DÉVELOPPEMENTS THÉORIQUES ET MÉTHODES NUMÉRIQUES POUR LES ANALYSES COMPARATIVES DE GÉNOMES ET PROTÉOMES BIAISÉS. Application à la comparaison des génomes et protéomes de Plasmodium falciparum et d'Arabidopsis thaliana

Olivier Bastien

▶ **To cite this version:**

Olivier Bastien. DÉVELOPPEMENTS THÉORIQUES ET MÉTHODES NUMÉRIQUES POUR LES ANALYSES COMPARATIVES DE GÉNOMES ET PROTÉOMES BIAISÉS. Application à la comparaison des génomes et protéomes de Plasmodium falciparum et d'Arabidopsis thaliana. Biochimie [q-bio.BM]. Université Joseph-Fourier - Grenoble I, 2006. Français. NNT : . tel-00080245

HAL Id: tel-00080245

https://theses.hal.science/tel-00080245v1

Submitted on 15 Jun 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Joseph Fourier – Grenoble I

THESE

Pour obtenir le grade de
DOCTEUR DE L'UNIVERSITÉ JOSEPH FOURIER
Discipline : Biologie

Présentée par

Olivier Bastien

Le 21 Avril 2006

DÉVELOPPEMENTS THÉORIQUES ET MÉTHODES NUMÉRIQUES POUR LES ANALYSES COMPARATIVES DE GÉNOMES ET PROTÉOMES BIAISÉS

Application à la comparaison des génomes et protéomes de *Plasmodium falciparum* et d'*Arabidopsis thaliana*

Directeur de thèse : Éric Maréchal

Composition du Jury :

Président :	Johannes Geiselmann
Rapporteur :	Jean-Loup Risler
Rapporteur :	Emmanuel Douzery
Examinateur :	Jean-Paul Comet
Examinateur :	Jean-Jacques Codani
Examinateur :	Éric Maréchal

Thèse préparée au sein du Laboratoire de Physiologie Cellulaire Végétale
CEA-Grenoble

A Delphine,

Remerciements

Je remercie chaleureusement Monsieur Jacques Joyard pour m'avoir accueilli dans le Laboratoire de Physiologie Cellulaire Végétale. Merci également à Madame Marilyn Vantard et Monsieur Norbert Rolland pour m'avoir renouvelé leur confiance après avoir pris la co-direction du laboratoire.

Je remercie vivement Monsieur Jean-Jacques Codani, ancien directeur de recherche à l'INRIA et cofondateur de Gene-IT, pour avoir accepté de soutenir cette thèse dans le cadre d'une convention Cifre et m'avoir apporté une aide logistique inestimable.

J'adresse mes plus vifs remerciements à Messieurs Hans Geiselmann, Professeur à l'Université Joseph Fourier de Grenoble, Jean-Loup Risler, Directeur de recherche au CNRS à Evry, Emmanuel Douzery, Professeur à l'Université de Montpellier et Jean-Paul Comet, Maître de Conférences à l'Université d'Évry Val d'Essonne pour l'honneur qu'ils me font en acceptant de juger cette thèse.

Je remercie du fond du cœur Monsieur Eric Maréchal pour m'avoir encadré dans cette thèse. Je le remercie chaleureusement pour son soutien sans faille dans les nombreuses difficultés que j'ai pu rencontrer, soutien sans lequel bien peu de choses auraient été rendues possibles. Comment ne pas également évoquer ici nos nombreuses discussions philosophiques et scientifiques ainsi que les nombreux conseils qu'il m'a prodigués.

Je remercie du fond du cœur Madame Sylvaine Roy pour sa collaboration et ses nombreux conseils, depuis mon DEA jusqu'à ce jour.

Un grand merci va naturellement à Monsieur Philipe Ortet, "Mister PechPee", pour tout ce que nous avons partagé ensemble : collaborations, discussions, e-mails endiablés et franches rigolades.

Tout ceci n'aurait pas eu la même saveur sans mon irremplaçable partenaire de bureau, j'ai cité Madame Juliette Jouhet.

Merci aussi à Madame Maryse Block pour ses conseils, son humanité et la très grande rigueur scientifique qu'elle inspire.

Un grand merci à Sylvain Lespinats pour nos discussions toujours agréables, productives et génératrices d'amitié. Je pense ici à Monsieur Bernard Fertil pour sa collaboration et sa gentillesse.

Merci aussi à Jacques Bourguignon, Jean-Emmanuel Sarry, Pascaline Le Lay, Corinne Rivasseau, Florent Villier, Jean-Christophe Aude, Karine Métayer, Cordelia Bisanz et Cyrille Botté pour les travaux que nous avons menés ensemble et l'ambiance dans laquelle ils se sont déroulés. Je remercie évidemment à cette occasion Gilles Curien avec qui nos discussions ne sont pas prêtes de finir. Je remercie aussi Jean-Michel Pabiot sans qui cette thèse n'aurait jamais vu le jour.

Je pense aussi ici à Mohamed Barakat et Thierry Heulin pour l'excellente collaboration que nous avons initiée et la toujours excellente humeur dans laquelle ils nous baignent à chaque nouvelle rencontre méditerranéenne.

J'exprime toute ma sympathie à tous les membres du laboratoire PCV.

Je n'oublie pas les ex-PCV Samuel Jabrin, Stéphane Miras, Pierre D'Hérin, et Alexandra Kraut devenus depuis mes amis et qui ont plus que contribué à la bonne humeur qui a dominé cette thèse.

Enfin, merci à Delphine, ma mère, Claude et Renée pour m'avoir supporté, encouragé et soutenu.

Sommaire

Introduction

Introduction

I. Le Paludisme, une maladie infectieuse difficile à combattre

A. Etat de la pandémie dans le monde

Le *paludisme*, ou *malaria*, est un fléau mondial qui touche de 350 à plus de 500 millions d'êtres humains et qui tue chaque année de 1,5 à 2,7 millions de personnes à travers le monde, essentiellement en Afrique, dont 1 million d'enfants de moins de cinq ans. Avec le SIDA, le paludisme est l'un des principaux facteurs de mortalité au sein des populations d'Afrique, d'Asie du sud-est et d'Amérique latine ; il porte une part importante dans l'appauvrissement continu de ces populations. (Brierley, 2005; World Malaria Report, 2005)[1]. L'Europe connaît des cas de paludisme dits d'importation. En France, en 1999, plus de 7 000 cas on été rapportés dont une vingtaine de décès : 95 % des cas ont été contractés lors de voyage en Afrique subsaharienne. La majorité des cas survient chez des personnes n'ayant pas suivi de prophylaxie.

Plus de 3 milliards de personnes vivent dans des zones à risque (**Figure 1**).

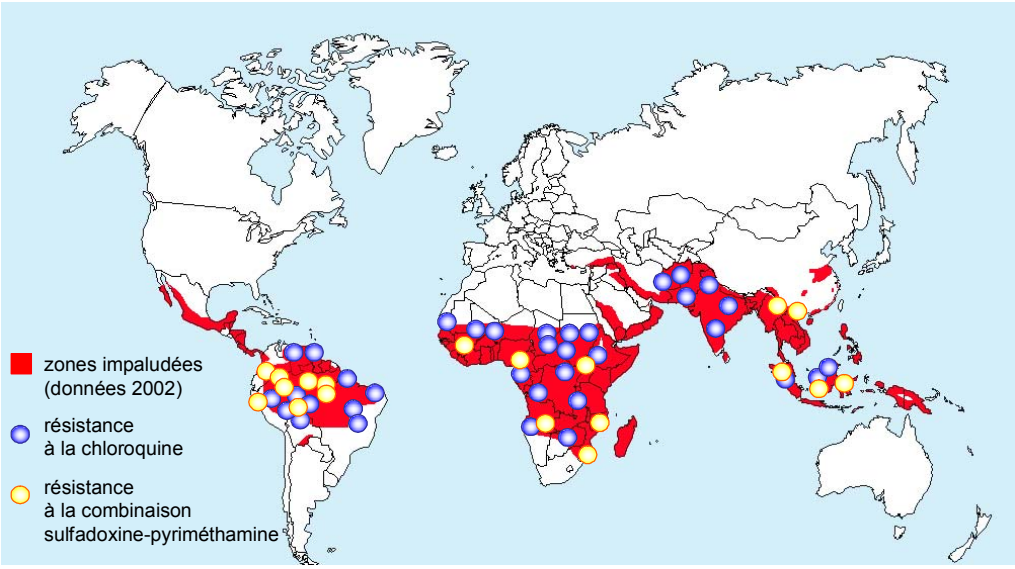


Figure 1: zones impaludées dans le monde. Les foyers de résistances aux traitements par la chloroquine et les associations sulfadoxine-pyriméthamine sont répartis sur l'ensemble des zones impaludées. Adapté de Ridley (2002).

[1] Le rapport mondial sur le paludisme, édité en 2005 par l'Organisation Mondiale de la Santé et de l'Unicef est accessible via internet sur le site du consortium Roll Back Malaria, http://rbm.who.int/wmr2005.

B. Cycle parasitaire de *Plasmodium falciparum* chez l'anophèle et l'être humain

Le paludisme, décrit par Hippocrate[2] au IV^{ème} siècle avant notre ère, était répandu en Europe dans les zones marécageuses, et on pensait alors que le mauvais air en était la cause[3]. Au XVIII^{ème}, Lancisi suggère que le paludisme est dû à un poison des marais transmis par les moustiques qui inoculent de "mauvaises humeurs dans le sang". A la fin du XIX^{ème} siècle, le médecin militaire Alphonse Laveran démontre la nature parasitaire de l'affection en détectant des "éléments pigmentés dans les globules rouges des malades atteints de fièvres palustres, qui se présentent sous formes de croissant, de sphères, de flagelles", objets qu'il nomme *Oscillaria malariae*. Donald Ross (prix Nobel de Médecine 1907) montre que le paludisme pouvait être transmis par des moustiques. En même temps, Grassi, Bastianelli et Bignami (1899) décrivent le cycle complet de développement de *Plasmodium falciparum*, *Plasmodium vivax* et *Plasmodium malariae* (pour revue, Desowitz, 1991).

Le paludisme correspond à la phase de propagation de parasites du genre *Plasmodium* chez l'être humain. Plus de 80% des cas mortels sont dus à une infection par *Plasmodium falciparum*. L'infection est véhiculée par des moustiques, essentiellement *Anopheles gambiae*. La propagation et la multiplication asexuée dans l'organisme s'effectuent tout d'abord dans les cellules du foie (phase hépatique), puis dans les globules rouges (phase érythrocytaire). Après ingestion de sang par un moustique femelle, la phase de reproduction du *Plasmodium* se déroule dans l'épithélium digestif de l'insecte vecteur. Puis le parasite migre vers les glandes salivaires de l'Anophèle, préparant ainsi une nouvelle infection (**Figure 2**).

C. Evolution de la résistance à la chimioprophylaxie[4] et nécessité de découvrir de nouvelles cibles pour des traitements thérapeutiques

Il y a plus de 350 ans, les missionnaires jésuites en Amérique du Sud font connaître les propriétés antipaludiques de l'écorce de quinquina. La quinine, un alcaloïde végétal toxique qui en est extrait, est le seul traitement connu par les européens et les américains jusqu'au milieu du XX^{ème} siècle. Elle devint introuvable durant la Deuxième Guerre mondiale. Il fallait donc mettre au point de toute urgence un nouvel antipaludique. Un produit synthétique appartenant aux composés dits amino-4-quinoléines avait été mis au point par une société pharmaceutique allemande, en 1934. Ce produit acquis par les Américains en 1943 a permis de mettre au point un dérivé efficace, la chloroquine.

---

[2] Hippocrate (460-377 av. J-C) décrit deux cas dans le livre des Epidémies. Le premier malade, Erasinus "fut pris d'une forte fièvre après le souper ; la nuit fut troublée. Premier jour, tranquillité pendant la journée, souffrance pendant la nuit. Deuxième jour, tout s'aggrava ; hallucination pendant la nuit. (…) Quatrième jour, malaise extrême ; point de sommeil pendant la nuit, rêves et discours (...). Cinquième jour, le matin il était calme et avait sa pleine raison, mais avant le milieu de la journée, il fut saisi d'un violent transport ; (…) ; les urines se supprimèrent. Il mourut vers le coucher du soleil. Chez ce malade, les accès fébriles furent jusqu'à la fin avec sueurs ; les hypochondres étaient gonflés, tendus et douloureux ; (…) ; il éprouva beaucoup de convulsions avec sueurs aux approches de la mort."

[3] Le terme *paludisme* dérive du français ancien *palud*, issu du latin *palus*, signifiant *marais*; le terme *malaria* dérive de l'italien *mala aria* qui signifie *mauvais air*.

[4] La prophylaxie est l'ensemble des mesures qui visent à limiter la propagation d'une maladie infectieuse, voire l'éradiquer. La chimioprophylaxie est l'ensemble des traitements faisant usage de molécules médicamenteuses (drogues).

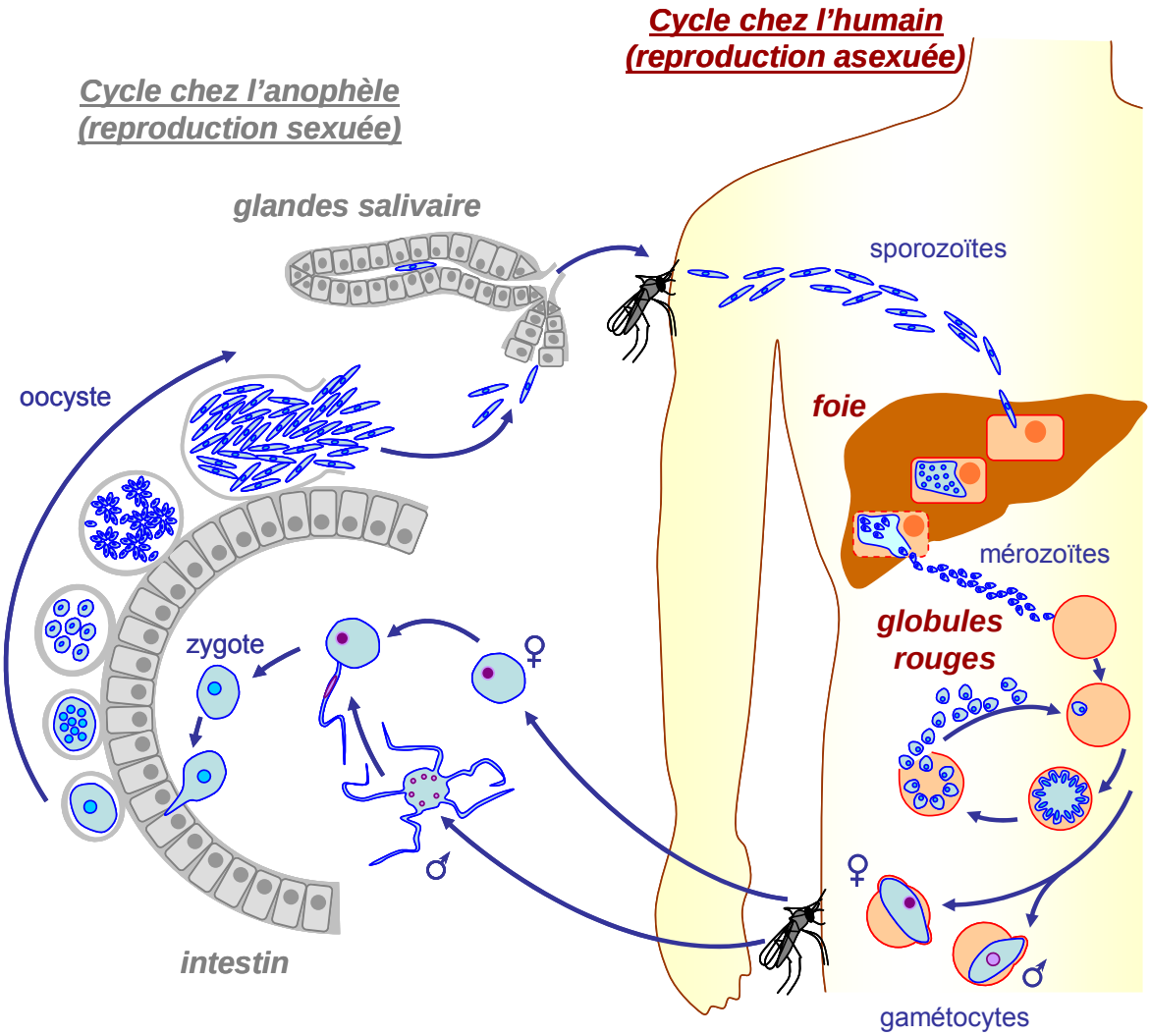


Figure 2: cycle parasitaire de *Plasmodium falciparum*. Le cycle des parasites du genre *Plasmodium* est complexe et comporte deux étapes essentielles : un cycle asexué chez l'humain, et un cycle sexué chez le moustique. L'être humain est infecté lors d'une piqûre par un moustique anophèle femelle, qui lui injecte le parasite sous forme de sporozoïte. Celui-ci migre rapidement, *via* la circulation sanguine, vers le foie. Il se divise très activement dans les cellules hépatiques et donne naissance, en quelques jours, à des dizaines de milliers de nouveaux parasites, les mérozoïtes. Puis, les parasites pénètrent et se multiplient à l'intérieur des globules rouges. Lorsque ces derniers éclatent, les mérozoïtes libérés infectent de nouveaux globules rouges. En parallèle, des cellules sexuées mâles et femelles (gamétocytes) se forment dans le sang du sujet infecté. Lorsqu'un moustique pique un tel sujet, il ingère ces gamétocytes, qui se transforment en gamètes. La fécondation dans l'intestin de l'anophèle engendre un oeuf (zygote), qui se différencie enfin en sporozoïte dans les glandes salivaires du moustique. Un nouveau cycle peut alors commencer.

L'Organisation mondiale de la santé (OMS) a lancé des programmes d'éradication du paludisme à l'échelle mondiale au milieu des années 1950, avec tentatives de dé-moustiquation et traitements massifs des populations humaines. Dès 1961, des souches de *Plasmodium falciparum* résistantes à la chloroquine sont apparues à cause de son utilisation excessive et probablement de doses insuffisantes (voir **Figure 1**). À ce moment-là, il n'y avait pas de médicament susceptible de traiter ces formes de paludisme résistantes à la chloroquine sauf l'antipaludique le plus ancien, la quinine (pour revue, Desowitz, 1991).

