Nouvelles approches en génomique comparative et bio-informatique structurale : à la recherche de relations séquence-structure-fonction.

Karsten Suhre

HABILITATION

A

DIRIGER DES RECHERCHES

Nouvelles approches en génomique comparative et bio-informatique structurale : à la recherche de relations séquence-structure-fonction.

présentée et soutenue publiquement le 26/10/2004

à l'Université Aix-Marseille II

par

Karsten Suhre

devant le jury composé de

Prof. Pedro Maldonado Coutinho	RAPPORTEUR
Prof. Philippe Derreumaux	RAPPORTEUR
Dr. Jorge Navaza	RAPPORTEUR
Dr. Richard Giégé	PRÉSIDENT
Prof. Jean-Michel Claverie	EXAMINATEUR

Remerciements

J'aimerais commencer par remercier Jean-Michel CLAVERIE, qui m'a accueilli au sein de son laboratoire et à qui je dois une bonne partie de ma formation à la bio-informatique.

Merci à Pedro MALDONADO COUTINHO, Philippe DERREUMAUX, Jorge NAVAZA et Richard GIÉGÉ d'avoir accepté de participer au jury et d'évaluer ce travail.

Merci à Chantal ABERGEL pour m'avoir initié au mystères de la résolution de structures protéiques par cristallographie, et également à Hiroyuki OGATA, Stéphane AUDIC et Cédric NOTREDAME pour de nombreuses discussions que nous avons eues au cours de ces dernières années, discussions qui m'ont beaucoup aidées pour avancer dans mon travail.

Merci aussi à François ENAULT, Jean-Baptiste CLAUDE et Ghislain BIDAUT pour m'avoir donné l'occasion de m'investir dans leurs travaux de thèse et les interaction très stimulantes qui en résultaient, et également à Sandra JEUDY pour avoir suscité mon intérêt dans le phasage des données cristallographiques par modélisation par homologie.

Merci à toute l'équipe de l'IGS pour l'accueil très chaleureux et le soutien qu'elle m'a apporté tout au long de ces dernières trois années, aussi bien du côté administratif et scientifique que du côté personnel.

Merci à Yves-Henri SANEJOUAND de l'ENS à Lyon et à Vincent GÉLI et Christophe De La ROCHE Saint ANDRÉ du LISM/IBSM pour des collaborations très stimulantes et fructueuses.

Table des matières

Liste des figures

Liste des tableaux

Chapitre 1 Introduction

1.1 Curriculum Vitæ

Mon parcours de chercheur se caractérise par une forte interdisciplinarité que l'on pourrait esquisser comme suit : début en physique théorique et en mathématiques numérique, passage à la chimie de l'atmosphère et à la météorologie, suivies d'une expérience dans l'industrie (ingénierie des automobiles), puis reconversion à la biologie structurale et fonctionnelle. Voyons alors plus en détail les étapes de mon parcours scientifique :

Après avoir été formé à la physique théorique (bac+5) et à la mathématique numérique (bac+2), grâce à des études universitaires poursuivies en Allemagne aussi bien qu'en Grande-Bretagne, j'ai procédé à un premier changement thématique en préparant une thèse dans le domaine de la chimie atmosphérique et de la météorologie, thèse d'abord financée par une Bourse du Gouvernement Français, puis par une bourse européenne de niveau postdoctoral Marie-Curie. Suite à la soutenance de ma thèse, qui a eu lieu seulement deux ans et trois mois après le début de mes recherches dans ce domaine, j'ai été directement recruté au CNRS pour développer la modélisation couplée physico-chimique à méso-échelle autour des questions liées à la pollution atmosphérique et au changement climatique.

Au bout de cinq ans d'activité en tant que chercheur au CNRS dans le département des Sciences de l'Univers, on m'a proposé un poste intéressant dans l'industrie automobile allemande. Pendant les deux années qui ont suivi, j'ai donc participé au management du projet de développement de la nouvelle Mercedes CLK décapotable, apprenant ainsi pleinement l'importance cruciale que jouent les trois mots clés « coût – échéances – qualité » dans la réussite d'une entreprise privée. Malgré l'acquis de toute une série de nouvelles compétences reliées à ces trois contraintes, mes tâches quotidiennes me faisaient regretter la rigueur que j'ai connue dans mon passé de chercheur. Comme les challenges de l'ère post-génomique me passionnaient depuis longtemps, j'ai saisi l'occasion pour oser, lors de ma réintégration au CNRS, un troisième changement thématique fondamental : à savoir la reconversion à la bio-informatique, domaine dans lequel je présente aujourd'hui ce mémoire d'HDR.

depuis 02/02	**Chargé de Recherche CNRS** Laboratoire Information Génomique & Structurale, Marseille.
01/00 – 01/02	**Ingénieur de Projet** Wilhelm Karmann GmbH, Osnabrück, Allemagne.
10/94 – 12/99	**Chargé de Recherche CNRS** Laboratoire d'Aérologie, Toulouse.
10/93 – 09/94	**Postdoc** (boursier Marie-Curie) Laboratoire d'Aérologie, Toulouse.
01/92 – 04/94	**Thèse de doctorat** (boursier du Gouvernement Français) Université Paul Sabatier, Toulouse.
10/88 – 04/89	**Department of Applied Mathematics** (niveau B. Sc.) University of Hull, Angleterre.
04/89 – 09/91, 10/86 - 09/88	**Assistent estudiantin** (46h/mois) Centre de calcul universitaire, Université d'Osnabrück, Allemagne.
10/84 – 09/91	***« Diplom Physik »*** (bac+5) et « ***Vordiplom Mathematik »*** (bac+2) Université d'Osnabrück, Allemagne.

Tableau 1 Curriculum vitæ

1.2 Production scientifique / administration de la recherche

Publications. Pendant les dix dernières années, j'ai publié en tout 38 papiers dans des revues à comité de lecture (rang A), dont 24 en Sciences de l'Univers, incluant notamment un papier en premier auteur dans la revue *Nature* (Suhre et al., 1997). Depuis ma reconversion à la biologie, j'ai co-signé 14 publications (4 en premier auteur).

Administration. En ce qui concerne mes activités dans l'administration de la recherche, je tiens à souligner mon rôle d'investigateur principal dans le projet CloudyColumn (4^{ième} PCRD de la Commission Européenne) et de coordinateur au niveau international de la partie modélisation IntegModel du projet ACE-2 (regroupant 5 projets du 4^{ième} PCRD), ainsi que ma participation à différents projets nationaux financés par le « Programme National de Chimie Atmosphérique (PNCA) », sans oublier les nombreux comités de coordination liés à des expériences d'envergure internationale et autres comités d'évaluation, groupes de réflexion au niveau national et conseils de laboratoire.

Encadrement. Quant à l'encadrement d'autres chercheurs, j'ai déjà eu l'occasion de pouvoir diriger une thèse en toute responsabilité, et ceci grâce à une dérogation à titre individuel, accordée par l'Université Paul Sabatier. J'ai également co-encadré cinq autres thèses ainsi qu'un chercheur postdoctoral en SDU et je participe actuellement à l'encadrement de trois thèses en SDV.

Enseignement. Tout au long de ma carrière, j'ai saisi toute occasion qui se présentait pour participer à l'enseignement. Comme exemple le plus récent, citons mon intervention dans le cadre du D.E.A. *Bio-informatique Biologie Structurale et Génomique* (BBSG) et de la maîtrise de biochimie à l'Université de la Méditerranée à Marseille, voir comme exemple le plus ancien, les travaux dirigés (TD) et travaux pratiques (TP) en programmation FORTRAN et Pascal ainsi que les cours de LaTeX, enseignements que j'ai dispensés déjà au cours de mes études universitaires, et ceci depuis 1987.

1.3 Compétences

Le point central de mes recherches est la combinaison de l'approche théorique et de la simulation numérique, tout en confrontant mes résultats théoriques à une validation expérimentale directe. J'ai débuté mon activité de recherche dans le domaine de la physique quantique, et ceci au cours de ce qu'on appelle en Allemagne le « *Diplomarbeit* », travail universitaire qui correspond au D.E.A. français et qui représente une activité de recherche d'une durée d'un an. Au cours de mes études universitaires de physique, j'ai simultanément acquis des connaissances en mathématiques en faisant mon « *Vordiplom* » (équivalent au niveau D.E.U.G.) de mathématiques ; suivi par un séjour à l'Université de Hull en Grande-Bretagne afin

d'y suivre les cours de mathématiques appliquées du niveau « *B. Sc*. » (équivalent au niveau licence français) dans le « *Department of Applied Mathematics* ».

Quant à mes compétences en informatique, j'aimerais souligner que dès ma deuxième année d'études universitaires en Allemagne, j'ai travaillé en parallèle dans le centre de calcul universitaire où on m'avait confié des tâches de programmation ainsi que d'assistance aux utilisateurs. Puis, tout au long de mes recherches au CNRS, l'ordinateur a constitué mon principal outil de travail, ce qui s'est traduit par des simulations lourdes en météorologie sur de grands calculateurs vectoriels (Cray et Fujitsu de l'IDRIS et de Météo France) ainsi que sur des machines parallèles (Thinking Machines CM5, Silicon Graphics et cluster Linux). La modélisation couplée de la chimie atmosphérique et de la météorologie m'a amené à développer des approches pragmatiques à la paramétrisation et à la prévision de processus atmosphériques complexes.

En ce qui concerne le management de projets, j'ai eu l'occasion d'acquérir des compétences en gestion et en communication à travers ma participation au pilotage de différents projets internationaux, en particulier en tant que coordinateur des activités de modélisation dans le cadre des grandes expériences de mesures ACE (IGAC Aerosol Characterisation Experiments: Bates et al., 1998; Huebert et al., 2003; Raes et al., 2000). Quant à mes années passées dans l'industrie automobile, milieu très axé « coût – échéances – qualité », cette expérience dans le secteur privé m'a permis de me familiariser avec une approche complètement différente de celle de la recherche scientifique, à savoir un contrôle total des paramètres clés d'un projet (budget, calendrier, cahier des charges, etc.) par l'intermédiaire d'audits et de revues de projets en continu sous le contrôle du client (Figure 1).

Ainsi, l'ensemble de ces expériences quelque peu disparates m'est aujourd'hui très profitable pour mes recherches en bio-informatique, domaine qui lui même est par sa nature un domaine d'une extrême interdisciplinarité.

Figure 1 Mercedes CLK décapotable (A209)

1.4 Thèmes de recherche antérieurs

Bien que ce mémoire soit consacré à faire état de mes contributions à la recherche en biologie, je tiens à évoquer brièvement mes travaux de recherche en Sciences Physiques et en Sciences de l'Univers, ceci me permettant de mettre en évidence les concepts communs aux différents domaines dans lesquels j'ai été actifs, ainsi que la logique inhérente à mon parcours de chercheur transdisciplinaire.

1.4.1 Sciences Physiques

En préparant mon diplôme de physique (sous la direction du professeur J. E. Roberts) dans le domaine de la théorie des quanta du champ, j'ai travaillé pendant plus d'un an à temps plein sur des preuves de théorèmes liés à la description algébrique du comportement des électrons relativistes dans un champ électromagnétique extérieur. C'est la seul période dans ma carrière de chercheur au cours de laquelle l'ordinateur n'occupait pas une place centrale dans ma démarche, et ce qui m'a permis ainsi d'apprendre sur le tas la force d'un raisonnement purement algébrique. Le résultat de ces travaux se résume à la preuve d'un théorème qui s'écrit :

$$\mathcal{R}(\mathcal{O}) = \mathcal{R}(\mathcal{O}')^{t'}$$

Le cadre de ce mémoire ne me permet malheureusement pas de détailler les implications de ce résultat (pour plus de détails voire Suhre, 1991).

1.4.2 Sciences de l'Univers

Mes activités au Laboratoire d'Aérologie à Toulouse (1991-1999) se divisent en trois volets qui se chevauchent. Premièrement, j'ai développé tout au long de cette période le module de chimie atmosphérique du modèle communautaire Méso-NH, qui est le modèle communautaire de météorologie à méso-échelle développé conjointement par Météo France et le Laboratoire d'Aérologie (Lafore et al., 1998). Deuxièmement, j'ai appliqué et validé ce modèle dans l'étude de divers processus impliqués dans le changement du climat ainsi que dans la pollution atmosphérique. Ceci se traduisait notamment par ma participation active à des campagnes de mesures expérimentales qui s'organisaient principalement dans le cadre des expériences ACE (IGAC Aerosol Characterisation Experiments: Bates et al., 1998; Huebert et al., 2003; Raes et al., 2000). Troisièmement, je me suis impliqué dans l'étude des processus d'échange de masses d'air entre la troposphère et la stratosphère (à environ 8-12 km altitude) en me basant sur les données issues du projet MOZAIC (*Measurement of OZone, Water Vapor, Carbon Monoxide and Nitrogen Oxides by Airbus-In-Service AirCraft*), qui est un projet de mesure d'ozone, de vapeur d'eau, de monoxyde de carbone et d'oxydes nitriques à bord des avions de ligne A340 (Marenco et al., 1998). Par la suite, je développerai brièvement mes principaux résultats dans les trois domaines évoqués ci-dessus.

Cycle du soufre dans la couche limite marine

Le diméthyl de soufre (DMS) est un produit métabolique du phytoplancton marin et constitue la principale source naturelle de soufre dans l'atmosphère. Il est lié à la régulation du climat par une boucle de rétroaction complexe qui fait intervenir les différentes composantes chimiques dans l'atmosphère, ainsi que les aérosols de soufres, les gouttelettes de nuages et le rayonnement solaire. Cette boucle de rétroaction climatique, également connue sous le nom d'« hypothèse DMS-nuages-climat » (ou CLAW selon les initiales de ses auteurs, Charlson et al., 1987), est considérée comme étant la première réalisation physique du modèle de Gaya (Lovelock, 1997). Par la modélisation couplée dynamique-chimie du DMS dans un

modèle de météorologie, j'ai mis en évidence le rôle de la turbulence atmosphérique dans l'oxydation du DMS (Suhre and Rosset, 1994a) ainsi que l'importance que jouent les embruns marins dans l'oxydation en phase hétérogène du dioxyde de soufre (Suhre et al., 1995). Afin de pouvoir résoudre les équations différentielles qui décrivent l'évolution temporelle des différents composants chimiques dans un modèle tri-dimensionnel, j'ai été également amené à développer un nouveau algorithme de résolution de systèmes différentiels raides (Suhre and Rosset, 1994b). Cette étude du cycle de soufre qui à la base a été mon sujet de thèse (Suhre, 1994), a été par la suite reprise sous ma responsabilité par une étudiante en thèse, Céline Mari (Mari et al., 1998; Mari et al., 1999).

Chimie atmosphérique en phase gaz, nuage et aérosol

La chimie atmosphérique en phase gaz, aérosol et nuage en interaction avec des processus météorologiques de turbulence, de convection, d'advection et des processus de formation de nuages est à la base de nombreuses questions liées à la pollution de l'air et au changement climatique. Dans le but de pouvoir aborder ces questions à l'échelle du processus, j'ai développé en étroite collaboration avec Météo France, le modèle numérique MésoNH-chimie. C'est un des premiers modèles muni d'une intégration complète des processus de chimie de l'atmosphère dans un code de météorologie à méso-échelle (Lafore et al., 1998). On se sert encore aujourd'hui de ce modèle dans de nombreuses études sur la pollution atmosphérique (Tulet et al., 2003; Tulet et al., 2002) ainsi que sur des processus de chimie atmosphérique plus fondamentaux, comme par exemple la convection profonde dans les tropiques (Mari et al., sous presse).

Un fort engagement dans la communauté internationale, communauté qui s'est formée autour des expériences ACE a constitué le pivot de ma démarche de validation et d'application de ce modèle. Dans le cadre de ces expériences, mes travaux de recherche ont donné lieu à de nombreuses publications importantes, à savoir pour la première fois une validation complète des simulations couplées dynamique-chimie-aérosols-nuages par confrontation directe aux données recueillies pendant les expériences de type Lagrangiennes d'ACE-1 et d'ACE-2 (Suhre et al., 2000a; Suhre et al., 2000b; Suhre et al., 1998). Egalement dans le cadre d'ACE, j'ai supervisé une

étude axée sur les interactions entre le rayonnement solaire et les nuages (Matthijsen et al., 1997; Matthijsen et al., 1998). De plus, j'ai activement participé à l'analyse des données expérimentales d'ACE, en fournissant d'une part des trajectographies numériques originales (Johnson et al., 2000a; Johnson et al., 2000b; Siems et al., 2000; Wood et al., 2000) et d'autre part en adaptant les données météorologiques issues du modèle de prévision du centre européen de météorologie à moyenne échelle (ECMWF) à l'étude des observations de la turbulence dans la couche limite marine (Wang et al., 1999a; Wang et al., 1999b).

En collaboration avec d'autres chercheurs du Laboratoire d'Aérologie (notamment par l'encadrement de thèses), j'ai contribué aux paramétrisations des aérosols dans des modèles de processus (Bedos et al., 1996; Cohard et al., 2000; FassiFihri et al., 1997) ainsi qu'à la réduction de schémas réactionnels chimiques (Crassier et al., 2000), travaux qui étaient toujours destinés à une utilisation ultérieure dans des modèles tri-dimensionnels couplé chimie-dynamique.

Echanges troposphère-stratosphère dans les tropiques

La découverte de pics d'ozone dans la haute troposphère tropicale faite par des avions de ligne A340 MOZAIC ("ozone-rich transients", Suhre et al., 1997, article fourni) dont l'origine a été récemment établie (Zahn et al., 2002), m'a amené à m'intéresser aux échanges entre la troposphère et la stratosphère dans les tropiques. C'est en collaboration avec un autre chercheur du Laboratoire d'Aérologie (Jean-Pierre Cammas) que j'ai alors développé un nouvel axe de recherche autour de ce thème, qui est toujours d'actualité au laboratoire(Baray et al., 2003; Cammas et al., 1998).

1.5 Thèmes de recherche développés en biologie

Si j'ai quitté le domaine des Sciences de l'Univers à une époque où j'étais bien établi dans la communauté internationale, c'était en partie pour des raisons purement personnelles. Cependant, ce choix a également été motivé par le désir d'apprendre et de connaître un domaine jusqu'alors complètement inconnu de moi. C'est pour ces raisons que j'ai dans un premier temps décidé de tenter ma chance dans l'industrie automobile. Pourtant, au bout de deux années, j'aspirais à retourner vers la recherche.

J'ai donc renoué avec ma carrière scientifique, en procédant par la même occasion à une reconversion vers les Sciences de la Vie.

Sachant que mes projets trouveraient leur dénominateur commun dans les objectifs centraux du laboratoire IGS, j'ai décidé, au moment de ma reconversion à la bio-informatique d'aborder en parallèle la génomique fonctionnelle et la biologie structurale. Le résumé de l'ensemble de mes travaux en biologie qui suivra a pour but de présenter l'interconnexion qui existe entre les différents thèmes que j'ai développés, tandis que je les décrirai en détails dans les chapitres suivants.

A ce point il convient d'introduire brièvement le projet de génomique structurale à haut débit ASG (« Après Séquençage Génomique »), conjointement mené par les laboratoires IGS et AFMB en collaboration avec Aventis. L'objectif du projet ASG est de déterminer un maximum de structures de protéines d'*Escherichia coli* par cristallographie, en visant des protéines dont la fonction est encore non-déterminée. En total, environ 300 gènes ont été sélectionnés sur un critère de conservation de séquence au travers du monde bactérien, le but ultime étant l'identification de nouvelles cibles antibiotiques à spectre large (Abergel et al., 2003). C'étaient souvent des questions concrètes liées au projet ASG qui ont éveillé mon intérêt dans des problématiques que je vais maintenant détailler.

Comme entrée en matière et pour démarrer ma reconversion, j'ai réévalué un critère de discrimination au niveau génomique du protéome de micro-organismes vivant à des températures ambiantes (mésophiles) ou élevées (thermophiles et hyper-thermophiles). En profitant du nombre croissant de génomes bactériens complètement séquencés, nombre qui est actuellement en croissance exponentielle, j'ai pu généraliser les conclusions des travaux antérieurs effectués par Cambillau et Claverie (Cambillau and Claverie, 2000 ; Suhre and Claverie, 2003, article fourni en annexe). Puis, je me suis intéressé à des méthodes destinées à établir des hypothèses sur la fonction d'un gène, sans qu'il y ait une homologie de séquence avec un gène déjà caractérisé. Ici encore, le nombre croissant de génomes bactériens rend ce genre de méthodes de plus en plus efficace, et donc particulièrement intéressant pour une application pratique. En co-encadrant la thèse de François Enault, nous avons alors développé une méthode qui est basé sur la co-évolution et sur la conservation de la

co-localisation sur le chromosome de deux (ou plusieurs) gènes dans des génomes plus ou moins proches (Phydbac: Enault et al., 2003a; Enault et al., 2003b; Enault et al., 2004). Afin de compléter cette approche, j'ai développé moi-même une base de données entièrement dédiée aux événements de fusion de gènes (FusionDB: Suhre and Claverie, 2004, article fourni en annexe). Le binôme Phydbac – FusionDB est aujourd'hui disponible sous forme de serveurs web et permet aux projets de génomique structurale (en particulier le projet ASG de l'IGS) d'associer une fonction putative à des protéines dont la structure vient d'être résolue, permettant ainsi une analyse orientée de cette structure dans un contexte fonctionnel. Ces hypothèses étant alors validées expérimentalement si nécessaire.

La modélisation numérique ayant été mon point fort lors de mes travaux en Sciences de l'Univers, j'ai décidé de tirer profit autant que possible de ce passé. En ce qui concerne la méthodologie, on peut considérer la dynamique moléculaire comme représentant l'équivalent en biologie de la dynamique atmosphérique en météorologie. Je me suis alors intéressé aux simulations en dynamique moléculaire en utilisant les programmes *Amber* et *Gromacs* dans le cadre des sujets issus du projet ASG. Or, bien que ces simulations aient été fort intéressantes en tant que telles, elles n'ont apporté que peu de réponses aux questions concrètes qui intéressent le laboratoire (c.-à-d. apprendre plus sur la fonction de ces protéines). En particulier, le temps de calcul de ces programmes est prohibitif et ne permet guère une simulation explicite des processus qui nous intéressent. Je me suis alors orienté vers des approches de modélisation plus pragmatiques, à savoir la modélisation par homologie et l'analyse en modes normaux.

Les liens forts qui existent entre l'équipe de bio-informatique et celle de cristallographie à l'IGS m'ont également amené à m'intéresser à des questions de caractère méthodologique, notamment à la résolution de structures par remplacement moléculaire. Mon travail de ces deux dernières années dans ce domaine a abouti à l'élaboration de protocoles d'utilisation de modèles homologues et de modèles perturbés en modes normaux pour le phasage de données cristallographiques par remplacement moléculaire (Suhre and Sanejouand, 2004b, article fourni en annexe).
Grâce à ces protocoles, j'ai pu résoudre le problème de phasage pour les protéines *YecD* d'*Escherichia coli* (PDB-id 1j2r) et la *TyrRS* de *Mimivirus* (affinement en

cours). Dans ces deux cas, ni le remplacement moléculaire, ni les méthodes de phasage de type MIR et MAD n'ont donné de résultat. Motivé par ce succès, nous avons alors mis au point des serveurs web pour le remplacement automatique en utilisant des modèles par homologie CaspR (travail de thèse de J.-B. Claude, Claude et al., 2004) et pour l'analyse en ligne des mouvement protéiques en modes normaux ElNémo (collaboration avec Y.-H. Sanejouand, ENS Lyon, Suhre and Sanejouand, 2004a, article fourni en annexe)

Simultanément à ces deux axes de recherche développés dans les domaines de la génomique fonctionnelle et de la biologie structurale, j'ai participé aux projets suivants :

- le séquençage et l'analyse du génome de la bactérie *Tropheryma whipplei* (Raoult et al., 2003) ;
- l'analyse des données de puces à ADN issues de la comparaison d'une souche de levure sauvage et d'un mutant en conditions méiotiques (Sollier et al., 2004) ;
- le développement de protocoles d'alignements multiples séquences-structures (Notredame and Suhre, sous presse; O'Sullivan et al., 2004; Poirot et al., 2004)

Les deux premiers projets faisait appel à mes compétences en analyse de données, notamment en « clustering » et en analyse statistique, compétences qui étaient également au centre de mes travaux dans le domaine de la génomique fonctionnelle. Quant au troisième projet, il est fortement lié aux questions posées par la modélisation par homologie. Il s'agit donc là de trois projets qui ont tous un rapport direct avec les deux principaux axes de recherche que j'ai développées depuis ma reconversion à la biologie, à savoir la génomique fonctionnelle et la biologie structurale, axes que j'ai abordés d'un point de vue de bio-informaticien « reconverti ». Voici le moment d'aborder tour à tour les détails qui caractérisent ma démarche, concernant « La génomique fonctionnelle » dans le Chapitre 2, puis « La biologie structurale » dans le Chapitre 3 et finalement « Autres projets en bio-informatique » au Chapitre 4.

Chapitre 2 La génomique fonctionnelle

Le thème central de ce chapitre est le « data mining » dans le grand nombre de génomes bactériens aujourd'hui disponible, nombre qui a encore doublé dans les deux dernières années. La comparaison de différents génomes complètement séquencés permet de rechercher un maximum d'informations afin de mieux comprendre le fonctionnement des gènes de ces micro-organismes, voire les propriétés des micro-organismes tout entier (mode de vie, métabolisme, résistance, etc.). Ma démarche s'inscrit dans cette perspective : la première question posée étant la suivante : peut on trouver la signature au niveau génomique d'un facteur discriminant les bactéries vivant à température ambiante (« mésophiles ») de celles qui sont capables de résister à des températures élevées (« thermophiles »), voire extrêmes (« hyper-thermophiles »). Le second projet avait pour but de développer des méthodes qui permettraient d'annoter des gènes de fonction inconnue sur la base de leur profil de co-évolution et de conservation de leur co-localisation sur le chromosome (« Phydbac »). Je me suis enfin intéressé à l'analyse des événements de fusion de gènes, qui donne des éléments complémentaires aux méthodes de la phylogénomie (« FusionDB »).

2.1 Analyse du protéome : les thermophiles

Dans une étude précédente, Cambillau et Claverie (2000) ont démontré que les protéines issues d'un organisme hyper-thermophile sont plus riches en résidus chargés, au prix d'un nombre de résidus polaires (non-chargés) réduit. Une différence importante des proportions entre ces deux types de résidus est selon cette étude une signature adéquate (bien qu'empirique) du protéome d'un organisme hyper-thermophile, relation nommée *CvP-bias* par les auteurs. Depuis cette étude, le nombre des génomes microbiens a plus que doublé, soulevant la question de la validité de cette relation face à cette nouvelle source d'information. En incluant ces nouveaux organismes, j'ai re-évalué ce critère en me basant sur les génomes de 9 bactéries et archébactéries adaptées à un mode de vie à caractère thermophile, de 9 hyper-thermophiles et de 53 mésophiles, tout en cherchant à identifier d'autres corrélations entre l'hyper-thermostabilité et la composition chimique du protéome. Les résultats de nos analyses (Suhre and Claverie, 2003) confirment la validité du *CvP-bias*. En plus, j'ai pu démontrer que le *CvP-bias* est un critère optimal dans un sens purement

statistique, c.-à-d. qu'il correspond à la composante principale qui discrimine les micro-organismes hyper-thermophiles des mésophiles (Figure 2). En parallèle, j'ai évalué deux autres critères qui ont récemment été proposés comme corrélés à l'hyper-

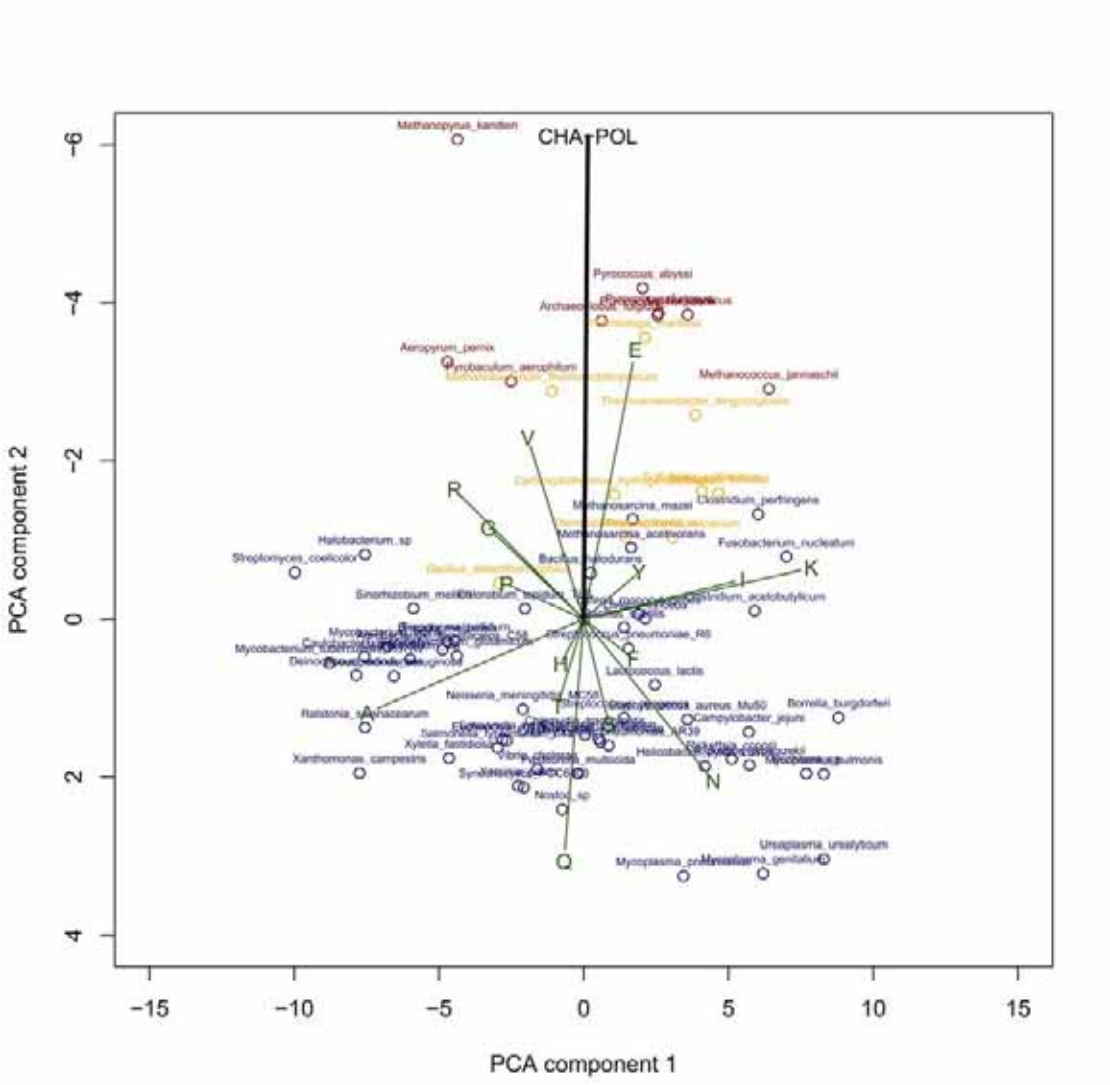

Figure 2 Biplot de l'analyse en composantes principales des thermophiles

Cette figure montre la stratification des micro-organismes en fonction de leur mode de vie (mésophiles – bleu ; thermophiles – orange ; hyperthermophiles – rouge), stratification parfaitement décrite par la deuxième composante de l'analyse en composantes principales (PCA-2). Le vecteur CHA-POL (*CvP bias*) se trouve parallèle à l'axe PCA-2, ce qui indique que le critère du *CvP bias* est en effet un critère naturel – dans un sens purement statistique – de discrimination entre des génomes de caractère mésophile, thermophile et hyper-thermophile. Il est également intéressant de noter que la PCA-1 correspond au contenu en nucléotides G+C des micro-organismes, ce qui se traduit par un positionnement presque parallèle à l'axe PCA-1 des acides aminés qui sont codés soit des codons riches en G+C (par exemple Alanine, lettre 'A' sur le graph), soit par des codons riches en A+T (par exemple Lysine, K).

thermostabilité : le point iso-électrique (pI) moyen du protéome et l'index statistique des di-nucléotides (Kawashima et al., 2000). Mes résultats montrent que seul le *CvP-bias* permet de discriminer clairement les organismes hyper-thermophiles des organismes mésophiles à l'échelle génomique.

2.2 Phylogénomie : PhydBac

Note : Les résultats présentés dans ce chapitre ont été obtenus dans le cadre du D.E.A. et puis de la thèse de François Enault (thèse en cours à l'IGS).

Les méthodes utilisant des profils phylogéniques sont destinées à identifier des liens fonctionnels entre deux protéines d'un même génome en se basant sur l'hypothèse que des protéines impliquées par exemple dans une même voie métabolique, ou celles constituant un complexe multi-protéique, ont une forte probabilité d'avoir évolué de manière corrélée. Ce paradigme a été implémenté pour la première fois par Pellegrini *et al.* (1999) et a été ensuite amélioré par Zheng *et al.* (2002) et Bilu et Linial (2002). Selon, Pellegrini *et al.* (1999) un profil phylogénique d'un gène d'un génome de référence est défini comme une série de bits où chaque bit représente la présence ou l'absence d'un gène homologue dans un génome de comparaison. On considère que deux gènes ont co-évolué si leurs profils phylogéniques ne diffèrent qu'au maximum dans une position, c.-à-d. si leur distance d'« Hamming » est inférieure ou égale à un. Une limitation majeure de cette approche est le choix arbitraire d'un seuil d'homologie à partir duquel deux gènes sont considérés comme orthologues. La méthode que nous avons développée ici évite l'utilisation d'un tel seuil, le concept de relations d'orthologie entre gènes n'étant pas utilisé explicitement.

Nous proposons de remplacer les profils phylogéniques binaires de Pellegrini *et al.* (1999) par des fonctions à valeurs réelles, définies par des scores de BLAST (Altschul et al., 1997) normalisés entre chaque gène du génome de référence et son meilleur « match » dans un génome donné. De plus, nous proposons de normaliser chaque colonne de la matrice de profils phylogéniques de manière à ce que chaque organisme ait le même poids dans les calculs, indépendamment de sa place dans la phylogénie par rapport au génome de référence. La distance de « Hamming » est remplacée par une (quasi-) distance *d*, faisant intervenir la co-variance *c* des profils phylogéniques

normalisés (d=1-|c|). En utilisant la base de données Ecocyc (Karp et al., 2002) comme référence, nous avons évalué différentes méthodes de calcul d'une telle distance phylogénique entre deux gènes à partir de leurs profils de co-évolution, notre critère de sélection étant le nombre d'enzymes identifiées comme liées par leur profile phylogénique qui appartiennent en même temps à une même voie métabolique d'Ecocyc. La meilleure des méthodes que nous avons évaluées augmente le nombre d'enzymes identifiées comme fonctionnellement liées de 25% par rapport à la méthode originale (distance de "Hamming" de Pellegrini et al., 1999). Nous avons également montré que la fraction des faux positifs selon cette méthode est inférieure à 20%.

Afin de compléter l'approche, nous avons introduit et validé une procédure d'annotation automatique de gènes de fonction inconnue en nous basant sur l'annotation de leurs plus proches voisins phylogéniques pour lesquels des informations sont disponibles dans la base de données MultiFun (Serres and Riley, 2000). En utilisant cette base de données comme référence, nous trouvons que la moitié des 3122 attributions automatiques de fonction qui peuvent être faites avec une p-value de 10^{-11} correspondant à des annotations originales de MultiFun pour des gènes de fonction connue (Enault et al., 2003a). L'autre moitié de ces prédictions sont des prédictions originales, prêtes à être testées. Afin de rendre ces résultats facilement accessibles aux utilisateurs, nous avons développé une interface pour le web, disponible à l'adresse suivante : http://igs-server.cnrs-mrs.fr/phydbac/ (Enault et al., 2003b).

Plus récemment, nous avons encore augmenté le potentiel de Phydbac en ajoutant une évaluation de la conservation de la co-localisation de deux gènes sur le chromosome au calcul de la distance. Prises ensemble avec l'augmentation d'un facteur deux du nombre de génomes considérés, ces modifications ont permis d'augmenter de 27% le pouvoir prédictif de Phydbac par rapport à la version précédente (Enault et al., 2004). De nouvelles fonctionnalités (profiles de consensus, « shopping cart », comparaison aux annotation de COG, ...), ainsi qu'un nombre important de génomes de référence (incluant tous les pathogènes majeurs ainsi que des agents potentiels de bio-terrorisme) complètent maintenant la présentation de Phydbac et en font un utilitaire précieux pour les bio-analystes (voir Figure 3).

