing. *EMBO J*, **17**, 5783–5795.
5. Newby,M.I. and Greenbaum,N.L. (2002) Sculpting of the spliceosomal branch site recognition motif by a conserved pseudouridine. *Nat. Struct. Biol.*, **9**, 958–965.
6. Segault,V., Will,C.L., Sproat,B.S. and Luhrmann,R. (1995) In vitro reconstitution of mammalian U2 and U5 snRNPs active in splicing: Sm proteins are functionally interchangeable and are essential for the formation of functional U2 and U5 snRNPs. *EMBO J.*, **14**, 4010–4021.
7. Yang,C., McPheeters,D.S. and Yu,Y.T. (2005) Ψ35 in the branch site recognition region of U2 small nuclear RNA is important for pre-mRNA splicing in *Saccharomyces cerevisiae*. *J. Biol. Chem.*, **280**, 6655–6662.
8. Veldman,G.M., Klootwijk,J., de Regt,V.C., Planta,R.J., Branlant,C., Krol,A. and Ebel,J.P. (1981) The primary and secondary structure of yeast 26S rRNA. *Nucleic Acids Res.*, **9**, 6935–6952.
9. Decatur,W.A. and Fournier,M.J. (2002) rRNA modifications and ribosome function. *Trends Biochem. Sci.*, **27**, 344–351.
10. Ofengand,J. (2002) Ribosomal RNA pseudouridines and pseudouridine synthases. *FEBS Lett.*, **514**, 17–25.
11. King,T.H., Liu,B., McCully,R.R. and Fournier,M.J. (2003) Ribosome structure and activity are altered in cells lacking snoRNPs that form pseudouridines in the peptidyl transferase center. *Mol. Cell*, **11**, 425–435.
12. Liang,X.H., Liu,Q. and Fournier,M.J. (2007) rRNA modifications in an intersubunit bridge of the ribosome strongly affect both ribosome biogenesis and activity. *Mol. Cell*, **28**, 965–977.
13. Meroueh,M., Grohar,P.J., Qiu,J., SantaLucia,J.,Jr., Scaringe,S.A. and Chow,C.S. (2000) Unique structural and stabilizing roles for the individual pseudouridine residues in the 1920 region of *Escherichia coli* 23S rRNA. *Nucleic Acids Res.*, **28**, 2075–2083.
14. Mengel-Jorgensen,J., Jensen,S.S., Rasmussen,A., Poehlsgaard,J., Iversen,J.J. and Kirpekar,F. (2006) Modifications in *Thermus thermophilus* 23 S ribosomal RNA are centered in regions of RNA-RNA contact. *J. Biol. Chem.*, **281**, 22108–22117.
15. Sumita,M., Desaulniers,J.P., Chang,Y.C., Chui,H.M., Clos,L.,2nd and Chow,C.S. (2005) Effects of nucleotide substitution and modification on the stability and structure of helix 69 from 28S rRNA. *RNA*, **11**, 1420–1429.
16. Ejby,M., Sorensen,M.A. and Pedersen,S. (2007) Pseudouridylation of helix 69 of 23S rRNA is necessary for an effective translation termination. *Proc. Natl. Acad. Sci. USA*, **104**, 19410–19415.
17. Badis,G., Fromont-Racine,M. and Jacquier,A. (2003) A snoRNA that guides the two most conserved pseudouridine modifications within rRNA confers a growth advantage in yeast. *RNA*, **9**, 771–779.
18. Hamma,T. and Ferre-D'Amare,A.R. (2006) Pseudouridine synthases. *Chem. Biol.*, **13**, 1125–1135.
19. Kiss,T. (2002) Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell*, **109**, 145–148.
20. Decatur,W.A. and Fournier,M.J. (2003) RNA-guided nucleotide modification of ribosomal and other RNAs. *J. Biol. Chem.*, **278**, 695–698.
21. Tang,T.H., Bachellerie,J.P., Rozhdestvensky,T., Bortolin,M.L., Huber,H., Drungowski,M., Elge,T., Brosius,J. and Huttenhofer,A. (2002) Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proc. Natl. Acad. Sci. USA*, **99**, 7536–7541.
22. Rozhdestvensky,T.S., Tang,T.H., Tchirkova,I.V., Brosius,J., Bachellerie,J.P. and Huttenhofer,A. (2003) Binding of L7Ae protein to the K-turn of archaeal snoRNAs: a shared RNA binding motif for C/D and H/ACA box snoRNAs in Archaea. *Nucleic Acids Res.*, **31**, 869–877.
23. Charpentier,B., Muller,S. and Branlant,C. (2005) Reconstitution of archaeal H/ACA small ribonucleoprotein complexes active in pseudouridylation. *Nucleic Acids Res.*, **33**, 3133–3144.
24. Baker,D.L., Youssef,O.A., Chastkofsky,M.I., Dy,D.A., Terns,R.M. and Terns,M.P. (2005) RNA-guided RNA modification: functional organization of the archaeal H/ACA RNP. *Genes Dev.*, **19**, 1238–1248.
25. Muller,S., Charpentier,B., Branlant,C. and Leclerc,F. (2007) A Dedicated Computational Approach for the Identification of Archaeal H/ACA sRNAs. *Methods Enzymol.*, **425**, 355–387.
26. Thebault,P., de Givry,S., Schiex,T. and Gaspin,C. (2006) Searching RNA motifs and their intermolecular contacts with constraint networks. *Bioinformatics*, **22**, 2074–2080.
27. Hamma,T., Reichow,S.L., Varani,G. and Ferre-D'Amare,A.R. (2005) The Cbf5-Nop10 complex is a molecular bracket that organizes box H/ACA RNPs. *Nat. Struct. Mol. Biol.*, **12**, 1101–1107.
28. Manival,X., Charron,C., Fourmann,J.B., Godard,F., Charpentier,B. and Branlant,C. (2006) Crystal structure determination and site-directed mutagenesis of the *Pyrococcus abyssi* aCBF5-aNOP10 complex reveal crucial roles of the C-terminal domains of both proteins in H/ACA sRNP activity. *Nucleic Acids Res.*, **34**, 826–839.
29. Rashid,R., Liang,B., Baker,D.L., Youssef,O.A., He,Y., Phipps,K., Terns,R.M., Terns,M.P. and Li,H. (2006) Crystal structure of a Cbf5-Nop10-Gar1 complex and implications in RNA-guided pseudouridylation and dyskeratosis congenita. *Mol. Cell*, **21**, 249–260.
30. Liang,B., Xue,S., Terns,R.M., Terns,M.P. and Li,H. (2007) Substrate RNA positioning in the archaeal H/ACA ribonucleoprotein complex. *Nat. Struct. Mol. Biol.*
31. Li,L. and Ye,K. (2006) Crystal structure of an H/ACA box ribonucleoprotein particle. *Nature*, **443**, 302–307.
32. Wu,H. and Feigon,J. (2007) H/ACA small nucleolar RNA pseudouridylation pockets bind substrate RNA to form three-way junctions that position the target U for modification. *Proc. Natl Acad. Sci. USA*, **104**, 6655–6660.
33. Jin,H., Loria,J.P. and Moore,P.B. (2007) Solution structure of an rRNA substrate bound to the pseudouridylation pocket of a box H/ACA snoRNA. *Mol. Cell*, **26**, 205–215.
34. Ofengand,J. and Bakin,A. (1997) Mapping to nucleotide resolution of pseudouridine residues in large subunit ribosomal RNAs from representative eukaryotes, prokaryotes, archaebacteria, mitochondria and chloroplasts. *J. Mol. Biol.*, **266**, 246–268.
35. Kirpekar,F., Hansen,L.H., Rasmussen,A., Poehlsgaard,J. and Vester,B. (2005) The archaeon *Haloarcula marismortui* has few modifications in the central parts of its 23S ribosomal RNA. *J. Mol. Biol.*, **348**, 563–573.
36. Del Campo,M., Recinos,C., Yanez,G., Pomerantz,S.C., Guymon,R., Crain,P.F., McCloskey,J.A. and Ofengand,J. (2005)