L'histoire de la recherche de nouveaux traitements est caractérisée par deux grandes difficultés, d'une part l'incapacité à développer un vaccin, d'autre part l'apparition de formes

résistantes aux molécules antipaludiques. Dans les années 1970, la combinaison de sulfadoxine et pyriméthamine se substitue ou complète les traitements par la chloroquine, mais des foyers de résistance apparaissent, conduisant à la résurgence de paludisme difficile, voire impossible à traiter (voir **Figure 1**). Dans ces zones, le sulfadoxine-pyriméthamine a été remplacé essentiellement par la méfloquine dans les années 1980. Le développement très rapide de la résistance à cette nouvelle drogue a conduit à rechercher de nouvelles molécules, et à évaluer l'efficacité d'anciens traitements jusqu'alors non exploités à échelle globale, en particulier l'artémisine (Farooq et Mahajan, 2004; Baird 2005; Wright 2005)

L'artémisinine (ou qinghaosu), dérivé d'une plante, *Artemisia annua*, est employé depuis plus de deux mille ans en Chine pour traiter les fièvres associées au paludisme. L'histoire contemporaine de l'artémisinine commence pendant la guerre du Vietnam lorsque l'armée nord-vietnamienne construit tout un réseau de souterrains. Comme ces tunnels récupéraient l'eau de pluie, les moustiques transporteurs du paludisme se reproduisaient dans l'eau stagnante. Le problème prit une telle ampleur, que l'armée nord-vietnamienne perdit plus de soldats par le paludisme que par les armes. Les Nord-vietnamiens reçurent l'aide de chercheurs militaires chinois, qui sélectionnèrent l'armoise annuelle comme traitement curatif. Un procédé d'extraction du principe actif, l'artémisinine, une lactone sesquiterpénique contenant un radical peroxyde, est mis au point en 1972. Au cours de ces vingt dernières années, la sécurité et l'efficacité de l'artémisinine et de ses dérivés semi-synthétiques ont été établies. L'artémisinine est donc rapidement devenue un traitement clé du paludisme. Sa popularité s'est particulièrement développée dans le sud-est asiatique et l'Afrique où la maladie est devenue résistante à presque tous les autres antipaludéens (pour revue Bray, et al., 2005; Woodrow et al., 2005). La monothérapie avec l'artémisinine donnant un taux assez élevé de résurgence de la maladie et de sérieuses inquiétudes quant au développement d'une résistance, l'OMS recommande de l'utiliser en association avec d'autres antipaludéens efficaces.

Certains nouveaux composés sont actuellement en développement (par exemple Vial et al., 2004). Parmi les contraintes importantes, il est essentiel que la cible biologique du traitement soit originale par rapport aux processus biologiques ciblés par les traitements existants, et que le coût de production soit faible. Quelques dizaines de projets sont actuellement pris en charge par des partenariats industriels (Nwaka, 2005).

A ce jour, seuls la quinine et les dérivés de l'artémisinine semblent épargnés des phénomènes de résistance. Face à une pandémie grave, contre laquelle la chimioprophylaxie se caractérise par l'apparition rapide de formes de résistances, et en absence de vaccin efficace, un effort soutenu de recherche est engagé pour comprendre la biologie du parasite *Plasmodium* afin d'identifier de nouvelles cibles pour des traitements nouveaux.

II. *Plasmodium falciparum*, un protiste du phylum Apicomplexa, muni d'un plaste acquis par endosymbiose secondaire

A. Caractéristiques générales des cellules des parasites apicomplexes

Le phylum des Apicomplexa constitue un groupe ancien et diversifié de protistes parasites comprenant plus de 4000 espèces. Certains sont des agents d'infections humaines et vétérinaires majeures telles que le paludisme (*Plasmodium*), la toxoplasmose (*Toxoplasma*), la coccidiose des ovins et des gallinacés (*Eimeria*), la cryptosporidiose (*Cryptosporidium*), la babesiose ou Texas water fever (*Babesia*), la theileriose ou East Coast fever (*Theileria*) et des maladies émergentes telles que la néosporose (*Neospora*).

Malgré la grande variété de cycles parasitaires, les apicomplexes partagent des structures subcellulaires uniques résumées sur la **Figure 3**. Un complexe apical, qui a donné le nom au phylum, a été conservé au cours de l'évolution. La forme cellulaire invasive (ou zoïte) comprend trois organites sécrétoires (les micronèmes, les rhoptries et les granules denses), dont le contenu est déchargé séquentiellement au cours de l'invasion. Après pénétration dans la cellule hôte, le complexe apical contribue à l'élaboration d'une poche membranaire qui sépare les parasites de la cellule hôte, la vacuole parasitophore.

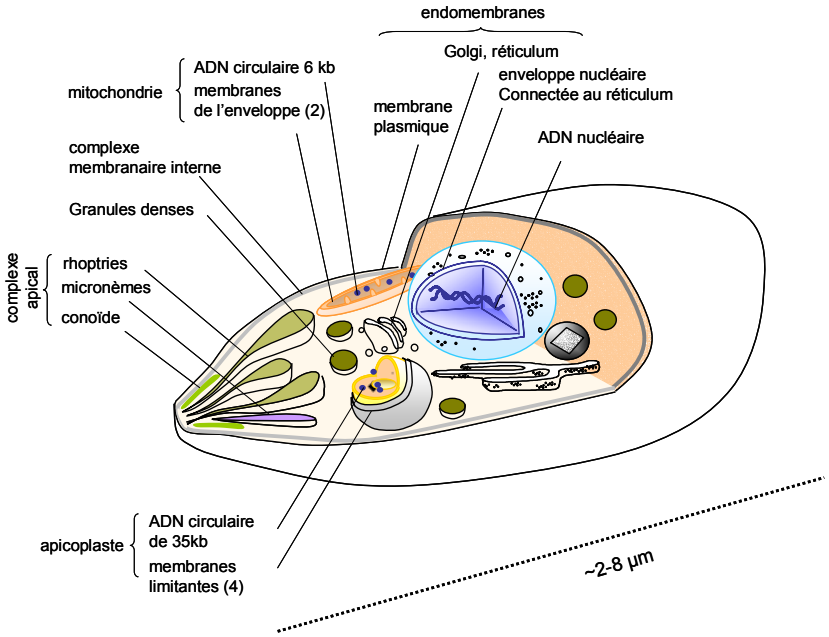


Figure 3: Caractéristiques générales d'une cellule de parasite apicomplexe. Les zoïtes (formes cellulaires invasives) sont des cellules fortement polarisées. Trois organites sécrétoires vésiculaires, les rhoptries, micronèmes et granules denses participent à l'invasion et à l'élaboration de la vacuole parsitophore. Comme dans les cellules végétales, en plus du noyau et de la mitochondrie, un *plaste vestigial*, ou apicoplaste contient de l'ADN. A la différence des chloroplastes simples des cellules végétales, l'apicoplaste est entouré d'un système membranaire additionnel de nature endosomale. Adapté de Bisanz et al. (2006).

La majorité des parasites apicomplexes étudiés à ce jour (*Plasmodium*, *Toxoplasma*, *Babesia*, etc.) présentent deux organites contenant des éléments génétiques extra-chromosomaux : d'une part une mitochondrie et d'autre part un chloroplaste vestigial, l'apicoplaste.

B. Présence d'un plaste vestigial chez *Plasmodium falciparum*

La découverte d'un plaste vestigial non-photosynthétique chez certains apicomplexes était inattendue, et il fallut plusieurs années avant que la communauté scientifique appréhende que ces parasites pouvaient partager un ancêtre commun avec certaines algues.

L'organite connu aujourd'hui sous le nom d'apicoplaste était clairement visible dès 1965 sur des micrographies d'*Eimeria* (Scholtyseck et Piekarski, 1965), sans qu'on identifie sa parenté avec le plaste ; il était qualifié alors de mitochondrie ou de vacuole digestive.

L'ADN de cet organite fut décrit pour la première fois par Gutteridge et al. (1971) chez *Plasmodium knowlesi* comme un ADN satellite mineur de faible densité dans des séparations sur gradient de chlorure de césium. A cette époque, aucune caractéristique d'ADN plastidial n'était encore reconnue, et cette molécule était qualifiée d'ADN mitochondrial de 35 kbases (Gutteridge et al., 1971; Chance et al., 1972). Un ADN circulaire similaire fut décrit chez *Plasmodium lophurae* (Kilejan, 1975). Un doute quant à son hypothétique origine mitochondriale survint avec l'identification d'un ADN de 6 kbases, présentant les caractéristiques d'un génome de mitochondrie chez *P. yoelii*, *P. chabaudi*, *P. berghei*, *P. falciparum*, *P. knowlesi* et *P. cynomolgi* (Vaidya et Arasu, 1987). Les études moléculaires de l'ADN circulaire de 35 kbases indiquèrent une parenté avec les génomes plastidiaux des algues et des cellules de plantes (Wilson et al., 1991; Gardner et al., 1991; Wilson et al., 1996; Wilson et Williamson, 1997), avec en particulier une organisation des gènes pour les sous-unités du complexe ribosomal sous forme de deux secteurs répétés inversés, typiques des ADN chloroplastiques (**Figure 4**).

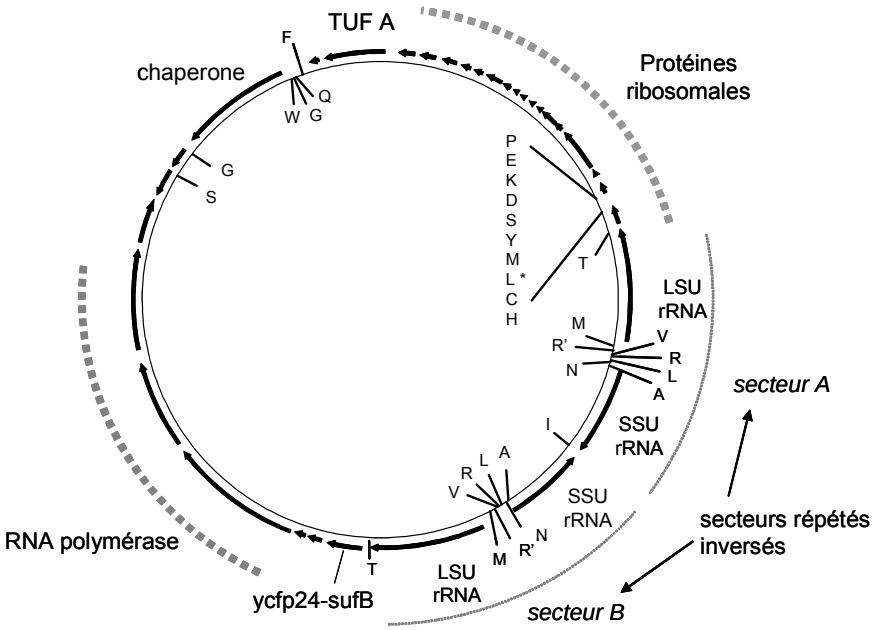


Figure 4: Organisation génétique de l'ADN de 35 kbases de l'apicoplaste. L'organisation génétique de l'ADN de type plastidial porté par l'apicoplaste est très compacte. les gènes sont indiqués par des flèches dans le sens de transcription. Chaque ARNt est indiqué par le code à une lettre de l'amino acide correspondant. Les gènes codant pour la grande sous-unité (LSU) et la petite sous-unité (SSU) du complexe ribosomal sont organisés en secteurs répétés inversés, typiques des ADN chloroplastiques. Figure adaptée de Bisanz et al. (2006).

Cette molécule d'ADN de 35 kbases représente le plus petit génome plastidial connu, portant ~60 gènes, sans un seul gène ou pseudogène photosynthétique. Par technique d'hybridation *in situ* avec une sonde correspondant à un ARN ribosomique, il a été possible de démontrer sa localisation dans la structure sub-cellulaire identifiée précédemment comme une vacuole limitée par de multiples membranes, l'apicoplaste (McFadden et al., 1996; Köhler et al., 1997). La **Figure 5** montre une micrographie de l'apicoplaste de *Plasmodium falciparum* en contact étroit avec la mitochondrie.

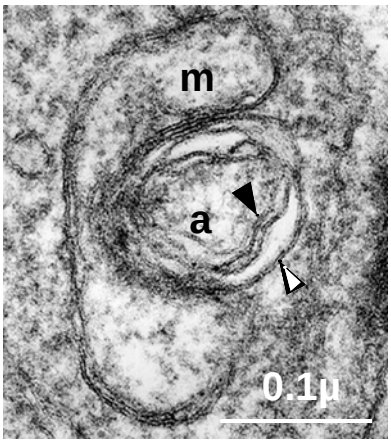


Figure 5. Une structure cellulaire végétale dans la cellule de *Plasmodium*, l'apicoplaste, un plaste vestigial. Détail d'une cellule parasitaire observé par microscopie électronique à transmission. Il existe un seul apicolaste (a) par cellule limité par de multiples membranes (flèches). (m), mitochondrie. Le nombre de membranes limitant le plaste est difficile à déterminer chez *Plasmodium falciparum* ; le consensus actuel est que l'apicoplaste est limité par 4 membranes, et que les variations observées sont liées à des difficultés de fixation d'un matériel biologique de taille particulièrement faible (pour revue Bisanz et al., 2006).

A l'exception notable des parasites du genre *Cryptosporidium*, la majorité des parasites apicomplexes étudiés à ce jour possède un apicoplaste, contenant un ADN circulaire dont l'organisation génétique est particulièrement conservée.

C. Origine du plaste chez *Plasmodium falciparum* par un processus d'endosymbiose secondaire

Les chloroplastes simples sont limités par deux membranes. L'origine de l'apicoplaste, limité par plus de deux membranes, est donc recherchée suivant un processus d'endosymbioses multiples au cours de l'évolution.

L'acquisition d'organites contenant de l'ADN, *i.e.* les chloroplastes et les mitochondries, a permis de construire les eucaryotes (cellules à noyau vrai) dans toute leur diversité, selon un processus d'inclusion (engulfing) d'une alpha-protéobactérie, ancêtre des mitochondries et d'une cyanobactérie, ancêtres de chloroplastes. Selon Archibald et Keeling (2002) et Palmer (2003), les données biochimiques et moléculaires actuelles sont en faveur d'une endosymbiose primaire unique à l'origine de l'ensemble des plastes (en particulier, l'ensemble des plastes décrits à ce jour contient un ADN circulaire caractérisé par une grande zone répétée inversée codant pour les ARN ribosomiques). Du point de vue de la compartimentation, les deux membranes limitant les plastes chez les endosymbiontes primaires (l'enveloppe) sont héritées de la cyanobactérie Gram-négative ancestrale (**Figure 6A**).

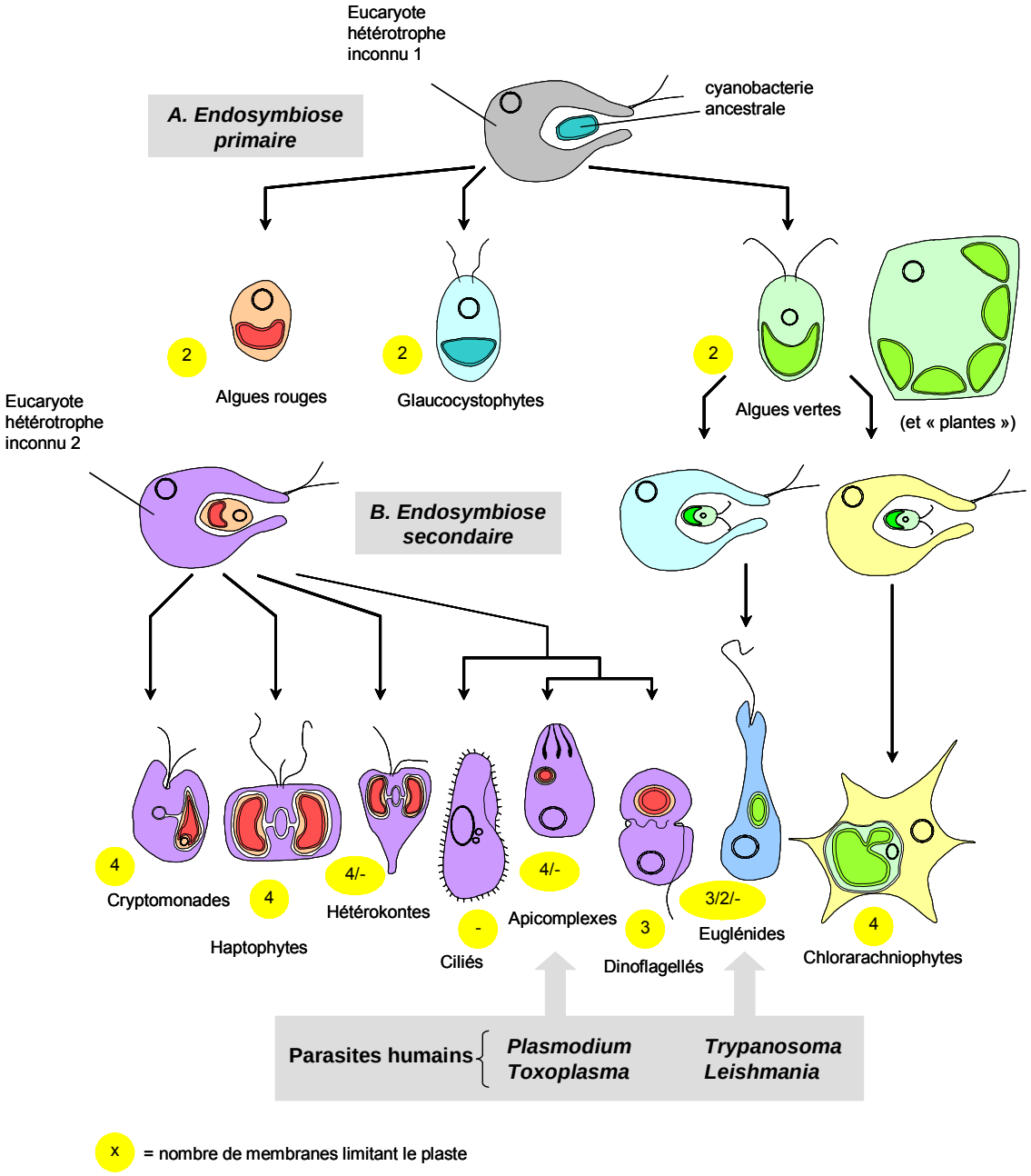


Figure 6. Schéma possible pour l'origine et l'évolution de l'ensemble des eucaryotes contenant des plastes, suivant des événements d'endosymbioses primaire et secondaire. L'endosymbiose est un événement rare dans l'Evolution. **(A) Endosymbiose primaire.** Une endosymbiose primaire unique entre un eucaryote hétérotrophe ancestral inconnu (1, en gris) et une cyanobactérie a conduit aux trois lignées primaires porteuses de plastes. **(A) Endosymbiose secondaire.** Deux autres événements d'endosymbiose secondaire impliquant deux types d'algues vertes différentes et des hôtes non reliés dans l'Evolution, ont conduit aux Euglénides (en bleu) et Chlorarachniophytes (en jaune). Une endosymbiose unique entre une algue rouge et un hôte hétérotrophe a conduit à l'ensemble des autres algues (en violet). La perte de la photosynthèse a eu lieu dans plusieurs taxons de ces groupes. Dans le cas des Ciliés, c'est le phylum entier qui n'est plus photosynthétique. Chez les Ciliés, on ne sait pas s'il reste des plastes cryptiques, comme c'est le cas chez les Apicomplexes. Les parasites apicomplexes tels que *Plasmodium* ou *Toxoplasma* ainsi que les parasites trypanosomatidae tels que *Trypanosoma* ou *Leishmania* sont donc issus d'une endosymbiose secondaire: la partie "végétale" de ces parasites, héritée de l'algue ancestrale, représente une cible pour de nouveaux types de traitements de type herbicide-antiparasitaire. D'après Archibald et Keeling (2002).