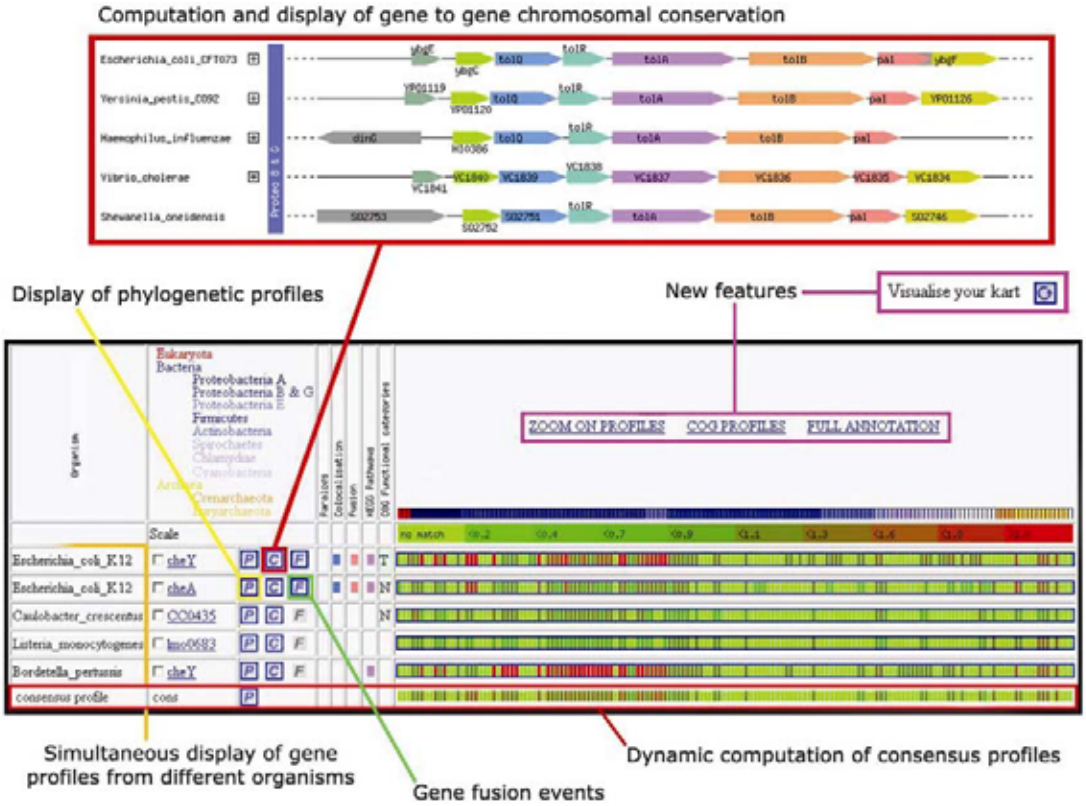

Figure 3 Exemple de sorties de Phydbac

En haut : présentation de la conservation de la co-localisation de plusieurs gènes sur le chromosome de différentes bactéries; en bas : un exemple de profils phylogénétiques ; la présence des différents gènes (lignes) dans les génomes (colonnes) est indiquée par des couleurs allant du vert (pas de match) au rouge (match parfait) ; les accès aux différentes fonctionnalités de Phydbac par des hyper-liens sont annotés sur la figure.

2.3 Fusion de gènes : FusionDB

Des événements de fusion de gènes contiennent de précieuses informations pour la prédiction par exemple d'interactions physiques entre deux protéines ou de leur implication dans les mêmes réseaux métaboliques ou régulateurs (Galperin and Koonin, 2000; Sali, 1999). On appelle ces évènements également les *Pierres de Rosette* de la génomique fonctionnelle en raison de leur rôle de chaînon manquant dans l'identification des relations entre gènes. A l'origine de ces fusions est soit une mutation délétère d'un codon STOP entre deux gènes déjà voisins sur un chromosome (voire dans un opéron), soit un réarrangement chromosomique incluant éventuellement une duplication partielle de certains gènes. Des informations tirées de

ces événements de fusion de gènes peuvent être combinées avec d'autres méthodes de la génomique comparative, classiquement des méthodes basées sur la conservation des profils phylogéniques et la conservation de la co-localisation chromosomique entre deux gènes comme celles décrites plus haut (voire également Enright and Ouzounis, 2001; Marcotte, 2000; Marcotte et al., 1999). Un certain nombre de base de données, accessibles sur Internet, comme par exemple AllFuse (Enright and Ouzounis, 2001), String (von Mering et al., 2003) et Predictome (Mellor et al., 2002) implémentent déjà cette idée. Néanmoins, la plupart de ces bases de données se limitent à la définition d'un événement de fusion de gènes par une paire de « matchs » de BLAST, adjacents et non-chevauchant, entre deux gènes issus d'un génome de référence et un ORF (« open reading frame ») dans un génome cible, sans pour autant apporter l'information contextuelle requise pour une analyse plus approfondie d'un tel événement. Bien que les recherches dans ces bases de données puissent donner des points de départ pour le développement de nouvelles hypothèses, leur taux de faux positifs peut être très élevé (en particulier dans les cas où des gènes ont évolués par duplication et où l'identification d'une véritable orthologie entre deux gènes est par conséquent relativement difficile). De plus, l'utilisateur de ces bases de données se trouve souvent laissé seul, face à la tache de rassembler les informations complémentaires requises pour une analyse complète de ces cas de fusion.

C'est dans cette optique que j'ai développé FusionDB (Suhre and Claverie, 2004, article fourni) une base de données dédiée spécifiquement aux événements de fusion de gènes. Afin d'améliorer l'identification d'orthologie entre deux gènes, je m'appuie sur une définition plus stricte, à savoir un critère de « *mutual best-match* » entre les deux gènes non-fusionés du génome de référence et l'ORF de fusion du génome cible (Tatusov et al., 1997). L'utilisation de ce critère réduit drastiquement le nombre de faux positifs en dépit d'une augmentation du taux de faux négatifs. Afin de remédier à cette limitation, j'ai étendu la notion d'événement de fusion entre deux gènes d'un génome de référence à celle d'une fusion entre deux « *Clusters of Orthologous Groups* » (COGs) (Tatusov et al., 2001), regroupant tous les événements de fusions détectés entre deux gènes appartenant à une paire de COGs dans tous les génomes possibles.

L'analyse de ces événements de fusions entre deux COGs permet l'étude de ces fusions dans leurs contextes phylogéniques, et ceci en se basant sur des alignements multiples et sur la reconstructions d'arbres phylogénétiques (Figure 4). Des questions spécifiques concernant l'histoire d'un événement de fusion de gènes, telles que « Un événement de fusion donné a t-il eu lieu une ou plusieurs fois dans l'évolution ?» ou « Des processus plus complexes comme le transfert horizontal de gènes, la fission ou la dégradation de gènes sont ils impliqués ?» peuvent être adressées en se servant de l'information fournie par FusionDB. De manière plus générale, l'extension du concept de fusion vers des événements de fusion de COGs permet également d'élucider des tendances de fusions dans un contexte génomique et d'adresser des questions comme « Quel type de gènes a la plus grande tendance à fusionner, et pourquoi ? » (Tableau 2).

De cette manière, FusionDB complète notre serveur web PhydBac (décrit ci-dessus) basé sur la même philosophie : fournir des informations détaillées permettant une analyse approfondie des interactions ou des liens fonctionnels potentiels entre deux protéines. FusionDB contient environ 20,000 événements de fusions. 1355 différentes paires de COGs impliquées dans au moins un événement de fusion ont ainsi été identifiées. Comme pour Phydbac, j'ai développé une interface web pour FusionDB disponible à l'adresse suivante : http://igs-server.cnrs-mrs.fr/FusionDB/ (Figure 4)

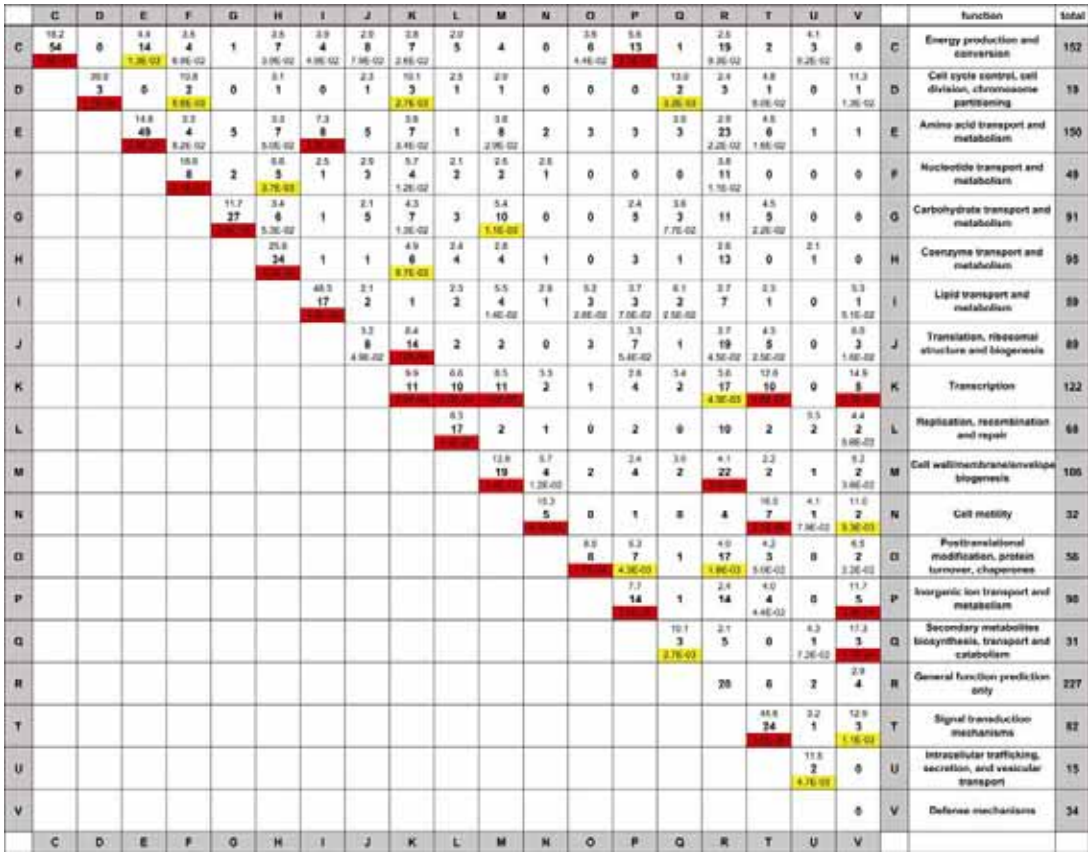

Tableau 2 Evénements de fusion entre différents COGs

Cette matrice présente le nombre d'événements de fusion identifiés entre deux COGs, où chaque COG appartient à une classe fonctionnelle spécifique (abrégées par des lettres, par ex. K pour transcription). Le nombre total d'événements de fusion identifiés pour un couple de COGs est indiqué dans les cellules de cette matrice. Dans le cas où ce nombre est significativement élevé par rapport à ce que l'on attend au hasard, la spécificité (p-value) ainsi que l'augmentation relative de ce nombre sont données en bas et en haut de la cellule, respectivement. Des combinaisons de COGs ayant des p-values inférieures à 0.001 sont surlignés en rouge. On identifie une forte tendance pour que la fusion se produise entre deux gènes d'une même classe fonctionnelle, mais on trouve également des cas de fusion entre gènes de classes différentes ayant des p-values significatives (par ex. gènes liés à la motilité (classe N) et à la transduction de signaux (classe T)).

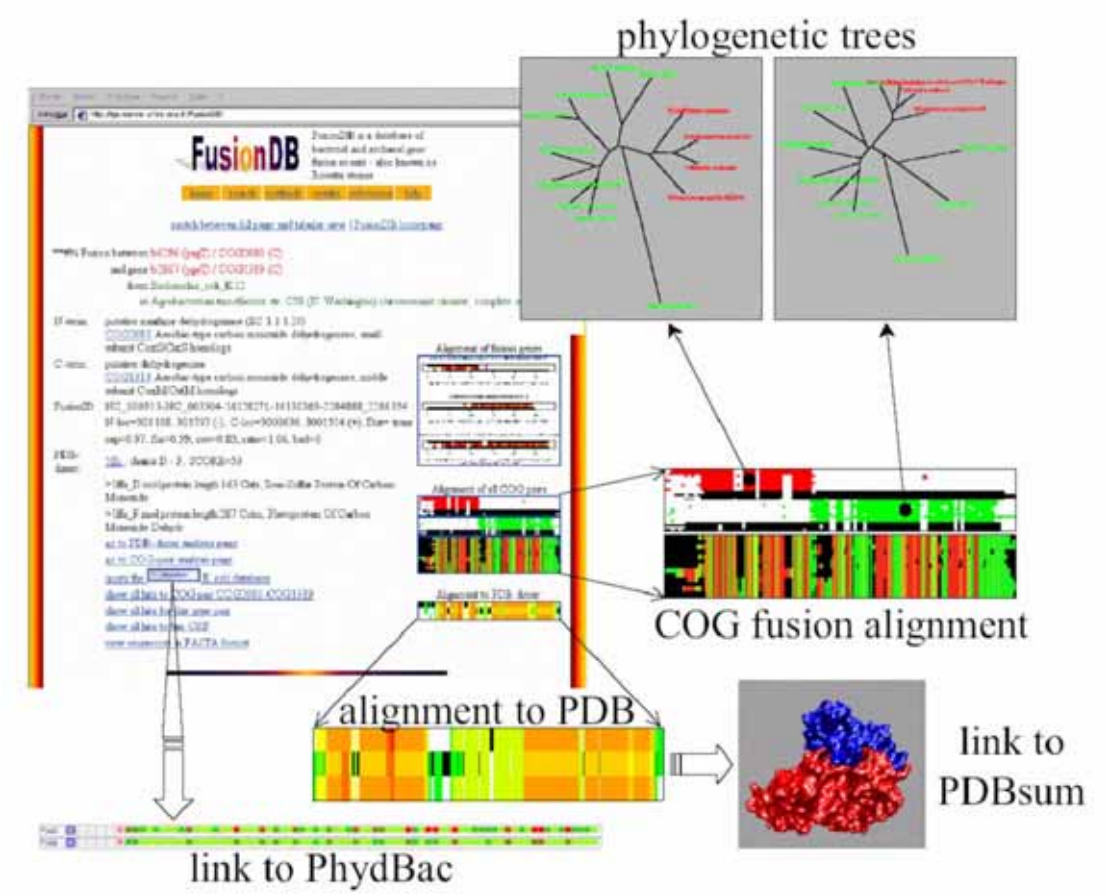

Figure 4 Exemple de sorties de FusionDB

Cette figure présente une page de résultats sur un exemple de FusionDB (en haut à gauche). A partir d'une telle page, l'utilisateur peut accéder à des informations décrivant de manière détaillée chaque événement de fusions : un alignement multiple entre tous les homologues des gènes en fusion (N- et C-terminal) et incluant tous les gènes de fusion identifiés par FusionDB (« *COG fusions alignment* »), des arbres phylogénétiques, un lien direct vers Phydbac, un alignement contenant des structures 3D lorsque la structure des deux gènes en fusion est connue sous forme multimérique, ainsi qu'un nombre d'informations supplémentaires caractéristique de chaque cas particulier.

Chapitre 3 La biologie structurale

Ce chapitre décrit le deuxième grand axe de recherche que j'ai développé dans le domaine de la biologie. Ici, je m'intéresse plus particulièrement à l'utilisation de la bio-informatique pour la biologie structurale, à savoir la modélisation par homologie et l'analyse en modes normaux des structures tri-dimensionnelles d'une protéine donnée. Ces deux méthodes permettent de réaliser des hypothèses fonctionnelles sans être obligé de passer par la modélisation explicite en dynamique moléculaire. Dans le contexte du premier projet de génomique structurale du laboratoire (projet ASG), j'ai notamment développé des méthodologies permettant de résoudre des cas difficiles de remplacement moléculaire basées sur l'utilisation de modèles par homologie et/ou perturbés en modes normaux. Comme dans mes travaux en génomique comparative, j'ai également veillé ici à mettre mes développement à la disposition d'une communauté plus large par l'intermédiaire de serveurs web accessible sur internet.

3.1 Modélisation par homologie et remplacement moléculaire

Après m'être intéressé de manière plutôt générale à la dynamique moléculaire, puis à la modélisation par homologie du fait de leur similitude avec la modélisation de la dynamique et de la chimie atmosphérique, et m'étant familiarisé en parallèle avec les méthodologies numériques de la cristallographie, j'ai rapidement réalisé le potentiel que représentaient les modèles par homologie pour le remplacement moléculaire (RM) en tant que modèles de recherche (« *template* ») généralisés, surtout dans les cas où le RM standard ne donnait pas de résultats concluants. Ce n'est pas forcement une idée très originale, mais c'est à ma connaissance la première fois que cette méthode aie été utilisée de manière systématique pour résoudre des cas réels. L'idée de base est de générer un grand nombre de modèles perturbés du « *template* » original et de tous les tester afin de rechercher une solution en RM. Comme la qualité du modèle dépend sensiblement de la qualité de l'alignement entre la séquence de la cible et la structure du « *template* », je me suis également intéressé aux alignements multiples, notamment aux méthodes qui permettent de combiner l'information de structure avec celle de séquence. Or, un des meilleurs programmes dans ce domaine, le programme T-Coffee, est développé par Cédric Notredame, chercheur à l'IGS. Sa présence au laboratoire nous a amené à collaborer dans le but de développer et d'améliorer des

protocoles d'alignement structures-séquences adaptés à la modélisation par homologie (Notredame and Suhre, sous presse, voir Chapitre 4.2).

La protéine *YecD* d'*Escherichia coli*, protéine issue du projet de génomique structurale ASG du laboratoire IGS, a été la première structure résolue par cette approche (PDB-id 1j2r). Il s'agit d'une protéine de 199 acides aminés qui cristallise dans le groupe d'espace $P2_12_12$ avec 4 molécules par unité asymétrique. Le cristal diffracte à 1.3 Å de résolution. Le RM classique, ainsi que toutes les tentatives de phasage par les méthodes MIR et MAD étant restés vaines, j'ai entrepris d'utiliser des modèles par homologie afin de résoudre sa structure. Le protocole qui a permis la résolution de la structure d'*YecD* par RM est détaillé ci-dessous :

- Identification de « *templates* » en utilisant des serveurs de « *threading* », en particulier Metaserver http://bioinfo.pl/meta/ (Ginalski et al., 2003) et Fugue http://www-cryst.bioc.cam.ac.uk/fugue/ (Shi et al., 2001) :
    - *1nba* (25% d'identité de séquence),
    - *1im5* (20% d'identité de séquence).
- Identification de plusieurs homologues d'*YecD* en effectuant une recherche d'homologie dans la banque de données SwissProt à l'aide du logiciel BLAST (Altschul et al., 1997).
- Alignement multiple structures-séquences par T-Coffee (incluant *YecD*, *1nba*, *1im5*, et les homologues d'*YecD ;* voire Figure 6) en utilisant la méthode d'alignement structure-séquence Fugue (Shi et al., 2001) ainsi que la méthode d'alignement structure-structure SAP (Taylor and Orengo, 1989) – (voir également Notredame and Suhre, sous presse).
- Génération d'un ensemble de modèles (30 dans ce cas) par modélisation par homologie avec MODELLER (Fiser and Sali, 2003; Sali et al., 1995; Sanchez and Sali, 2000) en appliquant une perturbation aléatoire de 4 Å aux modèles initiaux (Figure 7).
- Excision des résidus ayant un faible score d'alignement sur la base du CORE index de T-Coffee afin d'éliminer les parties du modèle dont la structure modélisée est peu fiable (réduction du bruit dans l'identification des solutions en RM)

- Définition des facteurs de température (B-factors) à des valeurs fixes de 20 $Å^2$.
- Recherche de solutions en RM à l'aide du logiciel AMoRe (Navaza, 2001) avec les 9 meilleurs modèles, sélectionnés sur la base de leur fonction objective de MODELLER (à noter que celle-ci montre souvent une bonne corrélation avec la qualité du modèle évaluée par Procheck (Laskowski et al., 1996)). L'indicateur d'une solution « prometteuse » en RM est tout d'abord donné par le contraste existant entre le coefficient de corrélation de la ou des meilleures solutions en rotation et le reste du « peloton », puis par une augmentation successive de ce coefficient à chaque étape de la translation.
- Affinement avec CNS (affinement en corps rigide, recuit simulé et minimisation d'énergie) afin d'évaluer la convergence d'une solution putative (Brunger et al., 1998). L'indicateur finale d'une solution (avant de regarder les cartes de densité électronique) est une réduction sensible du facteur R_{free} entre les différentes étapes de cet affinement.

Le modèle ayant produit la solution se caractérise par une succession d'augmentations du coefficient de corrélation (cc) d'AMoRe lors des quatre étapes de translation (15.1 / 18.3 / 20.0 / 21.2), aboutissant à une corrélation de 27.3 et un facteur R de 52.5% lors de la dernière étape d'AMoRe correspondant à un affinement en corps rigide. Un premier cycle d'affinement avec CNS (minimisation et recuit simulé à 2.7Å) donne un modèle présentant un R_{work} de 41% et un R_{free} de 49%, ainsi qu'une carte de densité électronique permettant une construction non-ambiguë de la molécule entière (à l'exception des 10 premiers résidus en N-terminal désordonnés dans le cristal). En partant de ce résultat, j'ai pu affiner le modèle jusqu'à obtenir une structure finale à 1.3 Å avec un R_{work} de 14.5% et un R_{free} de 16.5% (Figure 5).

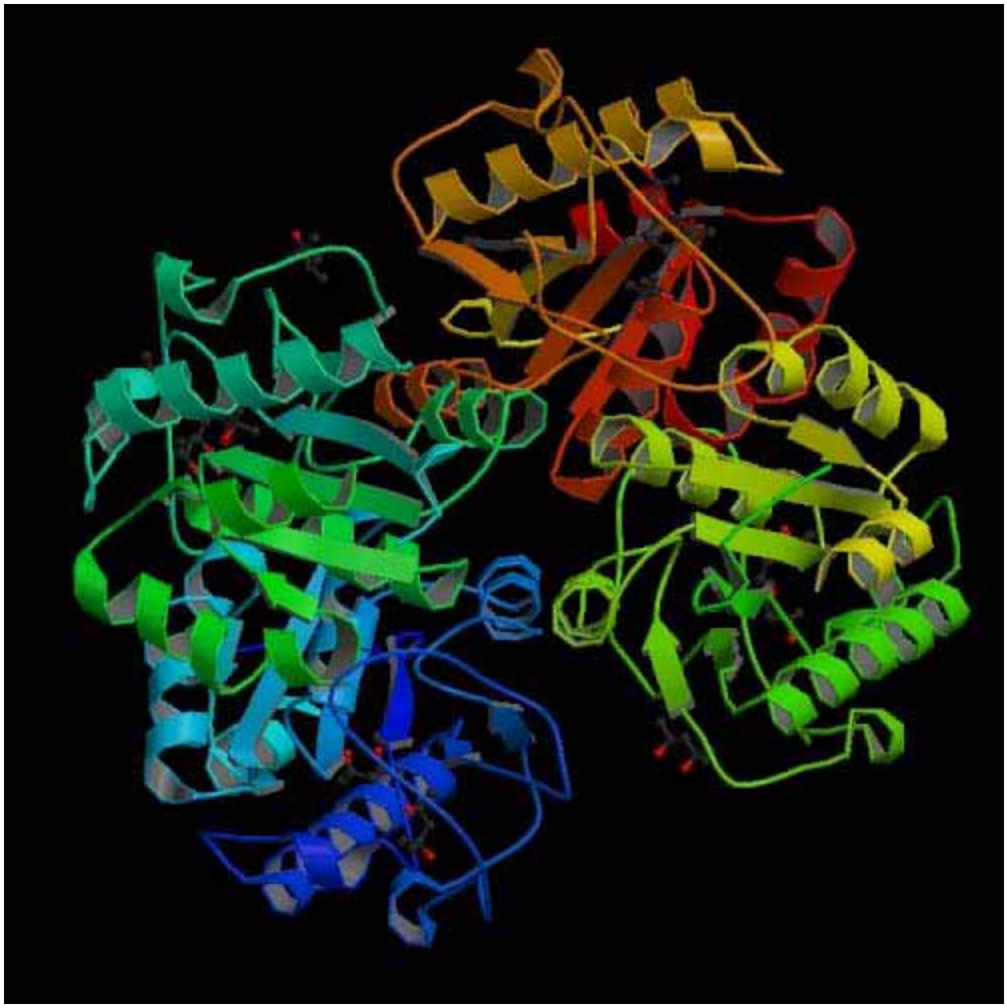

Figure 5 Structure d'*YecD* comme (PDB code 1j2r)

La protéine *YecD* est constituée d'un planché béta parallèle central, entouré des deux côtés par des hélices alpha. Dans le cristal, *YecD* est tétramérique sous forme d'une paire de dimères. Une grande boucle au milieu d'un monomère établi en effet un contact extensif avec la seconde molécule du dimère. Sous cette boucle se trouve le site actif présumé. Il lui manque néanmoins un résidu de la triade catalytique, qui est autrement parfaitement conservé dans ses homologues structuraux, y compris la présence d'une conformation cis-peptidique. Deux molécules 2-methyl-2,4-pentanediol (MPD) occupent la cavité du site actif, suggérant une molécule ayant à peu près deux fois cette taille comme substrat naturel possible d'*YecD*. Tenant compte des propriétés chimiques et des orientations de ces deux molécules de MPD, une molécule de S-Adenosyl-L-methionine (SAM) pourrait par exemple bien remplir la cavité. Des simulations en dynamique moléculaire confirment cette possibilité. Des essais de co-cristallisation *YecD*-SAM sont en cours.

Avant de m'attaquer à *YecD*, la méthode décrite ci dessus m'a déjà permis de résoudre la structure d'une autre protéine issue du projet ASG, le produit du gène *YggV* d'*E. coli*. Après un premier affinement par CNS le modèle d'*YggV* présentait déjà un R_{work} de 36.1% et un R_{free} de 39.8%. Dans ce cas, le RM classique ne

permettait pas non plus de résoudre la structure, et ceci malgré la disponibilité d'un homologue structural présentant 33% d'identité de séquence avec *YggV*. La structure a été résolue en parallèle par la méthode MAD, validant ainsi ma méthode (le RMSD entre tous les atomes de type C-alpha de ce modèle et ceux du modèle issu de la carte MAD et déposé à la PDB est de 0.66Å). Il s'agit donc du second cas réel résolu en utilisant les modèles par homologie.

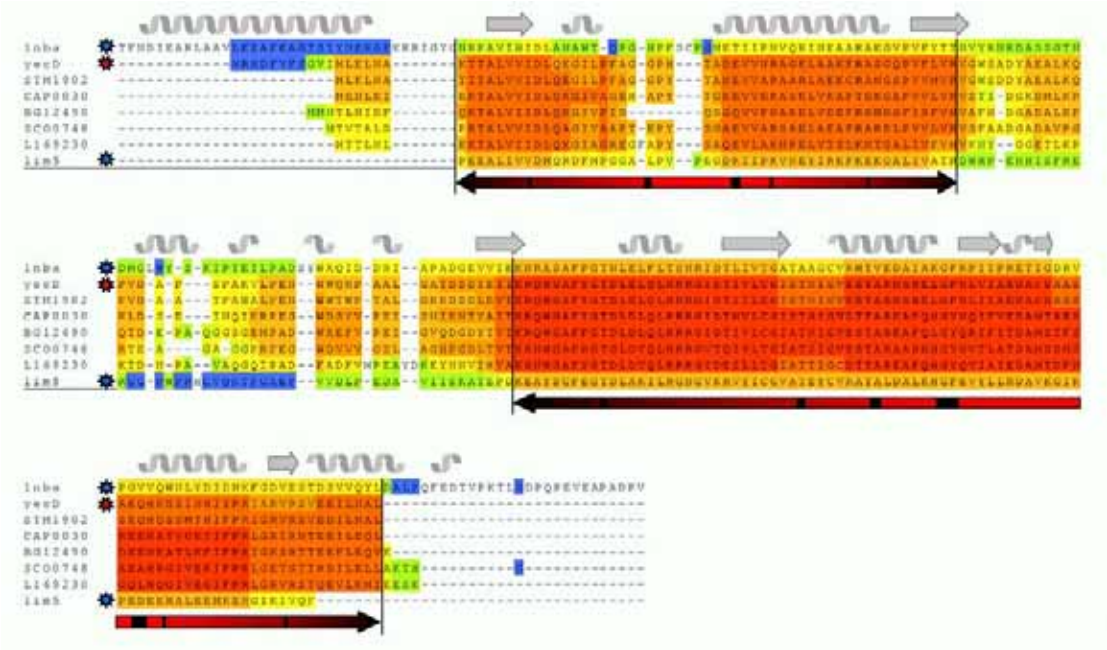

Figure 6 Alignement multiple structure-séquence d'*YecD*

Cet alignement multiple d'*YecD* (étoile rouge), généré par T-Coffee, incorpore les informations issues de l'alignement structurale des deux modèles *1nba* et *1im5* (étoiles bleues), des alignements générés par le « *threader* » Fugue entre toutes les séquences et les deux structures, ainsi que des alignements globaux et locaux entre toutes les séquences. Le CORE index (fond de couleur des résidus) donne la consistance des différents alignements (bleu/vert=faible ; jaune/rouge=fort) et sert d'indicateur de fiabilité de l'alignement. Les zones ayant un CORE index trop faible ont été excisées des modèles avant le remplacement moléculaire (les flèches rouges en-dessous de l'alignement indiquent les parties retenues).

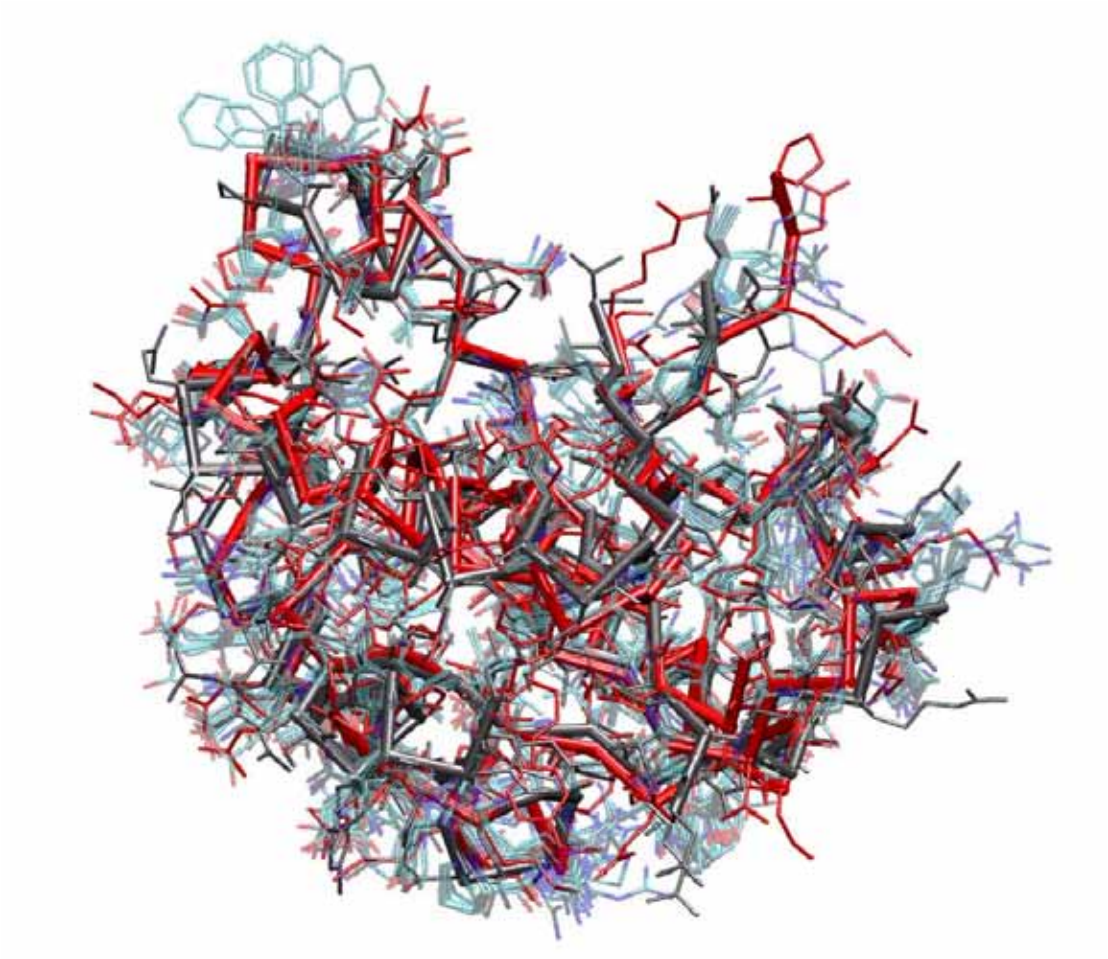

Figure 7 Ensemble de modèles par homologie d'*YecD* crée par MODELLER

Superposition des deux « *templates* » *1nba* (rouge) et *1im5* (gris) et des neuf meilleurs modèles d'*YecD* (sélectionnés en fonction de leur fonction objective de MODELLER). La modélisation de ces modèles est basée sur l'alignement de la Figure 6. Les parties de faible fiabilité ont été excisées et ne sont pas représentées ici.

3.2 Remplacement moléculaire automatisé : CaspR

Note : Les travaux relatifs à CaspR font partie du D.E.A. et de la thèse de Jean-Baptiste Claude.

Afin de déterminer si les deux cas de succès en remplacement moléculaire avec des modèles par homologie présentés ci-dessus (*YecD* et *YggV*) représentaient des cas isolés, ou si ma méthode pouvait se généraliser, j'ai posté une enquête au « *CCP4 bulletin board* ». Par ce moyen, j'ai alors appris qu'au moins deux autres structures protéiques ont pu être résolues en utilisant des modèles par homologie dans des

situations ou le RM classique ne donnaient aucune solution. Il s'agit des protéines suivantes :

- L-hydantoinase (L-hyd) d'*Arthobacter aurescens* (Abendroth et al., 2002, PDB-id 1gkr)
- E2 DNA-binding domain (HPV-31) du *human papillomaviorus serotype 31* (Bussiere et al., 1998, PDB-id 1a7g)

Motivé par cette information, j'ai entrepris, en collaboration avec Chantal Abergel de l'équipe de cristallographie de l'IGS, le développement d'un serveur automatisé et accessible par Internet pour le remplacement moléculaire en utilisant des modèles basés sur la modélisation par homologie. L'implémentation de ce serveur, dénommé CaspR, est réalisée dans cadre de la thèse de Jean-Baptiste Claude. Je me limite ici à résumer l'état de développement de ce serveur.

CaspR est aujourd'hui opérationnel et accessible à la communauté des cristallographes du monde entier (Claude et al., 2004). La Figure 8 montre l'implémentation de mon protocole de RM dans CaspR. Dans sa version de base, l'utilisateur n'a qu'à soumettre ses données cristallographiques dans un format standard (format MTZ de CCP4), la séquence de sa protéine et quelques séquences homologues (afin de guider l'alignement multiple), ainsi que l'identificateur PDB d'un ou plusieurs modèles structuraux. CaspR se charge alors de manière complètement automatique de toutes les étapes du RM par modélisation par homologie : génération de l'alignement multiple structures-séquences par T-Coffee, modélisation par homologie avec MODELLER, excision des boucles mal alignées en fonction du CORE index de T-Coffee, remplacement moléculaire en utilisant AMoRe et pré-affinement avec CNS. En sortie, CaspR produit un résumé visuel pour chaque étape correspondant aux différents programmes utilisés, décrivant clairement l'avancement et le résultat de la recherche (en particulier s'il y a convergence du facteur R lors de l'étape ultime d'affinement). En cas de besoin, CaspR propose un contrôle plus avancé, voir complet, du job par l'intermédiaire de pages d'entrée en modes avancé et expert, permettant à un utilisateur averti de paramétrer les différentes étapes du processus.