Number, position, and significance of the pseudouridines in the large subunit ribosomal RNA of *Haloarcula marismortui* and *Deinococcus radiodurans*. *RNA*, **11**, 210–219.
37. Massenet,S., Ansmant,I., Motorin,Y. and Branlant,C. (1999) The first determination of pseudouridine residues in 23S ribosomal RNA from hyperthermophilic Archaea *Sulfolobus acidocaldarius*. *FEBS Lett*., **462**, 94–100.
38. Klein,R.J., Misulovin,Z. and Eddy,S.R. (2002) Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc. Natl Acad. Sci. USA*, **99**, 7542–7547.
39. Zago,M.A., Dennis,P.P. and Omer,A.D. (2005) The expanding world of small RNAs in the hyperthermophilic archaeon *Sulfolobus solfataricus*. *Mol. Microbiol*., **55**, 1812–1828.
40. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.
41. Besemer,J., Lomsadze,A. and Borodovsky,M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res*., **29**, 2607–2618.
42. Gautheret,D., Major,F. and Cedergren,R. (1993) Modeling the three-dimensional structure of RNA using discrete nucleotide conformational sets. *J. Mol. Biol*., **229**, 1049–1064.
43. Gautheret,D. and Lambert,A. (2001) Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J. Mol. Biol*., **313**, 1003–1011.
44. Charbonnier,F., Erauso,G., Barbeyron,T., Prieur,D. and Forterre,P. (1992) Evidence that a plasmid from a hyperthermophilic archaebacterium is relaxed at physiological temperatures. *J Bacteriol*, **174**, 6103–6108.
45. Chomczynski,P. and Sacchi,N. (1986) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem*., **162**, 156–159.
46. Bakin,A. and Ofengand,J. (1993) Four newly located pseudouridylate residues in *Escherichia coli* 23S ribosomal RNA are all at the peptidyltransferase center - analysis by the application of a new sequencing technique. *Biochemistry*, **32**, 9754–9762.
47. Motorin,Y., Muller,S., Behm-Ansmant,I. and Branlant,C. (2007) Identification of modified residues in RNAs by reverse transcription-based methods. *Methods Enzymol*., **425**, 21–53.
48. Charpentier,B., Fourmann,J.B. and Branlant,C. (2007) Reconstitution of archaeal H/ACA sRNPs and test of their activity. *Methods Enzymol*., **425**, 389–405.
49. Schattner,P. (2002) Searching for RNA genes using base-composition statistics. *Nucleic Acids Res*., **30**, 2076–2082.
50. Roovers,M., Hale,C., Tricot,C., Terns,M.P., Terns,R.M., Grosjean,H. and Droogmans,L. (2006) Formation of the conserved pseudouridine at position 55 in archaeal tRNA. *Nucleic Acids Res*., **34**, 4293–4301.
51. Muller,S., Fourmann,J.B., Loegler,C., Charpentier,B. and Branlant,C. (2007) Identification of determinants in the protein partners aCBF5 and aNOP10 necessary for the tRNA: Ψ55-synthase and RNA-guided RNA:Ψ-synthase activities. *Nucleic Acids Res*.
52. Gurha,P., Joardar,A., Chaurasia,P. and Gupta,R. (2007) Differential roles of archaeal box H/ACA proteins in guide RNA-dependent and independent pseudouridine formation. *RNA Biol*, **4**, 101–109.
53. Dennis,P.P., Omer,A. and Lowe,T. (2001) A guided tour: small RNA function in Archaea. *Mol. Microbiol*., **40**, 509–519.
54. Harms,J., Schluenzen,F., Zarivach,R., Bashan,A., Gat,S., Agmon,I., Bartels,H., Franceschi,F. and Yonath,A. (2001) High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. *Cell*, **107**, 679–688.
55. Yusupov,M.M., Yusupova,G.Z., Baucom,A., Lieberman,K., Earnest,T.N., Cate,J.H. and Noller,H.F. (2001) Crystal structure of the ribosome at 5.5 A resolution. *Science*, **292**, 883–896.
56. Branlant,C., Krol,A., Machatt,M.A., Pouyet,J., Ebel,J.P., Edwards,K. and Kossel,H. (1981) Primary and secondary structures of *Escherichia coli* MRE 600 23S ribosomal RNA. Comparison with models of secondary structure for maize chloroplast 23S rRNA and for large portions of mouse and human 16S mitochondrial rRNAs. *Nucleic Acids Res*., **9**, 4303–4324.
57. Juzumiene,D.I. and Wollenzien,P. (2001) Arrangement of the central pseudoknot region of 16S rRNA in the 30S ribosomal subunit determined by site-directed 4-thiouridine crosslinking. *RNA*, **7**, 71–84.
58. Bullard,J.M., van Waes,M.A., Bucklin,D.J., Rice,M.J. and Hill,W.E. (1998) Regions of 16S ribosomal RNA proximal to transfer RNA bound at the P-site of Escherichia coli ribosomes. *Biochemistry*, **37**, 1350–1356.
59. Noon,K.R., Bruenger,E. and McCloskey,J.A. (1998) Posttranscriptional modifications in 16S and 23S rRNAs of the archaeal hyperthermophile *Sulfolobus solfataricus*. *J. Bacteriol*., **180**, 2883–2888.
60. Watanabe,Y. and Gray,M.W. (2000) Evolutionary appearance of genes encoding proteins associated with box H/ACA snoRNAs: cbf5p in *Euglena gracilis*, an early diverging eukaryote, and candidate Gar1p and Nop10p homologs in archaebacteria. *Nucleic Acids Res*., **28**, 2342–2352.
61. Behm-Ansmant,I., Urban,A., Ma,X., Yu,Y.-T., Motorin,Y. and Branlant,C. (2003) The *Saccharomyces cerevisiae* U2 snRNA:pseudouridine-synthase Pus7p is a novel multisite-multisubstrate RNA: Ψ-synthase also acting on tRNAs. *RNA*, **9**, 1371–1382.
62. Ganot,P., Caizergues-Ferrer,M. and Kiss,T. (1997) The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes Dev*., **11**, 941–956.