Selon ce schéma, l'endosymbiose primaire auraient donné naissance à trois grands groupes d'eucaryotes contenant des plastes simples :

- la ***lignée verte des endosymbiontes primaires*** (Viridiplantae) chez lesquels la chlorophylle *a* est associée à la chlorophylle *b*. Ce groupe rassemble les "algues vertes" (Chlorophyta) telles que *Chlamydomonas reinhardtii* et les "plantes" (Streptophyta) telles qu'*Arabidopsis thaliana*.
- la ***lignée rouge des endosymbiontes primaires*** chez lesquels la chlorophylle *a* est associée à la phycobiline. Ce groupe rassemble uniquement des organismes aquatiques, les "algues rouges" (Rhodophyta) telles que *Cyanidioschyzon merolae*.
- la ***lignée bleue des endosymbiontes primaires***, groupe de très faible biodiversité (Glaucocystophytes, une dizaine d'espèces répertoriées), chez lesquels la chlorophylle *a* est associée à deux pigments bleu-vert, la phycocyanine et l'allophycocyanine. Une des singularités étonnantes des glaucocystophytes, comprenant par exemple *Cyanophora paradoxa*, est la conservation autour du chloroplaste d'une paroi de peptidoglycanes héritée de celle de la cyanobactérie ancestrale. Le chloroplaste, très proche d'une cyanobactérie, est appelé cyanelle.

L'ensemble des organismes ayant hérité un plaste simple selon ce schéma est aujourd'hui rassemblé dans le règne des Archaeplastida (Adl et al., 2005).

Les scénarios d'endosymbioses secondaires, voire tertiaires pour certains groupes de Dinoflagellés, qui ont conduit des eucaryotes à inclure d'autres eucaryotes, sont extrêmement variés et encore discutés. Archibald et Keeling (2002) proposent une vue synthétique résumée sur la **Figure 6B**. Une conséquence importante de l'endosymbiose secondaire est la conservation de la membrane phagocytaire, connectée au flux endomembranaire, autour du procaryote inclus. Le plaste se trouve alors limité par trois ou quatre membranes.

Les protistes issus d'une endosymbiose secondaire seraient divisés en deux lignées selon l'origine de l'endosymbiose primaire qui apporte le plaste :

- la ***lignée verte des endosymbiontes secondaires*** chez lesquels le plaste serait hérité d'une algue verte. Ce groupe comprend d'une part les Euglénides, tels qu'*Euglena gracilis*, contenant un plaste limité de trois membranes. Les parasites tels que *Trypanosoma gambiense*, non photosynthétiques, ont divergé à partir d'un ancêtre commun aux Euglénides. Cette lignée comprend d'autre part les Chlorarachniophytes, tels que *Chlorarachnion reptans*, amibes rhizopodes dont le plaste est entouré de quatre membranes. Comme chez les Cryptomonades, les Chlorarachniophytes ont conservé un vestige du noyau de l'algue verte, le nucléomorphe, entre les deux membranes les plus externes et l'enveloppe du plaste.
- la ***lignée rouge des endosymbiontes secondaires*** chez lesquels le plaste serait hérité d'une algue rouge. Ce groupe très divers rassemble aujourd'hui les Cryptomonades, telles que *Guillardia theta*, qui ont conservé un nucléomorphe, vestige du noyau de l'algue rouge, entre les deux membranes les plus externes et l'enveloppe du plaste ; les Haptophytes, telles qu'*Emiliania huxleyi*, les Hétérokonthes, comprenant des algues

brunes telles que *Laminaria digitata*, ainsi que les oomycètes tels que *Phytophtora infestans* (précédemment classés parmi les champignons) ; les Ciliés, tels que *Paramecia aurelia*; les Dinoflagellés, tels que *Thalassiosira pseudonana*, dont le plaste est limité par trois membranes[5] ; et enfin les Apicomplexes.

Les eucaryotes sont donc issus de plusieurs lignées qui se sont combinées intimement selon plusieurs scénarios d'endosymbiose (**Figure 6A et B**). Du point de vue de la biologie cellulaire, le plaste est donc un organite qui peut être simplement entouré de deux membranes -*plastes primaires*- ou de trois ou quatre membranes -*plastes secondaires*-. La classification des Apicomplexa est donc une question délicate.

D. Question difficile de la classification des apicomplexes

La classification des eucaryotes unicellulaires a été remaniée à de nombreuses reprises (pour revue, Scamardella, 1999). Les unicellulaires ont tout d'abord été simplement séparés en protozoaires (animaux unicellulaires, von Siebold, 1848) et protophytes (végétaux unicellulaires), reflétant à l'échelle des protistes la subdivision proposée par Carl von Linné (1735) des grands règnes (deux règnes du vivant, *regnum Animale*, *regnum Vegetabile* et le règne minéral, *regnum Lapideum*). La création d'un troisième règne vivant propre aux unicellulaires a été proposée par Hoggs (1860), sous la forme du *regnum Primigenum* des "prototisca" (littéralement les "premiers êtres de la création"), concept de classification qui s'accompagne d'une hiérarchisation des unicellulaires comme des êtres archaïques, inférieurs, et précurseurs des organismes pluricellulaires. Cette vision de groupe "inférieur" se retrouve jusque dans les années 1950 (Luyet, 1950). Il est resté courant de présenter les unicellulaires comme des organismes précédant les métazoaires, dans une vision logique de complexification du vivant au cours de l'évolution.

Le naturaliste allemand Ernst Haeckel introduit le terme de *Protista*, signifiant les "êtres primordiaux", attirant l'attention sur le fait que "tous ces organismes de rang inférieurs (...) ne montrent pas d'affinité déterminée d'un côté ou de l'autre, ou possèdent des caractères animaux et végétaux unifiés et mélangés" (Haeckel, 1866). Ce règne contenant aussi les bactéries, est présenté par Haeckel entre les plantes et les animaux (voir **Figure 7**). Une sub-division des protozoaires proposée par Bütschli (1880) en Sarcodina (les organismes amiboïdes), Mastigophora (espèces flagellées), Infusoria (espèces ciliées) et Sporozoa (groupe d'unicellulaires parasites dans lequel *Plasmodium falciparum* est classé) reste encore utilisée, surtout chez les non-protistologues. Selon cette classification, les Apicomplexa sont aussi appelés Sporozoaires, du fait de la division du zygote à l'intérieur d'une structure résistante limitée par une paroi (ce type de division est appelé sporogonie) produisant de très nombreuses cellules infectieuses, appelées sporozoïtes (voir **Figures 2 et 3**). Copeland propose en 1947 de classer les bactéries et les cyanobactéries, être anucléés (procaryotes)

[5] Certains Dinoflagellés contiennent des plastes limités seulement de deux membranes. Des formes parasites non chlorophylliennes de Dinoflagellés existent. Des endosymbioses complexes sont recensées dans certains sous-groupes de Dinoflagellés, avec par exemple échange d'un plaste de lignée rouge contre un plaste de lignée verte.

dans un quatrième règne, celui des Monères (Copeland, 1947). Une décennie plus tard, Whittaker ré-introduit l'idée d'un règne pour les champignons (Whittaker, 1959).

Dans ces classifications, les Apicomplexa sont rassemblés avec les Dinoflagellés et les Ciliés dans le phylum des Alvéolés, c'est-à-dire munis de sacs membranaires ("alvéoles", appelées aussi le "complexe membranaire interne", voir **Figures 3**) sous la membrane plasmique.

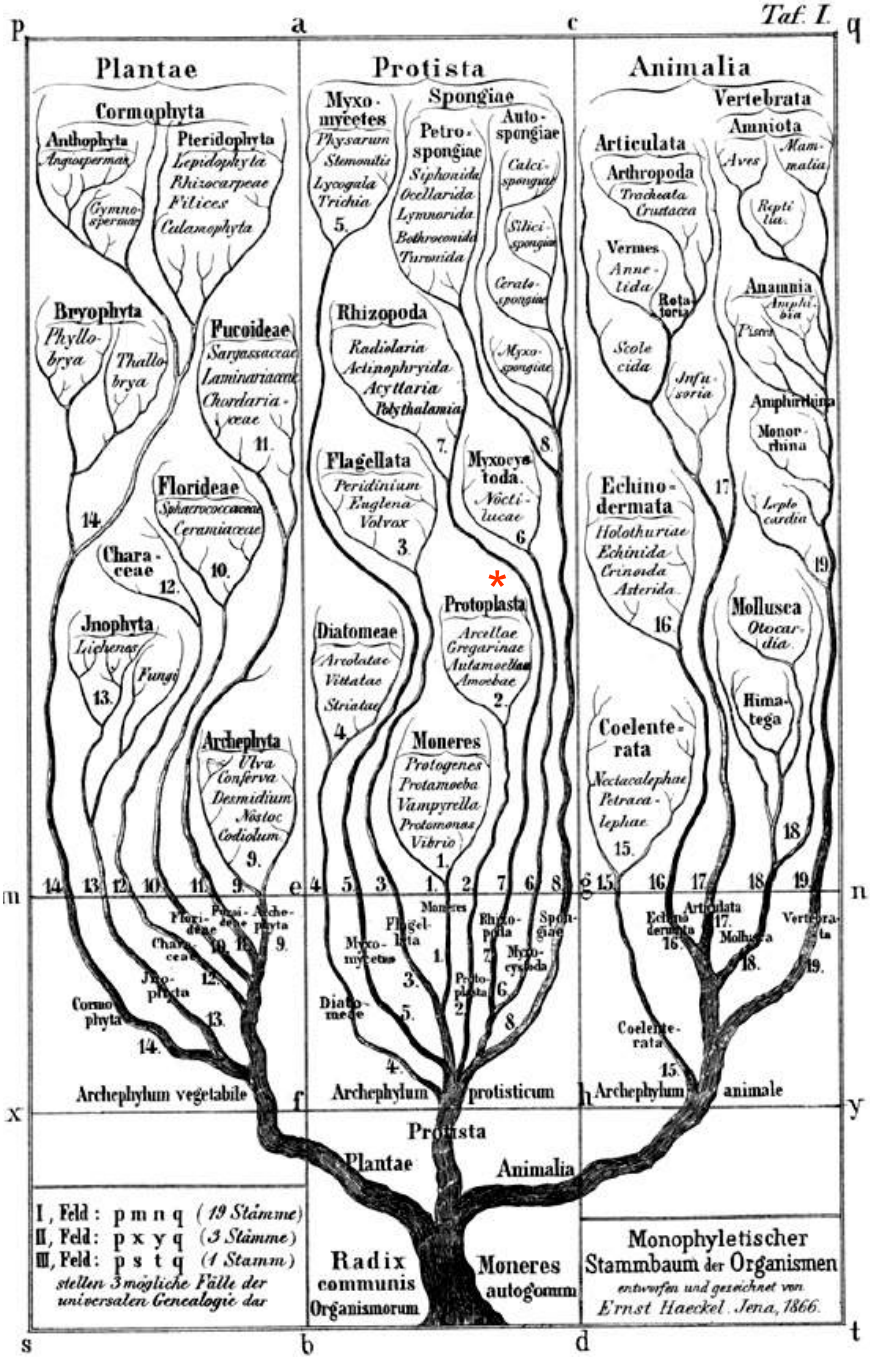


Figure 7: les trois règnes selon Haeckel (1866). Dans cette représentation tirée de Haekel (1866), les protistes sont à la charnière des animaux et des végétaux. Le signe (*) indique le groupe dans lequel certains Apicomplexes sont classés. Cette vision fonde les classifications ultérieures dans lesquelles les protistes sont rassemblés, malgré leur diversité phylogénique.

La suite du débat sur la classification des protistes est caractérisée par l'apport des analyses sub-cellulaires par microscopie électronique, des informations moléculaires ainsi que la prise en compte des concepts de la théorie endosymbiotique et la possibilité de considérer des phylums comprenant à la fois des êtres unicellulaires et pluricellulaires. L'unité du groupe des protistes devient difficile à maintenir. Le botaniste canadien Thomas Cavalier-Smith propose une classification en six règnes, tentant de concilier les données de phylogénie moléculaire et les scénarios d'endosymbiose (Cavalier-Smith, 1998), avec un règne des procaryotes (Bacteria), et cinq règnes eucaryotes (Protozoa, Animalia, Fungi, Plantae et Chromista), dans lesquels se dispersent les protistes (**Figure 8**).

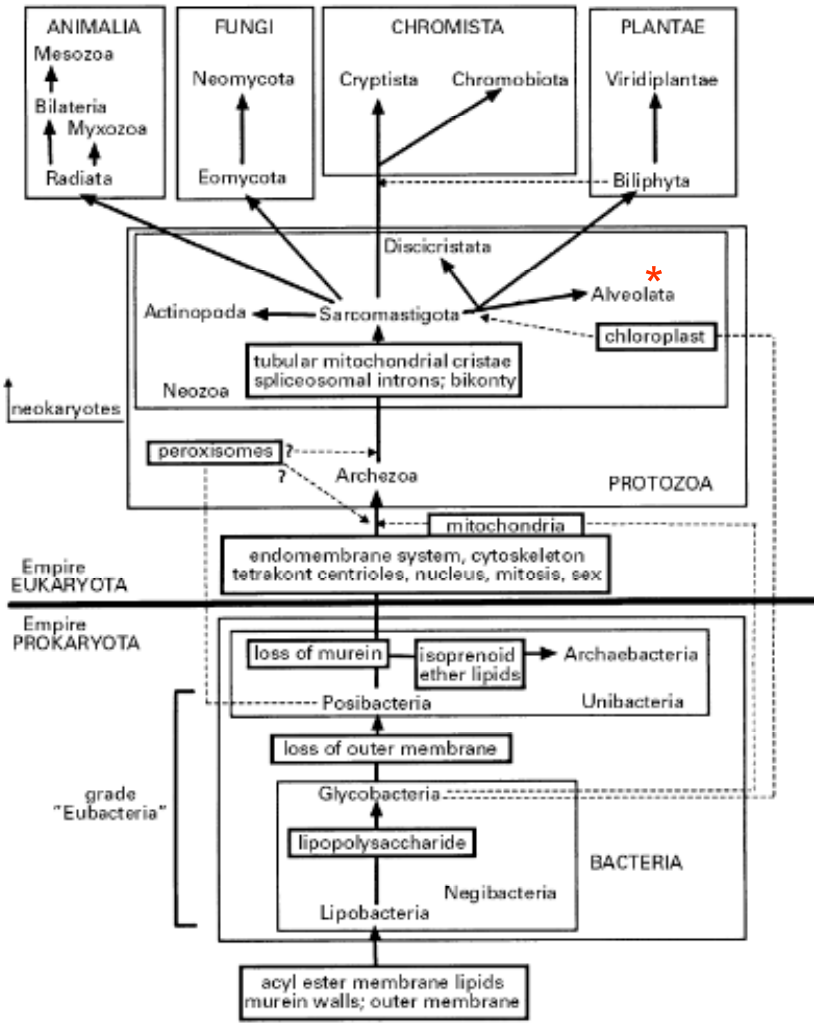


Figure 8: les six règnes selon Cavalier-Smith (1998). Dans cette représentation tirée de Cavalier-Smith (1998), les relations phylogénétiques sont postulées entre les règnes. Les quatre événements symbiogénétiques majeurs dans l'histoire du vivant sont indiqués par des flèches pointillées : (1) l'origine des mitochondries par endosymbiose avec une alpha-protéobactérie, (2) l'origine des peroxisomes à partir d'une posibactérie, (3) l'origine des plastes à partir d'une cyanobactérie et (4) l'origine des Chromistes par une endosymbiose secondaire conduisant à un plaste dans le lumen du réticulum plutôt que dans le cytosol. Le signe (*) indique le groupe dans lequel les Apicomplexes sont classés. Cette vision fonde les classifications ultérieures dans lesquelles une conciliation est recherchée avec les données de phylogénie moléculaire et les scénarios d'endosymbiose.

Une division en 3 domaines du vivant a par ailleurs été avancé par Carl Woese, à partir de données moléculaires (initialement les ARN ribosomiques) séparant Archae

(Archeobactéries), Bacteria (Eubactéries) et Eucarya (tous les eucaryotes) (Woese et al., 1990).

Cavalier-Smith a récemment affiné sa classification et propose un "arbre de la vie" (**Figure 9**, Cavalier-Smith, 2004), représentation synthétique dans laquelle un "eucaryote ancestral" serait un unicellulaire phagotrophe *unicilié*, issu d'un ancêtre bactérien par apparition simultanée du cytosquelette, du système endomembranaire, du noyau, des cils, et ayant réalisé très tôt une endosymbiose avec une alpha-protéobactérie à l'origine des mitochondries. Deux schémas d'organisation des cellules eucaryotes se seraient séparés, d'une part les Unichontes avec un cône de microtubules simples attaché à un unique centriole, et les Bichontes avec deux bandes de microtubules attachés à un centriole postérieur et un éventail de microtubules attaché à un centriole antérieur (**Figure 9**). Les protozoaires comprendraient dans ce schéma des êtres unichontes et des êtres bichontes.