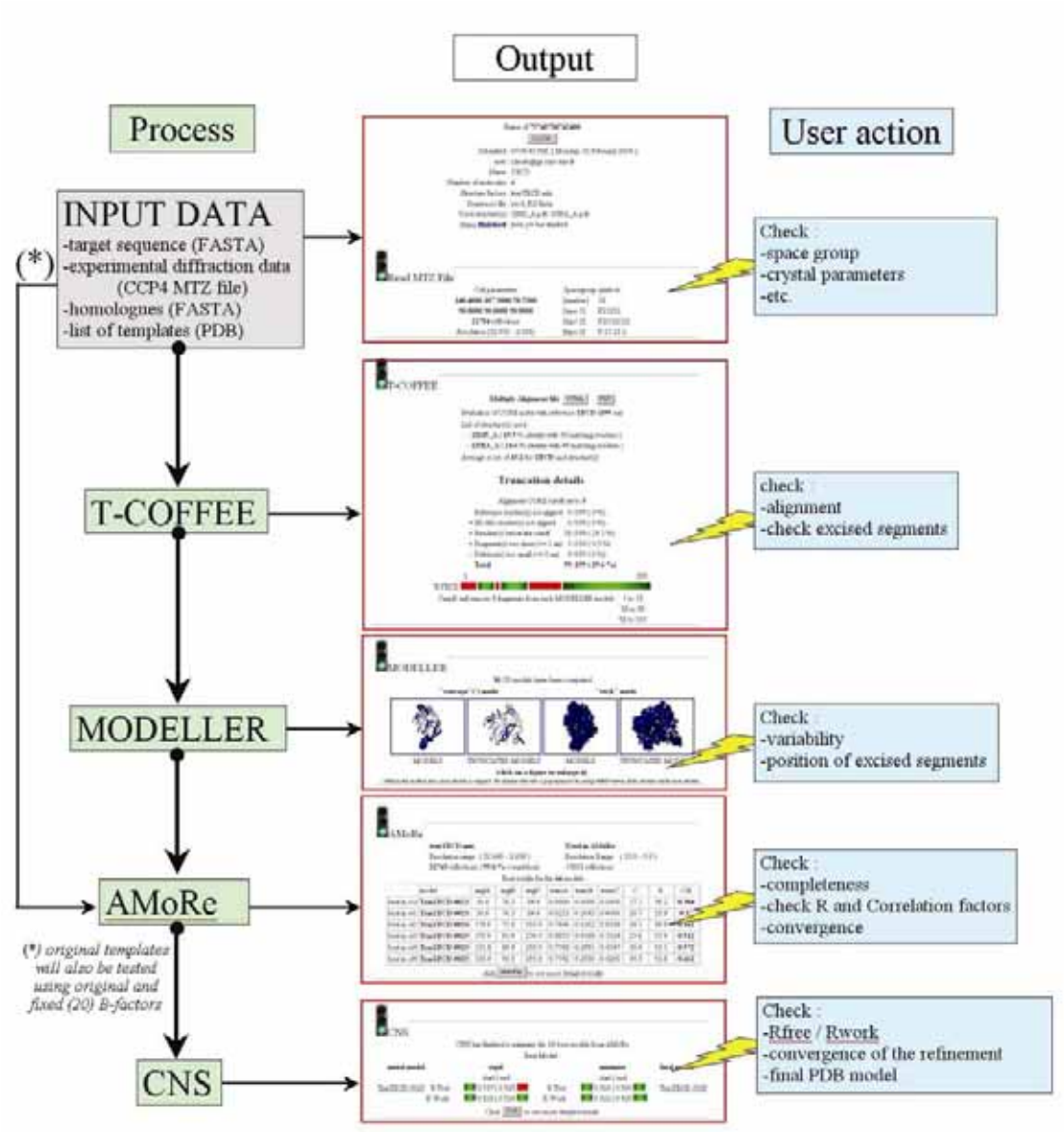

Figure 8 Flowchart CaspR

Cette figure présente des captures d'écran du serveur CaspR avec pour exemple le cas de référence *YecD* (milieu). Le protocole de remplacement moléculaire, tel que je l'ai initialement développé est présenté à gauche. Les différentes vérifications que l'utilisateur aura à effectuer sont indiquées à droite. Pour chaque étape, des pages de sorties plus détaillées sont accessibles (en particulier l'alignement T-Coffee et les résultats d'AMoRe et de CNS pour chaque modèle testé par CaspR).

A ce jour, CaspR a permis de résoudre (en plus de *YecD* et *YggV,* qui servent de benchmark) trois nouveaux cas tests « difficiles » (c.-à-d. des cas où le RM classique ne donnait pas de résultats) :

- *YahK* (protéine cible ASG)
- *1ajx* (HIV-1 protease ; voir également Tableau 3 plus bas)

- *1k6k* (E. coli hydrolase *ClpA*)

Depuis sa mise en accès libre sur le web, CaspR a également permis de résoudre plusieurs cas réels soumis par des utilisateurs externes. Il est à noter que pour les cas les plus « faciles », la solution proposée par CaspR est souvent meilleure que celle trouvée par RM classique à partir de la structure d'un homologue dans la PDB (jugée sur le facteur R après le pré-affinement par CNS). Ceci est particulièrement intéressant dans le cas de données à basse résolution où la qualité du modèle initial est décisive en regard du temps à investir dans la construction et l'affinement manuel de celui-ci.

Il est difficile d'identifier clairement les facteurs clés de la résolution des différents structures résolues à partir de modèles par homologie en RM. D'après mon expérience, il s'agit souvent d'une combinaison des points suivants :

- En appliquant une perturbation initiale aux modèles par homologie, et en en générant un assez grand nombre, on peut échantillonner un espace de conformations assez large pour accrocher la bonne solution.
- Si on utilise plus d'un « *template* » dans la construction de modèles par homologie (comme dans le cas d'*YecD*), on peut générer des chimères de celles-ci, c.-à-d. des modèles qui ne ressemblent plus à un « *template* » isolé mais à une combinaison de structures.
- En utilisant des modèles par homologie, on place les chaînes latérales correspondant à la cible sur les « *templates* ». Comme cette étape dépend de manière cruciale de la qualité de l'alignement multiple, le choix du programme d'alignement est essentiel.
- L'excision des boucles mal alignées (basée sur le CORE index de T-Coffee) s'est également avéré indispensable dans certains cas.

Ce qui est clair, est que plus on étend l'espace des conformations échantillonnées, plus on augmente les chances de trouver la solution en RM (Figure 9).

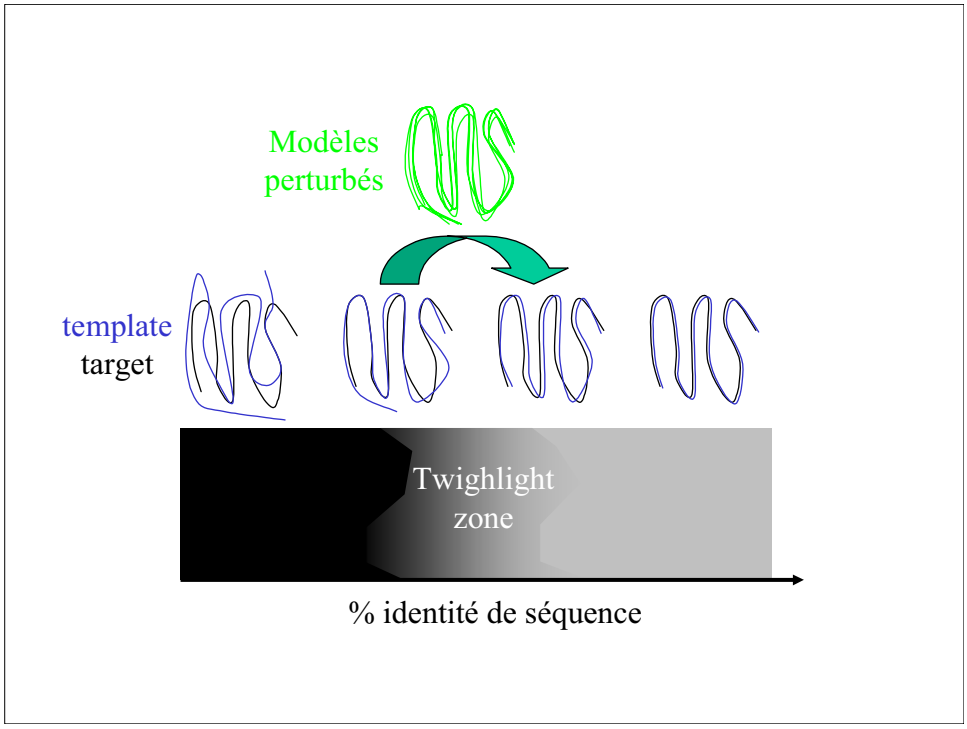

Figure 9 Sortir de la "twighlight zone" en utilisant des modèles par homologie

Ce schéma illustre l'idée de base de l'utilisation des modèles par homologie en remplacement moléculaire : générer un grand nombre de modèles perturbés à partir d'un ou plusieurs « *templates* » initial, qui ne donnent pas de solution (satisfaisante) en RM. En échantillonnant suffisamment l'espace conformationnel autour de ces « *templates* » on pourra éventuellement générer un modèle suffisamment proche de la « *target* » et ainsi obtenir une solution en RM (voir également le schéma présenté en Figure 12)

C'est en cherchant des méthodes permettant de construire des modèles pour le RM encore plus différents du « *template* » initial que ceux produit par la modélisation par homologie, que j'ai pensé à utiliser la technique d'analyse en modes normaux. Grâce à une collaboration avec Yves-Henri Sanejouand de l'Ecole Nationale Supérieure (ENS) à Lyon, j'ai pu exploiter cette voie plus en détail.

3.3 Analyse en modes normaux

Une des méthodes théoriques les plus adaptées à l'étude des mouvements collectifs d'une macro-molécule est l'analyse en modes normaux (AMN). Cette analyse mène à une expression de la dynamique d'une protéine en termes de variables collectives, à savoir les coordonnées en modes normaux (voir Tama, 2003 pour revue). Bien que les premières études en modes normaux aient été effectuées il y a plus de vingt ans par Go (1983) et Brooks (1983), elles restaient limitées à des protéines de taille

relativement modeste (moins de 100 résidus). Ce n'est que récemment, grâce à des méthodes plus avancées (Durand et al., 1994; Li and Cui, 2002; Marques and Sanejouand, 1995; Mouawad and Perahia, 1993; Tama et al., 2000), de descriptions simplifiés de la structure des protéines (Bahar et al., 1997; Hinsen, 1998; Tirion, 1996) ainsi que grâce aux ordinateurs qui devenaient de plus en plus puissants, que l'on a pu aborder l'analyse de systèmes macromoléculaires de plus en plus complexes, y compris les systèmes correspondants à des protéines trans-membranaires, le ribosome, ainsi que des capsides de virus entiers (Delarue and Sanejouand, 2002; Kim et al., 2003; Tama and Sanejouand, 2001; Tama et al., 2003).

En analysant les mouvements de 3800 protéines de structure connue dans deux conformations différentes (incluant des homologues), Krebs et al. (2002) ont démontré que plus de la moitié de ces mouvements peut être approchée par une perturbation appliquée dans la direction d'au maximum deux modes normaux de basse fréquence, un seul mode normal de basse fréquence étant souvent suffisant quand le caractère collectif du mouvement de la protéine est évident. Dans ce cas, il s'agit en règle générale d'un des trois modes de plus basse fréquence (Delarue and Sanejouand, 2002; Tama and Sanejouand, 2001). Ces résultats suggèrent fortement que le mouvement lié au changement de conformation d'une protéine (par exemple lors de la fixation d'un ligand) est soumis à une sélection évolutive, contraignant la protéine à suivre principalement un seul, voire un nombre limité de modes normaux de basse fréquence. Autrement dit, la séquence en acides aminés de la protéine aurait évolué de manière à ce que des barrières de faible énergie soient rencontrées quand la protéine est déformée en suivant les coordonnées du ou des modes normaux correspondants.

Une application importante des modes normaux est l'identification du changement conformationnel putatif d'une protéine lors de la fixation d'un substrat (Delarue and Sanejouand, 2002; Tama et al., 2000; Tama and Sanejouand, 2001). Plus récemment, l'AMN a également été utilisée dans l'étude de l'ouverture d'un canal membranaire (Valadie et al., 2003), dans l'analyse du mouvement structural du ribosome (Tama et al., 2003), dans la maturation d'une capside virale (Kim et al., 2003), dans la trans-conformation de la SERCA1 Ca-ATPase (Li and Cui, 2002; Reuter et al., 2003), dans les changements conformationels tertiaire et quaternaire de l'aspartate

transcarbamylase (Thomas et al., 1999) ainsi que dans l'analyse des mouvements de grandes protéines en général (Hinsen, 1998; Hinsen et al., 1999). L'AMN est souvent utilisée afin de prédire le type de changement conformationnel adopté par une protéine afin de remplir sa fonction. L'AMN peut également être utilisée pour vérifier si un changement conformationnel proposé sur la base de données expérimentales (de caractère non-structurale) est probable (un exemple récent est l'étude de l'ouverture d'un canal trans-membranaire, Valadie et al., 2003). L'AMN a été proposée comme outil de prédiction de mouvement de grande amplitude pouvant potentiellement améliorer la résolution finale de la reconstruction de particules individuelles par cryo-microscopie électronique (Brink et al., 2004). De plus, le fait que 50% des mouvements de protéines connus à ce jour puissent être correctement prédits avec seulement un ou deux modes normaux suggère une application possible de l'AMN en cristallographie, à savoir la génération de modèles structuraux perturbés dans le sens d'un ou de deux modes normaux, ensuite utilisés comme modèles de recherche en remplacement moléculaire. Nous avons démontré que cette approche permet de résoudre des problèmes difficiles de « phasage » où des modèles de recherche non-perturbés ne donnaient pas de solutions satisfaisantes (Suhre and Sanejouand, 2004b, voire plus bas).

L'AMN constitue donc un outil puissant avec un domaine large d'application en biologie structurale et notamment en cristallographie. J'ai ainsi développé un serveur web, ElNémo (pour **El**astic **Ne**twork **mo**del), accessible à l'adresse http://igs-server.cnrs-mrs.fr/elnemo/ (Suhre and Sanejouand, 2004a, article fourni). ElNémo permet d'aborder l'AMN de grosses protéines de 500 à 1000 résidus et plus en mode tout-atome, l'objectif du serveur étant en particulier le calcul d'un grand nombre de modèles de recherche pour le remplacement moléculaire (Figure 10).

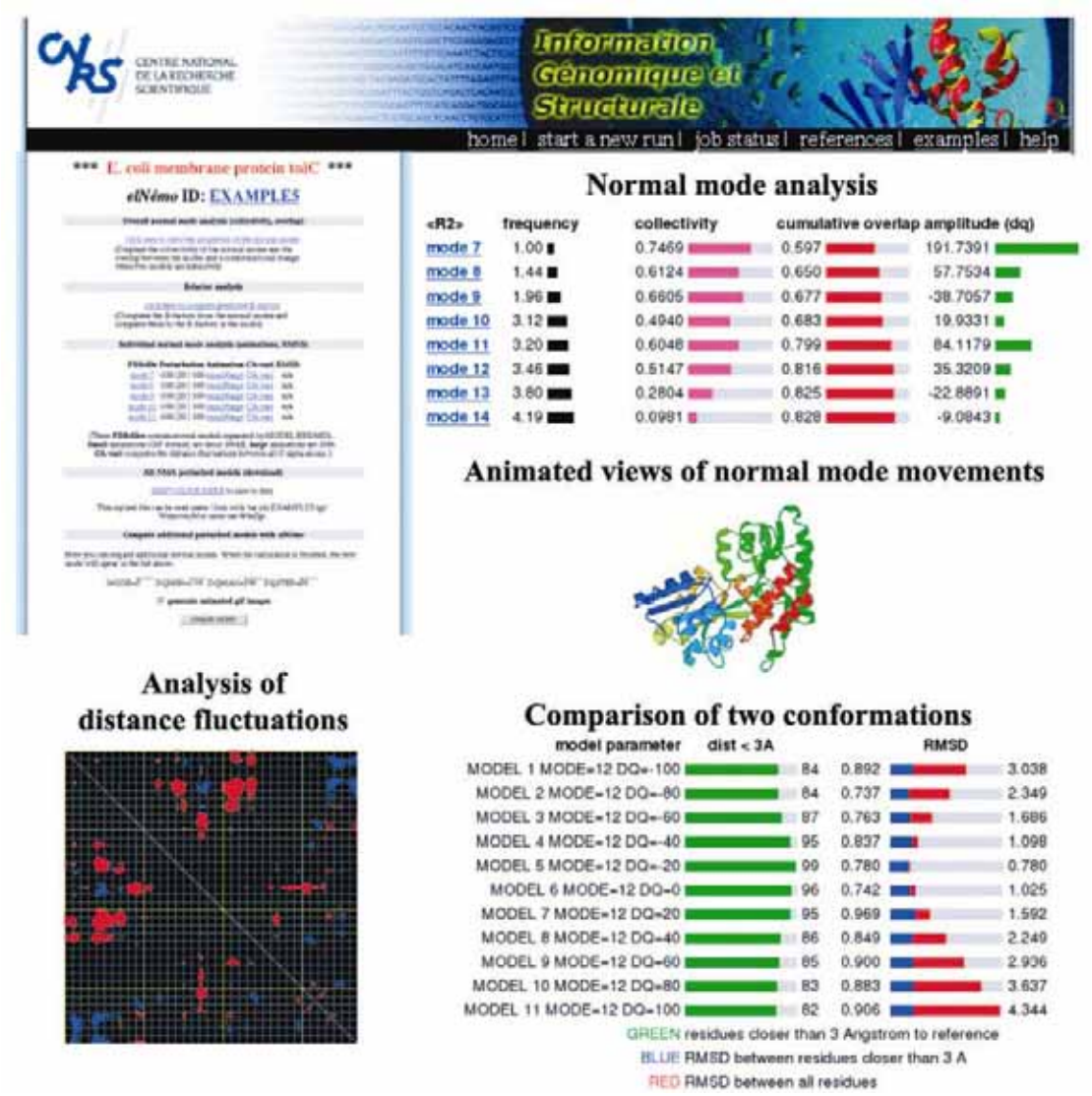

Figure 10 Le serveur AMN « *elNémo* »

Cette figure présente un extrait des différentes sorties d'ElNémo : à partir de la page principale (en haut à gauche), on peut accéder à l'analyse des modes normaux en fonction de leur fréquence (en haut à droite), à une comparaison de deux modèles homologues (calcul de RMSD) ainsi qu'à une analyse des variations de distance entre deux résidus. Chaque mode est également visualisé sous forme d'animation, générée avec *molscript* (Kraulis, 1991).

3.4 Les modes normaux en remplacement moléculaire

Le remplacement moléculaire (RM) est la méthode de choix, en termes de coût et de temps, dans la résolution de la structure tri-dimensionnelle d'une protéine par cristallographie. Multiples sont les facteurs gouvernant le succès ou l'échec du RM. A citer en particulier la qualité des données de diffraction à basse et à moyenne résolution ainsi que le nombre de molécules par unité asymétrique à positionner. Le facteur déterminant est néanmoins la disponibilité d'un modèle structural ayant une homologie forte avec la cible. Il est probable que tout cristallographe ait déjà vécu la situation décevante où le RM échoue en dépit d'un bon modèle de recherche, présentant une forte similitude de séquence (voir même une identité de séquence) avec la cible. Ayant finalement phasé les données de diffraction à l'aide de méthodes expérimentales plus coûteuses en temps et en argent, comme le remplacement isomorphe multiple (MIR) ou la diffraction anomale à multiples longueurs d'ondes (MAD), on trouve souvent dans ces cas « pathologiques » que la « nouvelle » structure présente un changement conformationnel important par rapport au modèle de recherche initial, ce qui explique *a posteriori* l'échec du RM. Pouvoir modéliser *a priori* les changements conformationnels les plus probables d'un modèle de recherche est alors d'un intérêt majeur, permettant par conséquence d'augmenter le nombre de structures cristallographiques pouvant être résolues par RM.

Comme je l'ai déjà évoqué plus haut, nous avons proposé l'analyse en modes normaux comme outil puissant d'anticipation des changements conformationnels les plus probables du modèle de recherche en RM et tester des modèles perturbés dans la direction d'un ou de plusieurs modes pour accrocher une solution potentielle en RM (Suhre and Sanejouand, 2004b, article fourni). La motivation d'une telle approche est la suivante : il y a presque vingt ans, on a découvert en utilisant des champs de force empiriques et une description de la protéine au niveau atomique, qu'un des mouvements parmi ceux de plus grande amplitude prédits par la théorie des modes normaux, (c.-à-d. un mouvement de basse fréquence), se compare souvent très bien avec le changement conformationnel lié à la fixation d'un ligand (Brooks and Karplus, 1983).

Inspirées par cette observation, nous avons proposé une approche de criblage, où un modèle de recherche en RM est perturbé en appliquant différentes amplitudes en suivant la direction d'un ou deux modes normaux de basse fréquence. On cherche alors des minima du facteur R (ou du coefficient de corrélation) en fonction de l'amplitude, soit après une simple étape de RM (par exemple avec AMoRe), soit après une deuxième étape correspondant à un cycle de minimisation d'énergie et/ou de recuit simulé supplémentaire (par exemple avec CNS). Afin d'évaluer le potentiel de cette idée, qui a déjà été considérée dans le passé, dans le cas isolé de la structure à basse résolution du l'actin F (Tirion et al., 1995), nous avons sélectionné des structures de protéines disponibles sous deux conformations différentes (Echols et al., 2003) et pour lesquelles les facteurs de structure ont été déposés dans la *Protein Data Bank* (PDB, Berman et al., 2000). Des protéines à deux domaines distincts qui sont interconnectés par une seul chaîne peptidique (comme par exemple la toxine diphtérique ou l'immunoglobuline) ont été exclues car la procédure de RM standard traiterait ce genre de problèmes par une approche de recherche à deux corps. Nous avons également omis les protéines dont le changement conformationnel est relativement faible, car ces cas ne représentent pas de véritable défi pour le RM. L'objectif est alors d'utiliser une des deux conformations de la protéine comme modèle de recherche en RM afin de résoudre la structure de l'autre.

Sur la base de trois exemples représentatifs, nous avons mis en évidence le potentiel de cette méthode. Dans les trois cas, le modèle de recherche original n'a pas donné de solution satisfaisante en RM, contrairement à l'un des modèles perturbés en modes normaux. Il s'agit de (1) la *maltodextrin binding protein* (Zanotti et al., 1992), (2) de la *HIV-1 protease* (Backbro et al., 1997), et (3) de la *glutamine binding protein* (Hsiao et al., 1996). Dans ces trois cas, l'application d'une perturbation suivant seulement un ou deux modes normaux de basse fréquence permettait de réduire le facteur R (R_{free}) d'au moins six points permettant de passer en dessous de la barre des 50% séparant typiquement le bruit de fond d'un signal physique (voire Tableau 2 pour plus de détails). La solution a été validé par inspection de la carte de densité électronique ainsi que par un affinement en introduisant des données à plus haute résolution (test de la convergence de l'affinement). Pour des raisons de simplicité nous avons choisi de limiter cette analyse à des cas présentant des homologies de séquence de 100% (identité). Ces résultats sont directement extrapolables à des situations plus générales

où la séquence de la cible et celle du modèle de recherche sont différentes. Dans ces cas, un nombre varié de protocoles peut être appliqué, incluant l'utilisation de modèles en tout-alanine ou de modèles basés sur la modélisation par homologie (comme décrit plus haut) à différentes étapes de la procédure.

	Maltodextrin-binding protein	HIV-1 protease	Glutamine-binding protein
Target	1omp	1ajx	1wdn
Template	1anf	1hhp	1ggg
No. residues	370	99	224
No. reflections†	5851	4280	3796
Completeness‡ (%)	97.9	78.9	97.1
Space group	$P1$	$P2_12_12_1$	$P2_12_12_1$
No. molecules	1	2	1
Best mode	Mode 7	Mode 11	Modes 7 + 8
Perturbation§	180	60	200 + 40
CC/R factor¶			
Target	86.0/20.9	73.2/31.3	72.1/29.9
Template	25.8/49.0	31.0/50.0	26.1/50.3
Best NMA	33.5/46.4	57.7/39.7	21.3/51.6
R_{work}/R_{free}††			
Target	16.5/23.3	35.4/38.6	25.9/35.9
Template	43.0/51.1	52.8/54.1	45.3/54.0
Best NMA	38.8/45.3	41.8/46.0	35.2/47.1

† Number of reflection theoretically available to a resolution of 3.2 Å. ‡ Completeness to a resolution of 3.2 Å. § Arbitrary units. Using Tirion's elastic network model, normal-mode frequencies as well as the corresponding unit for the displacements along a normal mode are defined through a scaling free factor, which was set to $k = 10$ in the present study. ¶ The CC and R factor of the best translation/rotation solution(s) found by *AMoRe* (Navaza, 1994) when using data to 3.2 Å resolution. †† Final R factor for the working and the test set after *CNS* (Brünger *et al.*, 1998) refinement using standard parameters to 3.2 Å resolution.

Tableau 3 Statistiques du remplacement moléculaire utilisant les modes normaux

Ce tableau résume les différents paramètres des trois cas de référence de notre étude (Tableau extrait de Suhre and Sanejouand, 2004b).

Il est intéressant de noter que les modes normaux ont déjà été utilisés par le passé pour affiner des facteurs d'agitations atomiques (B-facteurs) (Diamond, 1990; Kidera and Go, 1992). On observe en effet une bonne corrélation entre les B-facteurs calculés à partir des modes normaux et les B-facteurs cristallographiques (Bahar et al., 1997). Il est alors possible d'utiliser des B-facteurs calculés comme première approximation des valeurs expérimentales lors du RM, une possibilité intéressante que nous n'avons pas encore prospectée.

En somme, notre approche peut être considérée comme représentant une succession de perturbations du modèle de recherche initiale dans des directions différentes (mais toujours physiques), jusqu'à ce qu'on s'approche suffisamment de la cible pour qu'une solution soit trouvée par RM. Même dans des cas plus simples, où une

solution en RM existe, cette méthode peut être intéressante afin d'améliorer le modèle avant l'étape d'affinement manuelle, réduisant ainsi le temps à passer sur cette partie de la résolution d'une structure. Ceci est particulièrement intéressant en ce qui concerne des données à basse résolution. Jones (2001) a déjà discuté le potentiel de ce genre d'approche, à savoir l'utilisation de modèles issue de méthodes de reconnaissance de repliement, sans toutefois faire allusion aux modes normaux comme nous le présentons ici. Enfin, l'augmentation du rendement du RM aura un impact majeur sur le débit des différents projets de génomique structurale (Rupp et al., 2002). En nous basant sur les analyses de Krebs *et al.* (2002), nous estimons que la moitié des cas difficiles en RM, liés à des changements conformationnels du modèle de recherche, pourrait être résolu par l'implémentation de notre protocole.

Récemment, en utilisant des modèles perturbés en modes normaux en combinaison avec la modélisation par homologie, nous avons pu résoudre les deux premiers cas réels, dont un d'intérêt exceptionnel pour le laboratoire : la tyrosyl-tRNA synthétase (*MV2*) de *Mimivirus*. Un seul jeu de données natif, provenant d'un cristal unique de cette protéine à 2.6 Å de résolution a été obtenu à ce jour. La structure est actuellement en cours d'affinement. Le deuxième cas (*YahK* du projet ASG) correspond à une oxido-reductase de spécificité inconnue, pour laquelle un jeu de donnée à 1.6 Å a été obtenu. La structure d'*YahK* a été résolu en même temps par la méthode MAD, et nous avons ainsi pu comparer la carte de densité électronique obtenue par remplacement moléculaire et modes normaux. La Figure 11 montre la structure finale d'*YahK* obtenu par la méthode MAD dans la carte de densité électronique obtenu par AMN.

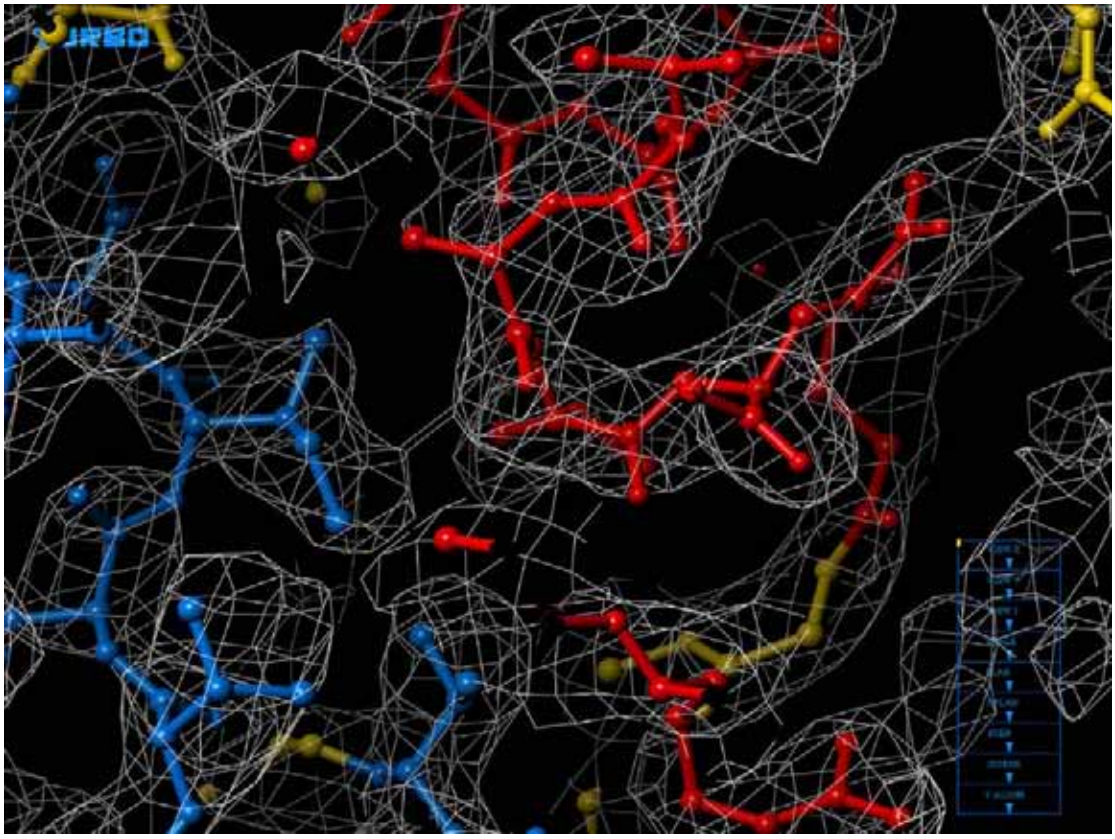

Figure 11 YahK dans la carte de densité électronique générée par NMA

Cette figure (générée avec Turbo, Roussel et al., 1990) présente le modèle final d'*YahK* construit à partir des données MAD. Ce modèle a été placé dans la carte de densité électronique générée à partir d'une solution en remplacement moléculaire avec AMN (modèle perturbé en appliquant deux modes normaux, ensuite converti en mode tout-Alanine et affiné automatiquement avec Refmac5 à 1.7Å de resolution).

En résumé, on peu dire que les méthodes que je viens d'exposer dans ce chapitre sont très complémentaires des efforts réalisés en parallèle dans le domaine du remplacement moléculaire (amélioration des programmes de RM). Pour visualiser ce concept (présenté en Figure 12) imaginons une structure cible (« *target* ») et un modèle de recherche (« *template* ») qu'on représente dans un espace fictif (défini par exemple par les 3N coordonnées des atomes de la protéine). On peut facilement s'imaginer que le succès du remplacement moléculaire dépend de la distance qui existe entre la « *target* » et le « *template* » dans cet espace fictif. On peut alors définir un « rayon de convergence » du remplacement autour de la « *target* ». Si le « *template* » se trouve à l'intérieur de ce rayon, alors le RM devrait donner une solution. Or, ce rayon de convergence du RM n'est pas une constante, mais il dépend principalement de deux paramètres : la puissance du programme de RM qu'on utilise

(AMoRe, Phaser, Molrep pour n'en citer que quelques-uns) et de l'habilité et de la persévérance de son utilisateur. Une amélioration du programme de RM ou une meilleure formation de son utilisateur peu alors augmenter ce rayon et donc le taux du succès en RM. L'utilisation des modèles par homologie ainsi que des modèles perturbés en modes normaux attaque le problème par l'autre côté : en perturbant la structure initiale du « *template* » on échantillonne l'espace conformationnel autour de celle-ci. Si alors l'espace échantillonné chevauche le rayon de convergence du RM on devrait pouvoir trouver une solution. Le fait de visualiser la problématique du RM sous cette forme (idéalisée) mène aux constats suivants :

- Plus on échantillonne l'espace conformationnel, plus on a de chances de trouver une solution en RM.
- L'utilisation des modes normaux représente un échantillonnage plus « dirigé » de cet espace que la modélisation par homologie, augmentant considérablement l'efficacité de la méthode.

Avec le nombre toujours croissant de structures disponibles dans la PDB et une augmentation toujours exponentielle des puissances de calcul des ordinateurs, il est probable que l'approche que j'ai décrite ici permettra dans le futur d'augmenter significativement le taux de succès du remplacement moléculaire (*YecD* et *MV2* en sont les deux premiers exemples).

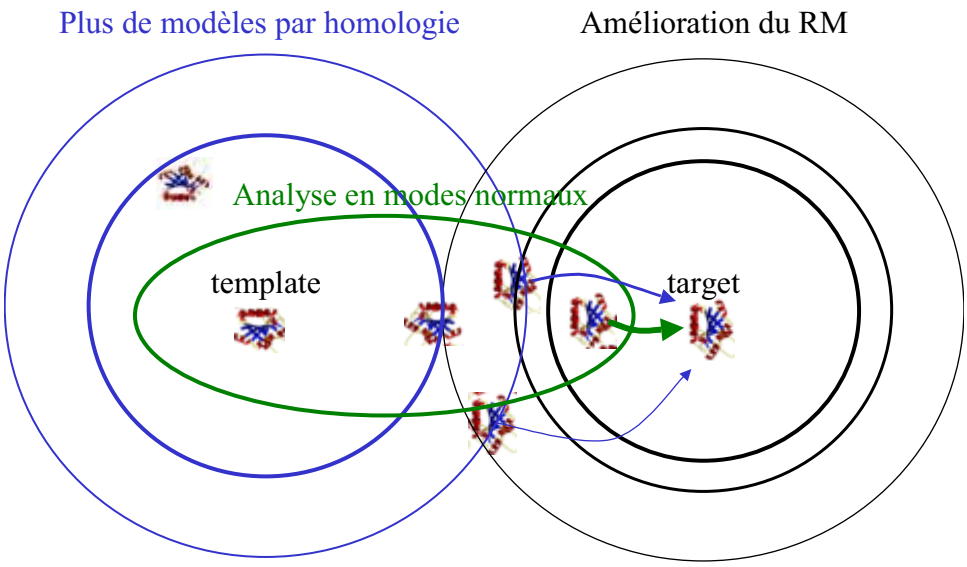

Figure 12 Amélioration du succès en RM

Ce schéma visualise la complémentarité entre les efforts d'amélioration des programmes de RM (cercles noirs) et les idées développées ici, à savoir la recherche d'une solution en RM dans un grand nombre de modèles perturbés (cercles bleus et vert).

Chapitre 4 Autres projets en bio-informatique

Comme je l'évoquais déjà plus haut, j'ai participé, parallèlement au développement de mes propres axes de recherche (Chapitre 2 et Chapitre 3), à d'autres projets en collaboration grâce à mes compétences en analyse statistique et en modélisation. Dans ce chapitre, je présenterai les trois projets les plus significatifs ayant tous donné lieu à des publications dans des journaux internationaux. Il s'agit du séquençage et de l'annotation du génome de la bactérie *Tropheryma whipplei* (section 4.1), du développement de protocoles d'alignement multiple structures-séquences, adaptés à la modélisation par homologie (section 4.2) et de l'analyse des données de puces à ADN d'un mutant de la levure *Saccaromyces cerevisiae* (section 4.3).

4.1 Séquençage de Tropheryma whipplei

L'IGS entretient depuis plusieurs années une collaboration étroite avec l'équipe de Didier Raoult (Unité des Rickettsies à l'hôpital de la Timone à Marseille). Lors de mon arrivée au laboratoire IGS, les deux équipes travaillaient sur le séquençage du génome de la bactérie *Tropheryma whipplei* (Raoult et al., 2003). *T. whipplei* est un pathogène intracellulaire à génome réduit (<1Mb), comparable aux familles des rickettsies, mycoplasmes, spirochètes, chlamydia et à d'autres parasites intracellulaires. C'est le premier génome réduit de la famille des bactéries Gram-positive et riche en G+C complètement séquencé. Nous avons ainsi pu aborder la question de l'adaptation d'une bactérie à une vie de parasite intracellulaire strict. Il est clair qu'une bactérie, une fois engagée dans cette voie et ayant perdu des gènes essentiels à une vie « autonome », peut se permettre de perdre encore plus de gènes alors devenus « superflus », et ceci à un rythme soutenu, quitte à ne garder qu'un ensemble minimal de gènes indispensables à la survie dans sa niche de parasite intra-cellulaire. Les mycoplasmes avec leurs génomes ultra-réduits représentent un exemple extrême de ce genre d'évolution.

Afin de mettre en évidence la (ou les) « stratégies » qui mènent à la vie de parasite intracellulaire strict, notre idée était de traiter les données des génomes des différents bactéries parasitaires selon la même approche que celles obtenues par des expériences de puce à ADN (Figure 13). Dans cette représentation, chaque colonne correspond à un micro-organisme et chaque ligne représente un gène ou une classe fonctionnelle.

La présence ou l'absence d'un gène ou d'une fonction dans une bactérie donnée est déterminée par le logiciel BLAST en utilisant deux génomes évolutionnairement distants comme référence, en l'occurrence *Escherichia coli* (Gram négative) et *Mycobacterium tuberculosis* (Gram positive). Le résultat de cette comparaison peut être présenté comme les données issues des expériences de puces à ADN, sous forme de matrices en couleurs dégradées allant du vert (pas de match) au rouge (match parfait). Une fois mis sous cette forme, ces données se prêtent à toutes les analyses initialement conçues pour les données des puces à ADN, notamment aux différentes méthodes de « clustering ». La Figure 13 présente un arbre issu d'un tel clustering. On remarque que les bactéries à génome réduit se regroupent en raison du grand nombre de fonctions qu'elles ont presque toutes perdues. Néanmoins, chaque clade de ces bactéries à génome réduit constitue à elle seul une branche quasiment indépendante dans le sous-arbre des bactéries à génome réduit.