Bibliographie

[1] Altman, S. Nobel lecture. enzymatic cleavage of RNA by RNA. *Biosci Rep* **10**, 317–37 (1990).

[2] Cech, T. Self-splicing and enzymatic activity of an intervening sequence RNA from tetrahymena (nobel lecture) .... *Angew. Chem. Int. Ed. Engl* (1990).

[3] Kruger, K. *et al.* Self-splicing rna : autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell* **31**, 147–57 (1982).

[4] Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. & Altman, S. The RNA moiety of ribonuclease p is the catalytic subunit of the enzyme. *Cell* **35**, 849–57 (1983).

[5] Fedor, M. J. & Uhlenbeck, O. C. Substrate sequence effects on "hammerhead" RNA catalytic efficiency. *Proc Natl Acad Sci USA* **87**, 1668–72 (1990).

[6] Uebayasi, M., Uchimaru, T., Sawata, S. & Shimayama, T. Theoretical and experimental considerations on the hammerhead ribozyme reactions : Divalent magnesium .... *J Org Chem* (1994).

[7] Lopez, X., Dejaegere, A., Leclerc, F., York, D. M. & Karplus, M. Nucleophilic attack on phosphate diesters : a density functional study of in-line reactivity in dianionic, monoanionic, and neutral systems. *The journal of physical chemistry B, Condensed matter, materials, surfaces, interfaces & biophysical* **110**, 11525–39 (2006).

[8] Pley, H. W., Lindes, D. S., DeLuca-Flaherty, C. & McKay, D. B. Crystals of a hammerhead ribozyme. *J Biol Chem* **268**, 19656–8 (1993).

[9] Pley, H. W., Flaherty, K. M. & McKay, D. B. Three-dimensional structure of a hammerhead ribozyme. *Nature* **372**, 68–74 (1994).

[10] Tuschl, T., Gohlke, C., Jovin, T. M., Westhof, E. & Eckstein, F. A three-dimensional model for the hammerhead ribozyme based on fluorescence measurements. *Science* **266**, 785–9 (1994).

[11] Scott, W. G., Murray, J. B., Arnold, J. R., Stoddard, B. L. & Klug, A. Capturing the structure of a catalytic RNA intermediate : the hammerhead ribozyme. *Science* **274**, 2065–9 (1996).