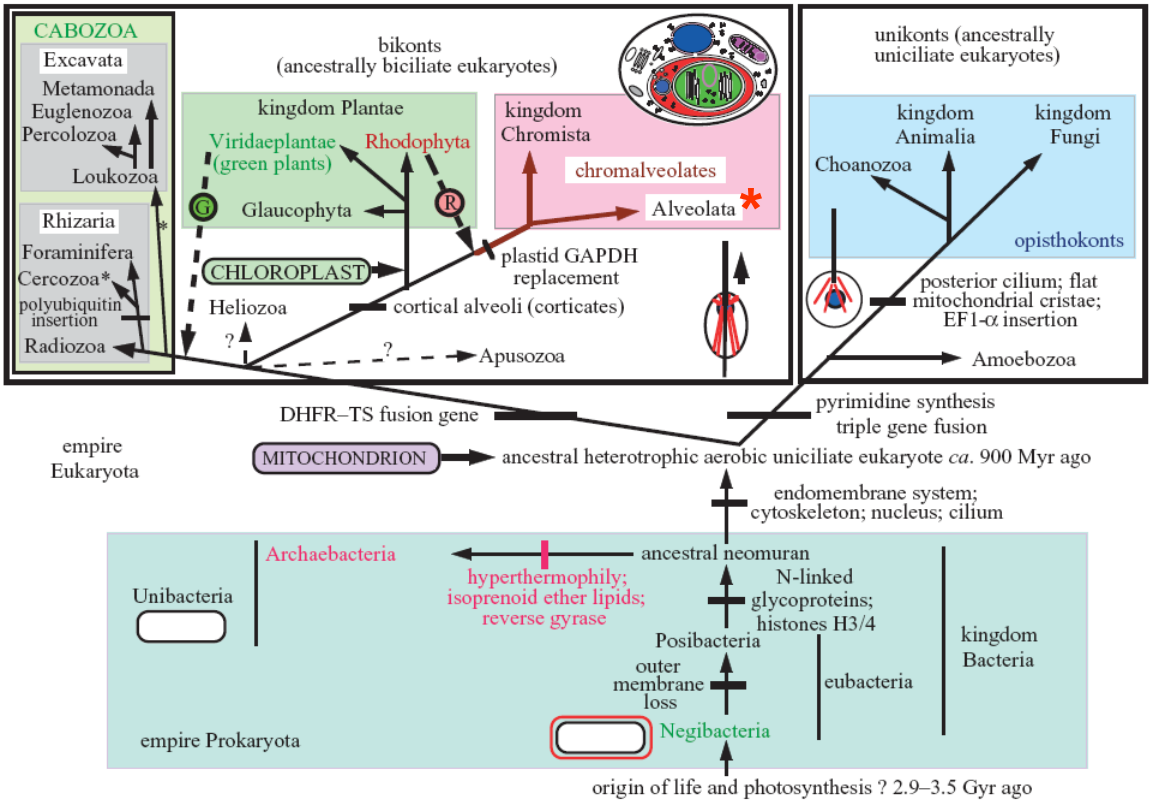


Figure 9: L'arbre de la vie selon Cavalier-Smith (2004). Cet arbre de la vie (Tree of Life), tiré de Cavalier-Smith (2004), est fondé sur des données moléculaires, ultra-structurales et paléontologiques. Contrairement aux suppositions classiques, la "racine" du vivant est proposée au niveau des eubactéries, probablement au niveau des Negibacteria, et non pas entre les eubactéries et les archéobactéries. La position de la "racine" des eucaryotes est proposée entre les Bichontes et les Unichontes. Le règne des Protozoaires n'est pas indiqué pour raison de simplicité de la représentation. Ce règne comprend quatre groupes majeurs (Alveolata, Cabozoa, Amoebozoa et Choanozoa) et le phylum Apusozoa. La figure indique les 4 schémas d'organisation des cellules vivantes (les Unibactéries avec une simple membrane plasmique, les Negibactéries limitées par deux membranes, les Unichontes avec un cône de microtubules simples attaché à un unique centriole, les Bichontes avec deux bandes de microtubules attachés à un centriole postérieur et un éventail de microtubules attaché à un centriole antérieur) ainsi que la cellule eucaryote la plus complexe dans le monde vivant, c'est-à-dire celle des Cryptophytes.

La classification s'est donc extrêmement sophistiquée et complexifiée et la vision des Protozoaires obscurcie. Alastair Simpson et Andrew Roger ont récemment proposé une représentation de ce qu'ils appellent les "vrais règnes des eukaryotes", obtenus en privilégiant les données moléculaires, en particulier des fusions de gènes marqueurs phylogénétiques. Cette classification se présente sous la forme d'un arbre séparant six groupes majeurs, Opisthokonta (rassemblant les unichontes monophylétiques), Amoebozoa, Plantae, Excavata, Rhizaria et Chromalveolata (groupe dans lequel se situent les Apicomplexa) (**Figure 10**).

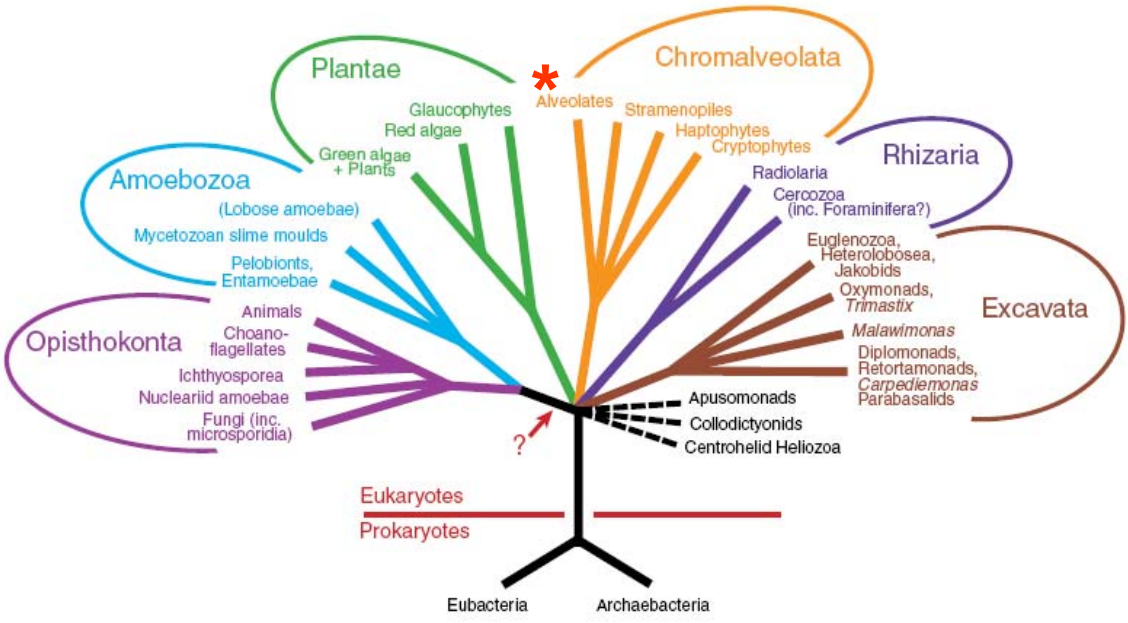


Figure 10: Les "vrais" règnes des Eucaryotes selon Simpson et Roger (2004). Ce schéma proposé par Simpson et Roger (2004), privilégie les informations moléculaires obtenues à partir d'un ensemble de séquences considérées comme des marqueurs phylogénétiques. La flèche indique la position possible de la racine, fondée sur l'analyse de fusion de gènes marqueurs. Le signe (*) indique le groupe dans lequel les Apicomplexes sont classés (dans le groupe des Alvéolés).

La nouvelle classification des eucaryotes, commandée par l'International Society of Protistologists, sous la coordination de Sina Adl (Adl et al., 2005) a été adoptée selon un schéma général dérivé des propositions de Cavalier-Smith et Simpson et Roger. Dans cette vision, les Protistes sont divisés en 4 règnes (*Amoebozoa*, à déplacement amiboïde à pseudopodes de forme variable ; *Rhizaria*, à pseudopodes fins ; *Chromalveolata* issus d'une endosymbiose secondaire avec un Archaeplastida ; *Excavata*, dotés d'un sillon nutritif) sur un total de 6 ! Les animaux n'ont plus le statut de règne, associés aux champignons vrais dans le règne des Opistochonthes (cellules dont le schéma d'organisation présente un cil simple). Les organismes contenant des plastes simples sont classés dans le règne des Archaeplastida (équivalent aux plantes, algues vertes, rouges et Glaucocystophytes). Le débat a été particulièrement vif sur la pertinence de séparer les groupes Chromalveolata et Excavata, et sur le positionnement des Apicomplexes dans le règne des Chromalveolata (Simonite, 2005).

Dans la nouvelle classification le phylum des Apicomplexa comprend uniquement des êtres parasitaires à l'exception des Colpodellida. A ce jour, la classification officielle du parasite malarial est la suivante: un Apicomplexa Aconoidasida (complexe apical incomplet -

conoïde absent- aux stades asexués mobiles) Haemosporida (zygote mobile avec conoïde, microgametes flagellés produits par schizogonie) (Adl et al., 2005).

E. Peut-on exploiter la face végétale de *Plasmodium falciparum* pour développer de nouveaux traitements thérapeutiques ?

La généalogie des apicomplexes reflète des héritages différents, suivant que les gènes ont été

- hérités de l'algue rouge qui a été ingérée au cours de l'endosymbiose secondaire
- ou hérités de l'eucaryote secondaire.

Suite aux événements symbiogénétiques primaires, les gènes de l'α-protéobactérie, ancêtre des mitochondries, et de la cyanobactérie, ancêtre des plastes, ont dans leur majorité disparu des génomes des organites. Une partie de ces gènes initiaux se retrouve dans le génome nucléaire des cellules eucaryotes. Par exemple, l'équipe de William Martin (Rujan et Martin, 2001; Martin et al., 2002; Martin 2003) a déterminé qu'environ 18% du génome d'*Arabidopsis* était d'origine cyanobactériale. Les protéines plastidiales codées par le noyau sont dotées de séquences d'adressage leur permettant d'être transférées vers l'organite. Certaines protéines héritées de la cyanobactérie ancestrale remplissent des fonctions analogues à celles documentées chez les procaryotes, d'autres remplissent de nouvelles fonctions propres à la cellule eucaryote.

Suite à l'endosymbiose secondaire qui a donné naissance au phylum des Apicomplexa, de nombreux gènes sont attendus dans le génome nucléaire avec un héritage de la cyanobactérie ancestrale (pour certaines fonctions conservées de l'apicoplaste), ainsi que du génome nucléaire de l'algue rouge.

Nous appelons dans ce mémoire l'ensemble de ce sous-génome algal, la "face végétale".

La caractéristique végétale la plus spectaculaire étant l'apicoplaste, il a été très vite proposé que cet organite soit une cible tout à fait intéressante pour de nouveaux traitements thérapeutiques de type herbicide (Fichera et Roos, 1997; pour revue McFadden et Roos, 1999; Maréchal et Cesbron-Delauw, 2001; Wiesner et Seeber, 2005; Bisanz et al., 2006). Le **Tableau 1** présente un ensemble de molécules thérapeutiques ciblant la face végétale de *Plasmodium* au niveau des processus ayant lieu dans l'apicoplaste.

Cibles générales	Molécules actives	Cibles spécifiques dans l'apicoplaste
Machinerie de transcription / traduction	lincosamides (clindamycine); macrolides (azithromycine); thiopeptides (thiostrepton, miccrococcine)	Initiation de la transcription
	chloramphénicol; kirromycine; amythiamicie A GE2270A; encyloxine IIa	Elongation de la traduction, facteur EF-Tu
	rifampicine, rifabutine	Sous-unité β de la RNA polymérase
	tétracycline	ribosome 70S
Réplication de l'ADN	quinolones et fluoroquinolones (ciprofloxacine, norfloxacine, trovofloxacine)	ADN gyrase
Synthèse des acides gras (acide gras synthase de type II)	herbicides aryloxyphénoxypropionate (fop) (haloxyfop, clodinafop, quizlofop, diclofop); fenoxaprop; tralkoxydim	Acétyl-CoA carboxylase (ACCase)
	thiolactomycine et analogues	β-cétoacyl ACP synthase II (FabF) et β-céto-ACP synthase III (FabH)
	cérulenine	
	triclosan, isoniazid	β-cétoacyl ACP synthase II (FabF)
		Enoyl-ACP réductase (FabI)
Synthèse des isoprénoïdes	fosmidomycine, FR-900098	DOXP réductoisomérase
Peptide déformylase	actinonine	Peptide déformylase

Tableau 1: Molécules potentiellement actives au niveau de processus essentiels se déroulant dans l'apicoplaste, ayant des propriétés antiparasitaires *in vitro* et/ou *in vivo*. D'après Wiesner et Seeber, 2005**.**

Les processus végétaux hérités de l'algue rouge ne se réduisent pas aux processus plastidiaux. De nombreuses autres voies biochimiques caractéristiques des plantes ont par ailleurs été mises en évidence (métabolisme du folate, transduction de signaux par les kinases CDPK, etc. pour revue, Maréchal et Cesbron-Delauw, 2001).

Lorsque j'ai débuté les travaux de thèse présentés dans ce mémoire, le génome malarial n'était pas séquencé, et il se posait alors la question de l'inventaire des gènes d'origine algale. Un certain nombre de difficultés apparaissaient comme des obstacles à l'établissement de cet inventaire, liées à une carence d'annotation de la majorité des gènes identifiés dans le génome malarial. Par exemple, dans le contexte des voies métaboliques lipidiques d'origine végétale chez le parasite, alors que Maréchal et al. (2002) avaient mesuré au laboratoire la synthèse de galactolipides chloroplastiques à partir de suspensions de cellules de *Plasmodium falciparum*, aucun homologue de gènes codant pour des enzymes de synthèses de galactolipides ne pouvait être détecté. L'absence apparente d'homologues de gènes codant pour des fonctions pour lesquelles on a par ailleurs des mesures expérimentales est documentée pour l'ensemble du métabolisme (McConkey et al., 2004). Nous avons supposé que l'échec de l'identification de nombreuses séquences homologues à des séquences connues, dans le génome de *Plasmodium falciparum* était lié à certaines singularités que présente ce génome.

III. Les singularités du génome de *Plasmodium falciparum*

A. Présentation générale du génome

Le génome de la souche 3D7 du parasite *Plasmodium falciparum* a été séquencé en 2002 (Gardner et al., 2002). Le génome de *Plasmodium falciparum* possède une taille de 23 mégabases (MB) et est organisé en 14 chromosomes classés par ordre croissant selon la taille, de 0.7MB à 3.4MB (**Tableau 2**).

La détection des gènes dans une séquence nucléique repose sur deux classes de méthodes complémentaires. La première comprend les méthodes de détection d'ORFs (*open reading frames* ou cadres ouverts de lecture) qui reposent sur certaines propriétés statistiques de l'ADN au niveau des régions codantes : la longueur des ORFs, la fréquence des codons, la composition en G+C, la présence de zones potentielles d'épissage, etc. Bien qu'elles soient dites *ab initio*, les méthodes de détection d'ORFs requièrent tout de même un jeu de gènes déjà caractérisés, sur lesquelles « apprendre » les critères qui déterminent la présence possible d'un gène. Dans le cas du génome de *Plasmodium*, *falciparum*, la méthode Glimmer a été corrigée afin de tenir compte de la structure particulière des gènes chez le parasite (Salszber et al., 1999). La deuxième classe comprend les méthodes comparatives dont l'objectif est de transférer les connaissances que l'on a acquises sur des séquences géniques (ou protéiques) précédemment caractérisées et répertoriées sur des séquences inconnues que l'on estime suffisamment proches (*homology based annotation*, pour l'annotation fondée sur l'homologie). La fiabilité de ces méthodes comparatives reposent 1) sur l'information partagée entre les séquences à annoter et 2) les connaissances (souvent de nature expérimentale) que l'on possède sur l'ensemble de séquences similaires, de génomes d'organismes plus ou moins proches dans l'évolution, disponibles dans les bases de données publiques. Les méthodes comparatives sont plus efficaces quand il existe des génomes proches de celui que l'on cherche à annoter déjà répertoriés dans les bases de données (McConkey et al., 2004).

L'application d'une combinaison de méthodes de détection d'ORFs, et de méthodes d'annotation fondée sur l'homologie a permis d'aboutir à la prédiction d'environ 5300 gènes chez *Plasmodium falciparum* (**Tableau 2**). Ce nombre de gènes est assez faible, proche de celui observé dans le génome de la levure. L'annotation du génome malarial, initiée par le consortium en charge du séquençage, s'est poursuivie grâce à la contribution de la communauté scientifique et est régulièrement mise à jour sur le site PlasmoDB[6], (version 5 en Février 2006, en relation avec le site ApiDB intégrant les informations génomiques pour d'autres Apicomplexes tels que ToxoDB, base de données génomiques de *Toxoplasma gondii*). Le site PlasmoDB permet d'accéder à l'ensemble des séquences géniques et à leurs identifiants, aux inférences fonctionnelles putatives ou établies pour chaque gène, et à un ensemble d'informations telles que l'expression des transcrits à différents stades parasitaires, la présence de domaines et motifs selon différentes méthodes de prédictions, etc.

[6] http://plasmodb.org/

	Plasmodium falciparum	*Saccharomyces cerevisiae*	*Arabidopsis thaliana*	*Homo sapiens*
Statistiques générales du génome				
Nombre de chromosomes	14	16	5	22 + X/Y
Estimation de la taille (bp)	22.853.764	12.495.682	115.409.949	3,272,187,692
%G+C moyen	19.4	38.3	34.9	41.0
Estimation du nombre de gènes	5268	5770	25498	22287
Longueur moyenne d'un gène	2283	1424	1310	1340
% génome codant	53	66	29	9
Annotation automatique fondée sur la similarité (homology based annotation)				
% de gènes présentant une similarité avec des séquences d'autres organismes (de fonction connue au moins partiellement) lors du séquençage initial	39 %	75 %	69 %	59 %
% de gènes ne présentant pas de similarité détectable avec des séquences d'autres organismes (de fonction connue ou non) lors du séquençage initial ("no BLASTP match to known proteins")	> 45 %	< 8%	< 20 %	<26 %
% de gènes de fonction totalement inconnue (= gènes présentant une similarité avec d'autres séquences de fonction inconnue + gènes sans similarité détectable avec des séquences d'autres organismes).	61 %	16 %	31 %	41 %
Caractéristiques moyennes des cadres de lecture				
Exons				
Nombre par gène	2.39	1.05	5.18	12.1
%(G+C)	23.7	40	45	48
Longueur moyenne	949	1356	253	111
Introns				
%(G+C)	13.5	36	34	40
Régions intergéniques				
%G+C	13.6	36	34	40

Tableau 2: Comparaison des génomes nucléaires de *Plasmodium falciparum*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana* et *Homo sapiens*. Les données présentées compilent des informations de Gardner et al. (2002) pour *Plasmodium falciparum*, Wood et al. (2001) pour la levure (complétées par les données accessibles via internet sur le site de la Comprehensive Yeast Genome Database[7]), The Arabidopsis genome initiative (2000) pour l'arabette, et International Human Genome Sequencing Consortium (2001) et Venter et al. (2001) pour l'humain (complétées par les données statistiques disponibles sur le site Ensembl[8]). Ces informations sont un état des lieux à la date de collecte des informations et sont continuellement mises à jour (en particulier en ce qui concerne le nombre de gènes).