Contrairement à notre hypothèse initiale, ces résultats montrent qu'il n'existe en fait pas une seule voie d'évolution à la vie de parasite intracellulaire strict, mais que chaque clade s'est en fait adapté à sa façon. L'idée généralement défendue selon laquelle cette adaptation représente un exemple d'évolution convergente serait alors à rejeter.

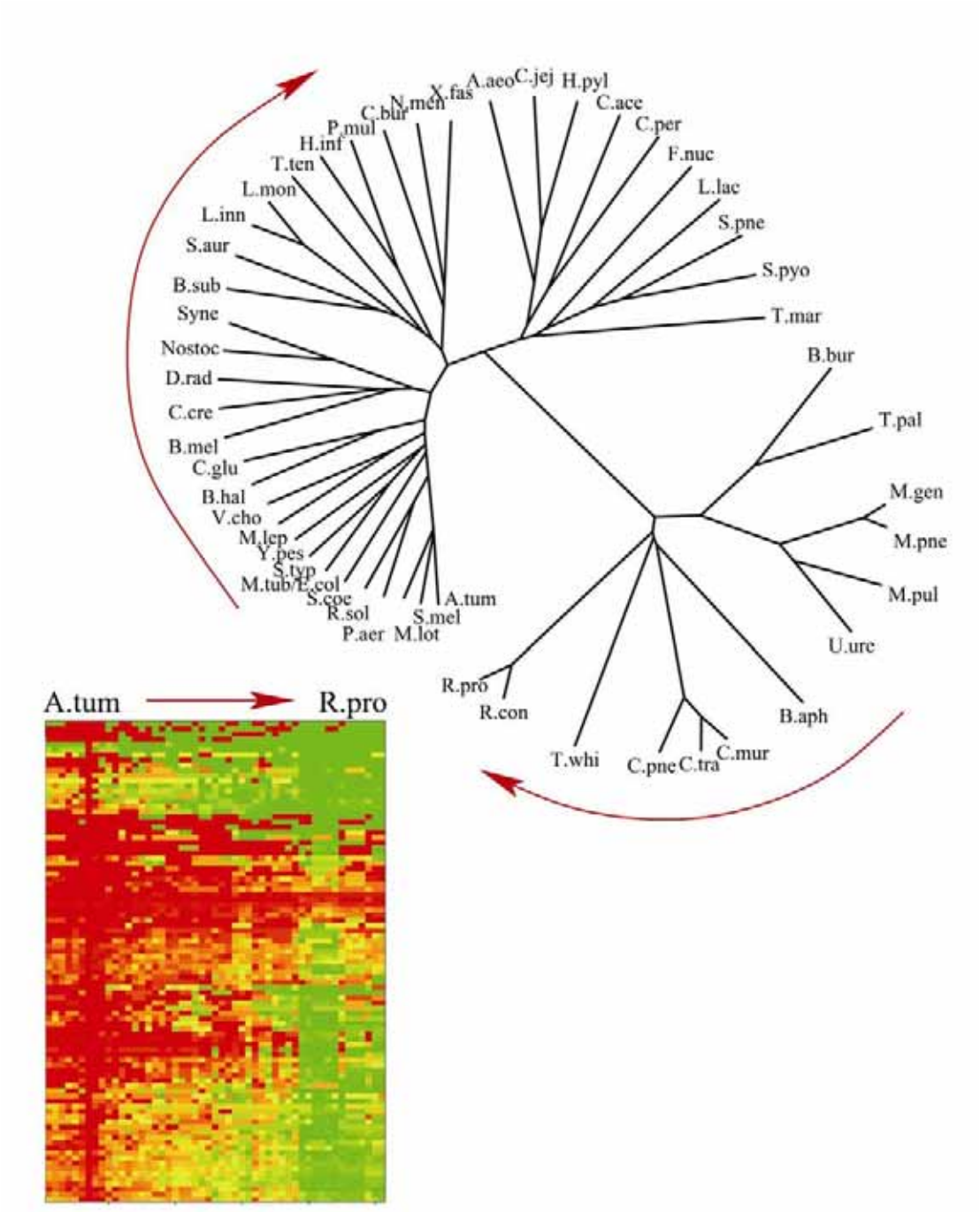

Figure 13 Puce à ADN virtuelle et arbre basé sur la co-présence de fonctions

Présentation sous forme de « puces virtuelle à ADN » du potentiel codant des bactéries (l'ordre est indiqué sur l'arbre par la flèche rouge, la colonne à gauche correspondant à *A. tumefaciens*, celle de droite à *R. prowazekii*) : chaque colonne de la « puce » correspond à une bactérie et chaque ligne représente une classe fonctionnelle (basée sur les annotation de *M. tuberculosis* par le Sanger Centre). Des couleurs chaudes indiquent que plus que la moitié des gènes présents à la fois dans E. coli et M. tuberculosis sont également identifiés dans la bactérie correspondante. Des couleurs vertes marquent la disparition d'une telle fonction. A partir de cette information, nous avons construit un arbre qui a la propriété intéressante de grouper les bactéries ayant un potentiel codant similaire. Il montre en particulier que chaque clade des bactéries à génome réduit a poursuivi sa propre voie vers le mode de vie de parasite intracellulaire strict.

4.2 Alignements multiples séquences-structures

Le programme d'alignement multiple T-Coffee, développé par Cédric Notredame à l'IGS est selon les dernières inter-comparaisons de programmes d'alignement un des meilleurs de sa classe (O'Sullivan et al., 2004). Permettant la construction d'alignements de séquences de haut qualité, T-Coffee a la particularité de pouvoir également incorporer des alignements issus d'autres algorithmes, permettant ainsi l'utilisation d'informations provenant d'outils d'alignement externes basés sur une comparaison structure-séquence (« *threader* ») et structure-structure (« *structure based alignments* »). L'information structurale étant plus conservée que l'information de séquence, ces alignements enrichis par l'information structurale sont particulièrement intéressants quand on doit aligner des séquences à faible taux d'homologie.

C'est notamment le cas des alignements requis pour la construction de modèles par homologie utilisés par notre méthode de remplacement moléculaire (voire section 3.1). Ceci m'a amené à collaborer avec Cédric Notredame afin de développer des protocoles adaptés à ce genre de problématiques (Notredame and Suhre, sous presse). Dans ce contexte, j'ai implémenté une interface T-Coffee pour l'alignement structure-structure en utilisant le logiciel *Lsqman* (Kleywegt, 1996), dont l'addition aux méthodes déjà présentes dans T-Coffee permet d'augmenter ses performances sur des cas de « benchmark » (O'Sullivan et al., 2004). Afin de rendre ce développement accessible à la communauté des biologistes, nous avons mis au point une interface web pour la version 3D de T-Coffee (Poirot et al., 2004), permettant aux utilisateurs de générer facilement en ligne des alignements multiples mélangeant structures et séquences et de visualiser les résultats avec une présentation incluant l'information structurale (Figure 14) : http://igs-server.cnrs-mrs.fr/Tcoffee/

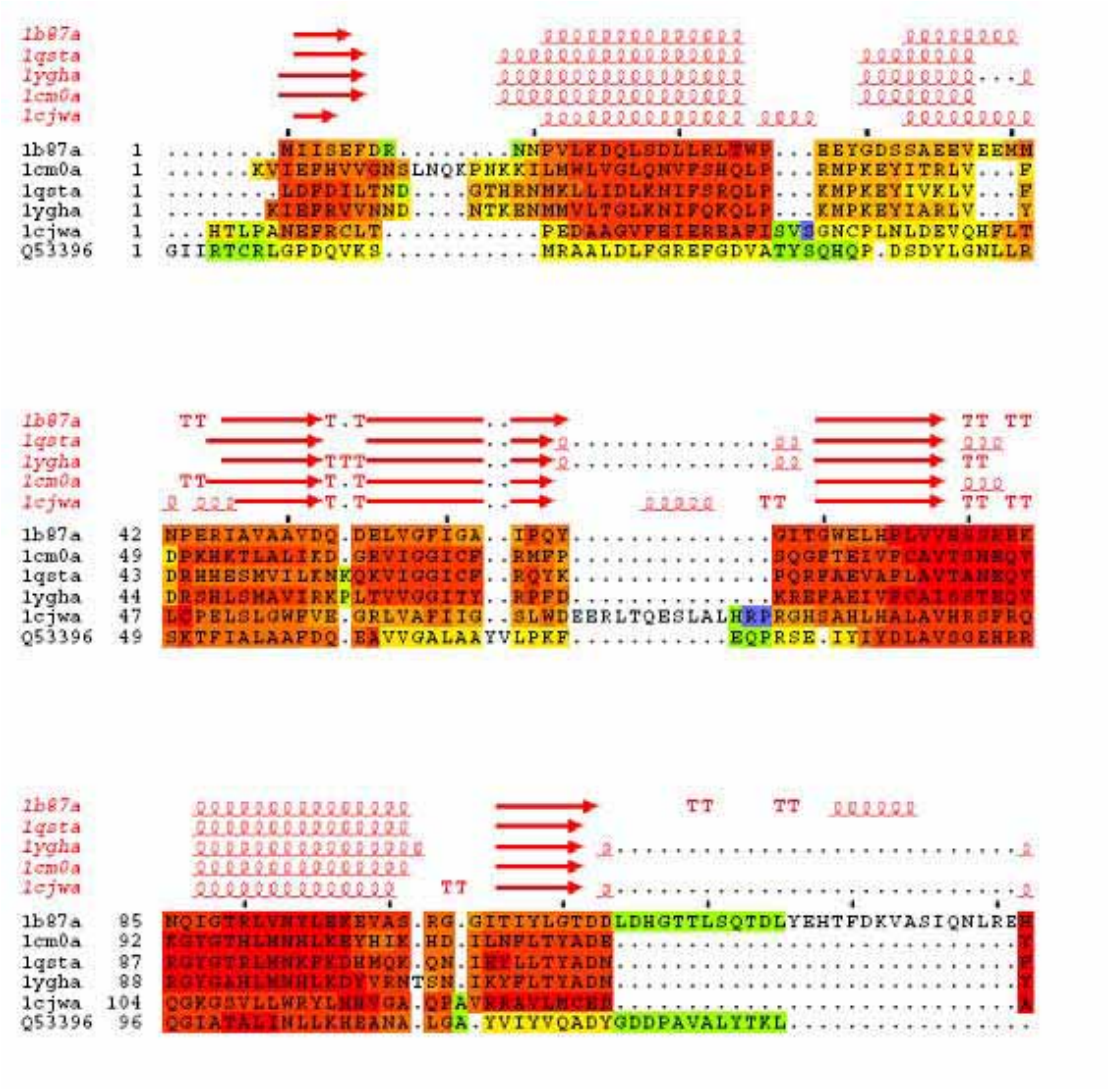

Figure 14 Exemple d'un alignement généré par le serveur [3DCoffee@igs]

3DCoffee utilise l'information structurale dans la génération de ses alignements multiples. Afin de visualiser un maximum d'informations fournies par le serveur dans une présentation synthétique, nous avons combiné la sortie classique du CORE index de T-Coffee avec une sortie générée par ESPript (Gouet et al., 1999), permettant ainsi de présenter en même temps la structure secondaire des protéines au dessus de l'alignement coloré selon le CORE index.

4.3 Analyse des données de puces à ADN

En collaboration avec l'équipe de Christophe de La Roche St.-André et de Vincent Géli du LISM (IBSM, Marseille), j'ai participé à l'étude du transcriptome d'un mutant (*set1Δ*) de délétion du gène *SET1* de la levure *Saccharomyces cerevisieae* en conditions méiotiques. La protéine Set1 est connue pour sa fonction de méthylase de la lysine 4 de l'histone H3, mais on la soupçonne également de jouer un rôle important dans le cycle cellulaire. Des anomalies concernant des homologues de *SET1* dans des eucaryotes supérieurs sont impliqués dans la génération de certains cancers. En comparant les données de puces à ADN issues de la souche sauvage en conditions méiotiques avec celles du mutant *set1Δ*, et en nous focalisant plus particulièrement sur les gènes liés à la méiose, nous avons pu mettre en évidence un effet différentiel de la perte de Set1 sur les gènes induits au cours de la méiose (« *middle genes* »), mais pas sur ceux qui sont induits au début (« *early genes* ») (Figure 15). Combinés avec des résultats expérimentaux supplémentaires, nos résultats montrent que Set1 joue donc un rôle dans le cycle cellulaire méiotique, et a également un rôle indépendant de son activité d'histone-méthylase (Sollier et al., 2004).

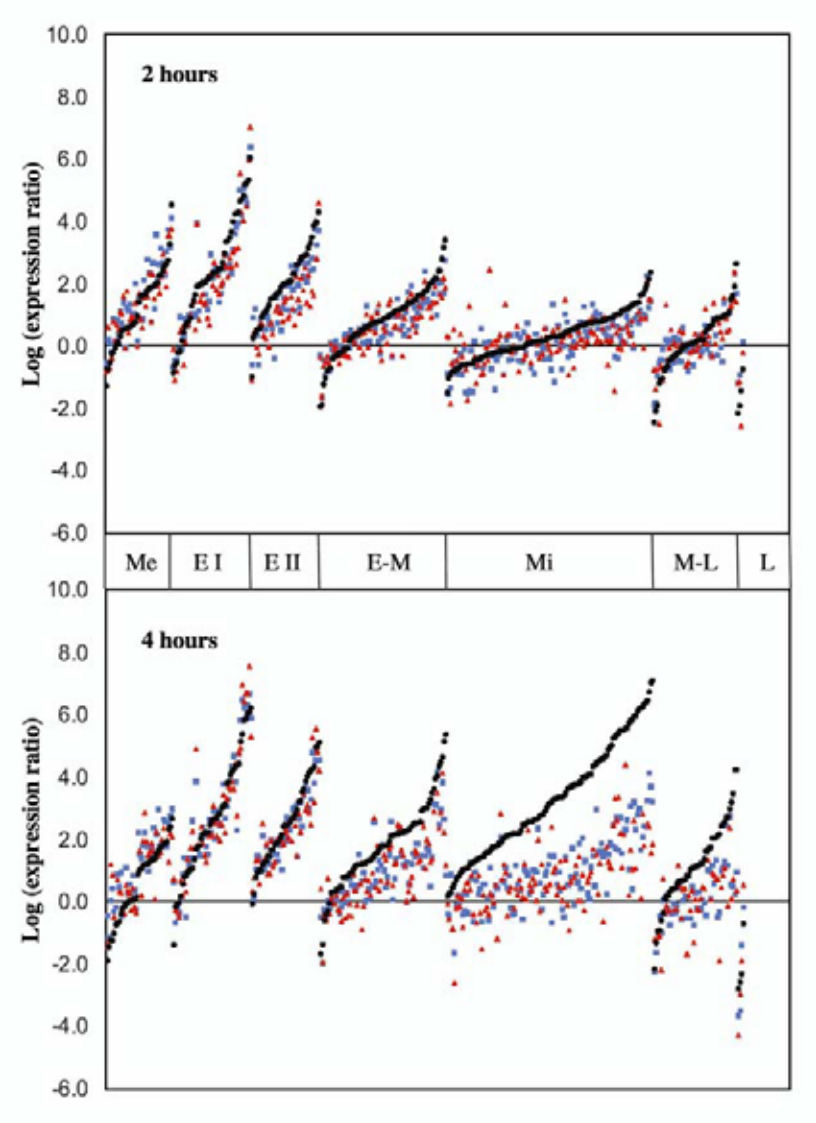

Figure 15 Expression des gènes spécifiques à la méiose

L'absence de méthylation de la lysine 4 de l'histone 3 (H3-K4) est corrélée à un défaut d'induction des gènes « middle » de la méiose. Cette figure présente le niveau d'expression de tous les gènes de la méiose (selon la classification de Chu et al., (1998): Me: metabolic, E I: early I, E II: early II, E-M: early-mid, Mi: middle, M-L: mid-late, L: late), 2 et 4 heures après l'induction de la sporulation de la souche sauvage (SK1, en noir) et des souches mutantes *swd3Δ* (bleu) et *set1Δ* (rouge). Dans chaque classe, et pour chaque point temporel, les gènes ont été ordonnés de gauche à droite selon le niveau d'expression de la souche sauvage.

Chapitre 5 Conclusions

Si l'on me demandait de chiffrer en une seule phrase (comme on a souvent l'habitude de le faire dans l'industrie privé) mes résultats scientifiques de ces deux et demi années passées à l'IGS, je citerais mes contributions à 14 publications en biologie, ma participation à la conception de 5 serveurs web (Phydbac, FusionDB, ElNémo, CaspR, 3D-Coffee@IGS) et à la résolution de 2 structures cristallographiques (*YecD*, *MV2*) ainsi que le développement de nouvelles méthodes de phasage des données de cristallographie, soit en passant par les modèles par homologie, soit en utilisant l'analyse en modes normaux.

Pour ne pas oublier complètement mon passé de physicien et de chimiste de l'atmosphère, je tiens à souligner qu'il existe, en dépit de ma re-conversion à la biologie, une forte continuité dans mes travaux d'un point de vue conceptuel. Cette continuité se matérialise aussi bien quand on compare les deux grands domaines de la science dans lesquels j'étais tour à tour actif lors de ma carrière au CNRS (c.-à-d. les Sciences de la Vie et les Sciences de l'Univers), que quand on regarde les interconnexions entre mes différents projets de recherche en biologie.

Quant aux points communs de mes travaux dans les Sciences de la Vie et les Sciences de l'Univers, je tiens à évoquer de nombreuses similitudes que je résume dans le tableau ci-dessous.

Sciences de l'Univers	**Sciences de la Vie**
Simulation météorologiques • paramétrisation de processus non-résolus par le modèle (turbulence, convection, flux en surface, …) • intégration numérique des équations de Navier-Stokes (algorithmes, stabilité numérique, conditions aux limites)	Dynamique moléculaire • paramétrisation des effets de la mécanique quantique par des champs de force empiriques • intégration numérique des équations de mouvement (algorithmes, stabilité numérique, conditions aux limites)

• assimilation de données pour définir les conditions initiales et afin d'analyser un état instantané Analyse de données issues de sources hétérogènes (*data mining*) • radiosondes (ballons) • radar, lidar, radiomètres • mesures aéroportées • observations par satellites, … Chimie atmosphérique • systèmes réactionnels avec des centaines de réactants • analyse et réduction du schéma réactionnel	• minimisation d'énergie pour éviter l'explosion du modèle au spin-up ou pour identifier un état stable Analyse de données issues de sources hétérogènes (*data mining*) • génomique et protéomique • transcriptome • structures cristallographiques • expériences en biochimie, ... Métabolisme • réseaux métaboliques avec des centaines de métabolites • analyse en modes élémentaires ou extrêmes

Tableau 4 Exemples de points communs de mes travaux en SDU et SDV

Le deuxième type de continuité conceptuelle que je tiens à évoquer se trouve du côté de mes recherches en SDV : afin de réussir pleinement ma reconversion à la bio-informatique, et d'éviter de n'être qu'un physicien « parachuté » dans un monde de biologistes dont il ne comprend pas les mystères, j'ai décidé de travailler, dans un premier temps, sur un grand nombre de sujets différents dans le but d'apprendre un maximum de ce domaine qui était nouveau pour moi. Pour le théoricien que je suis, deux moyens principaux se présentaient alors : procéder par la séquence ou par la structure, et ceci, si possible, soit en s'appuyant sur des méthodes déjà existantes, soit en inventant des méthodologies nouvelles. Dans mon cas, grâce à la présence à la fois d'une équipe de structuralistes (cristallographie) et de bio-informatique sur le site de l'IGS, j'ai eu l'occasion de « goûter » à tout un nombre d'approches différentes, toutefois réunies autour d'un thème central qui est l'identification de la fonction d'un gène, généralement issu du monde des organismes uni-cellulaire.

Pour finir, et pour parler un peu perspectives, j'aimerais souligner le fait que la présentation de cette HDR représente pour moi surtout l'achèvement de ma

reconversion dans le domaine de la bio-informatique. A court terme, il me reste un certain nombre de projets à mener à bien, projets que j'ai déjà partiellement démarrés et qui se basent sur des travaux que j'ai présentés dans ce mémoire.

En génomique, il s'agit de continuer ma participation aux nouveaux projets de séquençage, actuellement en cours en collaboration avec la Timone. Des premiers résultats ont déjà été obtenus dans le cadre du projet « *Rickettsia felis* », où l'objectif est de tirer un maximum d'information de ce génome concernant les différents aspects de sa pathogénie en le comparant au nombre déjà élevé d'autres génomes de rickettsies. Il y a donc des méthodologies d'inter-comparaison ciblées à développer, par exemple pour l'identification des séquences régulatrices et fonctionnelles. Ensuite, nous attend le génome de *Minibacterium timonensis*, une bactérie si petite qu'elle est capable de s'introduire profondément dans les chaînes de purification des eaux utilisées pour la dialyse hospitalière. Puis, il y aura le génome d'*Acinetobacter baumannii*, une souche des Acinetobacteries qui est résistante à de multiples antibiotiques et qui évolue dans le milieu hospitalier. Dans tout ces projets, le développement de nouveaux protocoles de génomique comparative sera au centre de mes recherches, avec l'objectif de mettre en relief les particularités dans le complément génétique de chaque organisme étudié, afin d'identifier éventuellement les gènes qui confèrent les caractéristiques de virulence ou de résistance particulier à ces micro-organismes.

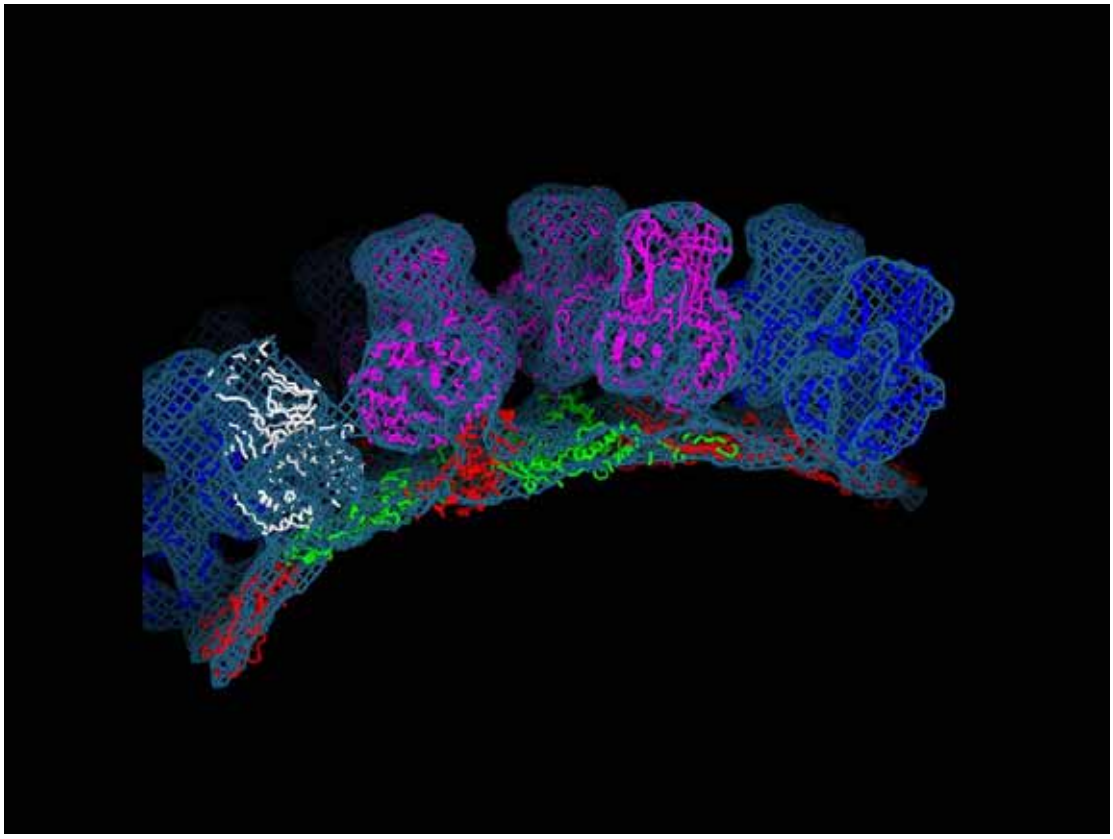

Figure 16 Capsid du rotavirus bovin

Carte de densité électronique obtenu par microscopie électronique de la capside du rotavirus bovin (« double layer protein », données du Laboratoire de Virologie Moléculaire Structurale, Gif s/Yvette). Les modèles des protéines VP2 (en vert et rouge) et VP6 (en bleu, blanc et magenta) ont été placés dans la carte à l'aide du logiciel URO (Navaza et al., 2002). La forme des molécules VP6 a été optimisée en appliquant une perturbation en modes normaux (3 modes de plus basse fréquence), permettant ainsi d'améliorer le coefficient de corrélation entre les modèles placés et les données de manière significative.

En ce qui concerne la biologie structurale, je continue l'exploration du potentiel des analyses en modes normaux. Deux applications me paraissent particulièrement intéressantes :

1. Le « docking » de protéines est un domaine en plein développement avec de nombreuses applications aussi bien fondamentales que médicales. Or, un grand nombre des cas difficiles de « docking », qui s'échappent encore à toute résolution numérique, concerne les protéines exhibant un changement conformationnel important lors de la formation du complexe (concept du « induced fit »). La recherche de solutions de « docking » en utilisant des modèles perturbés en modes normaux, calqué sur l'exemple de notre approche

en remplacement moléculaire, se propose alors comme une voie intéressante à explorer.

2. L'interprétation des données de microscopie électronique (ME) de complexes multi-protéiques, en utilisant des modèles atomiques de protéines issus de la cristallographie, est un domaine qui se développe de plus en plus, permettant ainsi l'étude d'objets biologiques de plus amples dimensions, l'exemple type étant des capsides virales et des virus tout entier. En se basant sur des programmes comme par exemple URO (Navaza et al., 2002), on cherche à placer des modèles atomiques de protéines dans des cartes de densité électronique issues de la ME ayant des résolutions de l'ordre de 15-25Å. L'approche est en fait très similaire aux méthodes utilisées en remplacement moléculaire. Or, en formant des complexes multiprotéiques, les protéines impliquées présentent souvent des changements de conformation très important comparé à leur structure non-complexée. L'AMN sera alors l'outil idéal pour modéliser ce changement sous la contrainte des données issues de la ME. Récemment, Tama et al. (2004) ont déjà montrer des premiers exemples théoriques d'une telle application en plaçant des protéines dans des cartes de densité électronique synthétiques. Brink et al. (2004) ont également utilisé les modes normaux, mais dans la reconstruction de particules isolées, également à partir de données de microscopie électronique. Dans la Figure 16 je présente des résultats (encore très préliminaires) d'une collaboration sur ce sujet que je viens d'entamer avec J. Navaza du Laboratoire de Virologie Moléculaire Structurale de Gif s/Yvette. Dans ce cas nous avons placé deux protéines de la capside du rotavirus bovin (VP2 et VP6) dans une carte de densité électronique de la ME, tout en améliorant le « fit » en utilisant les modes normaux. La filière URO – AMN que j'ai mis en place pour cet exemple pourra donc servir comme point de départ pour une exploration de plus grande amplitude du potentiel des modes normaux en microscopie électronique.

Finalement, dans une perspective visant le moyen à long terme, il me reste à définir un sujet qui me soit propre, tâche qui est surtout une question de choix et de priorités, étant donné le grand nombre de possibilités qu'offre l'actuel flux d'information génomique et structurale dans les banques de données à un bio-informaticien. Une possibilité intéressante serait une orientation vers la « *systems biology* », un sujet qui

est actuellement en pleine expansion et qui représente pour moi une forte ressemblance à la chimie atmosphérique par ces approches conceptuelles à l'analyse des flux de métabolites, la définition des schémas réactionnels, l'implémentations des algorithmes numériques, les techniques de simulations, etc. Alternativement, le « *data mining* » dans les banques de données structuraux me semble également un sujet très prometteur. Si les grands projets de génomique structurale continuent à produire des résultats à un rythme soutenu, comme c'est actuellement le cas, des méthodes de la génomique comparative pourraient bientôt être transformées en ce qu'on devrait appeler la « structuromique comparative ». Participer à cet exploit me paraît également un choix extrêmement intéressant.

Marseille, le 06 août 2004

Karsten Suhre.

Liste des références

Abendroth, J., Niefind, K., May, O., Siemann, M., Syldatk, C. and Schomburg, D. (2002) The structure of L-hydantoinase from Arthobacter aurescens leads to an understanding of dihydropyrimidinase substrate and enantio specificity. *Biochemistry*, **41**, 8589-8597.

Abergel, C., Coutard, B., Byrne, D., Chenivesse, S., Claude, J.B., Deregnaucourt, C., Fricaux, T., Gianesini-Boutreux, C., Jeudy, S., Lebrun, R., Maza, C., Notredame, C., Poirot, O., Suhre, K., Varagnol, M. and Claverie, J.M. (2003) Structural genomics of highly conserved microbial genes of unknown function in search of new antibacterial targets. *J Struct Funct Genomics*, **4**, 141-157.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-3402.

Backbro, K., Lowgren, S., Osterlund, K., Atepo, J., Unge, T., Hulten, J., Bonham, N.M., Schaal, W., Karlen, A. and Hallberg, A. (1997) Unexpected binding mode of a cyclic sulfamide HIV-1 protease inhibitor. *J Med Chem*, **40**, 898-902.

Bahar, I., Atilgan, A.R. and Erman, B. (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des*, **2**, 173-181.

Baray, J.L., S., B., D., D.R. and P., C.J. (2003) Dynamical study of a tropical cut-off low over South Africa, and its impact on tropospheric ozone. *Atmospheric Environment*, **37**, 1475-1488.

Bates, T.S., Huebert, B.J., Gras, J.L., Griffiths, F.B. and Durkee, P.A. (1998) International Global Atmospheric Chemistry (IGAC) project's first aerosol characterization experiment (ACE 1): Overview. *Journal of Geophysical Research-Atmospheres*, **103**, 16297-16318.

Bedos, C., Suhre, K. and Rosset, R. (1996) Adaptation of a cloud activation scheme to a spectral-chemical aerosol model. *Atmospheric Research*, **41**, 267-279.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235-242.

Bilu, Y. and Linial, M. (2002) The advantage of functional prediction based on clustering of yeast genes and its correlation with non-sequence based classifications. *J Comput Biol*, **9**, 193-210.

Brink, J., Ludtke, S.J., Kong, Y., Wakil, S.J., Ma, J. and Chiu, W. (2004) Experimental verification of conformational variation of human fatty acid synthase as predicted by normal mode analysis. *Structure (Camb)*, **12**, 185-191.

Brooks, B. and Karplus, M. (1983) Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc Natl Acad Sci U S A*, **80**, 6571-6575.

Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T. and Warren, G.L. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr*, **54 (Pt 5)**, 905-921.

Bussiere, D.E., Kong, X., Egan, D.A., Walter, K., Holzman, T.F., Lindh, F., Robins, T. and Giranda, V.L. (1998) Structure of the E2 DNA-binding domain from human papillomavirus serotype 31 at 2.4 A. *Acta Crystallogr D Biol Crystallogr*, **54**, 1367-1376.

Cambillau, C. and Claverie, J.M. (2000) Structural and genomic correlates of hyperthermostability. *J Biol Chem*, **275**, 32383-32386.

Cammas, J.P., Jacoby-Koaly, S., Suhre, K., Rosset, R. and Marenco, A. (1998) Atlantic subtropical potential vorticity barrier as seen by Measurements of Ozone by Airbus In-Service Aircraft (MOZAIC) flights. *Journal of Geophysical Research-Atmospheres*, **103**, 25681-25693.

Charlson, R.J., Lovelock, J.E., Andreae, M.O. and Warren, S.G. (1987) Oceanic Phytoplankton, Atmospheric Sulfur, Cloud Albedo and Climate. *Nature*, **326**, 655-661.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O. and Herskowitz, I. (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699-705.

Claude, J.B., Suhre, K., Notredame, C., Claverie, J.M. and Abergel, C. (2004) CaspR: a web server for automated molecular replacement using homology modelling. *Nucleic Acids Res*, **32**, W606-W609.

Cohard, J.M., Pinty, J.P. and Suhre, K. (2000) On the parameterization of activation spectra from cloud condensation nuclei microphysical properties. *Journal of Geophysical Research-Atmospheres*, **105**, 11753-11766.

Crassier, V., Suhre, K., Tulet, P. and Rosset, R. (2000) Development of a reduced chemical scheme for use in mesoscale meteorological models. *Atmospheric Environment*, **34**, 2633-2644.

Delarue, M. and Sanejouand, Y.H. (2002) Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model. *J Mol Biol*, **320**, 1011-1024.

Diamond, R. (1990) On the use of normal modes in thermal parameter refinement: theory and application to the bovine pancreatic trypsin inhibitor. *Acta Crystallogr A*, **46**, 425-435.

Durand, P., Trinquier, G. and Sanejouand, Y.H. (1994) A new approach for determining low-frequency normal modes in macromolecules. *Biopolymers*, **34**, 759.

Echols, N., Milburn, D. and Gerstein, M. (2003) MolMovDB: analysis and visualization of conformational change and structural flexibility. *Nucleic Acids Res*, **31**, 478-482.

Enault, F., Suhre, K., Abergel, C., Poirot, O. and Claverie, J.M. (2003a) Annotation of bacterial genomes using improved phylogenomic profiles. *Bioinformatics*, **19 Suppl 1**, i105-107.

Enault, F., Suhre, K., Poirot, O., Abergel, C. and Claverie, J.M. (2003b) Phydbac (phylogenomic display of bacterial genes): An interactive resource for the annotation of bacterial genomes. *Nucleic Acids Res*, **31**, 3720-3722.

Enault, F., Suhre, K., Poirot, O., Abergel, C. and Claverie, J.M. (2004) Phydbac2: improved inference of gene function using interactive phylogenomic profiling and chromosomal location analysis. *Nucleic Acids Res*, **32**, W336-W339.

Enright, A.J. and Ouzounis, C.A. (2001) Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol*, **2**, RESEARCH0034.

FassiFihri, A., Suhre, K. and Rosset, R. (1997) Internal and external mixing in atmospheric aerosols by coagulation: Impact on the optical and hygroscopic properties of the sulphate-soot system. *Atmospheric Environment*, **31**, 1393-1402.

Fiser, A. and Sali, A. (2003) Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol*, **374**, 461-491.

Galperin, M.Y. and Koonin, E.V. (2000) Who's your neighbor? New computational approaches for functional genomics. *Nat Biotechnol*, **18**, 609-613.

Ginalski, K., Elofsson, A., Fischer, D. and Rychlewski, L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, **19**, 1015-1018.

Go, N., Noguti, T. and Nishikawa, T. (1983) Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc Natl Acad Sci U S A*, **80**, 3696-3700.

Gouet, P., Courcelle, E., Stuart, D.I. and Metoz, F. (1999) ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics*, **15**, 305-308.

Hinsen, K. (1998) Analysis of domain motions by approximate normal mode calculations. *Proteins*, **33**, 417-429.

Hinsen, K., Thomas, A. and Field, M.J. (1999) Analysis of domain motions in large proteins. *Proteins*, **34**, 369-382.

Hsiao, C.D., Sun, Y.J., Rose, J. and Wang, B.C. (1996) The crystal structure of glutamine-binding protein from Escherichia coli. *J Mol Biol*, **262**, 225-242.

Huebert, B.J., Bates, T., Russell, P.B., Shi, G.Y., Kim, Y.J., Kawamura, K., Carmichael, G. and Nakajima, T. (2003) An overview of ACE-Asia: Strategies for quantifying the relationships between Asian aerosols and their climatic impacts. *Journal of Geophysical Research-Atmospheres*, **108**.

Johnson, D.W., Osborne, S., Wood, R., Suhre, K., Johnson, R., Businger, S., Quinn, P.K., Wiedensohler, A., Durkee, P.A., Russell, L.M., Andreae, M.O., O'Dowd, C., Noone, K.J., Bandy, B., Rudolph, J. and Rapsomanikis, S. (2000a) An overview of the Lagrangian experiments undertaken during the North Atlantic regional Aerosol Characterisation Experiment (ACE-2). *Tellus Series B-Chemical and Physical Meteorology*, **52**, 290-320.

Johnson, D.W., Osborne, S., Wood, R., Suhre, K., Quinn, P.K., Bates, T., Andreae, M.O., Noone, K.J., Glantz, P., Bandy, B., Rudolph, J. and O'Dowd, C. (2000b) Observations of the evolution of the aerosol, cloud and boundary-layer characteristics during the 1st ACE-2 Lagrangian experiment. *Tellus Series B-Chemical and Physical Meteorology*, **52**, 348-374.

Jones, D.T. (2001) Evaluating the potential of using fold-recognition models for molecular replacement. *Acta Crystallogr D Biol Crystallogr*, **57**, 1428-1434.

Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Collado-Vides, J., Paley, S.M., Pellegrini-Toole, A., Bonavides, C. and Gama-Castro, S. (2002) The EcoCyc Database. *Nucleic Acids Res*, **30**, 56-58.