[12] Blount, K. F. & Uhlenbeck, O. C. The structure-function dilemma of the hammerhead ribozyme. *Annual review of biophysics and biomolecular structure* **34**, 415–40 (2005).

[13] Nelson, J. A. & Uhlenbeck, O. C. Hammerhead redux : does the new structure fit the old biochemical data ? *RNA* **14**, 605–15 (2008).

[14] Martick, M. & Scott, W. G. Tertiary contacts distant from the active site prime a ribozyme for catalysis. *Cell* **126**, 309–20 (2006).

[15] Cochrane, J. & Strobel, S. Catalytic strategies of self-cleaving ribozymes. *Acc Chem Res* (2008).

[16] Emilsson, G. M., Nakamura, S., Roth, A. & Breaker, R. R. Ribozyme speed limits. *RNA* **9**, 907–18 (2003).

[17] Bevilacqua, P. C. & Yajima, R. Nucleobase catalysis in ribozyme mechanism. *Current opinion in chemical biology* **10**, 455–64 (2006).

[18] Breaker, R. R. *et al.* A common speed limit for RNA-cleaving ribozymes and deoxyribozymes. *RNA* **9**, 949–57 (2003).

[19] Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in caenorhabditis elegans. *Nature* **391**, 806–11 (1998).

[20] Moss, E. G. RNA interference : it's a small RNA world. *Curr Biol* **11**, R772–5 (2001).

[21] Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484–8 (2007).

[22] Kapranov, P., Willingham, A. T. & Gingeras, T. R. Genome-wide transcription and the implications for genomic organization. *Nat Rev Genet* **8**, 413–23 (2007).

[23] Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034–50 (2005).

[24] Hüttenhofer, A. *et al.* RNomics : an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J* **20**, 2943–53 (2001).

[25] Marker, C. *et al.* Experimental RNomics : identification of 140 candidates for small non-messenger RNAs in the plant arabidopsis thaliana. *Curr Biol* **12**, 2002–13 (2002).

[26] Lowe, T. M. & Eddy, S. R. tRNAscan-se : a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**, 955–64 (1997).

[27] Lowe, T. M. & Eddy, S. R. A computational screen for methylation guide snoRNAs in yeast. *Science* **283**, 1168–71 (1999).

[28] Rivas, E. & Eddy, S. R. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* **2**, 8 (2001).

[29] Laferriere, A., Gautheret, D. & Cedergren, R. An RNA pattern matching program with enhanced performance and portability. *Bioinformatics* (2002).

[30] Macke, T. J. *et al.* Rnamotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Research* **29**, 4724–35 (2001).

[31] Gautheret, D. & Lambert, A. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J Mol Biol* **313**, 1003–11 (2001).

[32] Lambert, A. *et al.* The erpin server : an interface to profile-based RNA motif identification. *Nucleic Acids Res* **32**, W160–5 (2004).

[33] Bourdeau, V., Ferbeyre, G., Pageau, M., Paquin, B. & Cedergren, R. The distribution of RNA motifs in natural sequences. *Nucleic Acids Research* **27**, 4457–67 (1999).

[34] Ferbeyre, G., Bourdeau, V., Pageau, M., Miramontes, P. & Cedergren, R. Distribution of hammerhead and hammerhead-like RNA motifs through the genbank. *Genome Res* **10**, 1011–9 (2000).

[35] Duarte, C. M., Wadley, L. M. & Pyle, A. M. RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Res* **31**, 4755–61 (2003).

[36] Klosterman, P. S., Tamura, M., Holbrook, S. R. & Brenner, S. E. Scor : a structural classification of RNA database. *Nucleic Acids Res* **30**, 392–4 (2002).

[37] Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 a resolution. *Science* **289**, 905–20 (2000).

[38] Wimberly, B. T. *et al.* Structure of the 30s ribosomal subunit. *Nature* **407**, 327–39 (2000).

[39] Klein, D. J., Schmeing, T. M., Moore, P. B. & Steitz, T. A. The kink-turn : a new RNA secondary structure motif. *EMBO J* **20**, 4214–21 (2001).

[40] Vidovic, I., Nottrott, S., Hartmuth, K., Lührmann, R. & Ficner, R. Crystal structure of the spliceosomal 15.5kD protein bound to a u4 snrna fragment. *Mol Cell* **6**, 1331–42 (2000).

[41] Marmier-Gourrier, N. *et al.* A structural, phylogenetic, and functional study of 15.5-kD/Snu13 protein binding on u3 small nucleolar rna. *RNA* **9**, 821–38 (2003).

[42] Rozhdestvensky, T. S. *et al.* Binding of l7ae protein to the k-turn of archaeal snoRNAs : a shared rna binding motif for C/D and H/ACA box snoRNAs in archaea. *Nucleic Acids Res* **31**, 869–77 (2003).

[43] Chao, J. A. & Williamson, J. R. Joint x-ray and NMR refinement of the yeast l30e-mRNA complex. *Structure* **12**, 1165–76 (2004).

[44] Felden, B. RNA structure : experimental analysis. *Curr Opin Microbiol* **10**, 286–91 (2007).

[45] Mooers, B. Crystallographic studies of dna and rna. *Methods* (2008).

[46] Ke, A. & Doudna, J. A. Crystallization of RNA and RNA-protein complexes. *Methods* **34**, 408–14 (2004).