[7] http://mips.gsf.de/proj/yeast/tables/

[8] http://www.ensembl.org/index.html

Lorsqu'elles sont comparées à des séquences issues d'organismes non-apicomplexes, la majorité des gènes annotés montre une plus forte similarité avec leurs homologues d'*Arabidopsis thaliana* (Gardner et al., 2002). Il est vraisemblable que 1) cette affinité végétale soit due à l'héritage algal dans le génome malarial, suite à l'événement d'endosymbiose secondaire décrit plus haut et 2) que le génome d'*Arabidopsis thaliana* soit apparu comme le plus proche car il était le seul génome complet de plante à l'époque et qu'aucun génome complet d'algue n'était disponible. Sur les ~5300 protéines codées par le génome malarial, 60 % n'ont aucune fonction assignée, même partielle. Ce cas est unique parmi les grands génomes eucaryotes séquencés à ce jour (**Tableau 2**). Trois raisons peuvent expliquer la non-annotation d'une séquence :

- soit la séquence est unique au parasite,
- soit la séquence est similaire à d'autres séquences de protéines répertoriées dans les bases de données publiques mais sans fonctions connues,
- soit la séquence ne présente aucune similarité avec aucune autre séquence précédemment répertoriée. Cette situation est attendue pour des protéines qui ont tellement divergé qu'il est impossible de retrouver leur famille d'appartenance par des méthodes comparatives (Gardner et al., 2002).

Nous avons supposé, tout comme d'autres équipes (McConkey et al., 2004; Callebaut et al., 2005), que le petit génome de *Plasmodium falciparum* n'était pas original au point d'avoir les deux tiers de son protéome qui lui soient uniques. McConkey et al. (2004) remarquent que seulement 8 % des protéines pour lesquelles il est possible d'attribuer une fonction seraient impliquées dans le métabolisme, pourcentage bien trop faible par comparaison avec les autres génomes de taille semblable (la levure présente 17% de protéines impliquées dans le métabolisme) et en regard des mesures expérimentales de voies métaboliques actives chez le parasite malarial.

Parmi les causes possibles 1) de la divergence extrême des protéines malariales et 2) de la mise en échec des méthodes automatiques de détection de similarité, la formidable richesse en A+T du génome de *Plasmodium falciparum* (82 % A+T) par rapport aux autres génomes (**Tableau 2**) a été avancée très tôt (Gardner et al., 2002).

B. Les biais compositionnels aux niveaux nucléiques et protéiques

Un paramètre important pour les méthodes de comparaisons de séquence est la distribution en acides aminés (voir chapitre bibliographique). L'influence du pourcentage en guanine et cytosine (%G+C, appelé dans la littérature "GC content") sur la composition en acide aminé fait l'objet d'une littérature abondante. Lobry (1997) a montré l'influence du %G+C en partant de 23490 protéines de 59 espèces bactériennes, au niveau des génomes entiers. L'évolution de la composition des séquences protéiques en fonction de l'augmentation du %G+C génomique, se caractérise par une diminution de la fréquence des acides aminés Ile, Phe, Lys, Tyr, Asn et Leu (I, F, K, Y, N et L en code 1 lettre, voir Rappels bibliographiques, **Tableau 3**) et une augmentation de la fréquence des acides aminés Gly, Pro, Ala et Arg (en code 1 lettre : G, P, A et R). Lobry (1997) a de plus montré que cette évolution de la

distribution des amino acides dans les protéines en fonction d'un %G+C croissant, était partiellement compensée par un biais dans l'usage de codons, représentant une forme de pression de conservation. Foster et al. (1997) ont confirmé ces observations pour les protéines codées par les mitochondries des organismes eucaryotes supérieurs. Foster et al. (1997) introduisent le formalisme résumé sur la **Figure 11** : ne considérant que les deux premières positions de codons dans les ORFs, les codons riches en G+C sont ceux qui codent majoritairement les acides aminés G, A, R et P, alors que les codons riches en A+T codent les acides aminés F, Y, M, I, N et K. L'acide aminé L est exclu car il est sur-représenté en raison de l'abondance des différents codons qui le codent.

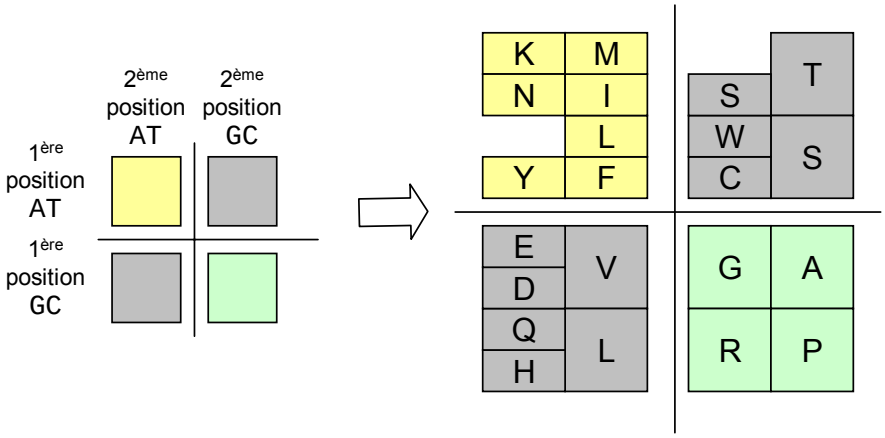


Figure 11 : partitionnement de la table des codons en fonction de la composition en A, T, C et G des premières et deuxièmes positions de codons. Cette partition implique une partition de l'ensemble des acides aminés, selon la répartition des codons qui sont à leurs origines.

Par convention les acides aminés caractérisés par la richesse significative de leurs codons en A+T sont dits appartenir à la série FYMINK, les acides aminés de la série GARP sont quant à eux caractérisé par la richesse de leurs codons en G+C. Nishizawa et Nishizawa (1998) ont montré que l'usage de Lysine (L) ou d'Arginine (R) dans les séquences était influencé par les fluctuations de %G+C à l'échelle des génomes de *Saccharomyces cerevisiae* et *Caenorhabditis elegans*. En se basant sur le formalisme de Foster et al. (1997), Singer et Hickey (2000) ont montré que le biais de composition en amino acides (richesse en FYMINK et appauvrissement en GARP quand le %G+C diminue) était aussi observé dans 21 génomes archéabactériens et bactériens. A partir de couples de séquence homologues de *Saccharomyces cerevisiae* et *Plasmodium falciparum* (des chromosomes 2 et 3 disponibles à cette époque), ces auteurs ont établi que le biais de composition en acides aminés semblait augmenter avec le pourcentage de divergence des séquences. Singer et Hickey (2000) ont également observé une différence de composition en G+C entre les positions de codons 1 et 2 (appelées positions non-synonymes) et la position 3 (appelée synonyme). Ce dernier résultat suggère que le biais est d'origine nucléique et non pas imposée par la nature de la fonction de la protéine, bien qu'il soit possible que l'environnement affecte la composition en nucléotides des génomes (Tekaia et al., 2002; Foerstner et al., 2005).

En ce qui concerne plus spécifiquement *Plasmodium falciparum*, Musto et al. (1995) ont montré, d'une part, que le taux en A+T était plus important sur la troisième position de codons, et d'autre part, que le biais était indépendant de la classe de fonction des protéines. En

effet, les gènes de ménage comme ceux des antigènes de surface sont soumis à la même contrainte de composition, bien qu'ils ne soient pas soumis à la même pression évolutive. Musto et al. (1995) ont postulé que ces observations étaient vraisemblablement générales dans le monde vivant, en particulier transférables au génome de *Staphylococcus aureus*, également extrêmement pauvre en G+C.

C. L'abondance de répétitions de faible complexité

Le génome de *Plasmodium falciparum* présente une singularité supplémentaire qui est une richesse extrême d'inserts de faible complexité, à l'intérieur des séquences protéiques (longs segments répétant un seul acide aminé, répétition de courts motifs, etc.) (exemple **Figure 12**). La caractérisation de ces segments et de leur déterminisme est un champ d'étude encore peu exploré.

```
Homme          SIVAIASYTGYNQEDSVIMNRSAVDRGFFRSVFYRSYKEQESKKGFDQEEVFEKPTRETC
Souris         SIVAIASYTGYNQEDSVIMNRSAVDRGFFRSVFYRSYKEQESKKGFDQEEVFEKPTRETC
Arabidopsis    AIVAISCYSGYNQEDSVIMNQSSIDRGFFRSLFFRSYRDEEKKMGTLVKEDFGRPDRGST
Pombe          AIVAILCYSGYNQEDSIIMNQASIDRGLFRSIFYRTYTDQEKKIGMTVMEEFERPVRSTT
Plasmodium     AIVAIMCYTGYNQEDSLIMNQSSIDRGLFRSVFYRTYTSEEKQQGSLIIESFEKPSVRVV
               :**** .*:*******:***::::***:***:*:*:* .:*.: *    * * :*

Homme          QGMRHAIYDKLDDDGLIAPGVRVSGDDVIIGKTVTLPENED----------------------
Souris         QGMRHAIYEKLDDDGLIAPGVRVSGDDVIIGKTVTLPENED----------------------
Arabidopsis    MGMRHGSYDKLDDDGLAPPGTRVSGEDVIIGKTTPISQDEA----------------------
Pombe          LRMKHGTYDKLEDDGLIAPGTRVSGEDIIIGKTAPIPLDHE----------------------
Plasmodium     KNLKRGDYTKLDDDGLIAPGIRVLGDDIIIGKVSPNIDDEDDIIIEKRNTSSSSIQIYNK
                 :::. * **:**** .** ** *:*:****. .   :.

Homme          ------------------------------------------------------------
Souris         ------------------------------------------------------------
Arabidopsis    ------------------------------------------------------------
Pombe          ------------------------------------------------------------
Plasmodium     DSISNNNSNNSNNNMNNMSNMSNMSNIRSSISSNLSFSSNIGSSNVLDTLPDSPINNTYN


Homme          -----ELESTNRRYTKR-------------------------------------------
Souris         -----ELESTNRRYTKR-------------------------------------------
Arabidopsis    -----QGQSS--RYTRR-------------------------------------------
Pombe          -----ELGQRTQLHAKR-------------------------------------------
Plasmodium     NNNNININSSSNNYSLHGAASVTSSTPSSTTIFSSGQTAGSSNSNTKYGTTIVSSTKDDT
                    :  .    :: :


Homme          -----------------DCSTFLRTSETGIVDQVMVTLNQEGYKFCKIRVRSVRIPQIG
Souris         -----------------DCSTFLRTSETGIVDQVMVTLNQEGYKFCKIRVRSVRIPQIG
Arabidopsis    -----------------DHSISLRHSETGMVDQVLLTTNADGLRFVKVRVRSVRIPQIG
Pombe          -----------------DVSTPLRSTESGIVDQVMVTTNQEGLKFVKVRMRSTRIPQIG
Plasmodium     EIPTLTISSTNVLKQYKKDCSLSLRSNENGVIDTVMLSSNSRGNKFAKVKVRSVRIPQIG
                                * *  ** .*.*::* *::: *  * :* *:::**.******


Homme          DKFASRHGQKGTCGIQYRQEDMPFTCEGITPDIIINPHAIPSRMTIGHLIECLQGKVSAN
Souris         DKFASRHGQKGTCGIQYRQEDMPFTCEGITPDIIINPHAIPSRMTIGHLIECLQGKVSAN
Arabidopsis    DKFSSRHGQKGTVGMTYTQEDMPWTIEGVTPDIIVNPHAIPSRMTIGQLIECIMGKVAAH
Pombe          DKFASRHGQKGTIGMTYRHEDMPFSAQGIVPDIIINPHAIPSRMTVAHLVECQLSKVSAL
Plasmodium     DKFASRHGQKGTIGITYRTEDMPFSSLGIFPDIIMNPHAVPSRMTIGHLVECLTGKVAAI
               ***:******** *: *  ****::  *: ****:****:*****:.:*:**  .**:*
```

Figure 22 : Exemple d'insert de faible complexité caractérisant les gènes de *Plasmodium falciparum*. Alignement de cinq secondes sous-unités de l'ARN polymérase de type II de *Plasmodium falciparum*, *Saccharomyces pombe*, *Arabidopsis thaliana*, *Mus musculus*, et *Homo sapiens*. Des blocks conservés sont interrompus par de longs segments de faible complexité.

Hormis l'enrichissement en amino acides de la série FYMINK (voir ci-dessus), le biais en A+T conduit à une sur-représentation des acides aminés hydrophiles (Verra et Hughes, 1999). Il serait donc attendu, dans l'hypothèse d'une corrélation avec le biais compositionnel que ces segments soient enrichis en amino acides hydrophiles. Dans le cas des antigènes de surface, la composition en acides aminés des séquences répétées semble toutefois différente de celle du reste du génome (Verra and Hughes, 1999), conduisant à une sous-représentation des acides aminés hydrophobes. Les segments protéiques de faible complexité chez *Plasmodium falciparum* sont souvent des inserts dont on ne connaît pas la fonction (il a été supposé que ceux-ci pouvaient avoir un rôle dans l'évasion à la réponse immunitaire (Verra and Hughes, 1999), sans que cela soit étayé plus tard).

La présence de segments de faible complexité semble plus importante dans les régions génomiques (et protéiques) ou la *pression de conservation fonctionnelle* est la moins forte (Pizzi and Frontali, 2001; Brocchieri, 2001). Se basant sur le fait que les segments de faible complexité nucléique sont autant présents dans les régions introniques que dans les exons, Xue and Forsdyke (2003) ont émis l'hypothèse que ces segments étaient peut-être déterminés par une pression exercée au niveau nucléique.

Cette singularité du génome de *Plasmodium falciparum* conduit à une augmentation de la taille des séquences de 10 à 20% par rapport aux séquences homologues d'autres organismes (Brocchieri, 2001; Xu et al., 2004). Outre le biais compositionnel, la présence d'inserts de faible complexité dans les gènes de *Plasmodium falciparum* est supposée rendre difficile la recherche de similarités entre ces séquences et celles d'autres organismes.

IV. Comment progresser dans la caractérisation génomique de *Plasmodium falciparum* en vue de contribuer à la lutte contre le paludisme ?

Comment résoudre l'énigme des gènes de fonctions inconnues dans le génome malarial ? Comment repérer ceux qui sont hérités de l'algue ancestrale impliquée dans l'événement endosymbiogénétique à l'origine des structures végétales du parasite ? Nous avons postulé, comme d'autres équipes, que les méthodes classiques d'analyses de séquences pouvaient être en limite de validité lorsque des séquences de compositions extrêmes étaient étudiées. Nous avons examiné dans ce contexte les méthodologies de comparaisons de séquences, en particulier les méthodes d'alignement deux à deux et les modèles statistiques correspondants (chapitre bibliographique).

Les travaux présentés dans ce mémoire comprennent :

- une étude fouillée des statistiques d'alignement, en particulier suivant le modèle de la *Z-value* (chapitre 1 des résultats),
- un ensemble de développements théoriques sur la comparaison des séquences dans un contexte statistique et phylogénique robuste lorsqu'on compare des séquences biaisées ou non (chapitre 2 des résultats),
- une recherche du déterminisme de la *Z-value* en liaison avec certains concepts de la biologie des systèmes (chapitre 3 des résultats),

- une analyse comparative des protéomes de *Plasmodium falciparum* et d'*Arabidopsis thaliana*, et un examen du rôle des matrices de substitution lors de l'alignement de séquences biaisées avec des séquences homologues non biaisées.

L'ensemble de ces travaux nous permet d'une part de proposer un ensemble d'améliorations des méthodologies existantes appliquées à l'analyse comparative de génomes biaisés. D'autre part, nos développements théoriques nous ont conduit à élaborer un modèle unificateur de concepts d'analyse de séquences, de phylogénie et de biologie des systèmes qui n'étaient pas encore corrélés théoriquement à ce jour.

Partie bibliographique

Partie bibliographique :
La comparaison de séquences protéiques

I. Introduction

En 2006, plus de 3 millions de séquences protéiques ont été obtenues par traduction automatique de séquences génomiques (plus de 100 milliards de bases d'ADN séquencées en 2006[1]) et ont été stockées dans les bases de données électroniques publiques. Les efforts collaboratifs de nombreux laboratoires et les progrès des technologies pour la génomique ont permis d'aboutir au séquençage de plus de 300 génomes et d'annoncer le séquençage de génomes de centaines d'autres espèces. La production de séquences protéiques est exponentielle et s'accélérera prochainement, grâce à une avancée technologique rapportée par Margulies et al. (2005), permettant un débit 100 fois plus élevé par rapport aux méthodes de séquençage classiques. Cette nouvelle technique permet par exemple de lire 25 millions de bases d'ADN (le génome complet de certains champignons), en quelques heures. La majorité des 3 millions de séquences protéiques actuellement stockées, et les millions à venir, ne seront pas confirmées expérimentalement, ni leurs fonctions analysées directement. La réduction de la diversité et de la complexité de ces grandes masses de données moléculaires, et pour plus de 90% de ces séquences, notre faculté de prédire des propriétés biologiques pertinentes, reposent donc sur des procédures d'analyses *in silico* et sur leur fiabilité (Ofran et al., 2005; Bastien et al., soumis).

L'information biologique très riche contenue dans une séquence protéique est issue, *en amont*, du déterminisme de l'enchaînement des acides aminés par traduction simple de segments génomiques (ADN, ARN) reflétant nécessairement une histoire d'événements génétiques d'un organisme ou d'une espèce, et *en aval*, de la contribution apportée par les amino acides à la fonction moléculaire de la protéine, dans une multitude de contextes physiologiques qui caractérisent les êtres vivants (Wu et al., 1974). Au cours de sa biosynthèse, une protéine est d'abord une succession d'acides aminés qui se replie spontanément (Anfinsen, 1973). Ce repliement peut se dérouler avec l'assistance de protéines chaperonnes, mais ce processus d'auto-assemblage peut être reconstitué *in vitro* par renaturation en absence de chaperonne, même pour des protéines membranaires (*e.g.* Nishyama et al., 2001). Un des premiers fondements de l'analyse des séquences repose donc sur le fait que l'enchaînement des résidus, ici les amino acides, contient les informations nécessaires et suffisantes à la très complexe structuration tridimensionnelle des protéines et de ce fait à leurs fonctions moléculaires (Wu et al., 1974).

---

[1] Ces données ont été compilées à partir des statistiques mises à disposition sur les sites internet de GenBank (http://www.ncbi.nlm.nih.gov/Genbank/index.html) et UniProt (http://www.ebi.uniprot.org/index.shtml) et en intégrant les informations des sites de séquençage des grands génomes procaryotes et eucaryotes.