Kawashima, T., Amano, N., Koike, H., Makino, S., Higuchi, S., Kawashima-Ohya, Y., Watanabe, K., Yamazaki, M., Kanehori, K., Kawamoto, T., Nunoshiba, T., Yamamoto, Y., Aramaki, H., Makino, K. and Suzuki, M. (2000) Archaeal adaptation to higher temperatures revealed by genomic sequence of Thermoplasma volcanium. *Proc Natl Acad Sci U S A*, **97**, 14257-14262.

Kidera, A. and Go, N. (1992) Normal mode refinement: crystallographic refinement of protein dynamic structure. I. Theory and test by simulated diffraction data. *J Mol Biol*, **225**, 457-475.

Kim, M.K., Jernigan, R.L. and Chirikjian, G.S. (2003) An elastic network model of HK97 capsid maturation. *J Struct Biol*, **143**, 107-117.
Kleywegt, G.J. (1996) Use of non-crystallographic symmetry in protein structure refinement. *Acta Crystallogr D Biol Crystallogr*, **52**, 842-857.
Kraulis, P.J. (1991) MOLSCRIPT : a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallography*, **24**, 946-950.
Krebs, W.G., Alexandrov, V., Wilson, C.A., Echols, N., Yu, H. and Gerstein, M. (2002) Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. *Proteins*, **48**, 682-695.
Lafore, J.P., Stein, J., Asencio, N., Bougeault, P., Ducrocq, V., Duron, J., Fischer, C., Hereil, P., Mascart, P., Masson, V., Pinty, J.P., Redelsperger, J.L., Richard, E. and de Arellano, J.V.G. (1998) The Meso-NH atmospheric simulation system. Part I: adiabatic formulation and control simulations. *Annales Geophysicae-Atmospheres Hydrospheres and Space Sciences*, **16**, 90-109.
Laskowski, R.A., Rullmannn, J.A., MacArthur, M.W., Kaptein, R. and Thornton, J.M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR*, **8**, 477-486.
Li, G. and Cui, Q. (2002) A coarse-grained normal mode approach for macromolecules: an efficient implementation and application to Ca(2+)-ATPase. *Biophys J*, **83**, 2457-2474.
Lovelock, J. (1997) A geophysiologist's thoughts on the natural sulphur cycle. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, **352**, 143-147.
Marcotte, E.M. (2000) Computational genetics: finding protein function by nonhomology methods. *Curr Opin Struct Biol*, **10**, 359-365.
Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. and Eisenberg, D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83-86.
Marenco, A., Thouret, V., Nedelec, P., Smit, H., Helten, M., Kley, D., Karcher, F., Simon, P., Law, K., Pyle, J., Poschmann, G., Von Wrede, R., Hume, C. and Cook, T. (1998) Measurement of ozone and water vapor by Airbus in-service aircraft: The MOZAIC airborne program, An overview. *Journal of Geophysical Research-Atmospheres*, **103**, 25631-25642.
Mari, C., Evans, M.J., Palmer, P.I., Jacob, D.J. and Sachse, G.W. (sous presse) Export of Asian pollution during two cold front episodes of the TRACE-P experiment. *Journal of Geophysical Research-Atmospheres*.
Mari, C., Suhre, K., Bates, T.S., Johnson, J.E., Rosset, R., Bandy, A.R., Eisele, F.L., Mauldin, R.L. and Thornton, D.C. (1998) Physico-chemical modeling of the First Aerosol Characterization Experiment (ACE 1) Lagrangian B - 2. DMS emission, transport and oxidation at the mesoscale. *Journal of Geophysical Research-Atmospheres*, **103**, 16457-16473.
Mari, C., Suhre, K., Rosset, R., Bates, T.S., Huebert, B.J., Bandy, A.R., Thornton, D.C. and Businger, S. (1999) One-dimensional modeling of sulfur species during the First Aerosol Characterization Experiment (ACE 1) Lagrangian B. *Journal of Geophysical Research-Atmospheres*, **104**, 21733-21749.
Marques, O. and Sanejouand, Y.H. (1995) Hinge-bending motion in citrate synthase arising from normal mode calculations. *Proteins*, **23**, 557-560.

Matthijsen, J., Suhre, K., Bechtold, P. and Rosset, R. (1997) The effect of fractional cloudiness on the oxidation of SO2. *Tellus Series B-Chemical and Physical Meteorology*, **49**, 343-356.
Matthijsen, J., Suhre, K., Rosset, R., Eisele, F.L., Mauldin, R.L. and Tanner, D.J. (1998) Photodissociation and UV radiative transfer in a cloudy atmosphere: Modeling and measurements. *Journal of Geophysical Research-Atmospheres*, **103**, 16665-16676.
Mellor, J.C., Yanai, I., Clodfelter, K.H., Mintseris, J. and DeLisi, C. (2002) Predictome: a database of putative functional links between proteins. *Nucleic Acids Res*, **30**, 306-309.
Mouawad, L. and Perahia, D. (1993) DIMB: diagonalization in a mixed basis. a method to compute low-frequency normal modes for large macromolecules. *Biopolymers*, **33**, 569-611.
Navaza, J. (2001) Implementation of molecular replacement in AMoRe. *Acta Crystallogr D Biol Crystallogr*, **57**, 1367-1372.
Navaza, J., Lepault, J., Rey, F.A., Alvarez-Rua, C. and Borge, J. (2002) On the fitting of model electron densities into EM reconstructions: a reciprocal-space formulation. *Acta Crystallogr D Biol Crystallogr*, **58**, 1820-1825.
Notredame, C. and Suhre, K. (sous presse) Computing multiple sequence/structure alignments with the T-Coffee package. *Current Protocols in Bioinformatics*.
O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D.G. and Notredame, C. (2004) 3DCoffee: Combining Protein Sequences and Structures within Multiple Sequence Alignments. *J Mol Biol*, **340**, 385-395.
Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, **96**, 4285-4288.
Poirot, O., Suhre, K., Abergel, C., O'Toole, E. and Notredame, C. (2004) 3DCoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment. *Nucleic Acids Res*, **32**, W37-W40.
Raes, F., Bates, T., McGovern, F. and Van Liedekerke, M. (2000) The 2nd Aerosol Characterization Experiment (ACE-2): general overview and main results. *Tellus Series B-Chemical and Physical Meteorology*, **52**, 111-125.
Raoult, D., Ogata, H., Audic, S., Robert, C., Suhre, K., Drancourt, M. and Claverie, J.M. (2003) Tropheryma whipplei Twist: a human pathogenic Actinobacteria with a reduced genome. *Genome Res*, **13**, 1800-1809.
Reuter, N., Hinsen, K. and Lacapere, J.J. (2003) Transconformations of the SERCA1 Ca-ATPase: a normal mode study. *Biophys J*, **85**, 2186-2197.
Roussel, A., Fontecilla-Camps, J.C. and Cambillau, C. (1990) CRYStallize: a crystallographic symmetry display and handling subpackage in TOM/FRODO. *J Mol Graph*, **8**, 86-88, 91.
Rupp, B., Segelke, B.W., Krupka, H.I., Lekin, T., Schafer, J., Zemla, A., Toppani, D., Snell, G. and Earnest, T. (2002) The TB structural genomics consortium crystallization facility: towards automation from protein to electron density. *Acta Crystallogr D Biol Crystallogr*, **58**, 1514-1518.
Sali, A. (1999) Functional links between proteins. *Nature*, **402**, 23, 25-26.
Sali, A., Potterton, L., Yuan, F., van Vlijmen, H. and Karplus, M. (1995) Evaluation of comparative protein modeling by MODELLER. *Proteins*, **23**, 318-326.
Sanchez, R. and Sali, A. (2000) Comparative protein structure modeling. Introduction and practical examples with modeller. *Methods Mol Biol*, **143**, 97-129.

Serres, M.H. and Riley, M. (2000) MultiFun, a multifunctional classification scheme for Escherichia coli K-12 gene products. *Microb Comp Genomics*, **5**, 205-222.
Shi, J., Blundell, T.L. and Mizuguchi, K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J Mol Biol*, **310**, 243-257.
Siems, S.T., Hess, G.D., Suhre, K., Businger, S. and Draxler, R.R. (2000) The impact of wind shear on observed and simulated trajectories during the ACE-1 Lagrangian experiments. *Australian Meteorological Magazine*, **49**, 109-120.
Sollier, J., Lin, W.K., Soustelle, C., Suhre, K., Nicolas, A., Geli, V. and Saint-Andre, C.D. (2004) Set1 is required for meiotic S-phase onset, double-strand break formation and middle gene expression. *Embo Journal*, **23**, 1957-1967.
Suhre, K. (1991) Das Feld am Punkt im äußeren elektromagnetischen Feld. *Diplomarbeit im Fachbereich Physik*. Universität Osnabrück, Allemagne.
Suhre, K. (1994) Modélisation couplée du transport et de la chimie du diméthyl de soufre dans la couche limite marine nuageuse. Impact climatique et étude de processus. *Thèse de doctorat*. Université Paul Sabatier, Toulouse, France.
Suhre, K., Andreae, M.O. and Rosset, R. (1995) Biogenic Sulfur Emissions and Aerosols over the Tropical South-Atlantic .2. One-Dimensional Simulation of Sulfur Chemistry in the Marine Boundary-Layer. *Journal of Geophysical Research-Atmospheres*, **100**, 11323-11334.
Suhre, K., Cammas, J.P., Nedelec, P., Rosset, R., Marenco, A. and Smit, H.G.J. (1997) Ozone-rich transients in the upper equatorial Atlantic troposphere. *Nature*, **388**, 661-663.
Suhre, K. and Claverie, J.M. (2003) Genomic correlates of hyperthermostability, an update. *J Biol Chem*, **278**, 17198-17202.
Suhre, K. and Claverie, J.M. (2004) FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. *Nucleic Acids Res*, **32 Database issue**, D273-276.
Suhre, K., Crassier, V., Mari, C., Rosset, R., Johnson, D.W., Osborne, S., Wood, R., Andreae, M.O., Bandy, B., Bates, T.S., Businger, S., Gerbig, C., Raes, F. and Rudolph, J. (2000a) Chemistry and aerosols in the marine boundary layer: 1-D modelling of the three ACE-2 Lagrangian experiments. *Atmospheric Environment*, **34**, 5079-5094.
Suhre, K., Johnson, D.W., Mari, C., Rosset, R., Osborne, S., Wood, R., Bates, T.S. and Raes, F. (2000b) A continental outbreak of air during the Second Aerosol Characterization Experiment (ACE 2): A Lagrangian experiment. *Journal of Geophysical Research-Atmospheres*, **105**, 17911-17924.
Suhre, K., Mari, C., Bates, T.S., Johnson, J.E., Rosset, R., Wang, Q., Bandy, A.R., Blake, D.R., Businger, S., Eisele, F.L., Huebert, B.J., Kok, G.L., Mauldin, R.L., Prevot, A.S.H., Schillawski, R.D., Tanner, D.J. and Thornton, D.C. (1998) Physico-chemical modeling of the First Aerosol Characterization Experiment (ACE 1) Lagrangian B - 1. A moving column approach. *Journal of Geophysical Research-Atmospheres*, **103**, 16433-16455.
Suhre, K. and Rosset, R. (1994a) Dms Oxidation and Turbulent Transport in the Marine Boundary-Layer - a Numerical Study. *Journal of Atmospheric Chemistry*, **18**, 379-395.
Suhre, K. and Rosset, R. (1994b) Modification of a Linearized Semiimplicit Scheme for Chemical-Reactions Using a Steady-State Approximation. *Annales Geophysicae-Atmospheres Hydrospheres and Space Sciences*, **12**, 359-361.

Suhre, K. and Sanejouand, Y.H. (2004a) ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res*, **32**, W610-W614.
Suhre, K. and Sanejouand, Y.H. (2004b) On the potential of normal-mode analysis for solving difficult molecular-replacement problems. *Acta Crystallogr D Biol Crystallogr*, **60**, 796-799.
Tama, F. (2003) Normal mode analysis with simplified models to investigate the global dynamics of biological systems. *Protein Pept Lett*, **10**, 119-132.
Tama, F., Gadea, F.X., Marques, O. and Sanejouand, Y.H. (2000) Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins*, **41**, 1-7.
Tama, F., Miyashita, O. and Brooks, C.L., 3rd. (2004) Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. *J Mol Biol*, **337**, 985-999.
Tama, F. and Sanejouand, Y.H. (2001) Conformational change of proteins arising from normal mode calculations. *Protein Eng*, **14**, 1-6.
Tama, F., Valle, M., Frank, J. and Brooks, C.L., 3rd. (2003) Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy. *Proc Natl Acad Sci U S A*, **100**, 9319-9323.
Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631-637.
Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D. and Koonin, E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res*, **29**, 22-28.
Taylor, W.R. and Orengo, C.A. (1989) Protein structure alignment. *J Mol Biol*, **208**, 1-22.
Thomas, A., Hinsen, K., Field, M.J. and Perahia, D. (1999) Tertiary and quaternary conformational changes in aspartate transcarbamylase: a normal mode study. *Proteins*, **34**, 96-112.
Tirion, M.M. (1996) Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Physical Review Letters*, **77**, 1905-1908.
Tirion, M.M., ben-Avraham, D., Lorenz, M. and Holmes, K.C. (1995) Normal modes as refinement parameters for the F-actin model. *Biophys J*, **68**, 5-12.
Tulet, P., Crassier, V., Solmon, F., Guedalia, D. and Rosset, R. (2003) Description of the Mesoscale Nonhydrostatic Chemistry model and application to a transboundary pollution episode between northern France and southern England. *Journal of Geophysical Research-Atmospheres*, **108**.
Tulet, P., Suhre, K., Mari, C., Solmon, F. and Rosset, R. (2002) Mixing of boundary layer and upper tropospheric ozone during a deep convective event over Western Europe. *Atmospheric Environment*, **36**, 4491-4501.
Valadie, H., Lacapcre, J.J., Sanejouand, Y.H. and Etchebest, C. (2003) Dynamical properties of the MscL of Escherichia coli: a normal mode analysis. *J Mol Biol*, **332**, 657-674.
von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P. and Snel, B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res*, **31**, 258-261.
Wang, Q., Lenschow, D.H., Pan, L.L., Schillawski, R.D., Kok, G.L., Prevot, A.S.H., Laursen, K., Russell, L.M., Bandy, A.R., Thornton, D.C. and Suhre, K. (1999a) Characteristics of the marine boundary layers during two Lagrangian

measurement periods 2. Turbulence structure. *Journal of Geophysical Research-Atmospheres*, **104**, 21767-21784.

Wang, Q., Suhre, K., Krummel, P., Siems, S., Pan, L.L., Bates, T.S., Johnson, J.E., Lenschow, D.H., Heubert, B.J., Kok, G.L., Schillawski, R.D., Prevot, A.S.H. and Businger, S. (1999b) Characteristics of marine boundary layers during two Lagrangian measurement periods 1. General conditions and mean characteristics. *Journal of Geophysical Research-Atmospheres*, **104**, 21751-21765.

Wood, R., Johnson, D., Osborne, S., Andreae, M.O., Bandy, B., Bates, T.S., O'Dowd, C., Glantz, P., Noone, K., Quinn, P.K., Rudolph, J. and Suhre, K. (2000) Boundary layer and aerosol evolution during the 3rd Lagrangian experiment of ACE-2. *Tellus Series B-Chemical and Physical Meteorology*, **52**, 401-422.

Zahn, A., Brenninkmeijer, C.A.M., Crutzen, P.J., Parrish, D.D., Sueper, D., Heinrich, G., Gusten, H., Fischer, H., Hermann, M. and Heintzenberg, J. (2002) Electrical discharge source for tropospheric "ozone-rich transients". *Journal of Geophysical Research-Atmospheres*, **107**.

Zanotti, G., Scapin, G., Spadon, P., Veerkamp, J.H. and Sacchettini, J.C. (1992) Three-dimensional structure of recombinant human muscle fatty acid-binding protein. *J Biol Chem*, **267**, 18541-18550.

Zheng, Y., Roberts, R.J. and Kasif, S. (2002) Genomic functional annotation using co-evolution profiles of gene clusters. *Genome Biol*, **3**, 121-129.

Chapitre 6 ANNEXES

6.1 Publications de l'auteur en Sciences de la Vie

Notredame, C. and Suhre, K. (sous presse) Computing multiple sequence/structure alignments with the T-Coffee package. *Current Protocols in Bioinformatics*.

Suhre, K. and Sanejouand, Y.H. (2004) ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res*, **32**, W610-W614.

Suhre, K. and Sanejouand, Y.H. (2004) On the potential of normal-mode analysis for solving difficult molecular-replacement problems. *Acta Crystallogr D Biol Crystallogr*, **60**, 796-799.

Suhre, K. and Claverie, J.M. (2004) FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. *Nucleic Acids Res*, **32 Database issue**, D273-276.

Sollier, J., Lin, W.K., Soustelle, C., Suhre, K., Nicolas, A., Geli, V. and Saint-Andre, C.D. (2004) Set1 is required for meiotic S-phase onset, double-strand break formation and middle gene expression. *Embo Journal*, **23**, 1957-1967.

Poirot, O., Suhre, K., Abergel, C., O'Toole, E. and Notredame, C. (2004) 3DCoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment. *Nucleic Acids Res*, **32**, W37-W40.

O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D.G. and Notredame, C. (2004) 3DCoffee: Combining Protein Sequences and Structures within Multiple Sequence Alignments. *J Mol Biol*, **340**, 385-395.

Enault, F., Suhre, K., Poirot, O., Abergel, C. and Claverie, J.M. (2004) Phydbac2: improved inference of gene function using interactive phylogenomic profiling and chromosomal location analysis. *Nucleic Acids Res*, **32**, W336-W339.

Claude, J.B., Suhre, K., Notredame, C., Claverie, J.M. and Abergel, C. (2004) CaspR: a web server for automated molecular replacement using homology modelling. *Nucleic Acids Res*, **32**, W606-W609.

Suhre, K. and Claverie, J.M. (2003) Genomic correlates of hyperthermostability, an update. *J Biol Chem*, **278**, 17198-17202.

Raoult, D., Ogata, H., Audic, S., Robert, C., Suhre, K., Drancourt, M. and Claverie, J.M. (2003) Tropheryma whipplei Twist: a human pathogenic Actinobacteria

with a reduced genome. *Genome Res*, **13**, 1800-1809.

Enault, F., Suhre, K., Poirot, O., Abergel, C. and Claverie, J.M. (2003) Phydbac (phylogenomic display of bacterial genes): An interactive resource for the annotation of bacterial genomes. *Nucleic Acids Res*, **31**, 3720-3722.

Enault, F., Suhre, K., Abergel, C., Poirot, O. and Claverie, J.M. (2003) Annotation of bacterial genomes using improved phylogenomic profiles. *Bioinformatics*, **19 Suppl 1**, i105-107.

Abergel, C., Coutard, B., Byrne, D., Chenivesse, S., Claude, J.B., Deregnaucourt, C., Fricaux, T., Gianesini-Boutreux, C., Jeudy, S., Lebrun, R., Maza, C., Notredame, C., Poirot, O., Suhre, K., Varagnol, M. and Claverie, J.M. (2003) Structural genomics of highly conserved microbial genes of unknown function in search of new antibacterial targets. *J Struct Funct Genomics*, **4**, 141-157.

6.2 Publications de l'auteur en Sciences de l'Univers

Tulet, P., Suhre, K., Mari, C., Solmon, F. and Rosset, R. (2002) Mixing of boundary layer and upper tropospheric ozone during a deep convective event over Western Europe. *Atmospheric Environment*, **36**, 4491-4501.

Wood, R., Johnson, D., Osborne, S., Andreae, M.O., Bandy, B., Bates, T.S., O'Dowd, C., Glantz, P., Noone, K., Quinn, P.K., Rudolph, J. and Suhre, K. (2000) Boundary layer and aerosol evolution during the 3rd Lagrangian experiment of ACE-2. *Tellus Series B-Chemical and Physical Meteorology*, **52**, 401-422.

Suhre, K., Johnson, D.W., Mari, C., Rosset, R., Osborne, S., Wood, R., Bates, T.S. and Raes, F. (2000) A continental outbreak of air during the Second Aerosol Characterization Experiment (ACE 2): A Lagrangian experiment. *Journal of Geophysical Research-Atmospheres*, **105**, 17911-17924.

Suhre, K., Crassier, V., Mari, C., Rosset, R., Johnson, D.W., Osborne, S., Wood, R., Andreae, M.O., Bandy, B., Bates, T.S., Businger, S., Gerbig, C., Raes, F. and Rudolph, J. (2000) Chemistry and aerosols in the marine boundary layer: 1-D modelling of the three ACE-2 Lagrangian experiments. *Atmospheric Environment*, **34**, 5079-5094.

Siems, S.T., Hess, G.D., Suhre, K., Businger, S. and Draxler, R.R. (2000) The impact of wind shear on observed and simulated trajectories during the ACE-1 Lagrangian experiments. *Australian Meteorological Magazine*, **49**, 109-120.

Johnson, D.W., Osborne, S., Wood, R., Suhre, K., Quinn, P.K., Bates, T., Andreae, M.O., Noone, K.J., Glantz, P., Bandy, B., Rudolph, J. and O'Dowd, C. (2000) Observations of the evolution of the aerosol, cloud and boundary-layer characteristics during the 1st ACE-2 Lagrangian experiment. *Tellus Series B-Chemical and Physical Meteorology*, **52**, 348-374.

Johnson, D.W., Osborne, S., Wood, R., Suhre, K., Johnson, R., Businger, S., Quinn, P.K., Wiedensohler, A., Durkee, P.A., Russell, L.M., Andreae, M.O., O'Dowd, C., Noone, K.J., Bandy, B., Rudolph, J. and Rapsomanikis, S. (2000) An overview of the Lagrangian experiments undertaken during the North Atlantic regional Aerosol Characterisation Experiment (ACE-2). *Tellus Series B-Chemical and Physical Meteorology*, **52**, 290-320.

Crassier, V., Suhre, K., Tulet, P. and Rosset, R. (2000) Development of a reduced chemical scheme for use in mesoscale meteorological models. *Atmospheric Environment*, **34**, 2633-2644.

Cohard, J.M., Pinty, J.P. and Suhre, K. (2000) On the parameterization of activation spectra from cloud condensation nuclei microphysical properties. *Journal of Geophysical Research-Atmospheres*, **105**, 11753-11766.

Wang, Q., Suhre, K., Krummel, P., Siems, S., Pan, L.L., Bates, T.S., Johnson, J.E., Lenschow, D.H., Heubert, B.J., Kok, G.L., Schillawski, R.D., Prevot, A.S.H. and Businger, S. (1999) Characteristics of marine boundary layers during two Lagrangian measurement periods 1. General conditions and mean characteristics. *Journal of Geophysical Research-Atmospheres*, **104**, 21751-21765.

Wang, Q., Lenschow, D.H., Pan, L.L., Schillawski, R.D., Kok, G.L., Prevot, A.S.H., Laursen, K., Russell, L.M., Bandy, A.R., Thornton, D.C. and Suhre, K. (1999) Characteristics of the marine boundary layers during two Lagrangian measurement periods 2. Turbulence structure. *Journal of Geophysical Research-Atmospheres*, **104**, 21767-21784.

Mari, C., Suhre, K., Rosset, R., Bates, T.S., Huebert, B.J., Bandy, A.R., Thornton, D.C. and Businger, S. (1999) One-dimensional modeling of sulfur species during the First Aerosol Characterization Experiment (ACE 1) Lagrangian B. *Journal of Geophysical Research-Atmospheres*, **104**, 21733-21749.

Suhre, K., Mari, C., Bates, T.S., Johnson, J.E., Rosset, R., Wang, Q., Bandy, A.R., Blake, D.R., Businger, S., Eisele, F.L., Huebert, B.J., Kok, G.L., Mauldin,

R.L., Prevot, A.S.H., Schillawski, R.D., Tanner, D.J. and Thornton, D.C. (1998) Physico-chemical modeling of the First Aerosol Characterization Experiment (ACE 1) Lagrangian B - 1. A moving column approach. *Journal of Geophysical Research-Atmospheres*, **103**, 16433-16455.

Matthijsen, J., Suhre, K., Rosset, R., Eisele, F.L., Mauldin, R.L. and Tanner, D.J. (1998) Photodissociation and UV radiative transfer in a cloudy atmosphere: Modeling and measurements. *Journal of Geophysical Research-Atmospheres*, **103**, 16665-16676.

Mari, C., Suhre, K., Bates, T.S., Johnson, J.E., Rosset, R., Bandy, A.R., Eisele, F.L., Mauldin, R.L. and Thornton, D.C. (1998) Physico-chemical modeling of the First Aerosol Characterization Experiment (ACE 1) Lagrangian B - 2. DMS emission, transport and oxidation at the mesoscale. *Journal of Geophysical Research-Atmospheres*, **103**, 16457-16473.

Cammas, J.P., Jacoby-Koaly, S., Suhre, K., Rosset, R. and Marenco, A. (1998) Atlantic subtropical potential vorticity barrier as seen by Measurements of Ozone by Airbus In-Service Aircraft (MOZAIC) flights. *Journal of Geophysical Research-Atmospheres*, **103**, 25681-25693.

Suhre, K., Cammas, J.P., Nedelec, P., Rosset, R., Marenco, A. and Smit, H.G.J. (1997) Ozone-rich transients in the upper equatorial Atlantic troposphere. *Nature*, **388**, 661-663.

Matthijsen, J., Suhre, K., Bechtold, P. and Rosset, R. (1997) The effect of fractional cloudiness on the oxidation of SO2. *Tellus Series B-Chemical and Physical Meteorology*, **49**, 343-356.

Martin, D., Tsivou, M., Bonsang, B., Abonnel, C., Carsey, T., SpringerYoung, M., Pszenny, A. and Suhre, K. (1997) Hydrogen peroxide in the marine atmospheric boundary layer during the Atlantic Stratocumulus Transition Experiment Marine Aerosol and Gas Exchange Experiment in the eastern subtropical North Atlantic. *Journal of Geophysical Research-Atmospheres*, **102**, 6003-6015.

FassiFihri, A., Suhre, K. and Rosset, R. (1997) Internal and external mixing in atmospheric aerosols by coagulation: Impact on the optical and hygroscopic properties of the sulphate-soot system. *Atmospheric Environment*, **31**, 1393-1402.

Bedos, C., Suhre, K. and Rosset, R. (1996) Adaptation of a cloud activation scheme to

a spectral-chemical aerosol model. *Atmospheric Research*, **41**, 267-279.

Suhre, K., Andreae, M.O. and Rosset, R. (1995) Biogenic Sulfur Emissions and Aerosols over the Tropical South-Atlantic .2. One-Dimensional Simulation of Sulfur Chemistry in the Marine Boundary-Layer. *Journal of Geophysical Research-Atmospheres*, **100**, 11323-11334.

Suhre, K. and Rosset, R. (1994) Modification of a Linearized Semiimplicit Scheme for Chemical-Reactions Using a Steady-State Approximation. *Annales Geophysicae-Atmospheres Hydrospheres and Space Sciences*, **12**, 359-361.

Suhre, K. and Rosset, R. (1994) Dms Oxidation and Turbulent Transport in the Marine Boundary-Layer - a Numerical Study. *Journal of Atmospheric Chemistry*, **18**, 379-395.

6.3 Thèses et mémoires de l'auteur

Suhre, K. (1994) Modélisation couplée du transport et de la chimie du diméthyl de soufre dans la couche limite marine nuageuse. Impact climatique et étude de processus. *Thèse de doctorat*. Université Paul Sabatier, Toulouse, France.

Suhre, K. (1991) Das Feld am Punkt im äußeren elektromagnetischen Feld. *Diplomarbeit im Fachbereich Physik*. Universität Osnabrück, Allemagne.

6.4 Article fourni « Thermophiles »

Suhre, K. and Claverie, J.M. (2003) Genomic correlates of hyperthermostability, an update. *J Biol Chem*, **278**, 17198-17202.

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 278, No. 19, Issue of May 9, pp. 17198–17202, 2003

Printed in U.S.A.

Genomic Correlates of Hyperthermostability, an Update*

Received for publication, February 6, 2003
Published, JBC Papers in Press, February 24, 2003, DOI 10.1074/jbc.M301327200

Karsten Suhre and Jean-Michel Claverie‡

From the Structural and Genomic Information Laboratory, UPR 2589-CNRS, Institute of Structural Biology and Microbiology, CNRS, Marseille, France

It has been shown (Cambillau, C., and Claverie, J. M. (2000) *J. Biol. Chem.* 275, 32383–32386) that a large difference between the proportions of charged *versus* polar (non-charged) amino acids (*CvP*-bias) was an adequate, if empirical, signature of the proteome of hyperthermophilic organisms (T_{growth} >80 °C). Since that study, the number of available microbial genomes has more than doubled, raising the possibility that the simple *CvP*-bias rule might no longer hold. Taking advantage of the new sequence data, we re-analyzed the genomes of 9 fully sequenced thermophiles, 9 hyperthermophiles, and 53 mesothermophile microorganisms to identify the genomic correlates of hyperthermostability on a wider data set. Our new results confirm that the C*v*P-bias previously identified on a much smaller data set still holds. Moreover, we show that it is an optimal criterion, in the sense that it corresponds to the most discriminating factor between hyperthermophilic and mesothermophilic microorganisms in a principal component analysis. In parallel, we evaluated two other recently proposed correlates of hyperthermostability, the proteome average pI and the dinucleotide statistical index (Kawashima, T., Amano, N., Koike, H., Makino, S., Higuchi, S., Kawashima-Ohya, Y., Watanabe, K., Yamazaki, M., Kanehori, K., Kawamoto, T., Nunoshiba, T., Yamamoto, Y., Aramaki, H., Makino, K., and Suzuki, M. (2000) *Proc. Natl. Acad. Sci.* 97, 14257–14262). We show that the *CvP*-bias is the sole criterion that is able to clearly discriminate hyperthermophile from mesothermophile microorganisms on a global genomic basis.

Although most organisms grow at temperatures ranging between 20 and 50 °C, several archaea and a few bacteria, such as *Pyrococcus* and *Aquifex*, have been found capable of withstanding temperatures close to or higher than 100 °C. Identification of the molecular basis of the increased thermostability of the proteins of such hyperthermophilic organisms is expected to help our understanding of protein folding as well as the design of enzymes retaining their activity at high temperatures (Ref. 1 and references therein). In a previous comparative study, Cambillau and Claverie (2) found that a large difference between the proportions of charged (Asp, Glu, Lys, Arg) *versus* polar (non-charged) (Asn, Gln, Ser, Thr) amino acids (abbreviated as *CvP*[1]-bias) was the most prominent signature of the hyperthermophilic life style at the proteome level. This global *CvP*-bias was reflected in the amino acid composition of the water-accessible residues computed from an analysis of the surface of 131 mesophilic *versus* 58 hyperthermophilic proteins.

Given the rapidly increasing number of fully sequenced microbial genomes (more than doubled since the initial study), inferences derived in the past from correlation studies on a limited set of examples are in constant danger of being proven wrong. Now the genomes of seven new thermophilic and hyperthermophile archaea and of three new thermophilic bacteria have been deciphered. Besides these thermophilic organisms, numerous new extremophile organisms, such as the halophilic archaea Halobacterium sp., the halophilic bacterium *Sinorhizobium meliloti*, and the alkaliphilic bacterium *Bacillus halodurans* have been added to the list, as well as a large number of mesophilic bacteria and archaea (Table I). With these new genomes from a wider evolutionary spectrum, any previously derived "life style" criterion is at risk of failure in the form of false positive (classifying a mesophile as hyperthermophile) or false negative (classifying an hyperthermophile as mesophile) predictions. In particular, the newly added extremophile sequence data allow us to investigate whether the bias in favor of charged rather than polar residues previously observed in the proteome of hyperthermophiles by Cambillau and Claverie (2) is truly specific to the adaptation in high temperatures or whether it could also be linked to other extreme environments.

This new body of sequence data also gives us an opportunity to reassess the conclusions of several other studies. By systematically comparing the archaeon *Thermoplasma volcanium* genomic sequence with seven other genomic sequences of archaea, all exhibiting higher optimal growth temperature (OGT), Kawashima *et al.* (3) identified a number of strong correlations between some characteristics of genome organization. For instance, they reported that the J_2 index (computed from the dinucleotide frequency, see "Materials and Methods") was increasing together with the OGT. They also noticed some characteristic changes in the distribution of the isoelectric points (pI) of the proteins: with increasing OGT, the fraction of the basic protein subset (pI >7) becomes larger and with it the genome-average pI.

In a different work, Kreil and Ouzounis (4) used hierarchical clustering and principal component analysis to identify the factors affecting the global amino acid composition of the predicted proteome of 6 thermophilic archaea, 2 thermophilic bacteria, 17 mesophilic bacteria, and 2 eukaryotic species. They concluded that the G + C content is indeed a dominant discriminating property but unrelated to any preference for a thermophilic lifestyle. They also noticed that thermophilic species could be identified by their global amino acid compositions alone, albeit without precisely defining the nature of this discriminating power.

In this study, we applied a similar approach to a much

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ To whom correspondence should be addressed. Tel.: 33-4-91-16-45-48; Fax: 33-4-91-16-45-49; E-mail: Jean-Michel.Claverie@igs.cnrs-mrs.fr.

[1] The abbreviations used are: *CvP*, charged *versus* polar; OGT, optimal growth temperature; PCA, principal component analysis; G + C, guanine + cytosine.

TABLE I

Properties of the 71 genomes analyzed here, ranked by decreasing CvP bias

Properties of the 71 genomes analyzed here, ranked by decreasing *CvP* bias (see "Materials and Methods" for details). Mesophiles (OGT, <55%) are highlighted in blue, thermophiles (OGT, <80°) in orange, and hyperthermophiles in red.

genome	kingdom	OGT	G+C%	PCA 2	CvP-bias	J2	average pI
Methanopyrus kandleri	A	98°C	61,2	-6,07	17,99	0,06	5,94
Pyrococcus abyssi	A	97°C	44,7	-4,18	15,54	0,17	7,35
Aquifex aeolicus	B	90°C	43,3	-3,85	15,41	0,21	7,56
Methanococcus jannaschii	A	85°C	31,3	-2,91	14,85	0,13	7,30
Pyrococcus horikoshii	A	95°C	41,9	-3,84	14,67	0,17	7,67
Pyrococcus furiosus	A	98°C	40,8	-3,88	14,57	0,18	7,45
Archaeoglobus fulgidus	A	82°C	48,6	-3,78	14,05	0,15	6,62
Thermotoga maritima	B	80°C	46,3	-3,56	13,23	0,18	6,87
Thermoanaerobacter tengcongensis	B	75°C	37,6	-2,58	12,09	0,14	7,13
Pyrobaculum aerophilum	A	98°C	51,4	-3,01	11,99	0,06	8,02
Methanobacterium thermoautotrophicum	A	65°C	49,5	-2,88	11,05	0,10	6,02
Aeropyrum pernix	A	90°C	56,3	-3,26	10,62	0,11	8,09
Clostridium perfringens	B	37°C	28,5	-1,32	9,39	0,12	6,49
Carboxydothermus hydrogenoformans	B	68°C	42,0	-1,57	9,24	0,15	7,46
Halobacterium sp	A	37°C	65,9	-0,81	9,17	-0,07	4,68
Fusobacterium nucleatum	B	37°C	27,2	-0,79	9,06	0,13	7,20
Bacillus stearothermophilus	B	55°C	52,7	-0,45	9,05	0,01	8,25
Sulfolobus tokodaii	A	80°C	32,8	-1,60	8,90	0,10	7,66
Sulfolobus solfataricus	A	78°C	35,8	-1,62	8,81	0,10	7,93
Streptomyces coelicolor	B	28°C	72,0	-0,59	7,99	-0,04	6,68
Methanosarcina mazei	A	37°C	41,5	-1,26	7,91	0,14	6,39
Thermoplasma volcanium	A	60°C	39,9	-1,03	7,50	0,06	6,95
Thermoplasma acidophilum	A	58°C	46,0	-1,03	7,43	0,04	6,76
Sinorhizobium meliloti	B	26°C	62,2	-0,14	7,35	-0,01	6,93
Caulobacter crescentus	B	30°C	67,2	0,48	6,99	-0,02	7,35
Pseudomonas aeruginosa	B	37°C	66,6	0,72	6,66	-0,04	6,84
Bacillus halodurans	B	30°C	43,7	-0,58	6,62	0,07	6,43
Brucella melitensis	B	37°C	57,2	0,29	6,58	-0,02	7,28
Methanosarcina acetivorans	A	39°C	42,7	-0,90	6,56	0,14	6,57
Chlorobium tepidum TLS	B	48°C	56,5	-0,13	6,42	0,01	7,38
Campylobacter jejuni	B	37°C	30,6	1,43	6,30	0,13	7,44
Clostridium acetobutylicum	B	37°C	30,9	-0,11	6,28	0,08	7,36
Mesorhizobium loti	B	26°C	62,5	0,51	6,20	-0,04	7,27
Listeria innocua	B	37°C	37,4	-0,01	6,16	0,06	6,14
Bacillus subtilis	B	30°C	43,5	0,10	6,07	0,06	6,59
Listeria monocytogenes	B	37°C	38,0	-0,06	6,00	0,06	5,97
Agrobacterium tumefaciens C58	B	30°C	59,0	0,39	5,93	-0,01	6,95
Streptococcus pneumoniae R6	B	37°C	39,7	0,37	5,52	0,10	6,45
Borrelia burgdorferi	B	35°C	28,2	1,25	5,51	0,14	8,37
Treponema pallidum	B	37°C	52,8	0,28	4,79	-0,03	8,17
Neisseria meningitidis MC58	B	37°C	51,5	1,15	4,63	0,00	7,37
Deinococcus radiodurans	B	30°C	66,6	0,71	4,32	-0,03	7,12
Ralstonia solanacearum	B	30°C	67,0	1,37	4,27	-0,11	7,59
Streptococcus pyogenes	B	37°C	38,5	1,25	4,20	0,08	6,71
Lactococcus lactis	B	30°C	35,3	0,83	4,11	0,11	6,62
Helicobacter pylori 26695	B	37°C	38,9	1,87	4,05	0,13	7,98
Mycoplasma pulmonis	B	37°C	26,6	1,97	3,97	0,18	8,63
Mycobacterium tuberculosis H37Rv	B	37°C	65,6	0,56	3,93	-0,08	6,85
Corynebacterium glutamicum	B	30°C	53,8	0,47	3,91	0,01	5,62
Mycobacterium leprae	B	37°C	57,8	0,35	3,83	-0,07	6,67
Xanthomonas campestris	B	26°C	65,1	1,95	3,43	-0,13	7,28
Staphylococcus aureus Mu50	B	37°C	32,9	1,27	3,42	0,00	6,73
Haemophilus influenzae	B	37°C	38,2	1,56	3,12	0,02	6,98
Salmonella typhi	B	37°C	51,9	1,53	2,82	-0,04	6,97
Escherichia coli K12	B	37°C	50,8	1,54	2,63	-0,04	6,81
Salmonella typhimurium LT2	B	37°C	52,2	1,63	2,59	-0,04	6,94
Rickettsia conorii	B	37°C	32,4	1,78	2,51	0,04	7,84
Pasteurella multocida	B	37°C	40,4	1,96	2,43	0,01	7,02
Buchnera sp	B	26°C	26,4	1,96	2,31	0,08	9,44
Rickettsia prowazekii	B	37°C	29,0	1,85	2,19	0,02	8,26
Vibrio cholerae	B	28°C	47,5	1,90	1,91	-0,01	6,83
Xylella fastidiosa	B	26°C	52,6	1,76	1,73	-0,06	7,71
Chlamydia trachomatis	B	37°C	41,3	1,47	1,66	0,14	7,12
Chlamydia muridarum	B	37°C	40,3	1,52	1,51	0,15	7,29
Ureaplasma urealyticum	B	37°C	25,5	3,04	1,43	0,06	8,30
Chlamydophila pneumoniae AR39	B	37°C	40,6	1,60	1,29	0,14	7,35
Yersinia pestis	B	37°C	47,6	2,11	1,07	-0,02	7,05
Mycoplasma genitalium	B	37°C	31,7	3,22	0,60	0,11	9,00
Synechocystis PCC6803	B	37°C	47,7	2,14	0,35	0,10	6,34
Mycoplasma pneumoniae	B	37°C	40,0	3,25	-0,21	0,07	8,57
Nostoc sp	B	30°C	41,3	2,41	-0,99	0,05	6,80

larger genomic data set. This allows us to show that the most discriminating factor between hyperthermophile and mesophile genomes remains the absolute difference between the frequency of charged and polar amino acid residues (*CvP*-bias). Furthermore, we show that a high J_2 index value is a necessary, but not sufficient, criterion for the recognition of a hyperthermophilic proteome, whereas the previously proposed genome-average pI is not a satisfactory correlate of hyperthermostability.