[47] Hennig, M., Williamson, J. R., Brodsky, A. S. & Battiste, J. L. Recent advances in RNA structure determination by NMR. *Current protocols in nucleic acid chemistry / edited by Serge L Beaucage [et al]* **Chapter 7**, Unit 7.7 (2001).

[48] Scott, L. G. & Hennig, M. RNA structure determination by NMR. *Methods Mol Biol* **452**, 29–61 (2008).

[49] Zhou, Z. H. Towards atomic resolution structural determination by single-particle cryo-electron microscopy. *Curr Opin Struct Biol* **18**, 218–28 (2008).

[50] Chartrand, P., Leclerc, F. & Cedergren, R. Relating conformation, mg2+ binding, and functional group modification in the hammerhead ribozyme. *RNA* **3**, 692–6 (1997).

[51] Hunsicker, L. M. & DeRose, V. J. Activities and relative affinities of divalent metals in unmodified and phosphorothioate-substituted hammerhead ribozymes. *J Inorg Biochem* **80**, 271–81 (2000).

[52] Hansen, M. R. *et al.* Identification and characterization of a novel high affinity metal-binding site in the hammerhead ribozyme. *RNA* **5**, 1099–104 (1999).

[53] Major, F. *et al.* The combination of symbolic and numerical computation for three-dimensional modeling of rna. *Science* **253**, 1255–60 (1991).

[54] Easterwood, T. R., Major, F., Malhotra, A. & Harvey, S. C. Orientations of transfer RNA in the ribosomal a and p sites. *Nucleic Acids Research* **22**, 3779–86 (1994).

[55] Gautheret, D., Major, F. & Cedergren, R. Modeling the three-dimensional structure of RNA using discrete nucleotide conformational sets. *J Mol Biol* **229**, 1049–64 (1993).

[56] Wollenzien, P., Juzumiene, D., Shapkina, T. & Minchew, P. Three dimensional model for the 16s ribosomal RNA that incorporates information for the mRNA track. *Nucleic Acids Symp Ser* 76–8 (1995).

[57] Leclerc, F., Cedergren, R. & Ellington, A. D. A three-dimensional model of the rev-binding element of hiv-1 derived from analyses of aptamers. *Nat Struct Biol* **1**, 293–300 (1994).

[58] Leclerc, F., Srinivasan, J. & Cedergren, R. Predicting RNA structures : the model of the RNA element binding rev meets the NMR structure. *Folding & design* **2**, 141–7 (1997).

[59] Ellington, A. D., Leclerc, F. & Cedergren, R. An RNA groove. *Nat Struct Biol* **3**, 981–4 (1996).

[60] Srinivasan, J., Leclerc, F., Xu, W., Ellington, A. D. & Cedergren, R. A docking and modelling strategy for peptide-RNA complexes : applications to BIV Tat-TAR and hiv Rev-RBE. *Folding & design* **1**, 463–72 (1996).

[61] Narlikar, G. J. & Herschlag, D. Mechanistic aspects of enzymatic catalysis : lessons from comparison of RNA and protein enzymes. *Annu Rev Biochem* **66**, 19–59 (1997).

[62] Doherty, E. A. & Doudna, J. A. Ribozyme structures and mechanisms. *Annual review of biophysics and biomolecular structure* **30**, 457–75 (2001).

[63] Zhao, Z.-Y. *et al.* Nucleobase participation in ribozyme catalysis. *J Am Chem Soc* **127**, 5026–7 (2005).

[64] Gordon, P. M., Fong, R. & Piccirilli, J. A. A second divalent metal ion in the group ii intron reaction center. *Chem Biol* **14**, 607–12 (2007).

[65] Stahley, M. R. & Strobel, S. A. Structural evidence for a two-metal-ion mechanism of group i intron splicing. *Science* **309**, 1587–90 (2005).

[66] Roychowdhury-Saha, M. & Burke, D. H. Distinct reaction pathway promoted by non-divalent-metal cations in a tertiary stabilized hammerhead ribozyme. *RNA* **13**, 841–8 (2007).

[67] ichi Nakano, S. & Bevilacqua, P. C. Mechanistic characterization of the hdv genomic ribozyme : a mutant of the c41 motif provides insight into the positioning and thermodynamic linkage of metal ions and protons. *Biochemistry* **46**, 3001–12 (2007).

[68] Ke, A., Zhou, K., Ding, F., Cate, J. H. D. & Doudna, J. A. A conformational switch controls hepatitis delta virus ribozyme catalysis. *Nature* **429**, 201–5 (2004).

[69] Boero, M., Tateno, M., Terakura, K. & Oshiyama, A. Double-metal-ion/single-metal-ion mechanisms of the cleavage reaction of ribozymes : First-principles .... *J. Chem. Theory Comput* (2005).

[70] Dahm, S. C., Derrick, W. B. & Uhlenbeck, O. C. Evidence for the role of solvated metal hydroxide in the hammerhead cleavage mechanism. *Biochemistry* **32**, 13040–5 (1993).

[71] Pontius, B. W., Lott, W. B. & von Hippel, P. H. Observations on catalysis by hammerhead ribozymes are consistent with a two-divalent-metal-ion mechanism. *Proc Natl Acad Sci USA* **94**, 2290–4 (1997).

[72] Steitz, T. A. & Steitz, J. A. A general two-metal-ion mechanism for catalytic rna. *Proc Natl Acad Sci USA* **90**, 6498–502 (1993).