Un postulat important de la théorie synthétique de l'évolution des systèmes vivants est que le processus de diversification a lieu au niveau de l'*information génétique* par mutations et recombinaisons (Mayr, 1964; 1997). Les séquences moléculaires de divers individus ou diverses espèces, qui ont évolué à partir d'une séquence ancestrale commune, se ressemblent structuralement. Dans cette expression néo-darwinienne, le maintien plus ou moins important de propriétés structurales, et donc la *ressemblance* plus ou moins grande que nous observons finalement, dépend non seulement des mutations/recombinaisons qui caractérisent la généalogie de chacune des séquences, mais aussi de la pression évolutive, exercée sur ces séquences dans un sens qui maintient une fonctionnalité biologique.

Dans ce mémoire, nous reprenons les définitions proposées par Wu et al (1974) pour les séquences protéiques et considérons comme :

- *similaires* des entités biologiques qui sont proches selon un critère de ressemblance structurale,
- *homologue*s des entités biologiques qui sont issues d'une histoire évolutive commune avérée ; dans un sens plus strict, deux séquences seront dites *homologues* si elles conservent aussi un certain degré de fonctionnalité biologique,
- *analogues* des entités biologiques qui partagent des propriétés fonctionnelles sans qu'elles aient nécessairement une histoire évolutive commune[2]

Lorsqu'on analyse une séquence biologique nouvelle, le processus d'inférence fonctionnelle le plus courant se base sur la *ressemblance statistiquement fondée*, avec des séquences déjà caractérisées dans des banques (« annotation basée sur l'homologie »). Il existe plusieurs manières d'annoter fonctionnellement un génome (pour revue, Brent 2005), soit manuellement par des annotateurs utilisant des logiciels d'analyse de séquences, soit de façon semi-automatique par des annotateurs avec des environnements présentant des synthèses de résultats obtenus avec les logiciels, soit de façon automatique par l'usage d'algorithmes. La plupart des premiers petits génomes complètement séquencés ont été annotés manuellement en effectuant une recherche d'homologie, sous l'oeil d'experts, pour chaque gène, avec une banque de séquences généraliste (Brent 2005). Un avantage de l'annotation manuelle est la qualité des annotations obtenues ; un défaut majeur est sa lenteur.

La recherche d'homologie par comparaison de séquences nécessite un critère ou un indice de ressemblance entre les séquences, permettant de plus de déduire une mesure possible du temps de divergence entre les espèces suivant les deux hypothèses suivantes : 1) il

[2] La distinction entre entités biologiques *homologues* et *analogues* repose sur une définition adoptée en anatomie comparative. Par exemple, les vertébrés tétrapodes ont des membres pentadactyles, appelés membres chiridiens, qui partagent un schéma d'organisation, et sont à ce titre *homologues*. Toutefois, ces membres n'ont pas nécessairement de relation fonctionnelle, car ils peuvent assurer des fonctions de préhension, de locomotion par la marche pour les vertébrés terrestres, ou par la nage pour les vertébrés aquatiques. En revanche, les ailes des insectes et des oiseaux sont *analogues*, sans partager de plan d'organisation ontogénique : l'une est une expansion du tégument, l'autre un membre avec un squelette osseux. En ce qui concerne les séquences génomiques, l'homologie repose sur l'hypothèse d'une généalogie partagée, à partir d'une séquence ancestrale commune (Wu et al., 1974). Une divergence fonctionnelle peut se produire au cours de l'évolution entre séquences homologues, conduisant à des familles de séquences pour lesquelles il est possible de retrouver une parenté structurale sans identité fonctionnelle. Une définition plus stricte de l'homologie pour les séquences moléculaires est actuellement proposée, exigeant *à la fois une similarité structurale et une analogie fonctionnelle*.

existe une relation entre le temps de divergence et l'indice de ressemblance, et 2) les deux séquences sont issues d'une même séquence dite ancestrale.

Les premières méthodes automatiques d'alignement de séquences ont été élaborées pragmatiquement, en associant mathématiques discrètes et informatique (algorithmique) (Ouzonis et Valencia, 2003). Aujourd'hui, l'alignement de séquences est une discipline majeure de la bioinformatique et une des techniques les plus fréquemment utilisées dans les laboratoires de biologie expérimentale. Ces techniques constituent le point de départ de nombreuses études globales telles que l'annotation automatique de génome (par exemple Gardner et al., 2002, pour le séquençage de *Plasmodium falciparum* qui fait l'objet de cette thèse), le clustering automatique des banques de séquences pour en réduire la diversité (par exemple Petryszak et al., 2005), les études analytiques d'un ensemble restreint de séquences candidates pour un processus biologique (par exemple Bisanz et al., 2006), la phylogénie moléculaire (pour revue, Brocchieri, 2001), etc.

Les séquences peuvent être comparées par alignements deux à deux, objet de ce chapitre bibliographique, ou par alignements multiples[3]. La recherche dans le domaine de l'*alignement de séquences deux à deux* peut être divisée en deux volets.

- Le premier concerne le *calcul de l'alignement* proprement dit. Il s'agit essentiellement d'un problème d'algorithmique qui peut être énoncé comme suit : "Etant donné deux séquences et une mesure de leur ressemblance, quel est l'alignement qui optimise cette mesure ?" L'idée d'optimisation (recherche d'un maximum ou d'un minimum) prolonge celle du principe de parcimonie connue en biologie évolutive (Lecointre et Le Guyader, 2002).
- Le second concerne les *statistiques d'alignements*. Un alignement permet une mesure d'un niveau optimal de ressemblance des deux séquences, selon la méthode et les paramètres qui ont servi pour calculer l'alignement. La question de la confiance que l'on peut avoir dans l'alignement se pose simplement selon l'énoncé suivant : ″Etant donné deux séquences, leurs alignements et une mesure de celui-ci, quelle était la probabilité d'observer cette mesure par chance ?″

Nous développons dans ce chapitre ces deux volets, d'une part en décrivant les principes et méthodes d'alignements de séquences, et d'autre part en présentant les méthodes d'estimation de la pertinence des alignements obtenus, principalement dans le contexte de la comparaison de deux séquences protéiques.

[3] Les méthodologies d'alignements multiples ne sont pas détaillées dans ce chapitre. A partir d'un lot de séquences ces méthodes visent à obtenir un alignement superposé respectant un ensemble de contraintes. Parmi les programmes effectuant ce calcul, sans être exhaustif, nous pouvons citer CLUSTALW (Thompson et al., 1994), MAP (Huang, 1994), MSA (Lipman et al., 1989).

II. Principe général du calcul de l'alignement de deux séquences

La structure primaire d'une séquence biologique (polymère protéique ou nucléique) peut être représentée classiquement comme une suite de caractères (représentant les unités monomériques, acides aminés ou bases). L'alphabet $\mathcal{A}$ utilisé dépend de la nature de la séquences, composé de quatre caractères pour les acides nucléiques ADN et ARN (respectivement {A, T, G, C} et {A, U, G, C}) et de vingt caractères pour les protéines (**Tableau 3**).

Code 3 lettres	Code 1 lettre	Acides aminés	Code 3 lettres	Code 1 lettre	Acides aminés
Phe	F	Phénylalanine	Ile	I	Isoleucine
Leu	L	lysine	Met	M	Méthionine
Ser	S	sérine	Thr	T	Thréonine
Tyr	Y	Tyrosine	Asn	N	Asparagine
Cys	C	Cystéine	Lys	K	Lysine
Trp	W	Tryptophane	Val	V	Valine
Pro	P	Proline	Ala	A	Alanine
His	H	Histidine	Asp	D	Acide aspartique
Gln	Q	Glutamine	Glu	E	Acide Glutamique
Arg	R	Arginine	Gly	G	Glycine

Tableau 3 : L'alphabet des acides aminés; codes à une et trois lettres pour chaque caractère

Par analogie avec le *mot*, une *séquence* est donc une concaténation de caractères de $\mathcal{A}$. Un ensemble $\mathcal{A}^+$ de *mots de longueur finie* sur un alphabet peut se définir selon les règles suivantes appliquées un nombre fini de fois: 1) tout élément de $\mathcal{A}$ appartient à $\mathcal{A}^+$, 2) si $i \in \mathcal{A}$ et $j \in \mathcal{A}^+$ alors $ij \in \mathcal{A}^+$. On adjoint à $\mathcal{A}^+$ le *mot vide*, noté Ø.

Pour passer de la notion de *mot* à celle de *segment de mot* (ou de *chaîne* à celle de *sous-chaîne*), on introduit les notions de *facteurs*, *suffixes* et *préfixes*. Dans $\mathcal{A}^+$, une séquence a est un facteur d'une séquence b s'il existe deux mots u et v tels que $b = uav$. On dit que u (respectivement v) est un *préfixe* (respectivement *suffixe*) de b lorsque $u \neq Ø$ (respectivement $v \neq Ø$). Un facteur de b est dit propre lorsqu'il est différent de Ø et de b.

Schématiquement, étant donné une mesure de proximité sur l'ensemble des caractères de l'alphabet $\mathcal{A}$, l'alignement de deux séquences consiste à maximiser le nombre de paires de

lettres identiques, ou proches entre les deux séquences (et/ou de minimiser le nombre de paires de caractères éloignées selon le critère de ressemblance) en respectant l'ordre dans chaque séquence. Il s'agit donc d'un *problème d'optimisation* (Pour une introduction à la notion d'optimisation sur les espaces vectoriels : Ciarlet, 1998). Les séquences étant considérées comme des suites de caractères ordonnées, un alignement est une suite de couples de lettres ordonnées appartenant chacune à l'alphabet $\mathcal{A}$, augmenté du symbole {-} qui symbolise le cas d'un caractère non apparié (nous appellerons alphabet étendu l'ensemble $\mathcal{A}^- = \mathcal{A} \cup \{-\}$), et respectant l'ordre initial dans chaque séquence.

Un exemple d'alignement de séquences protéiques *a* (*a* = NKVDRTGYK) et *b* (*b* = NKVDRYKV) est:

a	:	NKVDRTGYK-
b	:	NKVDR--YKV

Chaque paire $\begin{Bmatrix} N \\ N \end{Bmatrix}$, $\begin{Bmatrix} K \\ K \end{Bmatrix}$,... , $\begin{Bmatrix} G \\ - \end{Bmatrix}$, etc. peut être considérée comme un caractère de l'alphabet de l'alignement, composé à partir d'une partition de $\mathcal{A}^- \times \mathcal{A}^-$.

Deux caractéristiques des alignements doivent être soulignées. Tout d'abord, on ne cherche pas à aligner uniquement les acides aminés identiques mais aussi ceux-ci qui peuvent se ressembler selon un point de vue, par exemple physicochimique[4]. Cette ressemblance dépend du contexte dans lequel des résidus ont des caractères proches ou non (Kawashima et al., 1999). Par exemple, les amino acides prennent part différemment à la structuration des protéines suivant qu'elles sont dans des environnement hydrophiles ou hydrophobes et un même critère de proximité valable dans un cas ne l'est pas dans l'autre (Muller et al., 2001). La pertinence de l'alignement est donc contextuelle et nécessite la construction d'une mesure de ressemblance sur l'ensemble des acides aminés permettant d'évaluer une proximité pour calculer l'alignement dans le contexte d'étude donné.

Lorsqu'il s'agit de séquences reliées par un processus évolutif, l'alignement de deux résidus non-identiques suppose qu'un ensemble de *mutations* peut avoir conduit à la *substitution* des acides aminés à la position alignée. Une deuxième caractéristique est l'introduction du caractère {-} dans le processus d'optimisation, appelé indel (*insertion-délétion*).

[4] Les propriétés physicochimiques des acides aminés sont résumées sommairement sur le portail "Amino Acid Explorer" du National Center for Biotechnology Information, NCBI (http://www.ncbi.nlm.nih.gov/Class/Structure/aa/aa_explorer.cgi). Les propriétés physicochimiques des amino acides enrichies des corrélations qui ont pu être déterminées suivant des critères biologiques, physiologiques ou moléculaires, recensées dans la littérature, sont détaillées sur le portail "Amino Acid explorer" de l'University of Maryland Baltimore County, UMBC, http://www.evolvingcode.net:8080/aaindex/tools/spanningtree.xhtml, amélioration de la version originale de l'AAindex (Kawashima et al., 1999; http://www.genome.ad.jp/dbget/aaindex.html).

III. Les matrices de substitutions

Avant de présenter les mesures de ressemblance élaborées pour l'alignement des acides aminés, nous rappelons brièvement les différentes fonctions de proximités que l'on peut construire sur un ensemble d'objets.

A. Les fonctions de proximités: dissimilarité, distance et similarité

- La ***dissimilarité*** est une fonction $f : E \times E \rightarrow \Re^+$ telle que pour tout a et b, éléments de E les relations [1] et [2] sont respectées :

[1] $f(a,b) = 0 \Leftrightarrow a = b$

[2] $f(a,b) = f(b,a)$

- La ***distance*** est une fonction $f : E \times E \rightarrow \Re^+$ telle que pour tout a et b, éléments de E, les relations [1], [2] et [3] sont respectées:

[3] $\forall a,b,c \in E,\ f(a,c) \leq f(a,b) + f(b,c)$

La distance est donc une dissimilarité particulière, vérifiant l'inégalité triangulaire.

- La ***similarité*** est une fonction $f : E \times E \rightarrow \Re^+$ telle que pour tout a et b, éléments de E, les relations [4] et [5] sont respectées:

[4] $f(a,a) = \max_b f(a,b)$

[5] $f(a,b) = f(b,a)$

B. Les matrices de substitution sont des matrices de similarité

Le problème de l'optimisation d'un alignement de séquences peut être résolu par optimisation d'une *distance*, par exemple la distance de Hamming (pour deux séquences, nombre de positions pour lesquelles les caractères diffèrent) ou la distance de Levenstein (pour deux séquences, nombre minimal d'opérations élémentaires parmi les substitutions et les indels pour transformer l'une en l'autre). Une extension de cette dernière, appelée distance généralisée de Levenstein, consiste à associer à chacune des opérations unitaires un coût (pour revue: Noé, 2005). Si l'optimisation d'un alignement de séquences d'ADN peut être résolu par ce type d'approche, par exemple en minimisant la distance généralisée de Levenstein (Noé, 2005), l'optimisation d'alignement de séquences protéiques n'est pas résolu simplement en utilisant une *distance* comme mesure de proximité (Setubal et Meidanis, 1997).

Une première mesure de la *similarité* entre acides aminés comme critère de ressemblance a été proposée par Dayhoff et al. (1978) selon une approche mixte, empirique (déduite à partir d'alignements réels) et déterministe (suivant un modèle de divergence des protéines au cours de l'évolution). La famille de matrices de substitution dérivées de cette méthode est connue sous le nom de PAM (cf. p. 32). Se basant sur un modèle moins spéculatif quand à l'évolution des protéines, Henikoff et Henikoff (1992) ont construit une autre famille de matrices de substitution, nommée BLOSUM (cf. p. 33), famille de matrices la

plus utilisée à ce jour. Plusieurs autres modèles de construction de matrices existent (Risler et al. (1988); Gonnet et al., 1992; Jones et al., 1992). Après avoir présenté les matrices identités, nous détaillons dans ce paragraphe la construction des deux principaux types de familles de matrices (PAM et BLOSUM).

1. Matrice identité et matrice de distance dérivée

La matrice identité *Id* est la plus simple que l'on puisse construire. La similarité vaut 1 pour l'identité et 0 dès que l'acide aminé *i* est différent de l'acide aminé *j*. Cette similarité peut facilement être convertie en mesure de dissimilarité par la transformation [6], en notation matricielle :

[6] $Id - T = Dis$

avec *Id*, *T* et *Dis* $\in M_{20}(R)$, où $M_{20}(R)$ désigne l'ensemble des matrices carrées de dimension 20 ayant leurs valeurs dans l'ensemble des réels, *T* la matrice de transformation (1 sur la diagonale et –1 pour les autres valeurs) et *Dis* la matrice obtenue (0 sur la diagonale et 1 ailleurs).

La matrice *Dis* définit une *distance* (métrique), appelée *dis*, sur l'ensemble des acides aminés $\mathcal{A}$ et fait donc du couple ($\mathcal{A}$, *dis*) un espace métrique. Il s'agit d'un espace topologique pour lequel la topologie associée est discrète (Skandalis, 2001). Cette topologie présente certaines propriétés que l'on peut résumer comme suit : 1) chaque acide aminé est discernable dans l'espace des acides aminés et 2) chaque acide aminé est à égale distance de ceux qui lui sont distincts.

Dans cet espace, il n'y a donc pas de couples d'acides aminés distincts qui soient plus rapprochés que les autres. Pour une comparaison de séquences ayant des ressemblances structurales et fonctionnelles, cette propriété n'est évidemment pas concordante avec la biologie. En effet, sur le plan généalogique, les acides aminés sont disjoints selon des processus de mutations plus ou moins longs qui peuvent conduire de l'un à l'autre (distance de Levenstein). Cependant, une distance dérivée de la distance de Levenstein, en dehors de toute pression évolutive, ne reflète pas qu'au niveau fonctionnel, certains acides aminés, par exemple la lysine (K) et l'arginine (R), partagent des propriétés physico-chimiques qu'ils ne partagent pas avec d'autres résidus. La substitution de l'un par l'autre peut être favorisée par rapport aux autres substitutions, comme effet du processus de sélection naturelle. Dans ce contexte, l'information portée par le résidu K (respectivement R) (polarité, taille, nature du groupement chimique) est plus proche de celle de R (respectivement K).

La prise en compte des propriétés fonctionnelles des amino acides et de leur déterminisme génétique n'a pas été résolue en se fondant sur une fonction de distance, mais en estimant une valeur de similarité pour chaque couple donné d'acides aminés. L'obtention de cette valeur par un chemin théorique, c'est à dire en partant des connaissances physico-chimiques et génétiques, n'a pas été conclusive. Des approches empiriques, détaillées ci-dessous, partant d'alignements pré-établis, ont permis d'obtenir des valeurs de similarité. A notre connaissance, la topologie d'un espace des acides aminés munie d'une similarité, n'a pas été étudiée.

2. Les matrices PAM (Dayhoff et al., 1978)

Les matrices PAM (acronyme de "point accepted mutation", pour "mutation ponctuelle acceptée") ont été établies en 1978 par Margareth Dayhoff et ses collaborateurs, à partir d'un ensemble de séquences protéiques (cytochrome c, hémoglobine, myoglobine, etc.), recueillies dans l'Atlas of Protein Sequence and Structure-1969 et alignées par paires (Dayhoff et al., 1978). Les matrices PAM décrivent la probabilité qu'un acide aminé *i* soit muté en un autre *j* sur une période d'évolution donnée. Lorsqu'une matrice PAM se nomme PAM1, une mutation est considérée comme acceptée pour 100 acides aminés (1% de divergence). Pour obtenir une matrice PAM2, il suffit de multiplier PAM1 par elle-même. Pour obtenir une PAMX, il suffit de multiplier PAM1 X fois.