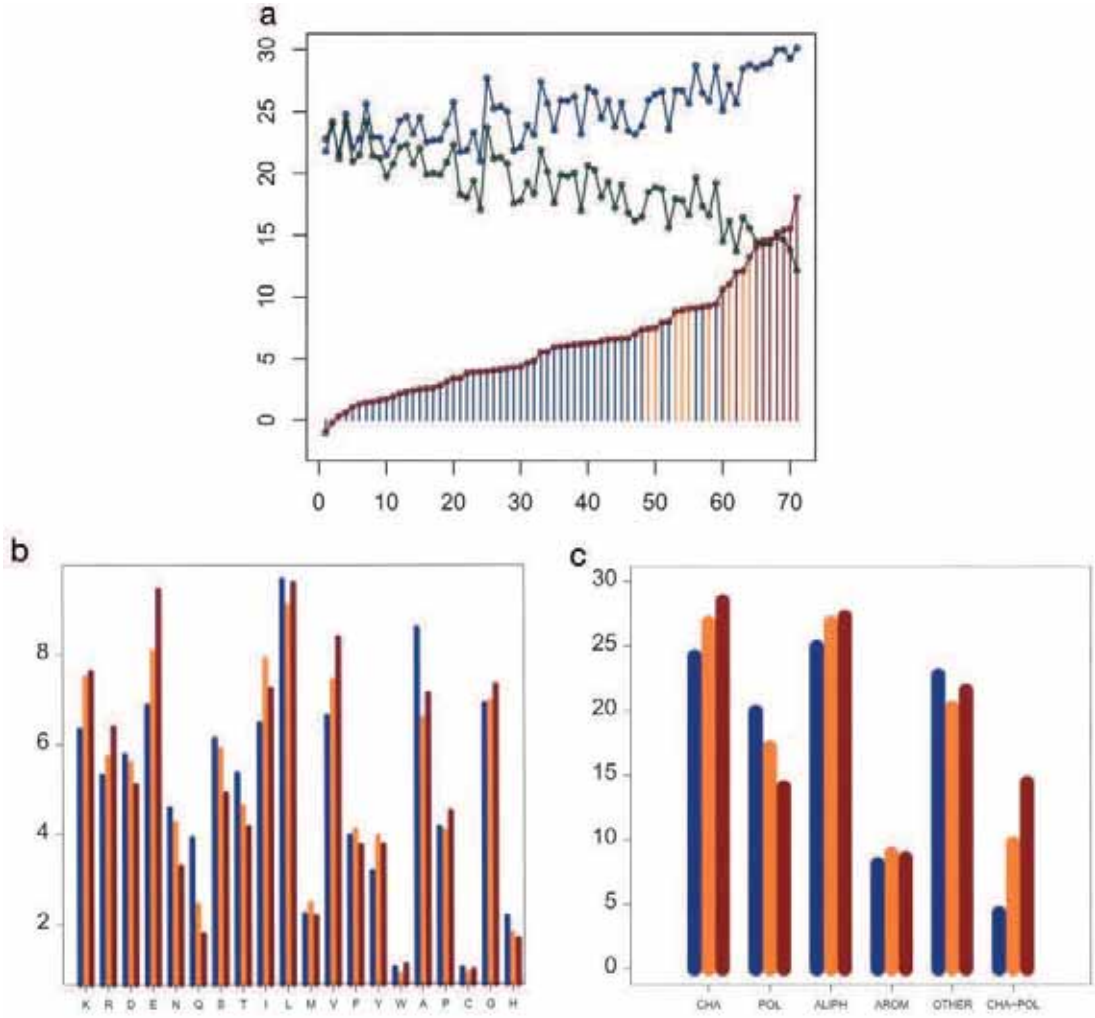

FIG. 1. *a*, plot of the percentages of charged amino acids (Asp, Glu, Lys, Arg, *blue*), polar non-charged amino acids (Asn, Gln, Ser, Thr, *green*), and of the difference of the two (*CvP*-bias, *red*). *Blue*, *orange*, and *red vertical lines* identify the mesophiles, thermophiles, and hyperthermophiles, respectively. *b*, plot of the percentages of the various amino acids in mesophiles (*blue*), thermophiles (*orange*), and hyperthermophiles (*red*). *c*, plot of the various amino acid classes; colors as in panel *b*.

MATERIALS AND METHODS

A total of 71 fully sequenced genomes was analyzed (Table I), including 53 mesophiles (<50 °C) (50 bacteria, 3 archaea), 9 thermophiles (50–80 °C) (4 bacteria, 5 archaea), and 9 hyperthermophiles (>80 °C) (1 bacteria, 8 archaea). 69 of those genome sequences were downloaded from GenBank™ and 2 unfinished genomes, *Carboxydothermus hydrogenoformans* and *Bacillus stearothermophilus*, from the Institute for Genome Research (TIGR, Rockville, MD) and from the University of Oklahoma, Norman, OK, respectively. The proteomes of the various organisms were defined according to the available open reading frame (ORF) annotation (when available in GenBank™). For the two unfinished genomes, the proteomes were defined as the subsets of ORFs longer than 200 amino acids. Transmembrane segments were then predicted using the simple algorithm of Kyte and Doolittle (5). For each organism, the "soluble" moiety of the proteome was then defined by discarding all proteins (ORFs) containing at least two predicted transmembrane segments. The subsequent statistical analyses were all performed on these predicted soluble proteins.

Following Kawashima *et al.* (3), the genome sequences were used to compute the J_2 index as: $J_2 = F_{YY} + F_{RR} - F_{YR} - F_{RY}$ where F_{YY} designates the relative frequency of dinucleotides pyrimidine (TT, TC, CT, CC), F_{RR} the relative frequency of purine dinucleotides (AA, AG, GA, GG), and F_{YR} and F_{RY} the corresponding mixed purine/pyrimidine combinations. Pure purine and pyrimidine dinucleotides were found to be more frequent than their mixed counterparts in hyperthermophilic Pyrococci (3).

The program "iep" from the European Molecular Biology Open Software Suite (EMBOSS) was used to calculate the theoretical pI value for every soluble protein and, from those, the average pI for every genome. Principal component analysis was performed using the statistical package *R*. 71 genomes from meso-, thermo-, and hyperthermophilic organisms were included. When multiple genome sequences from closely related strains or species were available, only one of them was included in the analysis to avoid introducing a potential statistical bias.

RESULTS

Hyperthermophiles Do Exhibit a Specific Proteome Composition Signature—The computed amino acid compositions for 9 fully sequenced hyperthermophilic and 53 mesophilic bacteria and archaea are presented in Table I and Fig. 1, *a–c*. The case of the 9 moderately thermophilic genomes will be discussed later. As previously observed by Cambillau and Claverie (Fig. 2 and a-c in Ref. 2) but now confirmed on a much larger data set, the proteins of hyperthermophiles exhibit a strong bias for the use of charged residues at the expense of polar residues. This includes strongly reduced frequencies for the thermolabile amino acid residues Gln and Asn, as also noted by Vieille *et al.* (6). Among other trends, the aliphatic residue Val appears to be preferred in hyperthermophiles, whereas the tiny non-polar non-charged residue Ala is avoided.

Principal component analysis (PCA) offers a way to analyze the data set more objectively and eventually identify more intricate relationships between amino acid frequencies and the adaptation of proteins to high temperature. Our PCA

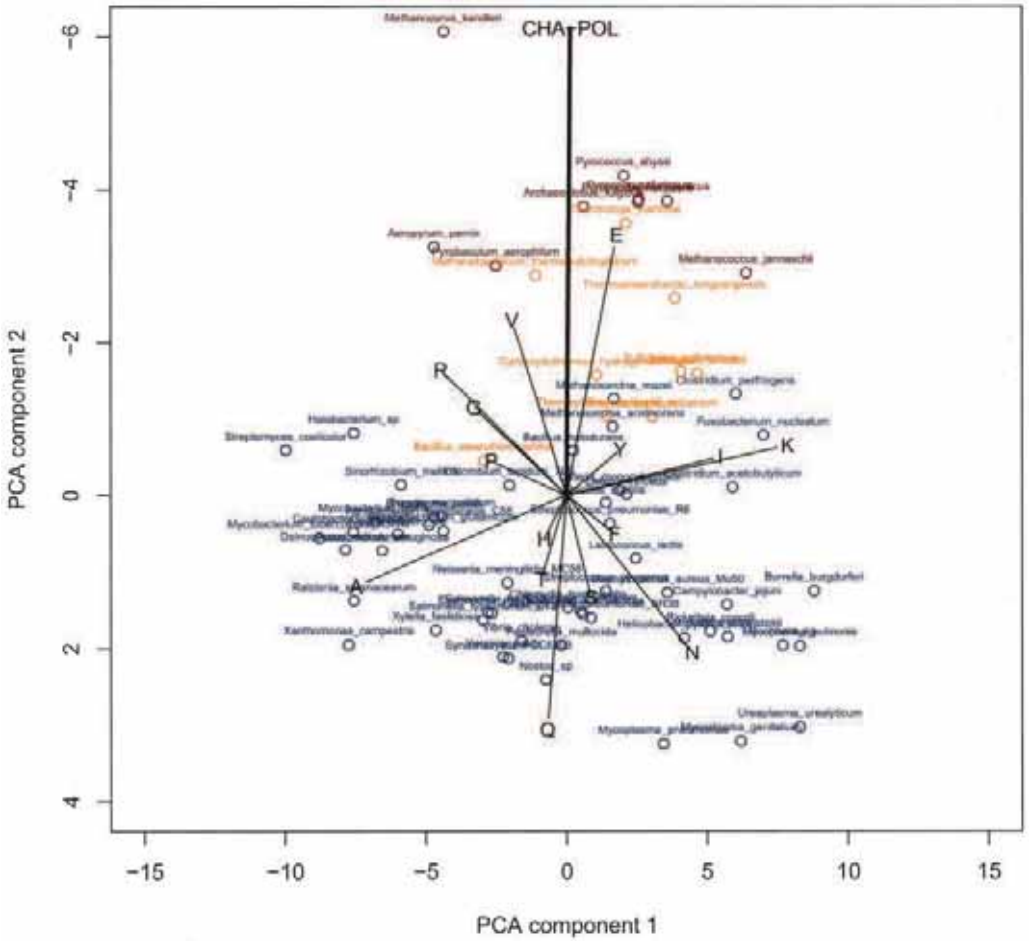

FIG. 2. **Biplot of a principal component analysis of the amino acid composition of all studied organisms.** *Blue dots* identify mesophiles; *orange dots* identify thermophiles; *red dots* identify hyperthermophiles. *Green vectors* represent the position of the different amino acids in the biplot (only vectors with significant contributions to PCA components 1 and 2 are drawn). The *black vector* represents $F_{CvP} = F_K + F_R + F_D + F_E - F_N - F_Q - F_S - F_T$.

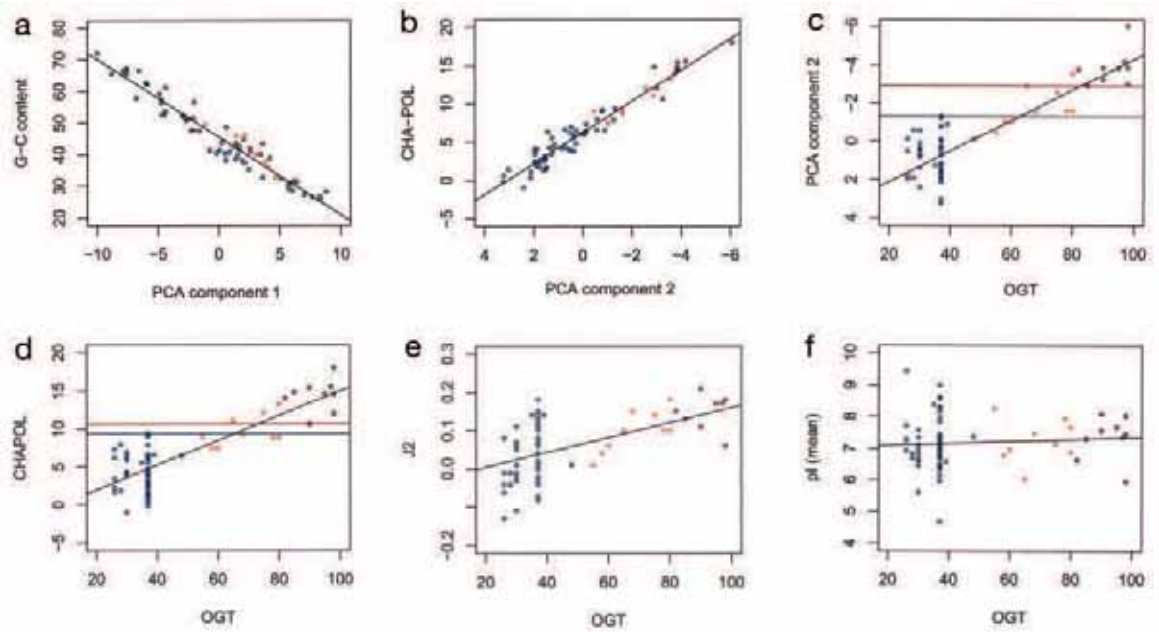

FIG. 3. **Correlation between various parameters.** *a*, correlation between G + C content and PCA component 1: $r^2 = 0.94$. *b*, correlation between *CvP*-bias and PCA component 2: $r^2 = 0.93$. *c*, correlation between PCA component 2 and OGT: $r^2 = 0.72$. *Red* and *blue lines* indicate the separation between the hyperthermophile organism with the lowest OGT and the mesophile with the highest OGT. *d*, correlation between *CvP*-bias and OGT: $r^2 = 0.68$. *e*, correlation between J_2 and OGT: $r^2 = 0.27$. *f*, correlation between average proteomic pI and OGT: $r^2 = 0.01$.

analysis (Fig. 2) is consistent with earlier results obtained with a much smaller data set by Kreil and Ouzounis (4); 85% of the data variance is accounted by the first two principal components. The first component (72% of the variance) strongly correlates with the G + C content of the genomes (Fig. 3*a*, $r^2 = 0.94$). The second component (13% of the variance) turns out to be the sole discriminating factor between meso- and hyperthermophilic lifestyles, correlating quite well with the optimal growth temperature (Fig. 3*c*, *OGT*, $r^2 = 0.72$). All other PCA components account for not more than 3.7% each of the variance.

The first PCA component being the G + C content, we had to find the biochemical interpretation of the second component. First, we found that this component exhibits a strong correlation with the previously defined (*CvP*-bias) (Fig. 3*b*, $r^2 = 0.93$). Second, once projected on the first two dimensions of PCA space, the (20-dimensional) vector representing the difference between the frequencies of charged and polar residues (CHA-POL) appears almost parallel to the second PCA dimension (Fig. 2, black vector). As shown in Fig. 3*d*, the *CvP*-bias also correlates with the OGT almost as well as the second PCA component ($r^2 = 0.68$ and 0.72, respectively). Finally, this parameter also successfully discriminates all hyperthermophilic microorganisms from all mesothermophilic ones. In conclusion, this analysis indicates that, among all possible combinations of G + C content and amino acid frequencies, the *CvP*-bias is a near optimal discriminating quantity to characterize the hyperthermophilic proteins from evolutionarily diverse microorganisms.

In contrast to the *CvP*-bias, other previously proposed quantities thought to characterize the hyperthermophilic life style do not fare well when confronted with this expanded proteome data set. Fig. 3, *e* and *f*, shows that the J_2 index, as well as the average pI, now fail to correctly discriminate between hyperthermophiles and mesophiles. In fact, both parameters correlate only weakly, if at all, with OGT ($r^2 = 0.27$ for J_2 and $r^2 = 0.01$ for average pI). However, all hyperthermophiles have a J_2 index greater than 0.06, so that a high J_2 index can be considered a necessary, but not a sufficient, criterion for the identification of a hyperthermophilic genome.

Moderately Thermophile Organisms—Because only a single moderate thermophilic organism had been sequenced (*Methanobacterium thermoautotrophicum*) at the time of the work of Cambillau and Claverie (2), little could be said about the properties of moderately thermophile organisms. Proteome data is now available for nine of them. Two (*Thermotoga maritima* and *Thermoanaerobacter tengcongensis*) have OGTs close to the threshold of 80 °C we somewhat arbitrarily used to define hyperthermophile organisms. It turns out that they both exhibit PCA2, *CvP*-bias, and J_2 values that could all allow their classification in continuity with previously defined hyperthermophiles.

This clear discrimination breaks down if we use an OGT of 75 °C as the hyperthermophilicity threshold. Down to this temperature, the value along the PCA2 coordinate remains a valid criterion, allowing two *Sulfolobus* species (OGT of 80 and 78 °C) to be classified in continuity with the other hyperthermophiles (PCA2 < -1.6 for OGT higher than 75 °C, and PCA2 > -1.3 OGT lower than 75 °C; see Fig. 3*c*). However, the straightforward, more biochemically meaningful *CvP*-bias criteria become invalidated by three mesophiles with OGT of 37 °C, *Fusobacterium nucleatum* (*CvP* = 9.06), *Halobacterium sp* (*CvP* = 9.17), *Clostridium perfingens* (*CvP* = 9.39), exhibiting larger values than *Sulfolobus* (*CvP* = 8.90 and 8.80).

DISCUSSION

Using a much larger data set including many new thermophile and mesophile whole genome sequences, this follow-up study confirms the previous suggestion that the global replacement of polar residues (Asn, Gln, Ser, Thr) by charged residues (Asp, Glu, Lys, Arg) is the dominant proteome characteristic of microorganisms adapted to hyperthermophilic growth condition (OGT >80 °C). This effect is observed for both bacteria and archaebacteria and thus is not a simple consequence of phylogenetic relationship.

Even though the strict correspondence between the highest *CvP*-bias and the highest OGT breaks down below 80 °C, this property globally remains a characteristic of all thermophilic (OGT >55 °C) microorganisms, as shown in Fig. 1*c* where the *CvP*-bias averaged over all mesothermophiles remains markedly higher than for mesophile organisms. The influence of other adaptive strategies (*e.g.* to high salinity or extreme pH environments), together with the phylogenetic affinities of certain mesophiles to thermophile organisms (7), probably contributes to weaken *CvP*-bias signal, allowing a few false positives to sneak in (such as *Clostridium perfringens* or *Streptomyces coelicolor*) (Table I.).

This study confirms that a strong *CvP*-bias is specifically associated with hyperthermophilic proteomes. This observation is consistent with the thermodynamic advantage resulting from the increased significance of coulomb interaction with the increasing temperature (as the dielectric constant of water decreases). The simultaneous increase of oppositely charged residues (mostly Arg, Lys, and Glu) further allows for more ion pairs to be formed at the surface of hyperthermostable proteins (2). This rationale involving the stability of proteins in a high temperature aqueous environment is also supported by our observation that the proteins predicted to be associated to the membrane (and thus designed for hydrophobic environments) exhibit a much less significant *CvP*-bias (data not shown). Finally, the fact that a large number of diverse genomes confirms a statistical trend previously inferred from a much smaller data set argues that increased ion-pair formation is both a significant physico-chemical factor and the preferred evolutionary pathway toward thermostable soluble proteins.

REFERENCES

1. Kumar, S., and Nussinov, R. (2001) *Cell. Mol. Life Sci.* **58,** 1216–1233
2. Cambillau, C., and Claverie, J. M. (2000) *J. Biol. Chem.* **275,** 32383–32386
3. Kawashima, T., Amano, N., Koike, H., Makino, S., Higuchi, S., Kawashima-Ohya, Y., Watanabe, K., Yamazaki, M., Kanehori, K., Kawamoto, T., Nunoshiba, T., Yamamoto, Y., Aramaki, H., Makino, K., and Suzuki, M. (2000) *Proc Natl. Acad. Sci.* **97,** 14257–14262
4. Kreil, D. P., and Ouzounis, C. A. (2001) *Nucleic Acids Res.* **29,** 1608–1615
5. Kyte, J., and Doolittle, R. F. (1982) *J. Mol. Biol.* **157,** 105–132
6. Vieille, C., Epting, K. L., Kelly, R. M., and Zeikus, J. G. (2001) *Eur. J. Biochem.* **268,** 6291–6301
7. Brochier, C., and Philippe, H. (2002) *Nature* **417,** 244

6.5 Article fourni « Base de données FusionDB »

Suhre, K. and Claverie, J.M. (2004) FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. *Nucleic Acids Res*, **32 Database issue**, D273-276.

Nucleic Acids Research, 2004, Vol. 32, Database issue ***D273–D276***
DOI: 10.1093/nar/gkh053

FusionDB: a database for in-depth analysis of prokaryotic gene fusion events

Karsten Suhre* and Jean-Michel Claverie

Information Génomique and Structurale, CNRS-UPR 2589, 31 chemin Joseph Aiguier, 13402 Marseille Cedex 20, France

Received August 6, 2003; Revised September 9, 2003; Accepted September 23, 2003

ABSTRACT

FusionDB (http://igs-server.cnrs-mrs.fr/FusionDB/) constitutes a resource dedicated to in-depth analysis of bacterial and archaeal gene fusion events. Such events can provide the 'Rosetta stone' in the search for potential protein–protein interactions, as well as metabolic and regulatory networks. However, the false positive rate of this approach may be quite high, prompting a detailed scrutiny of putative gene fusion events. FusionDB readily provides much of the information required for that task. Moreover, FusionDB extends the notion of gene fusion from that of a single gene to that of a family of genes by assembling pairs of genes from different genomes that belong to the same Cluster of Orthogonal Groups (COG). Multiple sequence alignments and phylogenetic tree reconstruction for the N- and C-terminal parts of these 'COG fusion' events are provided to distinguish single and multiple fusion events from cases of gene fission, pseudogenes and other false positives. Finally, gene fusion events with matches to known structures of heterodimers in the Protein Data Bank (PDB) are identified and may be visualized. FusionDB is fully searchable with access to sequence and alignment data at all levels. A number of different scores are provided to easily differentiate 'real' from 'questionable' cases, especially when larger database searches are performed. FusionDB is cross-linked with the 'Phylogenomic Display of Bacterial Genes' (PhydBac) online web server. Together, these servers provide the complete set of information required for in-depth analysis of non-homology-based gene function attribution.

INTRODUCTION

Gene fusion events have been proposed to represent valuable 'Rosetta stone' information for the identification of potential protein–protein interactions and metabolic or regulatory networks (1,2). More generally, information on gene fusion events can be combined with other non-homology-based approaches, such as phylogenomic profiling and identification of conserved chromosomal localization, to provide hypotheses for the characterization of proteins of unknown function (3–5). A number of web-based databases, such as AllFuse (5), STRING (6) and Predictome (7), implement this idea already. However, most of the available databases limit the definition of a gene fusion event to simple non-overlapping side-by-side BLAST (8) matches of two genes from a reference genome to a single open reading frame (ORF) in a target genome, but without providing much information for further in-depth analysis. Searches based on these databases give good starting points for hypothesis building, but the false positive rate may be quite high (in particular in cases where genes evolved through gene duplication and where the identification of gene orthology is hence difficult). The user is then left with the task of assembling the data required for more extensive case analysis.

Here we present a database that is based on a more strict definition of a gene fusion event, applying a mutual best match criteria [(9), see Fig. 1 and methods]. It drastically reduces the number of false positives, at the expense of a potentially similarly high number of false negatives. To recover from this drawback, gene fusion events between genes from different genomes that belong to the same Cluster of Orthologous Groups (COG) (10) are pulled together in what we call 'COG fusion events'. Analysis of these COG fusion events then allows for the investigation of gene fusion in its phylogenomic context, using multiple alignments and phylogenetic tree reconstruction. Questions on the history of individual gene fusion events, such as whether a particular event occurred only once or many times during evolution, or whether more complex processes such as horizontal gene transfer, gene fission and gene decay are involved may be addressed using the information provided by FusionDB. The extension to 'COG fusion events' also provides information on general gene fusion tendencies in a whole bacterial genomic context to address questions such as 'Which type of genes are most likely to fuse?' FusionDB thereby complements our phylogenetic profiling web server PhydBac (http://igs-server.cnrs-mrs.fr/phydbac/) (11), which is based on the same philosophy: providing detailed non-homology-based information for in-depth analysis of potential protein–protein interactions. FusionDB is thus complementary to the databases cited above (5–7).

*To whom correspondence should be addressed. Tel: +33 4 91 16 46 04; Fax: +33 4 91 16 45 49; Email: karsten.suhre@igs.cnrs-mrs.fr

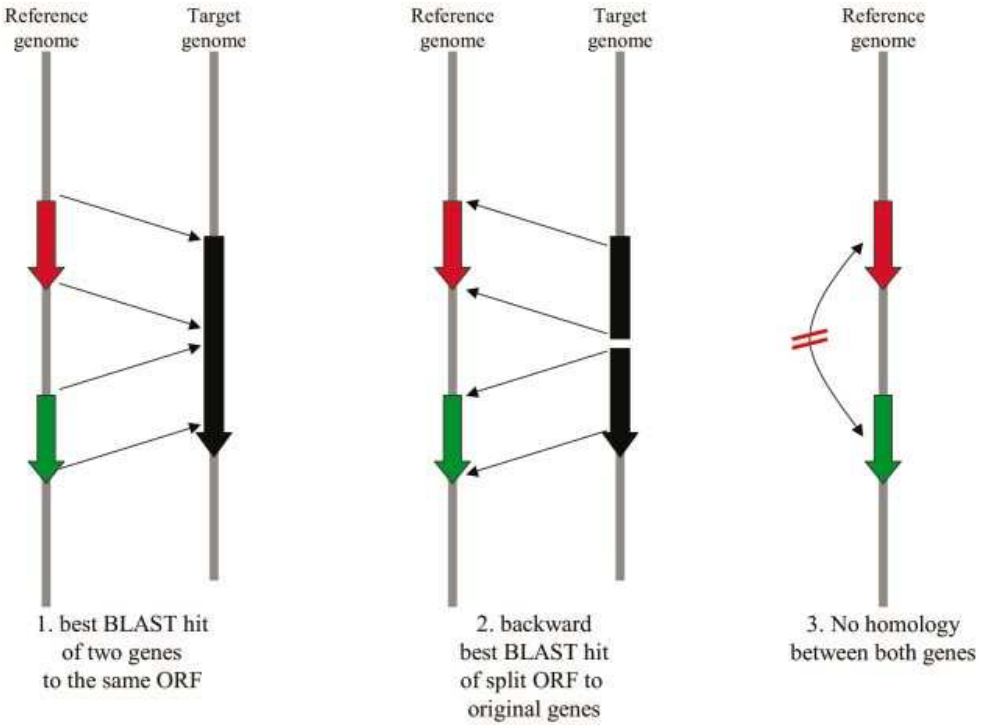

Figure 1. Criteria for a putative gene fusion event based on a mutual best match criteria (see text for details).

SOURCES OF GENOMIC DATA AND METHODS

All available 89 fully sequenced non-redundant bacterial and archaeal genomes (see http://igs-server.cnrs-mrs.fr/FusionDB/methods/ for a full list) were downloaded from NCBI RefSeq. Those genomes for which a COG annotation of their genes was available (51 genomes) were checked for putative gene fusion (PFE) events in all 89 genomes as follows: a PFE between two genes from a given reference genome in a given target genome is subject to three criteria (Fig. 1):

(i) Each of the two reference genes must match the same ORF in the target genome as their highest scoring BLAST hit. The overlap between the BLAST hits of both genes must not exceed 10% of the size of the smaller of the two target genes.

(ii) When split between the two BLAST hits, the two halves of the target ORF must match back to the original two reference genes as their best BLAST hit to the reference genome.

(iii) The reference genes must not be homologous to each other.

Note that the search for PFEs is done on the basis of the annotated genes from a given reference genome, but against all possible ORFs in the target genome (including overlapping ORFs). This increases the chances of finding a gene fusion event that might have been discarded by a human annotator. Every PFE is then subjected to a scoring scheme based on different evaluations of its pairwise and multiple (triple) alignments by calculating the following five scores.

(i) The separation index (sep) is a measurement of the mix between the domains from the two reference genes when they are placed in a triple alignment with the target ORF. This index varies between 0 (total mix) and 1 (complete separation).

(ii) The fusion index (fus) is the fraction of residues in the concatenated reference genes that have similar properties to their aligned counterparts in the target ORF. This index may vary between 0 (virtually no homology between the reference genes and the target ORF) and 1 (strong homology).

(iii) The gene coverage (cov) is the fraction of the two reference genes that is alignable with the target ORF in a triple alignment. This index varies between 0 (no relationship at all between the reference genes and the target ORF) and 1 (all domains of the reference genes have a counterpart in the target ORF).

(iv) The size ratio (ratio) between the size of the reference genes and the target ORF indicates possible domain gain or loss after the gene fusion event has occurred.

(v) The 'baditude' (bad) is the fraction of residues that are aligned between the reference genes when placed in a triple alignment with the target ORF. This index varies between 0 (both reference genes are evolutionarily unrelated) and 1 (both reference genes are homologues). A high 'baditude' is an indicator of genes with paralogous domains.

QUERYING THE DATABASE

FusionDB may be searched by gene name, gene annotation, gene function, COG identifier or simply by entering an amino acid sequence in FASTA format. Queries may be confined to specific reference and target genomes, and limits on the different scores can be imposed. Output in full-page mode contains visualization of the different alignments that were used for scoring, and in the case of gene pairs that both belong to a COG a special COG-analysis page is provided. This

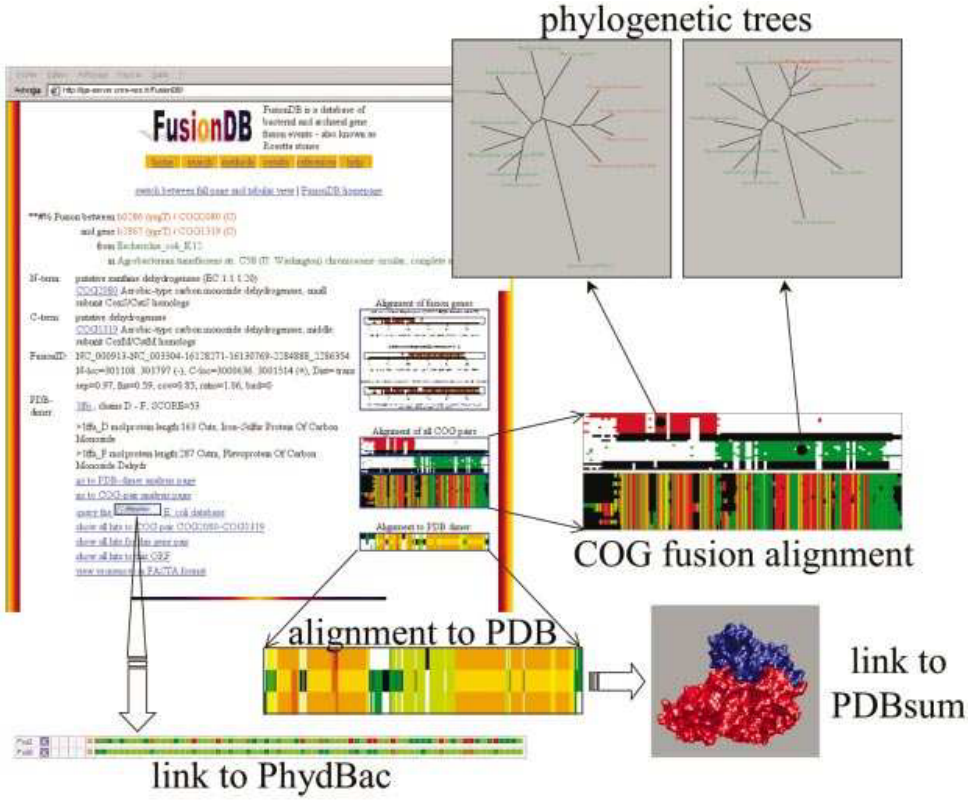

Figure 2. Screenshot of FusionDB full-page output for a query to COG2080 and examples of some related information that can be obtained through this page. PhydBac (http://igs-server.cnrs-mrs.fr/phydbac/) is the 'Phylogenomic Display of Bacterial Genes' online web tool. In the top of the 'COG fusion alignment', N- and C- terminal genes are presented in red and green, respectively, fusion ORFs are in black. The alignment of the merged genes with the fusion genes is presented below. A colour scale ranging from green over yellow to red represents the EMBOSS plotcon score for this 'merged alignment'. The 'phylogenetic trees' are based on the N- and the C-terminal 'COG fusion alignments', respectively. Genomes in which fusion events occurred are highlighted in red in the trees. The 'alignment to the PDB' is a representation of the T-Coffee alignment core index of the reference genes (top row), the fusion ORF (middle row) and the sequence of the heterodimer (bottom row), warmer colours indicating a higher confidence in the alignment quality. PDBsum (http://www.biochem.ucl.ac.uk/bsm/pdbsum/) is a database of the known 3D structures of proteins and nucleic acids.

COG-analysis page contains different types of multiple alignments and related phylogenetic trees, as well as information on related COG fusion events (networks) (Fig. 2). Extension of the research results, e.g. to all hits to a given fusion ORF is possible. In cases where a gene fusion event has a match to a heterodimer in the Protein Data Bank (PDB), a special PDB analysis page is available, providing a scored multiple alignment between the reference genes, the fusion gene and the sequences of the heterodimer in the PDB file. Output in tabulated mode or limitation to only the best hit for each gene pair may be requested if a large number of hits is expected. On each page a cross-link to PhydBac gives direct access to the phylogenetic profiles and eventual conserved chromosomal proximity of the two fusion genes.