[73] Torres, R. A., Himo, F., Bruice, T. C., Noodleman, L. & Lovell, T. Theoretical examination of mg(2+)-mediated hydrolysis of a phosphodiester linkage as proposed for the hammerhead ribozyme. *J Am Chem Soc* **125**, 9861–7 (2003).

[74] Leclerc, F. & Karplus, M. Two-metal-ion mechanism for hammerhead-ribozyme catalysis. *The journal of physical chemistry B, Condensed matter, materials, surfaces, interfaces & biophysical* **110**, 3395–409 (2006).

[75] Meltzer, P. S. Cancer genomics : small RNAs with big impacts. *Nature* **435**, 745–6 (2005).

[76] Bruneau, B. G. Developmental biology : tiny brakes for a growing heart. *Nature* **436**, 181–2 (2005).

[77] Cao, X., Yeo, G., Muotri, A., Kuwabara, T. & Gage, F. Noncoding RNAs in the mammalian central nervous system. *Annu Rev Neurosci* (2006).

[78] Pollard, K. S. *et al.* An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**, 167–72 (2006).

[79] Argaman, L. *et al.* Novel small RNA-encoding genes in the intergenic regions of escherichia coli. *Curr Biol* **11**, 941–50 (2001).

[80] Tycowski, K. T. & Steitz, J. A. Non-coding snorna host genes in drosophila : expression strategies for modification guide snoRNAs. *Eur J Cell Biol* **80**, 119–25 (2001).

[81] Tang, T.-H. *et al.* Identification of 86 candidates for small non-messenger RNAs from the archaeon archaeoglobus fulgidus. *Proc Natl Acad Sci USA* **99**, 7536–41 (2002).

[82] Besemer, J., Lomsadze, A. & Borodovsky, M. Genemarks : a self-training method for prediction of gene starts in microbial genomes. implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research* **29**, 2607–18 (2001).

[83] Besemer, J. & Borodovsky, M. Genemark : web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Research* **33**, W451–4 (2005).

[84] Washietl, S., Hofacker, I. L. & Stadler, P. F. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci USA* **102**, 2454–9 (2005).

[85] Hofacker, I. L. RNA consensus structure prediction with rnaalifold. *Methods Mol Biol* **395**, 527–44 (2007).

[86] Touzet, H. & Perriquet, O. Carnac : folding families of related RNAs. *Nucleic Acids Res* **32**, W142–5 (2004).

[87] Muller, S., Charpentier, B., Branlant, C. & Leclerc, F. A dedicated computational approach for the identification of archaeal H/ACA sRNAs. *Meth Enzymol* **425**, 355–87 (2007).

[88] Muller, S. *et al.* Combined in silico and experimental identification of the pyrococcus abyssi H/ACA sRNAs and their target sites in ribosomal RNAs. *Nucleic Acids Research* (2008).

[89] Charpentier, B., Fourmann, J.-B. & Branlant, C. Reconstitution of archaeal H/ACA sRNPs and test of their activity. *Meth Enzymol* **425**, 389–405 (2007).

[90] Mückstein, U. *et al.* Thermodynamics of RNA-RNA binding. *Bioinformatics* **22**, 1177–82 (2006).

[91] Muller, S. *et al.* Deficiency of the trnatyr :|*psiup*|35-synthase apus7 in archaea of the sulfolobales order is rescued by the H/ACA sRNA-guided machinery. *Nucleic acids research* in press (2008).

[92] Li, L. & Ye, K. Crystal structure of an H/ACA box ribonucleoprotein particle. *Nature* **443**, 302–7 (2006).

[93] Liang, B., Xue, S., Terns, R., Terns, M. & Li, H. Substrate RNA positioning in the archaeal H/ACA ribonucleoprotein complex. *Nat Struct Mol Biol* (2007).

[94] Dutheil, J., Pupko, T., Jean-Marie, A. & Galtier, N. A model-based approach for detecting coevolving positions in a molecule. *Mol Biol Evol* **22**, 1919–28 (2005).

[95] Manival, X. *et al.* Crystal structure determination and site-directed mutagenesis of the pyrococcus abyssi aCBF5-aNOP10 complex reveal crucial roles of the c-terminal domains of both proteins in H/ACA sRNP activity. *Nucleic Acids Res* **34**, 826–39 (2006).

[96] Charron, C. *et al.* The archaeal srna binding protein l7ae has a 3d structure very similar to that of its eukaryal counterpart while having a broader RNA-binding specificity. *J Mol Biol* **342**, 757–73 (2004).

[97] Walne, A. J. *et al.* Genetic heterogeneity in autosomal recessive dyskeratosis congenita with one subtype due to mutations in the telomerase-associated protein NOP10. *Hum Mol Genet* **16**, 1619–29 (2007).

[98] Saunders, L. R. & Barber, G. N. The dsRNA binding protein family : critical roles, diverse cellular functions. *FASEB J* **17**, 961–83 (2003).

[99] Faustino, N. A. & Cooper, T. A. Pre-mRNA splicing and human disease. *Genes Dev* **17**, 419–37 (2003).

[100] Wang, G.-S. & Cooper, T. A. Splicing in disease : disruption of the splicing code and the decoding machinery. *Nat Rev Genet* **8**, 749–61 (2007).

[101] Cáceres, J. F. & Kornblihtt, A. R. Alternative splicing : multiple control mechanisms and involvement in human disease. *Trends Genet* **18**, 186–93 (2002).

[102] Dredge, B. K., Polydorides, A. D. & Darnell, R. B. The splice of life : alternative splicing and neurological disease. *Nat Rev Neurosci* **2**, 43–50 (2001).