La détermination des valeurs de substitution repose sur l'observation des mutations dites "acceptées" dans les alignements, soit un total de 814 changements d'acides aminés observés. La table des changements d'acides aminés sert de base au calcul de *mutabilité relative des acides aminés* dans l'ensemble Σ des séquences. La mutabilité relative d'un acide aminé *i* (mut_i) est la probabilité que celui-ci change en un temps donné restreint. Elle se calcule pour deux séquences *a* et *b* comme le nombre total de mutations de l'acide aminé *i* rapporté à sa fréquence d'apparition dans les séquences *a* et *b*. La mutabilité relative de l'amino acide *i* pour un ensemble de séquences est alors défini comme la somme des mutabilités relatives sur l'ensemble des alignements :

[7] $$mut_i = \sum_{(a,b)\in\Sigma^2} mut_i(a,b)$$

La probabilité de mutation d'un acide aminé *i* en un autre amino acide, pour une période d'évolution donnée, est proportionnelle à sa mutabilité ($\eta.mut_i$). La probabilité M_{ii} de non-mutation de l'acide aminé *i* (valeurs de la diagonale) est donc donnée par

[8] $$M_{ii} = 1 - \eta.mut_i$$

Pour l'ensemble des vingt acides aminés, un indice de persistance *ps* est défini :

[9] $$ps = \sum_{i=1}^{20} \nu_i.M_{ii}$$

où ν_i est la fréquence de *i* dans l'ensemble des séquences Σ. L'indice *ps* peut-être interprété comme la probabilité pour une paire de séquences (*a*,*b*), prise (uniformément) au hasard dans Σ×Σ, et pour une position prise aux hasard, de rencontrer un acide aminé qui ne mute pas, quel qu'il soit. En posant *ps* = 0.99, alors une mutation est considérée comme acceptée pour 100 acides aminés (1 % de divergence) et il est possible de déduire la valeur de η.

Le calcul des coefficients non diagonaux de la matrice M_{ij} suit un modèle Bayesien, en considérant la probabilité $P_{ij/i}$ que l'acide aminé *i* mute en *j*, sachant que *i* a muté. $P_{ij/i}$ se déduit simplement de l'échantillon de séquences, comme le rapport du nombre de doublets {*i*,*j*} dans l'ensemble des alignements, par le nombre de doublets {*i*,*k*} où $k \neq i$. Les termes non diagonaux de la matrice sont donc :

[10] $M_{ij} = (1 - M_{ii}).P_{ij/i}$

La matrice *M* ainsi calculée correspond à un taux de mutation global de 1 % pour une période unitaire d'évolution. Pour pouvoir différencier une *mutation privilégiée* d'une *mutation obtenue au hasard*, une autre matrice nommée "Relatedness odds matrix" est construite de la façon suivante :

[11] $R_{ij} = \frac{\nu_i.M_{ii}}{\nu_i^2}$ si $i=j$ et $R_{ij} = \frac{\nu_i.M_{ij} + \nu_j.M_{ji}}{2.\nu_i.\nu_j}$ si $i \neq j$

Admettant que la matrice *R* est symétrique ($\nu_i.M_{ij} = \nu_j.M_{ji}$), la relation est simplifiée :

[12] $R_{ij} = \frac{M_{ij}}{\nu_i}$

La matrice finale (nommée matrice 'log-odds'), est obtenue en prenant le logarithme des valeurs de la matrice *R*.

[13] $PAM_{ij}=10.\log_{10}(R_{ij})$

Cette méthode suppose un modèle de l'évolution des protéines basé sur une probabilité de transition entre les acides aminés. Le modèle formulé est donc Markovien[5] et les matrices PAM dérivent directement des *matrices de transitions* (on dit aussi *de passage*) entre acides aminés. Le calcul original effectué par Dayhoff et al. (1978) a porté sur un ensemble de 1572 séquences protéiques regroupées en 34 familles. Cet ensemble de paires de séquences homologues avait été sélectionné pour avoir un degré de divergence faible et constant.

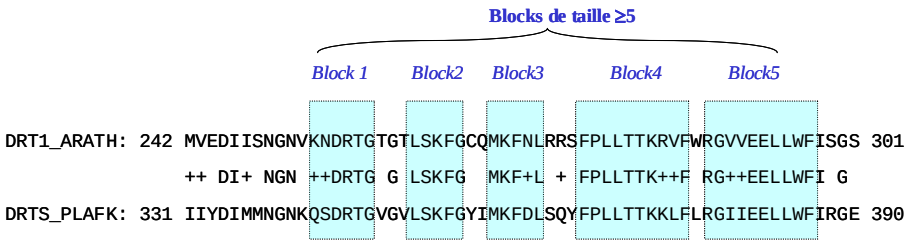
Les matrices PAM sont souvent délaissées pour plusieurs raisons. D'une part, elles supposent que tous les acides aminés sont également mutables. Par ailleurs, l'échantillon de données des matrices PAM récolté en 1978 est restreint comparativement à l'échantillon des matrices BLOSUM. Enfin, la plupart des séquences protéiques des matrices PAM sont biaisées, représentant majoritairement de petites protéines globulaires. Plusieurs améliorations portant sur l'estimation des fréquences de remplacement à partir d'alignements de séquences été proposées parmi lesquelles la méthode résolutive (Muller and Vingron, 2000; Muller et al., 2002) et la méthode du maximum de vraisemblance (Muller et al., 2002).

3. Les matrices BLOSUM (Henikoff et Henikoff, 1992)

Les matrices de la famille BLOSUM (BLOcks SUbstitution Matrix) ont été introduites par Steven et Jorja G. Henikoff en 1992, à partir d'un ensemble de séquences préalablement alignées par paires dans des *blocks* (Henikoff et Henikoff, 1992). Un block est défini comme une région sans gaps dans un alignement donné. Pour un alignement pertinent, un block est considéré comme une région d'homologie potentielle. Un pourcentage de clustering est défini par le maximum d'identité mesuré dans les alignements de blocks (voir **Figure 23**). Les

[5] Une chaîne de Markov est une collection d'états, où le passage d'un état à l'autre est associé à une probabilité.

matrices de la famille BLOSUM ont été calculées pour un pourcentage de clustering indiqué après le nom de la matrice (BLOSUM 62, pour un clustering de 62 %).



	Block1	Block2	Block3	Block4	Block5
taille des blocks	6	5	5	10	10
identité (%)	67	100	80	80	80
Block de clustering 80%	oui	-	-	-	-
Block de clustering 90%	oui	-	oui	oui	oui

Figure 23. Définition du block et du pourcentage de clustering. Sur cet exemple d'alignement de deux dihydrofolate réductase thymidilate synthases d'*Arabidopsis thaliana* (DRT1_ARATH) et de *Plasmodium falciparum* (DRTS_PLAFK), les zones de similarités entre les deux séquences sont indiquées entre les deux séquences : une lettre indique l'identité, le signe + indique une similarité forte. Les blocks (en bleu) sont les séries d'amino acides alignés sans gap dont la taille est supérieure à une limite donnée (ici 5). Le pourcentage d'identité d'un block correspond au pourcentage d'amino acides conservés dans le block. Le clustering de blocks conduit à éliminer les blocks de pourcentage d'identité supérieur à ce que l'on appelle le pourcentage de clustering. La figure montre deux exemples de clustering à 80 et 90 %.

Pour un échantillonnage Σ de blocks correspondant à un pourcentage de clustering, une matrice BLOSUM est simplement déduite à partir de la fréquence $\nu_{i,j}$ d'observation des acides aminés *i* et *j* alignés et des fréquences ν_i et ν_j des acides aminés dans l'échantillon.

[14] $$M_{ij} = \frac{\nu_{ij}}{2.\nu_i.\nu_j} \text{ si } i \neq j \text{ et } M_{ii} = \frac{\nu_{ii}}{\nu_i^2} \text{ pour la diagonale}$$

La matrice BLOSUM est donnée par :

[15] $$BLOSUM_{ij} = \frac{1}{\varpi}.\log(M_{ij})$$

où ϖ est une constante qui permet de fixer l'unité. Si ϖ = 1, l'unité est le *nat* (logarithme naturel) et si $1/\varpi = 2/\log(2)$, l'unité est le demi *bit* (unité couramment utilisée car elle permet de faire le lien avec la théorie de l'information et l'informatique théorique; Cover et Thomas, 1991).

Les données ayant servi à calculer les matrices de la famille BLOSUM ont été obtenues grâce au programme PROTOMAT qui permet de générer automatiquement un ensemble de blocks (plus de 2000 pour le calcul original de ces matrices) à partir de séquences alignées (Henikoff et Henikoff, 1991). La clusterisation des blocks ayant au moins

un certain pourcentage d'identité (par exemple 62% pour BLOSUM 62, la matrice actuellement la plus utilisée, toutes applications confondues) a permis d'obtenir des matrices correspondant à des degrés de divergence croissants, quand le pourcentage de clustering décroît.

4. Modèle général pour la comparaison des matrices de substitution

Une matrice est-elle meilleure qu'une autre ? La question est difficile et n'admet peut-être pas de réponse. La construction de matrices de substitution selon des méthodes empiriques dépend de la qualité de l'échantillon original de séquences, supposé suffisamment représentatif, utilisé pour le calcul. Il est difficile de démontrer la pertinence du choix initial de cet échantillon et on pourrait penser que la qualité des prédictions de substitutions, sur la base de ces matrices, soit variable. Des analyses comparatives montrent que les matrices fondées sur des alignements de séquences ou de structures et dont les termes sont calculés *directement* à partir des fréquences de résidus alignés (comme la famille BLOSUM par exemple) ont une performance pour la recherche dans les bases de données supérieure à celles basées sur des modèles d'évolution des protéines (Henikoff et Henikoff, 1993). Ceci est vrai tant pour l'entropie relative[6] des matrices que pour l'évaluation empirique de leur sensibilité. Le manque d'étude de la spécificité des matrices de substitution s'explique par la difficulté de construire un *gold standard*[7], l'hypothèse « deux séquences sont homologues » étant *toujours* une hypothèse (l'événement de spéciation, ou de duplication ne nous étant pas accessible).

Karlin et Altschul (1990) ont montré que toute matrice de substitution est implicitement une matrice de *log-odd-ratio*[8] si il existe une constante ϖ telle que $\sum_{i,j} \nu_i.\nu_j e^{-\varpi.s(i,j)} = 1$, avec ν_i et ν_j les fréquences des acides aminés dans l'échantillon et $s(i,j)$ le score de substitution de l'acide aminé i avec l'acide aminé j. Dans ces conditions, $s(i,j)$ est le logarithme du rapport de vraisemblance entre deux modèles en compétition :

- D'une part l'*hypothèse nulle* qui suppose que les amino acides alignés ne sont pas liés par l'évolution et donc que la probabilité d'observer la paire $\{i,j\}$ est le produit des probabilités individuelles de i et j.
- D'autre part l'*hypothèse alternative* qui suppose que les résidus sont liés par une substitution. La probabilité de la paire $\{i,j\}$ est donc mesurée par la fréquence de substitutions observées.

Le score est donné par l'équation [16] où ϖ est un facteur d'échelle qui détermine l'unité, $\nu_{i,j}$ est la fréquence observée de substitution entre i et j, et $\nu_i.\nu_j$ la fréquence d'apparition de la paire $\{i,j\}$ si celle-ci était due au hasard.

---

[6] L'entropie relative est une mesure qui estime le score moyen dans une matrice de substitution si on suppose que la fréquence de chaque score élémentaire est pondérée par la fréquence observée des paires d'acides aminés alignées (voir plus loin).

[7] Un "gold standard", pour "étalon or", est un cas reconnu comme vrai, qui peut servir de vrai positif pour des études statistiques

[8] Par définition, on appelle *odd-ratio* (traduction anglaise du rapport des côtes) le rapport de deux probabilités (Valleron, 1998).

[16] $$s(i,j) = \frac{1}{\varpi}.\log\left(\frac{\Pr(\{i,j\}\,|\,\text{hypothèse alternative})}{\Pr(\{i,j\}\,|\,\text{hypothèse nulle})}\right) = \frac{1}{\varpi}.\log\left(\frac{\nu_{ij}}{\nu_i.\nu_j}\right)$$

En choisissant $1/\varpi = 2/\log(2)$, le score est exprimé en demi-bits.

De manière formelle, pour toute matrice de substitution:

[17] $$\sum_{i,j} \nu_i.\nu_j e^{-\varpi.s(i,j)} = 1 \quad \Rightarrow \quad s(i,j) = \frac{1}{\varpi}\log\left(\frac{\nu_{ij}}{\nu_i.\nu_j}\right)$$

L'entropie relative *HR* est une mesure qui estime le score moyen dans une matrice de substitution si on suppose que chaque score élémentaire $s(i,j)$ est pondéré par ν_{ij}, la fréquence observée des paires d'acides aminés.

[18] $$HR = \sum_{i,j} \nu_{ij}.s(i,j)$$

L'entropie relative exprime donc le score moyen par paire de résidus dans un alignement de séquences « réellement » homologues. On conçoit facilement que plus des séquences sont divergentes ou éloignées dans l'évolution, plus cette quantité est faible. En généralisant les matrices de substitutions fondées sur des alignements empiriques comme des matrices de *log-odd-ratio*, Altschul (1991) montre que:

[19] $$HR = \sum_{i,j} \nu_{ij}.\log\left(\frac{\nu_{ij}}{\nu_i.\nu_j}\right)$$

Altschul (1991) observe sur la série de matrices PAM que *HR* décroît lorsque la valeur de PAM augmente, en concordance avec la supposition que les matrices PAM de fort indice (par exemple PAM250) sont censées représenter la mesure de similarité pour des séquences ayant plus fortement divergé que les matrices de faible indice (par exemple PAM20). Il est important de noter que *HR* est surtout une estimation théorique de la sensibilité des matrices de substitutions.

Benner et al. (1994) ont montré que les valeurs des matrices de substitution n'évoluent pas de la même façon en fonction du pourcentage de divergence caractérisant les séquences ayant servi à leurs constructions. Pour de faibles pourcentages de divergence, ce sont essentiellement le code génétique et le modèle de mutation aléatoire qui sont sous-jacents à l'estimation de la distance entre acides aminés. Pour de grands pourcentages de divergence, ce sont essentiellement les propriétés physico-chimiques qui influencent les valeurs de similarités dans les matrices (Benner et al., 1994).

IV. La recherche de l'alignement optimal de deux séquences

A. Principes du calcul de la similarité entre deux séquences

Disposant d'une mesure de proximité dans l'ensemble des acides aminés, comment construire une fonction de proximité pour l'ensemble des séquences ? Les algorithmes développés pour répondre à cette question sont fondés sur l'hypothèse d'une étendue de la

ressemblance entre séquences homologues le long de ces séquences (ou de segments) et sur l'additivité de la similarité, mesurée au niveau des acides aminés alignés.

Si on se donne deux séquences (*a* et *b*) et un alignement donné, un *score* est défini comme la somme des scores élémentaires des paires d'acides aminés alignés. Le calcul tient compte des *indels* {-} avec, dans le cas général, un score appelé *pénalité d'ouverture de gap*[9] (ou *gapo*) pour l'alignement d'un acide aminé avec un premier caractère {-}, et un score appelé *pénalité d'extension de gap* (*ou gape*) pour les acides aminés suivants poursuivant l'alignement avec des caractères {-}. Pour un alignement donné, le score est donc donné selon la formule générale :

[20] $$s_l(a,b) = \sum_l s(a_i,b_j) + \sum_g gapo + \sum_e gape$$

où *l* est le nombre de résidus *i* de la séquence *a* alignés par avec des résidus *j* de la séquence *b*, *g* est le nombre d'ouvertures de gap et *e* le nombre d'extensions de gap.

Par exemple, pour l'alignement de deux séquences protéiques *a* = NKVDRTGYK et *b* = NVDRYK par

```
a    :    NKVDRTGYK
b    :    N-VDR--YK
```

un score de l'alignement est donné par :

[21] $s_l(a,b) = s(\mathrm{N},\mathrm{N})+s(\mathrm{V},\mathrm{V})+s(\mathrm{D},\mathrm{D})+s(\mathrm{R},\mathrm{R})+s(\mathrm{Y},\mathrm{Y})+s(\mathrm{K},\mathrm{K})+ 2\ gapo + 1\ gape.$

Il est évident que généralement, l'ouverture et l'extension des gaps sont des pénalités, c'est-à-dire que $gapo < 0$ et $gape < 0$.

Pour deux séquences et un système de scores donnés, il existe plusieurs alignements possibles. On appelle par définition *similarité* entre la séquence *a* et la séquence *b*, le score maximum possible entre *a* et *b*, c'est à dire :

[22] $$s(a,b) = \max(s_l(a,b))$$

où s_l est le score pour un alignement *l* donné. L'alignement parmi tous les alignements possibles dont le score est maximal est appelé *l'alignement de a et b* au sens strict.

Les séquences biologiques sont définies comme des chaînes de caractères de longueurs finies. On appelle global un alignement où l'ensemble des acides aminés des deux séquences sont alignés. Un alignement local est peut être restreint à des segments de chaque séquence. Nous décrivons brièvement les méthodes principales développées pour déterminer ces alignements optimaux.

B. L'alignement global : algorithme Needleman-Wunsch

L'algorithme de Needleman-Wunsch a été développé pour déterminer le meilleur alignement global entre deux séquences *a* et *b*, de tailles *m* et *n* respectivement (Needleman et

[9] En anglais, *gap* signifie une *discontinuité*.

Wunsch, 1970). Le principe de cet algorithme consiste à calculer les scores maximaux d'alignements entre tous les préfixes[10] de *a* et *b*. On note:

$a^v=a_1a_2...a_v$ et $b^w=b_1b_2...b_w$ les préfixes de taille *v* et *w*,

$a^0=b^0=\varnothing$,

$S_{max}(v,w)$ le score maximal de l'alignement entre a^v et b^w,

$s(a_i,b_j)$ est le score élémentaire entre les lettres a_i et b_i,

d est le coût d'une indel (on ne considérera ici que le cas où il existe un coût unique de gap, c'est à dire où *d=gapo=gape*),

$s(a,b)=S_{max}(m,n)$ la similarité entre *a* et *b*.