By default, all queries are limited to a separation index (sep) of 0.6. This is found to be the most robust indicator of a 'true' gene fusion event (K. Suhre *et al.*, in preparation; see also FusionDB/results/). Note that the fusion index (fus) is dependent on the evolutionary distance between the reference and the target genome. Values of the gene coverage (cov) and the size ratio (ratio) that differ significantly from 1 are indicators of domains that have been lost or added in the process of evolution. Such cases should be inspected carefully. In some cases this can give rise to a high 'baditude' (bad) score when the added domains are homologous. If for a given query gene no fusion event is found, the user may try to extend the search to the COG family to which this gene belongs (or use the sequence search option, note also that genes with a high degree of paralogy in most genomes may not be identified as a fusion event). In situations where both genes of a PFE are associated with a COG and where several fusion events are identified by FusionDB, coherence between the phylogenetic trees of the N- and C-terminal genes as well as the history of the gene fusion can be used as indicators of 'real' fusion events and true functional orthology between the implicated genes in the different genomes. This kind of key information is not readily available on other existing database servers.

CONCLUDING REMARKS AND FUTURE PLANS

FusionDB presents significant additions to other gene-fusion-related databases. The extension of the concept of a gene fusion to a 'COG fusion' event and the application of a mutual best match criteria not only reduces the number of false positives, but also makes the use of gene fusion events as 'Rosetta stones' applicable at a genome-independent level, where the common gene pool of all prokaryotes is viewed as the sum of all identified (and still to be discovered) COGs. The

wealth of pre-calculated multiple alignments and phylogenetic trees will be welcomed by many biological analysts and annotators, as FusionDB currently covers ~20 000 potentially 'real' gene fusion events (having a separation index > 0.6), which correspond to 1355 different fused COG pairs. A more detailed analysis of these cases is underway (K. Suhre *et al.*, in preparation). FusionDB will be updated regularly as the number of publicly available fully sequenced genomes increases, and lower eukaryotes should be added in a future version. This will be particularly beneficial for obtaining more complete phylogenetic trees, which is still the best way to evaluate the 'reality' of the gene fusion events. Ultimately, FusionDB is designated to prioritize and record the experimental validity of the molecular or functional interaction of the genes involved in gene fusion events.

ACKNOWLEDGEMENTS

Sequence data was downloaded from NCBI RefSeq (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). Multiple alignments were computed using T-Coffee (http://igs-server.cnrs-mrs.fr/Tcoffee/) (12). Phylogenetic tree reconstruction was done using Phylip (http://evolution.genetics.washington.edu/phylip.html). Multiple alignments were scored with EMBOSS plotcon (http://www.hgmp.mrc.ac.uk/Software/EMBOSS/). Mode data for protein structures are from the PDB (http://www.rcsb.org/pdb/) (13).

REFERENCES

1. Galperin,M.Y. and Koonin,E.V. (2000) Who's your neighbor? New computational approaches for functional genomics. *Nat. Biotechnol.*, **18**, 609–613.
2. Sali,A. (1999) Functional links between proteins. *Nature*, **402**, 23–26.
3. Marcotte,E.M. (2000) Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol.*, **10**, 359–365.
4. Marcotte,E.M., Pellegrini,M., Thompson,M.J., Yeates,T.O. and Eisenberg,D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.
5. Enright,A.J. and Ouzounis,C.A. (2001) Functional associations of proteins in entire genomes via exhaustive detection of gene fusion. *Genome Biol.*, **2**, 341–347
6. von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.
7. Mellor,J.C., Yanai,I., Clodfelter,K.H., Mintseris,J. and DeLisi,C. (2002) Predictome: a database of putative functional links between proteins. *Nucleic Acids Res.*, **30**, 306–309.
8. Altschul,S.F., Madden,T.L., Schaeffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
9. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
10. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
11. Enault,F., Suhre,K., Poirot,O., Abergel,C. and Claverie,J.M. (2003) Phydbac (phylogenomic display of bacterial genes): an interactive resource for the annotation of bacterial genomes. *Nucleic Acids Res.*, **31**, 3720–3722.
12. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
13. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

6.6 Article fourni « Modes normaux et remplacement moléculaire »

Suhre, K. and Sanejouand, Y.H. (2004) On the potential of normal-mode analysis for solving difficult molecular-replacement problems. *Acta Crystallogr D Biol Crystallogr*, **60**, 796-799.

electronic reprint

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449
Editors: **E. N. Baker and Z. Dauter**

On the potential of normal-mode analysis for solving difficult molecular-replacement problems

Karsten Suhre and Yves-Henri Sanejouand

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

On the potential of normal-mode analysis for solving difficult molecular-replacement problems

Karsten Suhre[a]* and Yves-Henri Sanejouand[b]

[a]Information Génomique et Structurale (UPR CNRS 2589), 31 Chemin Joseph Aiguier, 13402 Marseille CEDEX 20, France, and [b]Laboratoire de Physique, Ecole Normale Supérieure, 46 Allées d'Italie, 69364 Lyon CEDEX 07, France

Correspondence e-mail: karsten.suhre@igs.cnrs-mrs.fr

Received 14 November 2003
Accepted 26 January 2004

Molecular replacement (MR) is the method of choice for X-ray crystallographic data phasing when structural data of suitable homologues are available. However, MR may fail even in cases of high sequence homology when conformational changes arising for example from ligand binding or different crystallogenic conditions come into play. In this work, the potential of normal-mode analysis as an extension to MR to allow recovery from such drawbacks is demonstrated. Three examples are presented in which screening for MR solutions with templates perturbed in the direction of one or two normal modes allows a valid MR solution to be found where MR using the original template failed to yield a model that could ultimately be refined. It has been shown recently that half of the known protein movements can be modelled by displacing the studied structure using at most two low-frequency normal modes. This suggests that normal-mode analysis has the potential to break tough MR problems in up to 50% of cases. Moreover, even in cases where an MR solution is available, this method can be used to further improve the starting model prior to refinement, eventually reducing the time spent on manual model construction (in particular for low-resolution data sets).

1. Introduction

Molecular replacement (MR) is the most cost-effective method for solving the three-dimensional structure of a new protein by X-ray crystallography. The success or failure of MR depends on a variety of factors, not least of which are the diffraction data quality in the mid- and low-resolution range and the number of molecules in the asymmetric unit cell that have to be placed. The decisive factor is the availability of a template of sufficiently high structural homology to the target. However, it is likely that most structural genomics projects have already encountered the frustrating situation where MR failed despite a 'good' template with high sequence similarity (or even identity) to the target. On the eventual phasing of the diffraction data by applying more time-consuming experimental methods such as multiple isomorphous replacement (MIR) and multiple-wavelength anomalous diffraction (MAD), it often turns out that the 'new' structure exhibits an important conformational change with respect to the original template, explaining *a posteriori* the failure of the MR attempt. *A priori* modelling of the most likely conformational changes that a given template might undergo would thus be of significant benefit and could eventually allow an increased number of crystal structures to be solved by MR.

Here, we propose normal-mode analysis (NMA) as a powerful tool for anticipating the most likely conformational changes of a given template and to screen its perturbed structures for possible MR solutions. The rationale behind this approach is as follows. It was noticed almost 20 years ago, using empirical force fields and a protein description at the atomic level, that one of the largest amplitude motions predicted by normal-mode theory for proteins, that is one of the lowest-frequency normal modes of motion, often compares well with their functional conformational change as observed by crystallographers upon ligand binding (Harrison, 1984; Brooks & Karplus, 1985). More recently, using much more simplified protein descriptions, namely elastic network models (Tirion, 1996; Bahar *et al.*, 1997; Hinsen, 1998), it was shown that of 3800 known protein motions more than half can be described well by perturbing the considered protein along the direction of at most two low-frequency normal modes (Krebs *et al.*, 2002), that is, by displacing the structure along two corresponding perpendicular directions of the configurational space. Moreover, when the collective character of the protein motion is obvious, a single low-frequency normal mode often proves to be sufficient and it is usually one of the three lowest-frequency modes (Tama & Sanejouand, 2001; Delarue & Sanejouand, 2002). Such results strongly suggest

Table 1
Overview of the crystallographic parameters, molecular-replacement results and refinement statistics.

	Maltodextrin-binding protein	HIV-1 protease	Glutamine-binding protein
Target	1omp	1ajx	1wdn
Template	1anf	1hhp	1ggg
No. residues	370	99	224
No. reflections†	5851	4280	3796
Completeness‡ (%)	97.9	78.9	97.1
Space group	$P1$	$P2_12_12_1$	$P2_12_12_1$
No. molecules	1	2	1
Best mode	Mode 7	Mode 11	Modes 7 + 8
Perturbation§	180	60	200 + 40
CC/R factor¶			
Target	86.0/20.9	73.2/31.3	72.1/29.9
Template	25.8/49.0	31.0/50.0	26.1/50.3
Best NMA	33.5/46.4	57.7/39.7	21.3/51.6
R_{work}/R_{free}††			
Target	16.5/23.3	35.4/38.6	25.9/35.9
Template	43.0/51.1	52.8/54.1	45.3/54.0
Best NMA	38.8/45.3	41.8/46.0	35.2/47.1

† Number of reflection theoretically available to a resolution of 3.2 Å. ‡ Completeness to a resolution of 3.2 Å. § Arbitrary units. Using Tirion's elastic network model, normal-mode frequencies as well as the corresponding unit for the displacements along a normal mode are defined through a scaling free factor, which was set to $k = 10$ in the present study. ¶ The CC and R factor of the best translation/rotation solution(s) found by *AMoRe* (Navaza, 1994) when using data to 3.2 Å resolution. †† Final R factor for the working and the test set after *CNS* (Brünger *et al.*, 1998) refinement using standard parameters to 3.2 Å resolution.

that protein movements between open and closed forms (*e.g.* without and with ligand) may actually be under selective pressure to follow mainly one or a few low-frequency normal modes of the protein.

This suggests a screening approach in which a given template is perturbed with varying amplitudes in the direction of one or two of its low-frequency normal modes. One would then look for minima in the resulting R values after standard MR, optionally followed by a simulated-annealing and/or an energy-minimization refinement step. To evaluate the potential of this idea, which has been considered once before in the case of the determination of the low-resolution structure of F-actin (Tirion *et al.*, 1995), we selected proteins for which the structures are known in two different conformations (Echols *et al.*, 2003) and for which structure factors have been deposited in the Protein Data Bank (PDB; Berman *et al.*, 2000). Proteins with two distinct domains connected by a single linker peptide (such as the diphtheria toxin or immunoglobin) were excluded, as the standard MR procedure would be to attempt a two-body search in these cases. Also omitted are proteins with only small conformational changes as these would most likely be of no challenge in MR. The idea is then to use one of the two protein models as an MR template in order to solve the structure of the other protein. Here, we present results for three representative cases in which the original template failed to give a proper solution but the perturbed model did: maltodextrin-binding protein (Zanotti *et al.*, 1992), HIV-1 protease (Backbro *et al.*, 1997) and glutamine-binding protein (Hsiao *et al.*, 1996). Although we limit this analysis to cases with 100% sequence homology for the sake of simplicity, the ideas presented in this paper can readily be applied to the more general situation where the sequences of the template and target are different. In that case a number of different protocols can be applied, including using all-alanine or homology-based models in different stages of the procedure.

2. Methods

Protein models (targets and templates) and the corresponding structure factors were downloaded from the Protein Data Bank. Targets were orientated onto the templates using rigid-body superposition. All atomic B factors were set to a value of 10 Å^2 prior to MR. The MR search was performed using the stand-alone version of *AMoRe* in fully automatic mode and data to a resolution of 3.2 Å (Navaza, 1994). Based on the best rotation/translation function(s) found by *AMoRe*, the models were refined to 3.2 Å resolution using the standard protocol implemented in *CNS*: reorientation with the 'realspace_transform.inp' script followed by initial B-factor and bulk-solvent corrections and two cycles of simulated annealing (only applied for the glutamine-binding protein, as discussed below) and coordinate and individual B-factor minimization (constrained torsion-angle dynamics) as implemented in the 'refine.inp' script. All parameters of these script were kept to their default values (Brünger *et al.*, 1998). Root-mean-square deviations (r.m.s.d.s) between the C$^\alpha$ atoms of two proteins and the fraction of C$^\alpha$ atoms that are closer than 3 Å were computed using *LSQMAN* from the *DEJAVU* package (Kleywegt, 1996). Normal modes were computed based on ideas from Tirion (1996) as implemented by Tama & Sanejouand (2001) using the RTB approach (Durand *et al.*, 1994; Tama *et al.*, 2000). All calculations were performed in all-atom mode applying a 5 Å cutoff in the definition of the elastic interactions. A web interface to these tools for automatic computation of NMA-perturbed structures for MR (including the cases presented here as pre-computed examples) is available at http://igs-server.cnrs-mrs.fr/elnemo/index.html. The Fortran source code of the software is available at http://ecole.modelization.free.fr/modes.html. Normal modes are numbered by increasing frequency values, the first six zero-frequency modes being the three trivial rigid-body translational and rotational modes. The lowest non-trivial normal mode is thus mode 7.

3. Results

The maltodextrin-binding protein is a prime example for the application of NMA to MR (see Table 1 for details). It is constituted of two domains of about equal size. Upon ligand binding, the maltodextrin-binding protein undergoes a hinge movement from its open towards its closed conformation. We assume the open form to be our MR target as structure factors are only available for this model. Rigid-body superposition between template and target allows fitting of the larger of the two domains, which corresponds to 59% of the residues. The r.m.s.d. between the C$^\alpha$ atoms belonging to this domain is 0.98 Å. An MR search using the closed form as a template allowed the identification of the correct rotation/translation function, but even after a refinement step (energy minimization) the final free R factor of the resulting model remained at the quite high value of $R = 51.1\%$. This is because of the unmatched second domain of the protein. By perturbing the structure of the protein, following the lowest-frequency mode (mode 7), we find that this movement captures the conformational change quite well. Rigid-body superposition of the target and the perturbed template shows that 91% of the residues of the perturbed template have a r.m.s.d. of 1.2 Å with respect to the C$^\alpha$ atoms of the target. The total r.m.s.d. for all C$^\alpha$ atoms is 1.4 Å, compared with 3.8 Å in the case of the unperturbed template. Application of standard MR and refinement with the optimal

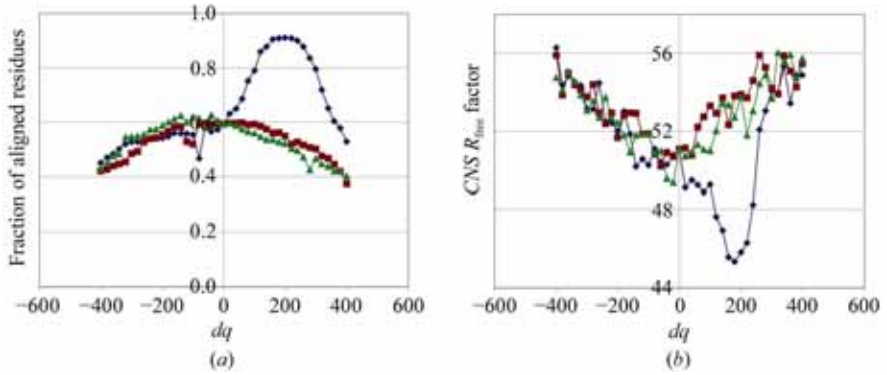

Figure 1
Maltodextrin-binding protein. Fraction of aligned residues, using *LSQMAN* and a cutoff distance of 3 Å (*a*) and *CNS* final free *R* factor (*b*), as a function of the perturbation *dq* (in arbitrary units) of the template along modes 7 (blue circles), 8 (red squares) and 9 (green triangles).

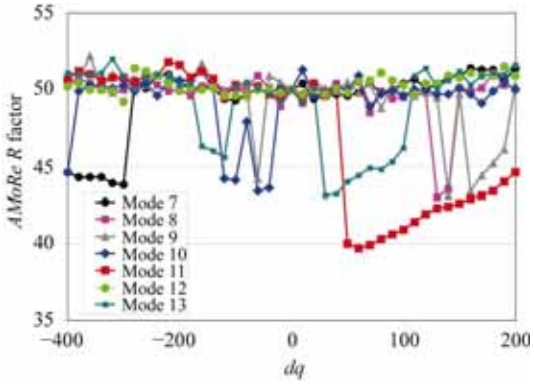

Figure 2
HIV-1 protease. *AMoRe R* factor as a function of the perturbation (in arbitrary units) of the template along one of its seven lowest-frequency normal modes.

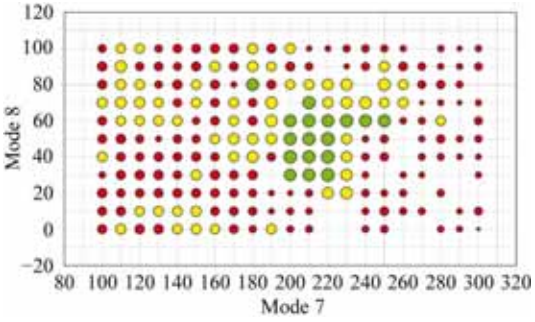

Figure 3
Glutamine-binding protein. *CNS* final free *R* factor as a function of the perturbation along modes 7 and 8; circle size is proportional to 65 − *R* factor; colour code: *R* factor better than 50, green; better than 55, yellow; other, red; missing circles in the screened range (100–300 arbitrary units for mode 7; 0–100 for mode 8; step size 10) indicate that *CNS* failed to refine the corresponding model.

perturbed model yields a final free *R* factor of 45.3%. An animated view of the predicted protein movement following its lowest-frequency mode is presented online (Suhre, 2004). A full analysis of the observed conformational changes between its opened and closed state is available *via* the molmovdb website (http://molmovdb.org).

This corresponds to a gain of almost 6% in the free *R* factor. Visual inspection of the corresponding electron-density map and further refinement to higher resolution shows that here a solution for the entire protein was found, whereas only 60% of the protein could be modelled using the original template. In Fig. 1, the result of the search for MR solutions is shown in which one of the three lowest normal modes (modes 7, 8 and 9) is used to apply the perturbation. In this case, the free *R* factor depends continuously on the perturbation and the optimum can already be identified after the MR step. Note that modes 8 and 9 do not yield any MR solution better than the original template.

Our second example, HIV-1 protease, pictures a case where two molecules per asymmetric unit have to be placed by MR. This is a relatively small protein with less than 100 residues and the conformational change between the two forms mainly concerns a single antiparallel β-sheet (see Suhre, 2004). Using one molecule of the original template, corresponding to the open non-liganded form of the protein, the correct position of the second molecule, with both molecules of the target being in the closed form, could not be found. Unsurprisingly, the final free *R* factor obtained with such a model is quite high, *R* = 55.1%. Application of an appropriate normal-mode perturbation along mode 11 eventually allows the identification of a suitable rotation/translation function which results in a sudden decrease of the *R* factor, the best model having a final free *R* factor of *R* = 46.0% (Fig. 2 and Table 1). Visual inspection of the corresponding electron-density maps shows that in this case NMA clearly yields a solution, whereas this is not the case when the original template is used. Note that some of the other modes also yield 'good' solutions.

As a third example, we chose glutamine-binding protein. This protein undergoes extensive conformational changes (hinge movement) between its open and closed forms. Rigid-body superposition of both models allows the fitting of a large domain that contains about 60% of the residues (134 of 224) with an r.m.s.d. of 0.8 Å, while the r.m.s.d. between all C^{α} atoms is as large as 5.3 Å. An MR attempt using the open form as a template to solve the structure of the closed form failed with a free *R* factor after *CNS* refinement (simulated annealing and energy minimization) of 54.0%. Similarly, screening for an MR solution using NMA perturbation along only one low-frequency mode also failed. Eventually, a bimodal scan yielded a solution when combining perturbations in the direction of the two lowest normal modes, the best model having a final free *R* factor of 47.1% (Fig. 3 and Table 1). The r.m.s.d. between all C^{α} atoms of this highest-scoring perturbed template and the target is 2.1 Å and more than 90% (202) of the residues have an r.m.s.d. of 1.6 Å to the target. However, in this case a simulated-annealing step was necessary for convergence of the refinement, because no optimal model was found at the level of the MR step, which is at variance with what was observed with our two previous examples. This particular bimodal screen took about 13 h of CPU time to complete on an Athlon 2400+ processor (MR with *AMoRe* and *CNS* refinement of 231 models).

Finally, we should also briefly mention two cases where the use of NMA for MR is not that efficient. In order to obtain better results with a perturbed model than with the original template to solve the structure of the closed form of citrate synthase (437 residues; PDB code 6csc) with its open form (PDB code 5csc) as a template, we had to screen three low-frequency normal modes; this yielded only a slightly improved free *R* factor of 44.5% compared with 45.2% obtained with the original template. However, as the original template already yields a 'good' solution, this may not be such a surprising result. Attempts to apply the same approach to the quite large lactoferrin protein (691 residues; PDB code 1cb6, open form, used as target; 1lcf, closed form, used as template) failed. In fact, successive

application of the optimal perturbation (computed by projecting the difference vector between the open and closed forms onto the normal modes) along the first ten normal modes steadily decreases the r.m.s.d. between all C^{α} atoms, but in the process the final free R factor goes through a maximum of R = 55.3% when the fourth mode is added, reaching a value of R = 45.4% only when the eleventh mode is also taken into account. If we compare this to the free R factor of R = 47.0% obtained with the unperturbed template, we conclude that lactoferrin falls into the class of cases where more than two low-frequency modes are needed in order to anticipate its conformational change.

4. Discussion

Molecular replacement is the most cost-effective method for solving the three-dimensional structure of a new protein by X-ray crystallography and it is thus the method of choice for structure determination. However, MR may fail even in cases of high sequence homology when conformational changes, *e.g.* arising from ligand binding or different crystallogenetic conditions, come into play (by failure of the MR approach we mean that no refinable model can be found and that additional experimental techniques are required to achieve phasing of the diffraction data). Here, we demonstrate the potential of normal-mode analysis as an extension to MR that allows recovery from such drawbacks. We have provided three examples where application of standard MR protocol did not allow solution of the structure of a protein in one conformation (open or closed), whereas using a template perturbed following one or two low-frequency modes allowed lowering of the final free R factor below the 'noise' limit of about 50% (also confirmed by visual inspection of the resulting electron-density maps and further refinement to higher resolution). Although we limited our analysis to cases where template and target have 100% sequence identity, this approach should also be applicable to templates with much lower sequence similarity. Ideally, one would then start by building a homology model, *e.g.* using programs such as *MODELLER* (Sali & Blundell, 1993). Such a protocol is already implemented in *CCP*4 (Collaborative Computational Project, Number 4, 1994), so that an extension of *CCP*4 including normal-mode perturbation between the homology-modelling step and MR can be envisaged. Note also that NMA has already been used in order to refine temperature factors (Diamond, 1990; Kidera *et al.*, 1992) based on the fact that atomic fluctuations computed by NMA are often found to be well correlated with crystallographic B factors (Bahar *et al.*, 1997). Therefore, adding B-factor values predicted by NMA to the perturbed templates could also prove useful.

Overall, our approach can be viewed as perturbing the original template structure in different (but still physically meaningful) directions until one of the new models comes close enough to the searched structure to identify an MR solution. Even in cases where an MR can be found, this method can be of interest to further improve the starting model for refinement, eventually reducing the time spent on manual construction. This should be particularly true when working with low-resolution data sets. Jones (2001) has discussed the potential of such fold-recognition models for MR, but without referring explicitly to normal-mode perturbations, as presented here. As previously mentioned, an increase in MR success rates would be especially valuable in the context of the ongoing high-throughput structural genomics projects (Rupp *et al.*, 2002).

It has been shown recently (Krebs *et al.*, 2002) that half of the known protein movements can be modelled by displacing the studied structure using at most two low-frequency normal modes. A screening procedure following ideas presented in this paper would be highly efficient and parallelizable on a cluster of Linux PCs. Thus, NMA analysis may prove able to break tough MR problems in up to 50% of cases.

Protein models and structure factors were obtained from the Protein Data Bank (PDB). The authors wish to thank J. Navaza for access to the latest version of his *AMoRe* program. We acknowledge free access to the *CNS* (Brünger *et al.*, 1998) and *DEJAVU* (Kleywegt, 1996) software packages. We are grateful to C. Abergel for her interest in this work and helpful discussions.

References

Backbro, K., Lowgren, S., Osterlund, K., Atepo, J., Unge, T., Hulten, J., Bonham, N. M., Schaal, W., Karlen, A. & Hallberg, A. (1997). *J. Med. Chem.* **40**, 898–902.
Bahar, I., Atilgan, A. R. & Erman, B. (1997). *Fold Des.* **2**, 173–181.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Brooks, B. & Karplus, M. (1985). *Proc. Natl Acad. Sci. USA*, **82**, 4995–4999.
Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Delarue, M. & Sanejouand, Y. H. (2002). *J. Mol. Biol.* **320**, 1011–1024.
Diamond, R. (1990). *Acta Cryst.* A**46**, 425–435.
Durand, P., Trinquier, G. & Sanejouand, Y. H. (1994). *Biopolymers*, **34**, 759–771.
Echols, N., Milburn, D. & Gerstein, M. (2003). *Nucleic Acids Res.* **31**, 478–482.
Harrison, R. W. (1984). *Biopolymers*, **23**, 2943–2949.
Hinsen, K. (1998). *Proteins*, **33**, 417–429.
Hsiao, C. D., Sun, Y. J., Rose, J. & Wang, B.-C. (1996). *J. Mol. Biol.* **262**, 225–242.
Jones, D. T. (2001). *Acta Cryst.* D**57**, 1428–1434.
Kidera, A., Inaka, K., Matsushima, M. & Go, N. (1992). *J. Mol. Biol.* **225**, 477–486.
Kleywegt, G. J. (1996). *Acta Cryst.* D**52**, 842–857.
Krebs, W. G., Alexandrov, V., Wilson, C. A., Echols, N., Yu, H. & Gerstein, M. (2002). *Proteins*, **48**, 682–695.
Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.
Rupp, B., Segelke, B. W., Krupka, H. I., Lekin, T., Schafer, J., Zemla, A., Toppani, D., Snell, G. & Earnest, T. (2002). *Acta Cryst.* D**58**, 1514–1518.
Sali, A. & Blundell, T. L. (1993). *J. Mol. Biol.* **234**, 779–815.
Suhre, K. (2004). *ElNémo Examples.* http://igs-server.cns-mrs.fr/elnemo/examples.html.
Tama, F., Gadea, F. X., Marques, O. & Sanejouand, Y. H. (2000). *Proteins*, **41**, 1–7.
Tama, F. & Sanejouand, Y. H. (2001). *Protein Eng.* **14**, 1–6.
Tirion, M. (1996). *Phys. Rev. Lett.* **77**, 1905–1908.
Tirion, M., ben-Avraham, D., Lorenz, M. & Holmes, K. C. (1995). *Biophys. J.* **68**, 5–12.
Zanotti, G., Scapin, G., Spadon, P., Veerkamp, J. H. & Sacchettini, J. C. (1992). *J. Biol. Chem.* **267**, 18541–18550.

6.7 Article fourni « Serveur web ElNémo »

Suhre, K. and Sanejouand, Y.H. (2004) ElNemo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Res*, **32**, W610-W614.

W610–W614 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh368

ElNémo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement

Karsten Suhre* and Yves-Henri Sanejouand[1]

Information Génomique & Structurale (UPR CNRS 2589), 31, chemin Joseph Aiguier, 13402 Marseille Cedex 20, France and [1]Laboratoire de Physique, Ecole Normale Supérieure, 46, allées d'Italie, 69364 Lyon Cedex 07, France

Received February 6, 2004; Revised and Accepted March 9, 2004

ABSTRACT

Normal mode analysis (NMA) is a powerful tool for predicting the possible movements of a given macromolecule. It has been shown recently that half of the known protein movements can be modelled by using at most two low-frequency normal modes. Applications of NMA cover wide areas of structural biology, such as the study of protein conformational changes upon ligand binding, membrane channel opening and closure, potential movements of the ribosome, and viral capsid maturation. Another, newly emerging field of NMA is related to protein structure determination by X-ray crystallography, where normal mode perturbed models are used as templates for diffraction data phasing through molecular replacement (MR). Here we present *ElNémo*, a web interface to the *El*astic *Ne*twork *Mo*del that provides a fast and simple tool to compute, visualize and analyse low-frequency normal modes of large macro-molecules and to generate a large number of different starting models for use in MR. Due to the 'rotation-translation-block' (RTB) approximation implemented in *ElNémo*, there is virtually no upper limit to the size of the proteins that can be treated. Upon input of a protein structure in Protein Data Bank (PDB) format, *ElNémo* computes its 100 lowest-frequency modes and produces a comprehensive set of descriptive parameters and visualizations, such as the degree of collectivity of movement, residue mean square displacements, distance fluctuation maps, and the correlation between observed and normal-mode-derived atomic displacement parameters (B-factors). Any number of normal mode perturbed models for MR can be generated for download. If two conformations of the same (or a homologous) protein are available, *ElNémo* identifies the normal modes that contribute most to the corresponding protein movement. The web server can be freely accessed at http://igs-server.cnrs-mrs.fr/elnemo/index.html.

INTRODUCTION

One of the best suited theoretical methods for studying collective motions in macromolecules is normal mode analysis (NMA), which leads to the expression of protein dynamics in terms of a superposition of collective variables, namely, the normal mode coordinates [see (1) for a review]. Though the first normal mode studies were performed as early as 20 years ago (2,3), they remained restricted to small-size proteins until more recently, when methodological advances (4–8), simplified protein descriptions (9–11), and ever faster computer systems allowed them to address increasingly large macromolecular systems, up to entire protein complexes, including the entire ribosome (12–14).

Noteworthy is that by analysing more than 3800 known protein motions, Krebs *et al*. (15) have shown that more than half of them can be approximated by applying a perturbation in the direction of at most two low-frequency normal modes of the considered protein. Moreover, when the collective character of the protein motion is obvious, a single low-frequency normal mode often proves to be enough, and it is usually one of the three lowest-frequency ones (12,13). Such results strongly suggest that protein movements between open and closed forms (e.g. with and without ligand) may actually be under selective pressure, so as to follow mainly one, or a few, low-frequency normal modes of the protein. In other words, amino-acid sequences may have evolved so that low-energy barriers are found when the protein is displaced along the few corresponding normal mode coordinates.

*To whom correspondence should be addressed. Tel: +33491164604; Fax: +33491164549; Email: karsten.suhre@igs.cnrs-mrs.fr

One major application of normal modes is the identification of potential conformational changes, e.g. of enzymes upon ligand binding (7,12,13). The method has also been used recently in the study of membrane channel opening (16), the analysis of structural movements of the ribosome (14), viral capsid maturation (17), transconformations of the SERCA1 Ca-ATPase (8,18), tertiary and quaternary conformational changes in aspartate transcarbamylase (19) and the analysis of domain motions in large proteins in general (11,20). NMA is most often used in order to try to guess what kind of conformational change a protein undergoes in order to fulfil its function, by analysing its lowest-frequency modes one after the other. It can also be used to check if a conformational change proposed on the basis of non-structural experimental data is likely to occur or not, as recently done in the case of membrane channel opening (16). As a tool able to predict large-amplitude motions, it has been suggested that it has the potential to improve the resolution of the final reconstructions of single particles from electron cryomicroscopy (21). Moreover, the fact that 50% of the observed protein movements can be accurately described by only one or two low-frequency normal modes prompts for an application of NMA in X-ray crystallography data phasing, i.e., to use normal mode perturbed models as templates in molecular replacement. We have shown that this approach allows to break difficult phasing problems where the original unperturbed template fails to yield a usable solution (22). NMA thus represents a powerful tool for a wide range of applications in structural biology and X-ray crystallography. We designed *ElNémo* as a comprehensive, but still easy-to-use interface for NMA. Particular emphasis was put on its ability to handle large protein systems with 500–1000 or more residues in an all-atom level of description, having in mind the generation of a great number of normal mode perturbed models as templates for MR.

METHODS

The details of NMA have been described elsewhere (7,16). Here we summarize the basic principles of the computations that are performed by *ElNémo*.

Normal mode calculation is based on the harmonic approximation of the potential energy function around a minimum energy conformation. This approximation allows the analytic solution of the equations of motion by diagonalizing the Hessian matrix (the mass-weighted second derivatives of the potential energy matrix). The eigenvectors of this matrix are the normal modes, and the eigenvalues are the squares of the associated frequencies. The protein movement can be represented as a superposition of normal modes, fluctuating around a minimum energy conformation. For proteins, the normal modes responsible for most of the amplitude of the atomic displacement are associated to the lowest frequencies. In order to avoid time-consuming energy minimizations, as well as the corresponding drift of the studied structure, a single-parameter Hookean potential is used, which was shown to yield low-frequency normal modes as accurate as those obtained with more detailed, empirical, force fields (9):

$$E_p = \sum_{d_{ij}^0 < R_c} c\left(d_{ij} - d_{ij}^0\right)^2$$

where d_{ij} is the distance between two atoms i and j, d_{ij}^0 is the distance between the atoms in the three-dimensional structure, c is the spring constant of the Hookean potential (assumed to be the same for all interacting pairs) and R_c is an arbitrary cut-off, beyond which interactions are not taken into account (the *ElNémo* default cut-off is 8 Å, but the user is free to change this setting; values of 10–13 Å are often used when only C-alpha atoms are taken into account). This approximation implies that the reference structure represents the minimum energy conformation. Moreover, all atom masses are set to the same fixed value in the kinetic energy term, as this approximation was shown to have little influence on the low-frequency modes. Therefore, only normalized frequencies are reported, the lowest non-trivial frequency being set to one. Note that there are always six zero frequencies (corresponding to the three overall rotations and three overall translations of the system), but more than six can be obtained if a group of atoms is at a distance larger than R_c from the others.

The building block approximation, also named RTB (for 'rotation-translation-block'), groups several residues into a single super-residue, the rigid-body rotations and translations of the super-residues being used as a set of new coordinates instead of the Cartesian ones (5). Tama *et al*. (7) have shown that this approximation has very little influence on the low-frequency modes. Due to this approximation it becomes possible to treat very large proteins in an all-atom level of description in reasonable computing time (*ElNémo* automatically determines the number of residues to be grouped together based on the number of residues in the protein, but the user may override this setting; for small proteins, each block contains a single residue). Note that for larger and larger proteins, the size of the domains involved in functional motions is expected to grow. Thus, when the system is large, the grouping of several residues into a single block is expected to have little impact on the lowest-frequency modes, which depend mainly on the overall shape of the system. Indeed, they can be captured at extremely high levels of coarse-graining (23) or by using low-resolution structural data (21).

Normal mode perturbed models are structural models in PDB (Protein Data Bank) format that correspond to the original reference structure, with a perturbation proportional to the corresponding normal mode applied to every atom. As normal modes define only the direction, but not the amplitude, of the conformational change of a protein, the user can specify a range of amplitudes (in arbitrary units) that will be used in the computation. Note that, due to the approximation of the atom–atom interactions by a harmonic potential, application of too large amplitudes may yield distorted structures and result in sterical clashes. However, it is necessary to specify amplitudes larger than those allowing for a fair comparison with B-factors, because the latter reflect atomic motions at room temperature, while in the present context the purpose of normal mode perturbation is to capture much larger amplitude motions.

Distance fluctuation maps highlight residue pairs i and j with the strongest variation in the distance between their C-alpha atoms in a given mode. In *ElNémo*, the top ranking (10%) distance fluctuations are coloured in blue (increase) and red (decrease). Flexible and rigid blocks, as well as their relative movements can be easily identified in such maps.

The degree of collectivity indicates the fraction of residues that are significantly affected by a given mode. For maximal collective movements the degree of collectivity tends to be a value of one, whereas for localized motions, where the normal mode movement only involves few atoms, the degree of collectivity approaches zero. While low-frequency normal modes are expected to have collective characters, especially those related to functional conformational changes of proteins (12), computed ones sometimes happen to be localized. In such cases, they correspond to motions of some extended parts of the system, as often observed in crystallographic protein structures for N- or C-termini. These motions are usually meaningless and can be ignored, though it is common practice, and probably safer, to remove such extended parts prior to the normal mode computation.

The overlap measures the degree of similarity between the direction of a chosen conformational change and the direction of a given normal mode. A conformational change is here defined by the difference vector between the reference structure and a second conformation of the same protein or that of a close homologue. *ElNémo* reports cumulative values for the square of the overlap, starting with the lowest-frequency non-trivial normal mode. Note that, because the normal modes form a basis, this cumulative sum reaches a value of one when it is computed over all modes. If the considered conformational change has a collective character, the cumulative sum usually reaches a value of 0.7–0.8 already within the 20–50 lowest-frequency modes (13,24). What makes NMA useful for predicting protein movements is the fact that in a large number of cases, one or two low-frequency normal modes, i.e. those with the highest overlap are enough for providing a fair description of the conformational change (1,12).

B-factors are computed from the mean square displacement $\langle R^2 \rangle$ of the first 100 lowest-frequency normal modes using the relationship $B = (8\pi^2/3)\langle R^2 \rangle$ and linear scaling to the observed B-factors in the reference structure as described in (7). Correlations between NMA and crystallographic B-factors are usually found to be >0.5–0.6 (13), while values >0.8 have been reported (1). Adjusting R_c can slightly improve such correlations. This probably reflects the fact that modifying R_c affects low-frequency densities (9). The comparison between computed and observed crystallographic B-factors provides a measure of how well the protein's flexibility in its crystal environment is described by the normal modes.