[103] O'Rourke, J. & Swanson, M. Mechanisms of RNA-mediated disease. *J Biol Chem* (2008).

[104] Galvão, R., Mendes-Soares, L., Câmara, J., Jaco, I. & Carmo-Fonseca, M. Triplet repeats, RNA secondary structure and toxic gain-of-function models for pathogenesis. *Brain Res Bull* **56**, 191–201 (2001).

[105] Orr, H. T. & Zoghbi, H. Y. Trinucleotide repeat disorders. *Annu Rev Neurosci* **30**, 575–621 (2007).

[106] Sobczak, K., de Mezer, M., Michlewski, G., Krol, J. & Krzyzosiak, W. J. RNA structure of trinucleotide repeats associated with human neurological diseases. *Nucleic Acids Research* **31**, 5469–82 (2003).

[107] Kaplan, S., Itzkovitz, S. & Shapiro, E. A universal mechanism ties genotype to phenotype in trinucleotide diseases. *PLoS Comput Biol* **3**, e235 (2007).

[108] Ranum, L. P. W. & Day, J. W. Myotonic dystrophy : RNA pathogenesis comes into focus. *Am J Hum Genet* **74**, 793–804 (2004).

[109] Napierała, M. & Krzyzosiak, W. J. Cug repeats present in myotonin kinase RNA form metastable "slippery" hairpins. *J Biol Chem* **272**, 31079–85 (1997).

[110] Jasinska, A. *et al.* Structures of trinucleotide repeats in human transcripts and their functional implications. *Nucleic Acids Research* **31**, 5463–8 (2003).

[111] Mankodi, A. *et al.* Muscleblind localizes to nuclear foci of aberrant RNA in myotonic dystrophy types 1 and 2. *Hum Mol Genet* **10**, 2165–70 (2001).

[112] Day, J. W. & Ranum, L. P. W. RNA pathogenesis of the myotonic dystrophies. *Neuromuscul Disord* **15**, 5–16 (2005).

[113] Cho, D. H. & Tapscott, S. J. Myotonic dystrophy : emerging mechanisms for dm1 and dm2. *Biochim Biophys Acta* **1772**, 195–204 (2007).

[114] Zaman, G. J. R., Michiels, P. J. A. & van Boeckel, C. A. A. Targeting rna : new opportunities to address drugless targets. *Drug Discov Today* **8**, 297–306 (2003).

[115] Xavier, K. A., Eder, P. S. & Giordano, T. RNA as a drug target : methods for biophysical characterization and screening. *Trends Biotechnol* **18**, 349–56 (2000).

[116] Ecker, D. & Griffey, R. RNA as a small-molecule drug target : doubling the value of genomics. *Drug Discov Today* **4**, 420–429 (1999).

[117] Hermann, T. & Westhof, E. RNA as a drug target : chemical, modelling, and evolutionary tools. *Curr Opin Biotechnol* **9**, 66–73 (1998).

[118] Mooers, B. H. M., Logue, J. S. & Berglund, J. A. The structural basis of myotonic dystrophy from the crystal structure of cug repeats. *Proc Natl Acad Sci USA* **102**, 16626–31 (2005).

[119] Shuker, S. B., Hajduk, P. J., Meadows, R. P. & Fesik, S. W. Discovering high-affinity ligands for proteins : Sar by NMR. *Science* **274**, 1531–4 (1996).

[120] Zeng, J. & Treutlein, H. R. A method for computational combinatorial peptide design of inhibitors of ras protein. *Protein Eng* **12**, 457–68 (1999).

[121] Zeng, J. *et al.* Design of inhibitors of ras–raf interaction using a computational combinatorial algorithm. *Protein Eng* **14**, 39–45 (2001).

[122] Hajduk, P. J. *et al.* NMR-based discovery of lead inhibitors that block dna binding of the human papillomavirus e2 protein. *J Med Chem* **40**, 3144–50 (1997).

[123] Pellecchia, M., Sem, D. S. & Wüthrich, K. NMR in drug discovery. *Nat Rev Drug Discov* **1**, 211–9 (2002).

[124] Yuan, Y. *et al.* Muscleblind-like 1 interacts with RNA hairpins in splicing target and pathogenic RNAs. *Nucleic Acids Res* **35**, 5474–86 (2007).

[125] Kino, Y. *et al.* Muscleblind protein, mbnl1/exp, binds specifically to chhg repeats. *Hum Mol Genet* **13**, 495–507 (2004).

[126] Dang, M. Y. I. M. K. T. S. M. T. T. Y., W. Solution structure of the zf-ccchx2 domain of muscleblind-like 2, isoform 1 (To be Published).

[127] Hudson, B. P., Martinez-Yamout, M. A., Dyson, H. J. & Wright, P. E. Recognition of the mRNA au-rich element by the zinc finger domain of tis11d. *Nat Struct Mol Biol* **11**, 257–64 (2004).

[128] Caflisch, A., Miranker, A. & Karplus, M. Multiple copy simultaneous search and construction of ligands in binding sites : application to inhibitors of hiv-1 aspartic proteinase. *J Med Chem* **36**, 2142–67 (1993).

[129] Böhm, H. J. Ludi : rule-based automatic design of new substituents for enzyme inhibitor leads. *J Comput Aided Mol Des* **6**, 593–606 (1992).

[130] Caflisch, A. & Karplus, M. Computational combinatorial chemistry for de novo ligand design : Review and assessment. *Perspectives in Drug Discovery and Design* (1995).

[131] Stultz, C. M. & Karplus, M. Dynamic ligand design and combinatorial optimization : designing inhibitors to endothiapepsin. *Proteins* **40**, 258–89 (2000).

[132] Leclerc, F. & Karplus, M. Mcss-based predictions of RNA binding sites. *Theoretical Chemistry Accounts : Theory, Computation, and Modeling (Theoretica Chimica Acta)* **101**, 131–137 (1999).

[133] Caflisch, A., Walchli, R. & Ehrhardt, C. Computer-aided design of thrombin inhibitors. *Physiology* (1998).

[134] Bagheri, B., Ilin, A., Tan, R., McCammon, J. & Briggs, J. University of houston brownian dynamics program user's guide and programmer's manual release 5.1. *University of Houston* (1989).

[135] Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems : application to microtubules and the ribosome. *Proc Natl Acad Sci USA* **98**, 10037–41 (2001).

[136] Lebars, I., Husson, C., Yoshizawa, S., Douthwaite, S. & Fourmy, D. Recognition elements in rrna for the tylosin resistance methyltransferase rlma(ii). *J Mol Biol* **372**, 525–34 (2007).

[137] Ritchie, D. W. Evaluation of protein docking predictions using hex 3.1 in capri rounds 1 and 2. *Proteins* **52**, 98–106 (2003).

[138] Lunde, B. M., Moore, C. & Varani, G. RNA-binding proteins : modular design for efficient function. *Nat Rev Mol Cell Biol* **8**, 479–90 (2007).

[139] Reddy, S. Y., Leclerc, F. & Karplus, M. Dna polymorphism : a comparison of force fields for nucleic acids. *Biophys J* **84**, 1421–49 (2003).

[140] Beautrait, A. *et al.* Multiple-step virtual screening using vsm-g : overview and validation of fast geometrical matching enrichment. *Journal of molecular modeling* **14**, 135–48 (2008).

Résumé

Mots-clés: ARN, ribozymes, RNomics, modélisation 3D, docking, conception de ligands.

La découverte des propriétés catalytiques des ARN a bouleversé nos conceptions en biologie moléculaire sur les rôles des ARN au cours de l'évolution, des origines de la vie jusqu'à maintenant. Du point de vue biochimique, on commence seulement à comprendre dans le détail la catalyse des ribozymes. L'enzymologie moléculaire offre des approches expérimentales pour étudier la catalyse par les ARN ; les méthodes de mécanique quantique permettent de modéliser les réactions qualitativement et quantitativement et d'essayer de comprendre l'origine du pouvoir catalytique des enzymes. C'est dans cette optique que nous avons modélisé la réaction chimique qui intervient dans le ribozyme à tête marteau qui est considéré comme un prototype dans la catalyse par les ARN car il est un des plus petits ribozymes naturels que l'on connaisse. Nous avons proposé le 1er modèle à 2 cations métalliques pour un ribozyme auto-clivable. A ce jour, le mécanisme réactionnel et les changements conformationnels associés à la catalyse ne sont pas complètement élucidés. Pour y contribuer, nous cherchons à comparer de façon théorique les catalyses de type "metallo-enzyme" et "nucléobase" ou hybrides.

Des découvertes récentes ont mis en évidence l'existence de nombreux ARN non-codant impliqués dans de nombreuses et diverses fonctions biologiques. Dans cette thématique, nous nous sommes focalisés sur un domaine du vivant encore peu étudié : les archaea. Nous avons développé et mis en œuvre une approche bio-informatique pour rechercher des gènes d'ARN non-codant dans ces génomes qui a été largement validée expérimentalement sur une classe particulière d'ARN non-codant que sont les ARN guides de modifications à boîtes H/ACA. Des développements permettant d'améliorer la sensibilité de la méthode sont prévus. D'autre part, les connaissances acquises sur les ARN à boîtes H/ACA ouvrent des perspectives dans la compréhension d'une maladie humaine (la dyskeratose) reliée à un dysfonctionnement associée à cette famille d'ARN non-codant chez l'homme. Le prolongement de cette thématique abordée à l'aide d'outils bio-informatiques s'appuiera sur l'utilisation de méthodes de modélisation 3D développées par ailleurs.

Les travaux menés sur les ribozymes et l'essor de la RNomics démontrent la versatilité de fonction et de structure des ARN. La biologie structurale a permis d'acquérir beaucoup de connaissances sur la structuration des ARN et leurs interactions. La conception assistée par ordinateur de ligands dirigée contre des ARN comme cibles thérapeutiques est dès lors possible. La modélisation 3D de macromolécules offre une approche possible d'étude des liens structure/fonction en l'absence de données structurales mais aussi d'étude théorique de la flexibilité conformationnelle qui joue un rôle primordial dans le repliement des ARN et leur interaction avec des ligands. Les travaux présentés ici s'inscrivent dans le cadre d'un développement d'approches de modélisation de complexes ARN/protéines et ARN/ligand adaptées au mode d'interaction étudié, selon que l'ARN est structuré ou non. Dans le premier cas, une approche de docking classique est utilisée où l'ARN peut être considéré comme une cible pour un ligand protéique ou autre. Dans le second cas, une approche de docking "par fragment" (SELEX *in silico*) inspirée des méthodes de conception de drogues est utilisée pour prédire la reconnaissance par des protéines de liaison à l'ARN et éventuellement concevoir d'autres ligands ARN. La méthode est en cours de validation et sera appliquée à des cibles d'intérêt impliquées dans les dystrophies myotoniques.