Root mean square distances (RMSD) between the normal mode perturbed models and a second (not necessarily sequence-identical) structure are computed by a rigid body superposition using the *lsqman* software (25). Reported are the RMSD between all C-alpha atoms of the two protein conformations, the number of C-alpha atoms that are closer than 3 Å in the rigid body superposition and the RMSD between those atoms only. These numbers can be used as a proxy for the overlap in the case of not 100% sequence-identical proteins.

USING THE WEB INTERFACE

The principal input to *ElNémo* is a protein model in PDB format (26). A numerical FORTRAN code, which is the heart of *ElNémo*, determines the corresponding interaction matrix for the elastic network model and computes its 100 largest eigenvalues and their eigenvectors (the normal modes). For each mode, its degree of collectivity of movement and the mean square displacement of all residues is output. The user may select the number of low-frequency modes for which normal mode perturbed models will be computed, specifying an amplitude range and increment (DQMIN, DQMAX, DQSTEP). The automatic generation of three-dimensional animated views of these modes from three different viewpoints (using Molscript; 27) can be requested. Distance fluctuation maps are also made available for all normal mode perturbed models. B-factors are derived from the mean square displacements of all atoms in the 100 lowest-frequency modes. When a second conformation of the same protein is submitted, *ElNémo* computes the degree of collectivity of motion for all normal modes and reports the contribution of each of the 100 lowest-frequency modes to the conformational change (amplitude). This option requires that both models have the same number of atoms and that the residues are numbered identically. If only a homologue of the reference protein (<100% sequence identity) is available in a different conformation, *ElNémo* computes the RMSD between the normal mode perturbed models and the homologous structure in order to identify the normal mode perturbations that best describe the associated protein movement (Figure 1).

Although *ElNémo* will probably produce useful results when using original (unprocessed) PDB files, some modifications of the input data are advisable. Note that *ElNémo* only reads the 'ATOM' record from the PDB file, and that water residues are ignored. Therefore, molecules that are coded as 'HETATM' need to be changed to 'ATOM' if they are to be accounted for in the normal mode calculation. Examples are seleno-methionine residues, haeme groups, nucleic acids and calcium ions. However, this conversion is not done automatically by *ElNémo* to avoid inclusion of crystallogenic agents, such as 2-methyl-2,4-pentanediol (MPD) and Tris, that are unrelated to the protein in its real environment. To prevent lumping of residues that are part of separate molecules into one RTB super-residue, different chain identifiers should be used. Alternate conformations and hydrogen atoms should be erased, as their presence will have only a minor influence on the results. More specific hints on how to best prepare an *ElNémo* job can be found on the help page.

The computation of the normal modes for small- to medium-size proteins (100–400 residues) in an all-atom level of description takes between 10 and 30 min when using default settings. Even very large proteins, such as the capsid-like protein lumazine synthase complex and the entire ribosome take no longer than some hours to complete (both molecules are constituted of more than 9000 residues). Preparation of additional normal mode perturbed models and animated views takes between 3 and 10 min per mode to complete for small- to medium-size proteins. However, this part of the computation becomes the most time-consuming part of an *ElNémo* run for larger proteins. Therefore, the user is advised to limit the number of normal mode perturbed models and visualizations in these cases to the first three to five modes at this stage and to request additional models and animations once the initial job is completed based on an analysis of the degree of collectivity of movement of the different modes. At that time, models that are perturbed using two

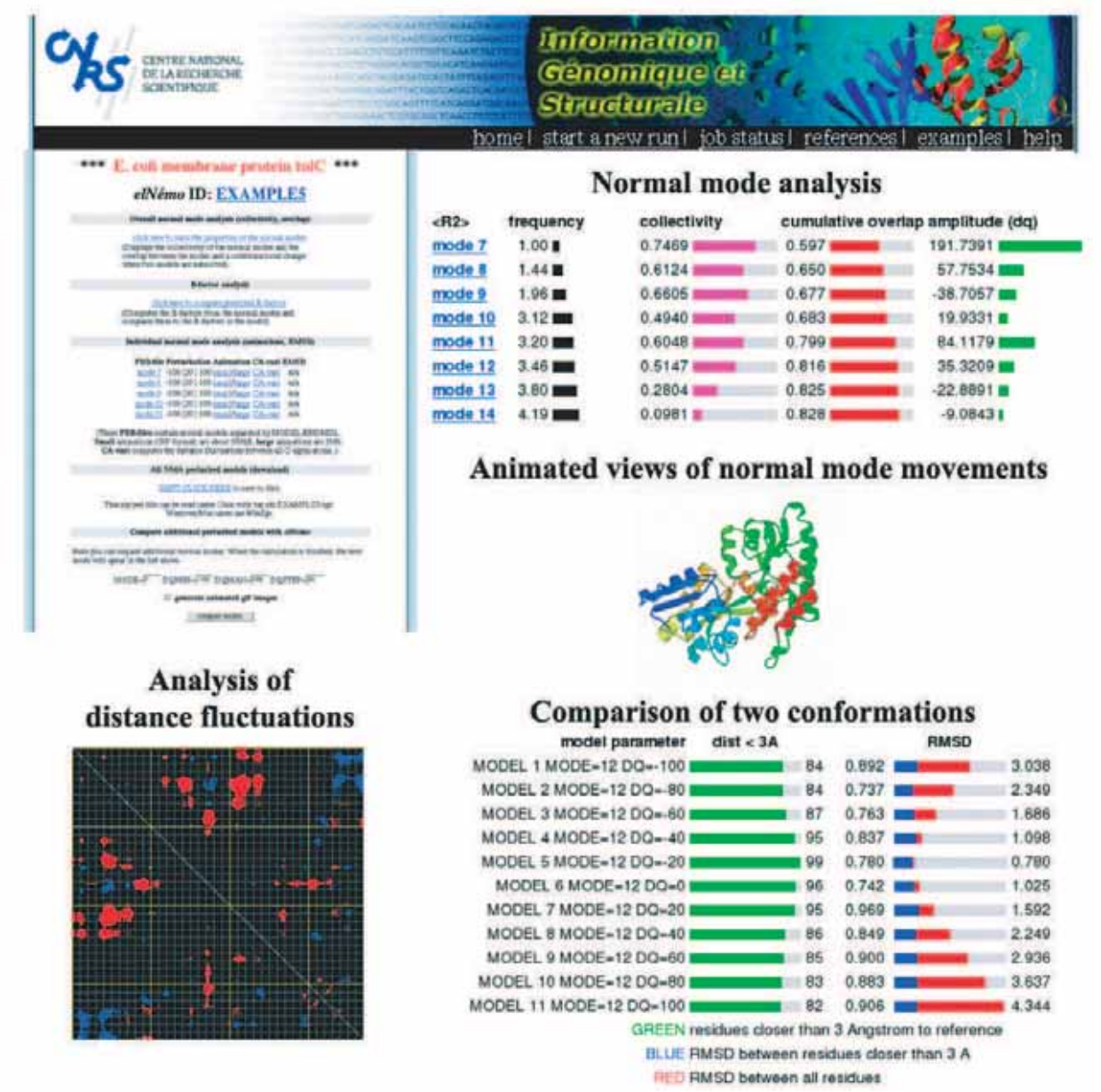

Figure 1. An example of a typical *ElNémo* output that is available for every run through the result page (top left). The normal mode analysis page (top right) displays the different properties of the first 100 lowest-frequency modes, i.e. their frequency, degree of collectivity of movement, mean square displacement $\langle R^2 \rangle$, overlap (if two conformations are available) and its corresponding amplitude. Three-dimensional animations from three orthogonal viewpoints are available in large and small sizes. Comparison of a normal mode perturbed structure and a second conformation in terms of RMSD and number of residues that are closer than 3 Å can be done (bottom right). Analysis of distance fluctuations between all CA atoms is presented in the form of a cross-plot, where red and blue dots indicate those residues for which the pairwise distance changes most significantly in the movement defined by a given mode. The result page also allows submission of normal mode calculations for new modes with varying amplitude ranges. The resulting normal mode perturbed models in PDB format can be downloaded for further processing (e.g. using VMD (28) to visualize the protein movements as presented on the *ElNémo* example page) or as templates for MR.

normal modes at the same time can also be requested (i.e. for use as MR templates). *ElNémo* jobs are processed in a linear batch queue on a first-come first-served basis. Notification of the user by email about the job status is available. A second, independent queue is used for the computation of additional modes. As walk-through examples, some of our recent applications of NMA to conformational changes of membrane channel proteins and to molecular replacement are presented in the example section of the *ElNémo* web server. The corresponding input and output is available under the job id EXAMPLE-n on the job-status page, so that the interested user can analyse and re-run the corresponding jobs.

FINAL REMARKS

NMA is a powerful tool for the study of protein movements and conformational changes, proven by the ever-growing

range of applications in different structural biology domains and more recently in X-ray crystallography as a source of templates for molecular replacement. Methodological and technical advances, i.e. the elastic network model and the RTB approximations, make NMA of whole viruses and of the entire ribosome possible. The *ElNémo* web server reflects the long-standing experience in this domain of the second author (Y.H.S.) and his co-workers. It makes the respective tools available to a wide community of potential NMA users, without exposing them to the need to gain in-depth understanding of the technique and its implementations. The availability of a comprehensive and easy-to-use dedicated NMA web server like *ElNémo* will therefore facilitate an even more widespread application of this interesting technique.

ACKNOWLEDGEMENTS

Molscript (27) and LSQMAN from the DéjàVu package (25) are used by the server. VMD (28) was used to generate some of the examples.

REFERENCES

1. Tama,F. (2003) Normal mode analysis with simplified models to investigate the global dynamics of biological systems. *Protein Pept. Lett.*, **10**, 119–132.
2. Go,N., Noguti,T. and Nishikawa,T. (1983) Dynamics of a small globular protein in terms of low-frequency vibrational modes. *Proc. Natl Acad. Sci. USA*, **80**, 3696–3700.
3. Brooks,B. and Karplus,M. (1983) Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. *Proc. Natl Acad. Sci. USA*, **80**, 6571–6575.
4. Mouawad,L. and Perahia,D. (1993) DIMB: diagonalization in a mixed basis: a method to compute low-frequency normal modes for large macromolecules. *Biopolymers*, **33**, 569–611.
5. Durand,P., Trinquier,G., and Sanejouand,Y.H. (1994) A new approach for determining low-frequency normal modes in macromolecules. *Biopolymers*, **34**, 759.
6. Marques,O. and Sanejouand,Y.H. (1995) Hinge-bending motion in citrate synthase arising from normal mode calculations. *Proteins*, **23**, 557–560.
7. Tama,F., Gadea,F.X., Marques,O. and Sanejouand,Y.H. (2000) Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins Struct. Funct. Genet.*, **41**, 1–7.
8. Li,G. and Cui,Q. (2002) A coarse-grained normal mode approach for macromolecules: an efficient implementation and application to Ca(2+)-ATPase. *Biophys. J.*, **83**, 2457–2474.
9. Tirion,M.M. (1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, **77**, 1905–1908.
10. Bahar,I., Atilgan,A.R. and Erman,B. (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des.*, **2**, 173–181.
11. Hinsen,K. (1998) Analysis of domain motions by approximate normal mode calculations. *Proteins*, **33**, 417–429.
12. Tama,F. and Sanejouand,Y.H. (2001) Conformational change of proteins arising from normal mode calculations. *Protein Eng.*, **14**, 1–6.
13. Delarue,M. and Sanejouand,Y.H. (2002) Simplified normal mode analysis of conformational transitions in DNA-dependant polymerases: the Elastic Network Model. *J. Mol. Biol.*, **320**, 1011–1024.
14. Tama,F., Valle,M., Frank,J. and Brooks,C.L. III (2003) Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy. *Proc. Natl Acad. Sci. USA*, **100**, 9319–9323.
15. Krebs,W.G., Alexandrov,V., Wilson,C.A., Echols,N., Yu,H. and Gerstein,M. (2002) Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. *Proteins*, **48**, 682–695.
16. Valadie,H., Lacapcre,J.J., Sanejouand,Y.H. and Etchebest,C. (2003) Dynamical properties of the MscL of *Escherichia coli*: a normal mode analysis. *J. Mol. Biol.*, **332**, 657–674.
17. Kim,M.K., Jernigan,R.L., Chirikjian,G.S. (2003) An elastic network model of HK97 capsid maturation. *J. Struct. Biol.*, **143**, 107–117.
18. Reuter,N., Hinsen,K. and Lacapere,J.J. (2003) Transconformations of the SERCA1 Ca-ATPase: a normal mode study. *Biophys. J.*, **85**, 2186–2197.
19. Thomas,A., Hinsen,K., Field,M.J. and Perahia,D. (1999) Tertiary and quaternary conformational changes in aspartate transcarbamylase: a normal mode study. *Proteins*, **34**, 96–112.
20. Hinsen,K., Thomas,A. and Field,M.J. (1999) Analysis of domain motions in large proteins. *Proteins*, **34**, 369–382.
21. Brink,J., Ludtke,S.J., Kong,Y., Wakil,S.J., Ma,J. and Chiu,W. (2004) Experimental verification of conformational variation of human fatty acid synthase as predicted by normal mode analysis. *Structure*, **12**, 185–191.
22. Suhre,K. and Sanejouand,Y.H. (2004) On the potential of normal mode analysis for solving difficult molecular replacement problems. *Acta Cryst. D*, **60**, 796–799.
23. Doruker,P., Jernigan,R.L. and Bahar,I. (2002) Dynamics of large proteins through hierarchical levels of coarse-grained structures. *J. Comput. Chem.*, **23**, 119–127.
24. Perahia,D. and Mouawad,L. (1995) Computation of low-frequency normal modes in macromolecules: improvements to the method of diagonalization in a mixed basis and application to hemoglobin. *Comput. Chem.*, **19**, 241–246.
25. Kleywegt,G.J. (1996) Use of non-crystallographic symmetry in protein structure refinement. *Acta Cryst. D*, **52**, 842–857.
26. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
27. Kraulis,P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallography*, **24**, 946–950.
28. Humphrey,W., Dalke,A. and Schulten,K. (1996) VMD – visual molecular dynamic. *J. Molec. Graphics*, **14**, 33–38.

6.8 Article fourni « MOZAIC »

(Comme une “preuve de ma vie antérieure” de chercheur des Sciences de l’Univers, je n’ai pu m’empêcher de rajouter un réprint de mon article décrivant des observations de l’expérience MOZAIC, observations qui ont récemment trouvées leurs explications dans les travaux de Zahn et al. (2002)).

Suhre, K., Cammas, J.P., Nedelec, P., Rosset, R., Marenco, A. and Smit, H.G.J. (1997) Ozone-rich transients in the upper equatorial Atlantic troposphere. *Nature*, **388**, 661-663.

Ozone-rich transients in the upper equatorial Atlantic troposphere

K. Suhre*, J.-P. Cammas*, P. Nédelec*, R. Rosset*, A. Marenco* & H. G. J. Smit†

** Laboratoire d'Aérologie (UMR CNRS/UPS 5560), OMP, 14, Avenue Edouard Berlin, 31400 Toulouse, France*
† Research Centre Jülich, Institute for Chemistry of the Polluted Atmosphere (ICG-2), PO Box 1913, 52425 Jülich, Germany

High concentrations of ozone are found in the Earth's stratosphere, but strong stratification suppresses efficient exchange of this ozone-rich air with the underlying troposphere. Upward transport of tropospheric trace constituents occurs mainly through equatorial deep convective systems. In contrast, significant downward transport of ozone-rich stratospheric air is thought to take place only outside the tropics by exchange processes in upper-level fronts associated with strong distortions of the tropopause[1]. Ozone within the tropical troposphere is assumed to originate predominantly from ground-based emissions of ozone precursors, particularly from biomass burning[2], rather than from a stratospheric source. Recent measurements of ozone in the upper troposphere in convective regions over the Pacific Ocean[3] indeed reveal near-zero concentrations. Here we present sharply contrasting observations: ozone-rich (100–500 parts per billion by volume) transients were frequently encountered by specially equipped commercial aircraft at a cruising altitude of 10–12 km (in the upper troposphere) in the vicinity of strong convective activity over the equatorial Atlantic Ocean. This strongly suggests that the input of stratospheric ozone into the troposphere can take place in the tropics. We suggest that this transport occurs either by direct downward movement of air masses or by quasi-isentropic transport from the extratropical stratosphere.

Five commercial Airbus 340 aircraft operated by Air France, Austrian Airlines and Lufthansa have been equipped with ozone, temperature and humidity instruments, as part of the 'Measurement of Ozone by Airbus In-Service Aircraft' (MOZAIC) project[4,5]. We report here an analysis of more than 100 flights across the Inter-Tropical Convergence Zone (ITCZ) over the tropical Atlantic Ocean between Europe and South America. During most of the Atlantic transect, ozone mixing ratios are well below 50 parts per billion by volume (p.p.b.v.) with only small fluctuations. But in one-third of the reported flights, uniquely within the 15° N–15° S equatorial latitude belt, we observe events where ozone mixing ratios suddenly increase for a short time period to over 100 p.p.b.v., and in some cases up to 500 p.p.b.v., at a horizontal scale of 5–80 km. A representative example is shown in Figs 1 and 2. At the same time, the aircraft enters a zone of turbulence of about the same horizontal extent, as the pressure–altitude records indicate a vertical displacement of the order of 10 m (Fig. 1d) and abrupt changes in wind speed and direction are observed (Fig 1e, f). At the same time as these high-ozone events, water mixing ratios usually increase drastically (Fig. 1b) and temperatures slightly decrease (Fig. 1c), but not systematically for the latter. Relative humidity >100% is sometimes observed during MOZAIC flights and particularly during these high-ozone events. This is not ascribed to artefacts; the extra water is probably provided by the evaporation of water droplets due to adiabatic compression and associated heating which cause a substantial temperature increase in the humidity-sensor housing at cruising speeds of Mach 0.8. Thus, the high humidity values must be interpreted qualitatively as indicating the presence of ice crystals during the sampling. The aircraft often deviates from its otherwise straight flight track during these high-ozone events, apparently to outflank an equatorial convective system. The presence of active deep convection has been identified on NOAA/AVHRR satellite images in most cases. Note that for safety reasons passenger aircraft never pass through the active part of tropical deep convective systems, but at a distance of 15–30 km around the convective cloud towers, thus passing under the anvil clouds and their precipitation streaks.

In 10% of all analysed flights across the ITCZ, an event with very high ozone mixing ratio (>150 p.p.b.v.) over a distance of some tens of kilometres has been observed, together with all the following: turbulence, decreasing temperature, rising humidity, changes in wind speed and direction, deviation of the aircraft trajectory from a straight line, and convective systems on NOAA/AVHRR images. In even more cases (30% of all flights), high ozone values (100–200 p.p.b.v.) have been found and correlated with some, but not always all, of the observations described above. These latter cases are likely to be the observation of residual perturbations some time after the event itself occurred, whereas the 10% of 'full' cases probably represent observations that took place shortly after the perturbations were caused. Here we consider only observations of 'full' cases made over the Atlantic Ocean, where MOZAIC data is most dense.

More than 5,000 MOZAIC flights in the upper troposphere (low-ozone regime) as well as in the lower stratosphere (high-ozone regime) show that the performance of the ozone analyser is not affected by the aircraft entering or leaving clouds (no significant ozone changes are observed), by turbulence (many cases of strong turbulence without change in ozone concentrations have been observed outside the tropics), or by icing of the ozone inlet (no significant changes of flow rates in the sampling line and ozone device have been observed). The rare erratic ozone values that are sometimes found consist in the worst case of only two successive records. On a large number of MOZAIC flights to South Africa and to Asia similar cases of high-ozone transients have been encountered when crossing the ITCZ, but never in middle latitudes. On one occasion one aircraft encountered the same ozone transient twice: on its flight towards Bangkok and again when heading back to Paris several hours later. On another occasion, two different aircraft flying from South America towards Europe sampled the same ozone transient within a 30 minute interval and at a horizontal separation of 10 km. Measurement artefacts can therefore be ruled out as a possible cause for the high ozone mixing ratios.

On a spatial scale the observed events are small (5–80 km) and sharply marked. The proximity of the observed high-ozone values to equatorial deep convective systems, as indicated by aircraft deviations, turbulence, satellite images and humidity measurements (indicating the presence of ice particles), suggest that the observed phenomena are linked to these convective systems. We can exclude any local surface sources and orography-induced vertical transport as all selected events occur over the open ocean. In the free troposphere photochemically produced ozone, even in plumes from biomass burning, never exceeds levels of 100–150 p.p.b.v. (refs 6, 7). Cloud-charging processes caused by the dynamics in heavy cumulonimbus clouds can be a local source of ozone[8]: here ozone is produced by point discharges from water droplets in intense electrical fields[9] or by electrical discharging of a lightning flash[10]. In some field experiments in the vicinity of electrified clouds, high-ozone transients have been observed in the lower troposphere[10,11]. If the ozone transients observed during MOZAIC flights originate from discharging processes in cumulonimbus clouds, then the ozone is probably produced in the core of these clouds while the aircraft is flying outside the core in the divergence zone of the deep convective region. The ozone would then originate from air masses being detrained from the upward-moving convective flow. At present it is not clear if the dynamics within the observed deep

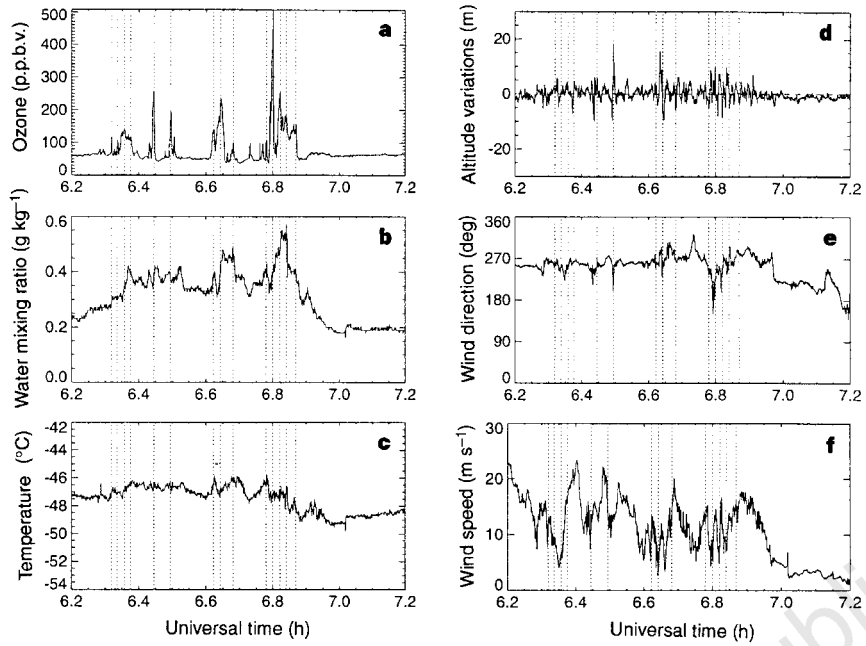

Figure 1 Time series of: **a**, ozone mixing ratio; **b**, water vapour mixing ratio; **c**, temperature; **d**, barometric flight altitude variations; **e**, wind direction; **f**, wind speed. These time series were measured on MOZAIC flight Paris–São Paulo on 21 November 1994. The time of the encounter of ozone maxima are marked by vertical dashed lines on all graphs. This time series represents one hour of measurements during which the aircraft travelled 860 km from 7° 20′ N, 27° 20′ W to 0° 30′ N, 30° 20′ W. (data acquisition every 4 s, resulting in a spatial resolution of about 1 km). Before and after the event shown here, ozone mixing ratios were always well below 100 p.p.b.v. in tropical latitudes. Instrumentation was as follows: **a**, a dual-beam ultraviolet absorption instrument, Thermo-Electron model 49-103 (detection limit 2 p.p.b.v., overall precision ±2%); **b** and **c**, an AERODATA AD-F8-88 instrument installed in a Rosemount housing; **d**, **e**, **f**, Airbus A340 on-board instruments. Documentation of ten representative cases is available on the ftp site ftpaero.aero.obs-mip.fr (file: /pub/salsa/mozaic/SUPP1.PS).

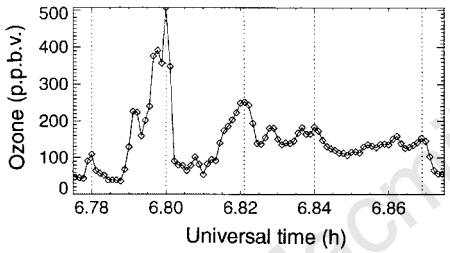

Figure 2 Expanded-scale view of part of the ozone time series in Fig. 1a. Data points, represented by diamonds, are recorded every 4 s which corresponds to a horizontal resolution of ~1 km. The locationof the aircraft at the middle of this time series was approximately 3° 0′ N, 28° 50′ W.

convective clouds encountered concurrently with the observed ozone transients during the MOZAIC flights were sufficiently strong to induce cloud-charging processes with electric fields high enough to produce significant amounts of ozone.

An alternative and more likely explanation for our observations is that the high ozone mixing ratios have a stratospheric origin. Direct vertical downward transport could, for example, move ozone-rich air from the equatorial stratosphere into the upper troposphere. Recent measurements[12,13] show that ozone mixing ratios high enough to cause the ozone transients measured by the aircraft occur in mixed layers formed by tropical cumulus anvil outflow[14] in the stratosphere at about 18 km altitude. The observation of overshoots of cumulonimbus tops into the stratosphere[15–17] provides further evidence that the entrainment of stratospheric air into an anvil outflow is dynamically possible. Precipitation streaks loaded with ice particles falling from the anvil[18–20] can then transport ozone-rich air through the tropopause region into the upper troposphere. Further evaporative cooling and stronger downdrafts[21] may then move the air rapidly down to the aircraft's flight level at 10–12 km altitude. This mechanism requires continuous evaporative cooling to reduce the potential temperature of the air parcel from its originally high stratospheric value of ~400 K to ~340–350 K at flight level. This can only be achieved by a continuous supply of ice particles growing in the anvil outflow, as proposed in Danielsen's stratospheric dehydration mechanism[18,19] with the largest particles falling close to the anvil outflow into the descending air parcel. The precise dynamical mechanism of that ozone downward transport still has to be established, and certainly needs more detailed sensing of the upper part of convective systems than available in the MOZAIC project. Another mechanism for vertical downward transport of stratospheric air into the troposphere has been proposed by Newell *et al.*[22]. Although this process operates on much larger horizontal and temporal scales, it might play an initiating role in transferring stratospheric air several kilometres downwards before the actual convective event moves it down to flight level.

Instead of tracing their origin in the equatorial stratosphere, the observed ozone transients could also be the result of horizontal transport of ozone-rich air from the extratropical stratosphere followed by vertical convective transport to flight level. The central Atlantic, being downstream of the eastern major trough over the United States and at the tail end of the corresponding storm tracks, is a favoured region of extratropical wave-breaking far equatorward of the subtropical jet stream. Fragmentation of extratropical stratospheric-air intrusions[23] into elongated (2,000–3,000 km) and slender (200 km) streamers could be consistent with MOZAIC observations. This quasi-isentropic mechanism does not require cooling of the transported air mass (the 340–350 K isentropic layer in the middle latitudes is situated in the lower stratosphere where ozone mixing ratios are as high as those observed in the transients), but it is not yet clear whether thin streamers can extend into the equatorial region without undergoing significant mixing with tropospheric air. Meridional cross-sections of a climatologically averaged height of the 340–350 K isentropic layer indicate that the MOZAIC aircraft fly below this layer in the subtropics and enter it only in the equatorial latitude belt, in agreement with the observed locations of the ozone transients. Newell *et al.*[24] and Browell *et al.*[25] report ozone layers of stratospheric origin in the upper tropical troposphere, but with maximum ozone mixing ratios well below 100 p.b.b.v. The difficulty in determining whether this mechanism is at the origin of the observed ozone transients is that in such a long-range isentropic transport the air gradually loses its original

stratospheric characteristics by radiative processes (which act on the temperature and hence the potential vorticity). Although this air may keep its initial chemical composition, the loss of its stratosphere characteristics makes it difficult to distinguish from tropospheric air when using basic weather analysis charts.

The ozone transients encountered relatively often on the reported MOZAIC flights is a new finding that contrasts strongly with the near-zero ozone concentrations in about the same part of the upper troposphere observed by Kley *et al.*[3], emphasizing the enormous variability of equatorial tropospheric ozone concentrations. We have put forward three possible explanations of the observed phenomena, without however being able to decide between them on the basis of current knowledge. Formation of ozone by cloud-charging processes at the observed scale (5–80 km) in the upper troposphere has so far not been observed, favouring a dynamical explanation. The possible existence of a mechanism of transport of stratospheric air into the upper equatorial troposphere is most intriguing and reflects our still poor knowledge of the dynamics of the upper part of the equatorial troposphere, particularly with regard to mass transfer between stratosphere and troposphere during deep convection and extratropical baroclinic wave-breaking extending into the equatorial belt. Furthermore, the existence of the ozone transients implies that the ozone budget of the equatorial troposphere is less well understood than has been assumed. □

Received 24 March; accepted 19 June 1997.

1. Holton, J. R. *et al.* Stratosphere-troposphere exchange. *Rev. Geophys.* **33,** 403–439 (1995).
2. Crutzen, P. J., Heidt, L. E., Krasnec, J. P., Polloc, W. H. & Seiler, W. Biomass burning as a source of atmospheric trace gases CO, H_2, N_2O, NO, CH_3Cl and COS. *Nature* **282,** 253–356 (1979).
3. Kley, D. *et al.* Observations of near-zero ozone concentrations over the convective Pacific: effects on air chemistry. *Science* **274,** 230–233 (1996).
4. Marenco, A., Nedelec, P., Thouret, V. & Grouhel, C. in *DLR-Mitteilungen* (eds Schumann, U. & Wurzel, D.) **94-06,** 26–31 (DLR, Ober Pfaffen Hofen, 1994).
5. Marenco, A. The MOZAIC programme. *EUROTRAC Newsl.* **17,** 2–7 (1996).
6. Andreae, M. O. *et al.* Influence of plumes from biomass burning on atmospheric chemistry over the equatorial and tropical South Atlantic during CITE-3. *J. Geophys. Res.* **99,** 12793–12809 (1994).
7. Delany, A. C., Haagensen, P., Walters, S., Wartburg, A. F. & Crutzen, P. J. Photochemically produced ozone in the emissions from large-scale tropical vegetation fires. *J. Geophys. Res.* **90,** 2425–2429 (1985).
8. Griffing, G. W. Ozone and oxides of nitrogen production during thunderstorms. *J. Geophys. Res.* **82,** 943–950 (1977).
9. Shlanta, A. & Moore, C. B. Ozone and point discharge measurements under thunderstorms. *J. Geophys. Res.* **77,** 4500–4510 (1972).
10. Orville, R. E. Ozone production during thunderstorms, measured by the absorption of ultraviolet radiation from lightning. *J. Geophys. Res.* **72,** 3557–3561 (1967).
11. Clarke, J. F. & Griffing, G. W. Aircraft observations of extreme ozone concentrations near thunderstorms. *Atmos. Environ.* **19,** 1175–1179 (1985).
12. Vömel, H., Oltmans, S. J., Kley, D. & Crutzen, P. J. New evidence for the stratospheric dehydration mechanism in the equatorial Pacific. *Geophys. Res. Lett.* **22,** 3235–3238 (1995).
13. Russel, P. B., Pfister, L. & Selkirk, H. B. The tropical experiment of the stratosphere-troposphere exchange project (STEP): science objectives, operations, and summary findings. *J. Geophys. Res.* **98,** 8563–8589 (1993).
14. Lilly, D. K. Cirrus outflow dynamics. *J. Atmos. Sci.* **45,** 1594–1605 (1988).
15. Roach, W. T. On the nature of the summit areas of severe storms in Oklahoma. *Q. J. R. Meteorol. Soc.* **93,** 318–336 (1967).
16. Burnham, J. Atmospheric gusts—a review of the results of some recent research at the royal aircraft establishment. *Mon. Weath. Rev.* **98,** 723–734 (1970).
17. Cornford, S. G. & Spavins, C. S. Some measurements of cumulonimbus tops in the pre-monsoon season in north-east India. *Meteorol. Mag.* **102,** 314–332 (1973).
18. Danielsen, E. F. A dehydration mechanism for the stratosphere. *Geophys. Res. Lett.* **9,** 605–608 (1982).
19. Danielsen, E. F. In situ evidence of rapid, vertical, irreversible transport of lower tropospheric air into the lower tropical stratosphere by convective cloud turrets and by large-scale upwelling in tropical cyclones. *J. Geophys. Res.* **98,** 8665–8681 (1993).
20. Knollenberg, R. G., Kelly, K. & Wilson, J. C. Measurements of high number densities of ice crystals in the tops of tropical cumulonibus. *J. Geophys. Res.* **98,** 8639–8664 (1993).
21. Sun, J., Braun, S., Biggerstaff, M. I., Fovell, R. G. & Houze, R. A. Jr Warm upper-level downdrafts associated with a squall line. *Mon. Weath. Rev.* **121,** 2919–2927 (1993).
22. Newell, R. E., Zhu, Y., Browell, E. V., Read, W. G. & Waters, J. W. Walker circulation and tropical upper tropospheric water vapor. *J. Geophys. Res.* **101,** 1961–1974 (1996).
23. Appenzeller, C., Davies, H. C. & Norton, W. A. Fragmentation of stratospheric intrusions. *J. Geophys. Res.* **101,** 1435–1456 (1996).
24. Newell, R. E. *et al.* Vertical fine-scale atmospheric structure measured from NASA DC-8 during PEM-West A. *J. Geophys. Res.* **101,** 1943–1960 (1996).
25. Browell, E. V. *et al.* Large-scale air mass characteristics observed over Western Pacific during summertime. *J. Geophys. Res.* **101,** 1691–1712 (1996).

Acknowledgements. We thank Air France A340 pilot De Boysson for eye-witness information from MOZAIC flights in the tropics; we also acknowledge discussions with many of our colleagues. On behalf of the MOZAIC programme, we thank Air France, Lufthansa, Austrian Airlines and Sabena, who agreed to carry the MOZAIC equipment free of charge. This work was supported by the European Commission DG XII and Centre National de la Recherche Scientifique (CNRS-INSU); satellite images were made available by NOAA-SAA.

Correspondence should be addressed to K.S. (e-mail: suhk@aero.obs-mip.fr). Additional information on the MOZAIC programme is available on http://www.cnrm.meteo.fr:8000/mozaic/

Conspecific sperm precedence in *Drosophila*

Catherine S. C. Price

Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, Illinois 60637, USA

Traits that influence the interactions between males and females can evolve very rapidly through sexual selection[1] or sexually antagonistic coevolution[2]. Rapid change in the fertilization systems of independent populations can give rise to reproductive incompatibilities between populations[3,4], and may contribute to speciation[5]. Here I provide evidence for cryptic reproductive divergence among three sibling species of *Drosophila* that leads to a form of postmating isolation. When a female mates with both a conspecific and a heterospecific male, the conspecific sperm fertilize the vast majority of the eggs, regardless of the order of the matings. Heterospecific sperm fertilize fewer eggs after these double matings than after single matings. Experiments using spermless males show that the seminal fluid of the conspecific male is largely responsible for this conspecific sperm precedence. Moreover, when two males of the same species mate sequentially with a female from a different species, a highly variable pattern of sperm precedence replaces the second-male sperm precedence that is consistently found within species. These results indicate that females mediate sperm competition, and that second-male sperm precedence is not an automatic consequence of the mechanics of sperm storage.

Whenever a female mates with more than one male during a single fertile period, the opportunity for sperm competition is created[6]. A male's fitness on mating will be greatly influenced by his ability to interfere with resident sperm or to defend his sperm against interference by subsequent males. Male traits affecting sperm competition are therefore expected to be under strong selection. Female fitness is also influenced by multiple mating[7], and females can be directly harmed by male adaptations for sperm competition[8], so females should be under strong selection to mediate sperm competition. Such strong selection sets up the opportunity for rapid divergence among populations in traits that influence sperm competition. Nevertheless, the possibility of competition-dependent postmating, prezygotic reproductive isolation has not to my knowledge been investigated in organisms in which sperm competition has been extensively studied, such as *Drosophila*.

Conspecific sperm precedence (CSP) has been documented previously in grasshoppers[9,10], crickets[11] and flour beetles[12]. It is not known, however, whether the reproductive isolation described in these species is strictly dependent upon sperm competition, as none of these studies directly compared the number of hybrids produced after a single observed heterospecific copulation with the number produced after one heterospecific and one conspecific copulation. Here I show that conspecific sperm preference among three species of *Drosophila* is a form of reproductive isolation that becomes apparent only after multiple matings, and is produced in large part by the seminal fluid of the conspecific male. Moreover, two independent experiments indicate that there is divergence among these species in traits that influence sperm competition.

The common outcome of sperm competition within most species of insects and birds is second-male precedence[13]. In *Drosophila*, females of many species mate with multiple males in nature, and in the laboratory the second male to mate typically fathers at least 85% of the offspring[14]. Although the mechanisms of sperm precedence in *Drosophila* remain unknown, the phenomenon appears to be influenced in large part by the seminal fluid of the second male[15].