Par convention, $S_{max}(\varnothing,\varnothing)=0$. $S_{max}(m,\varnothing)$ est le score maximal de l'alignement d'une séquence de longueur *m* avec la chaîne vide. On a donc $S_{max}(m,\varnothing)= S_{max}(\varnothing,m) =m.d$

Considérons maintenant un alignement de score maximal entre a^v et b^w. Cet alignement doit nécessairement se terminer par:

$$\begin{bmatrix} a_v \\ b_w \end{bmatrix}, \begin{bmatrix} a_v \\ - \end{bmatrix}, \text{ ou } \begin{bmatrix} - \\ b_w \end{bmatrix}$$

Dans le premier cas, les couples précédents de l'alignement constituent nécessairement un alignement optimal entre a^{v-1} et b^{w-1}. Dans les deux autres cas, les paires précédentes de l'alignement constituent un alignement optimal de a^{v-1} et b^w (respectivement a^v et b^{w-1}). Une équation de récurrence permet simultanément le calcul du score maximal et de l'alignement entre *a* et *b* :

[23] $$S_{\max}(m,n) = \max\begin{cases} S_{\max}(m-1,n-1) + s(a_m,b_n) \\ S_{\max}(m-1,n) - d \\ S_{\max}(m,n-1) - d \end{cases}$$

Cette méthode de calcul par *programmation dynamique*[11] peut être représentée sur une table, appelé *matrice de programmation dynamique* (**Figure 24**), qui représente la séquence *a* en ligne et la séquence *b* en colonne. La valeur de chaque cellule de la matrice est calculée à partir des trois cellules adjacentes comme décrit sur la **Figure 24**.

[10] voir définition du préfixe au point II de ce chapitre.

[11] La programmation dynamique est une méthode de résolution, pour les problèmes qui satisfont au principe d'optimalité de Bellman : "une sous-trajectoire d'une trajectoire optimale est elle-même optimale pour la fonction d'objectif restreinte aux trajectoires ayant pour origine celle de cette sous-trajectoire".

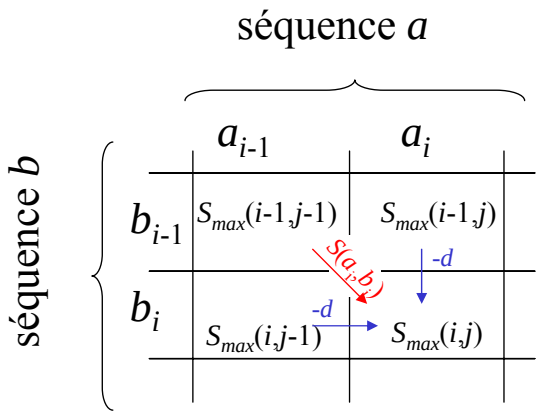


Figure 24: Calcul d'un alignement optimal de deux séquences par programmation dynamique. Représentation locale d'une matrice de programmation dynamique au voisinage de résidus a_i et b_j de deux séquences *a* et *b* pour lesquelles on cherche à obtenir l'alignement optimal.

Le calcul se déroule en deux étapes indépendantes. La première est une phase dite de *descente* où l'équation de récurrence permet de calculer et de sommer le score optimal. Durant cette phase, un pointeur est conservé dans chaque cellule, orienté vers la cellule antérieure qui a servi au calcul du score. Une fois cette étape achevée, la matrice de programmation ne possédera qu'une valeur maximale sommée, dite score maximal. Les pointeurs conservés permettent d'effectuer une phase dite *ascendante* permettant de générer l'alignement caractère par caractère.

La méthode décrite par Needleman et Wunsch (1970) conduit à des alignements globaux. Pour des raisons biologiques évidentes, il n'existe pas toujours une ressemblance entre deux séquences sur leur globalité. Des segments peuvent se ressembler comme dans le cas des protéines modulaires ou lorsque deux enzymes issues d'un ancêtre commun n'ont conservé une ressemblance qu'au niveau de la fixation d'un substrat. Des méthodes pour la recherche des homologies au niveau de segments, à l'intérieur des séquences, ont été développées, visant à optimiser des alignements locaux.

C. Les alignements locaux: algorithmes S-W, FASTA et BLAST

La méthode de Needleman et Wunsch, (1970) a été adaptée à la recherche d'alignements optimaux locaux par Smith et Waterman (1981). La modification majeure de l'algorithme S-W se situe au niveau de l'équation de récurrence en forçant la valeur de S_{max} à être supérieure ou égale à 0 :

[24] $$S_{\max}(m,n) = \max\begin{cases} S_{\max}(m-1,n-1) + s(a_m,b_n) \\ S_{\max}(m-1,n) - d \\ S_{\max}(m,n-1) - d \\ 0 \end{cases}$$

Autrement dit, si le meilleur alignement jusqu'à la position (*m*,*n*) aboutit à un score négatif, l'alignement est interrompu. Un nouvel alignement local est engagé à partir de cette position. L'alignement local maximal est donc borné à gauche et à droite par les premières cellules contenant une valeur nulle. Cet algorithme peut être complété par une fonction

insertion-délétion linéaire dont le but est de pénaliser moins fortement les extensions de gaps (Setubal et Meidanis, 1997). Pour une fonction *d* qui dépend de la longueur *l* des *gape* :

[25] $$d(l) = gapo + (l-1).gape$$

L'algorithme S-W est un algorithme exact dans le sens où il ne fait aucune approximation de la formule de récurrence lors du calcul de l'alignement et qu'il explore l'ensemble de l'espace défini par les deux séquences. La contrepartie de cette exactitude est que le temps de calcul croît avec la taille de cet espace, soit le produit des deux longueurs des séquences. Ce coût de calcul en limite l'usage pour le criblage des banques de données moléculaires de grandes tailles, malgré la parallélisation de S-W permettant de réaliser des comparaisons à l'aide de supercalculateurs ou de grilles (Glemet et Codani, 1997; Bastien et Maréchal, soumis).

Plusieurs heuristiques[12] ont été développées pour une plus grande rapidité de calcul, avec un compromis acceptable en terme d'optimalité de l'alignement obtenu. Les deux plus importants de ces algorithmes sont d'une part FASTA (Pearson et Lipman, 1988) et surtout BLAST (Altschul et al., 1990) qui est de loin le plus utilisé. Ce dernier fut spécifiquement développé pour la recherche d'alignements locaux optimaux entre une séquence requête et une base de données. Le principe de l'algorithme repose sur l'idée que de bons alignements doivent nécessairement contenir un petit segment où il y a un très bon score d'alignement. Ces petits segments, une fois détectés, servent de graines à partir desquels l'alignement est étendu. L'algorithme original de BLAST ne permettait pas d'insertion-délétion. Deux améliorations successives ont été apportées. La première, dans la version BLAST2, est une prise en compte des *gape* (Altschul et al., 1997). La seconde amélioration permet la recherche de similarités "non stringentes" par un processus itératif, nommé Psi-BLAST (Altschul et al., 1997), qui, au cours de chaque étape, construit une matrice position-spécifique (Gribskov et al., 1987).

V. Modèles statistiques pour l'évaluation de la pertinence d'un alignement de deux séquences

L'estimation de la significativité d'un alignement est un des problèmes les plus importants de la théorie des alignements de séquences biologiques. De cette estimation dépendent la pertinence biologique et pratique des méthodes de classifications automatiques, d'annotations et de reconstruction phylogénétiques. L'idée sous-jacente est que par rapport à une séquence donnée, dite *séquence requête*, l'ensemble des séquences se divise schématiquement en deux classes

- les séquences partageant un niveau significatif de similitude avec la séquence requête, du fait d'une parenté évolutive,
- les séquences qui ne sont pas apparentées.

[12] Le terme heuristique signifie littéralement "aide à la recherche". Ici, on appelle heuristique une approche utilisant des règles empiriques dans un domaine particulier, pour résoudre des problèmes trop complexes, voire impossible à résoudre par des méthodes exactes.

On considère que ces dernières sont peu différentes de séquences construites sans règles apparentes, appelées *séquences aléatoires*. Disposant d'une méthode d'alignement, et donc de mesure d'une proximité par le score d'alignement, quelle est la répartition des séquences de tout type par rapport à la requête ? Cette répartition permet elle de déterminer si une séquence dite *sujette*, caractérisée par un score d'alignement avec la séquence requête *s*, est une séquence significativement similaire ? La probabilité qu'une séquence sujette aléatoire présente un score d'alignement avec une séquence requête au dessus d'un seuil *s*, est appelé la *p-value*. Il existe essentiellement deux modèles pour évaluer cette *p-value*.

A. Le modèle de Karlin-Altschul

Samuel Karlin et Stephen F. Altschul ont proposé en 1990 une méthode d'estimation statistique de la significativité d'un alignement de plus grand score, méthode actuellement la plus utilisée (Karlin et Altschul, 1990). Considérant que le score calculé était un extrême parmi les scores possibles, Karlin et Alschul (1990) ont proposé un modèle dérivé des travaux d'Emil-Julius Gumbel (Gumbel, 1958) sur les lois régissant les valeurs extrêmes, en particulier pour les distributions de minima et de maxima (pour revue, Coles, 2001).

Ce modèle considère la répartition des scores entre deux séquences aléatoires *A* et *B* de longueurs *m* et *n* respectivement. Ces séquences sont construites par tirage indépendant de *m* et *n* résidus, avec une même distribution (modèle dit *iid,* pour identiquement et indépendamment distribuée). Par ce mode de construction de séquences aléatoires, le caractère *i* a la probabilité ν_i d'apparaître, calculée par la fréquence dans la séquence cible. On appelle score du plus grand segment (MSS, pour Maximal Segment Score) le score associé à l'alignement de plus grand score. La formule de Karlin-Altschul est une conséquence de la modélisation du nombre de régions présentant un score dépassant un certain seuil *s* par une distribution de Poisson (Pour une introduction aux distributions, Skorokhod, 2005). Brièvement, pour deux séquence aléatoires *A* et *B*, une distribution des acides aminés et une matrice de similarité, alors le nombre d'alignements sans gap distinct avec un score supérieur ou égal à *s* est approximativement distribué selon une loi de Poisson dont la moyenne est

[26] $$E(s) \approx K.m.n.\exp(-\lambda.s)$$

où λ et *K* peuvent être calculés à partir de la matrice de score et de la composition des séquences. *E*(*s*) est appelé la *E-Value*. En conséquence, la variable aléatoire correspondant au *score d'un alignement sans gap entre A et B* suit une loi de Gumbel (Loi des valeurs extrêmes de type I) à quatre paramètres :

[27] $$P(S(A,B) \le s) \approx \exp(-K.m.n.\exp(-\lambda.s))$$

L'expression $P(S(A,B)>s)$ correspond à la *p-value* définie plus haut. Lorsque *s* est suffisamment grand, la *E-value* et la *p-value* sont équivalentes (selon le développement de Taylor de exp(*x*) à l'ordre 1 permettant l'approximation : $\exp(-x) \underset{0}{\approx} 1-x$). Il est important de noter que cette distribution est asymptotique, c'est à dire valable lorsque *m* et *n* sont suffisamment grands. Le modèle admet donc deux restrictions pour être valide avec des séquences réelles :

- les distributions en acides aminés des deux séquences ne doivent pas être trop dissimilaires,
- les longueurs des deux séquences doivent être assez proches.

Vingron et Waterman (1994) ont montré qu'en fonction de la longueur des alignements avec gap, la croissance des MSS était linéaire pour de petites pénalités de gaps et logarithmique pour de grandes pénalités de gaps. Bien qu'aucune distribution, même asymptotique, n'ait été établie pour les scores d'alignements locaux avec gaps, il a été montré que la loi de Gumbel à deux paramètres

[28] $$P(S(A,B)\leq s)\approx\exp(-.\exp(-\frac{s-\theta}{\beta}))$$

était un modèle acceptable pour des alignements avec gaps globaux et locaux, les paramètres θ et β étaient estimés à partir de simulations (Pearson, 1998; Comet et al., 1999; Altschul et al., 2001; Webber and Barton, 2003).

B. Le modèle de la *Z*-value

Un méthode alternative proposée dès 1985 par Willian R. Lipman et David J. Lipman se fonde sur des simulations de type Monte-Carlo[13] pour estimer la significativité d'un alignement entre deux séquences réelles a et b (Lipman et Pearson, 1985). Un nombre C d'alignements est réalisé entre a et une séquence aléatoire B, dérivée de b par permutation (Fitch, 1983). Ces comparaisons aboutissent à l'estimation d'une moyenne empirique μ de la variable aléatoire $S(a,B)$ et de son écart-type σ (l'estimation de la moyenne et de l'écart-type empiriques sont indiqués par la notation ^). La *Z-value* est alors définie par la formule :

[29] $$Z(a,b^*)=\frac{s(a,b)-\hat{\mu}}{\hat{\sigma}}$$

où * désigne la séquence permutée et $s(a,b)$ le score d'alignement de a et b. La *Z-value* est issue de la *cote Z*, unité de mesure statistique qui permet d'exprimer une position dans une distribution, par rapport à la moyenne et l'écart type, en d'autres termes, l'étalement. Parce que cette unité de mesure tient compte de la moyenne et de l'étalement, son utilisation permet de ramener à une échelle commune des objets différents et, du même coup, de faciliter leur comparaison. Une expression en *cote Z*, permet des classements d'objets à la fois différents et équivalents. Dans le cas des scores d'alignements, il existe plusieurs façons d'estimer la *Z-value*, en tentant par exemple de corriger la dissymétrie des calculs de $Z(a,b^*)$ et $Z(b,a^*)$ (voir pour comparaison de différentes corrections proposées Bastien et Maréchal, soumis). Bien que le terme de *Z-score* ait été introduit, nous utiliserons le terme générique de *Z-value* dans ce mémoire.

[13] La méthode de Monte-Carlo peut être utilisée pour simuler des processus aléatoires. Des résultats obtenus lors de ces simulations, on peut déduire des solutions à des problèmes complexes difficiles à formaliser. La méthode de Monte-Carlo doit son nom à Nicholas Métropolis et Stanislaw Ulam utilisant cette méthode dans le cadre de calculs liés au projet Manhattan de conception des premières bombes atomiques. Son appellation date des années 1944 et fait référence à la principauté de Monaco et à son casino.

Pour une évaluation de la pertinence de la *Z-value* appliquée à l'analyse de séquences, Comet et al. (1999) ont effectué une comparaison tout-contre-tout du protéome de *Saccharomyces cerevisiae* par la méthode de Smith-Waterman, en ne retenant que les scores supérieurs ou égaux à 30. Ces auteurs ont ensuite calculé les *Z-values* des alignements retenus pour 20, 50, 100 et 200 permutations et observé que le calcul de $Z(a,b^*)$ était convergent et dépendait de la précision de l'estimation de μ et σ, par conséquent de C. En particulier, Comet et al. (1999) ont montré que l'écart type dépendait de la valeur de Z et décroissait en fonction de $\sqrt{C}$ suivant la formule :

[30] $$\sigma(Z) \approx \frac{1.26}{\sqrt{C}} Z$$

Dans la pratique, on utilise des valeurs de C allant de 100 à 1000 (Louis et al., 2001; pour revue Bastien et al., soumis). Comet et al. (1999) ont de plus montré que la loi asymptotique de la Z-value était indépendante de la taille et de la composition des séquences comparées. Un résultat important de cette étude est que la distribution des *Z-values* issues de ces comparaisons semble suivre une loi de Gumbel (Pearson, 1998; Comet et al., 1999). En se basant sur l'algorithme dit de clumping de Waterman et Vingron (1994), Bacro et Comet (2001) ont examiné cette propriété remarquable et montré que la loi des *Z-values* pouvait être approchée par une loi de Gumbel aux paramètres séquences-indépendents.

L'algorithme initial de Comet et al (1999) pour le calcul des *Z-values* fixe C, et rend l'écart-type variable d'une comparaison à l'autre. Partant de cette constatation Aude et Louis (2002) ont proposé un algorithme itératif de calcul de la *Z-value* dont l'objectif est la réduction du nombre de permutations nécessaires au calcul. Cette approche est basée sur le fait que l'équation précédente permet d'exprimer le nombre de permutations nécessaire pour obtenir un écart type donné :

[31] $$C = \left(1.6 \times \frac{Z}{\sigma(Z)}\right)^2$$

L'algorithme débute en calculant la *Z-value* à partir de 25 permutations. Si la *Z-value* est inférieure à 6, l'algorithme arrête le calcul, considérant que si la *Z-value* est distribuée normalement avec une moyenne de 6 et un écart-type de 1.5, la probabilité d'être au-dessus de 7.5 est supérieure à 90% (et donc la probabilité de manquer une *Z-value* significative est faible). Si la *Z-value* est entre 6 et 35, le calcul est itéré. Si la *Z-value* est supérieure à 35, C est fixé à 100. Plusieurs commentaires sur cette approche peuvent être formulées. D'une part la première étape de l'algorithme suppose une normalité de la distribution, hypothèse que l'on sait fausse. D'autre part la méthode ne fixe pas la variance pour des grandes valeurs de Z (exemple pour Z=400 et C=100, l'écart-type est égal à 50.4, soit une erreur relative de 12.6%). Ceci a pour conséquence de limiter l'exploitation de cet algorithme pour les études utilisant les valeurs relatives des *Z-values*.

Le calcul de la *Z-value*, bien que réputé meilleur pour la comparaison de séquences, est à ce jour sous-utilisé. Deux raisons essentielles expliquent ce faible usage, d'une part le coût de la simulation de Monte-Carlo qui nécessite une puissance de calcul importante, d'autre part l'implémentation des modèles statistiques directement dans les algorithmes de

comparaison (par exemple le modèle de Karlin-Altschul directement implémenté dans BLAST) qui favorise l'usage de ces statistiques au détriment du développement de méthodes alternatives. Pour notre étude, nous avons essentiellement exploité le modèle statistique de Lipman-Pearson de la *Z-value* et la méthode d'alignement de Smith-Waterman implémentés dans le logiciel polyvalent BioFacet, développé par la société Gene-IT[14] (Glemet et Codani, 1997).

VI. Quelles méthodes pour une analyse comparative de génomes biaisés ?

Il est difficile d'évaluer parmi les matrices de substitution, les méthodes de calcul d'alignement et les modèles statistiques, ceux qui sont les plus appropriés dans une situation extrême d'analyse comparative de séquences biaisées compositionnellement. Il est apparu à l'origine de ce travail que nous devions tenter d'examiner chacun de ces dispositifs séparément. Nous avons en particulier avancé dans la compréhension des matrices de substitutions en amont et des statistiques des scores d'alignement en aval, dans le cas d'une comparaison de séquences particulièrement divergentes. L'examen de la méthode de comparaison (en gros BLAST versus Smith-Waterman) était quant à lui plus difficile du fait de l'implémentation du modèle statistique de Karlin-Altschul au coeur de l'algorithme BLAST.

Pour un usage pratique, Comet et al. (1999) avaient montré la robustesse des statistiques de la *Z-value* vis-à-vis des longueurs et compositions des séquences. Un seuil empirique pour les *Z-values* a été déterminé entre 6 et 12, intervalle qualifié de *twilight-zone*, au dessous duquel un alignement est considéré comme peu fiable et au dessus duquel un alignement peut être considéré comme relevant (*i.e.* avec une forte probabilité de relation d'homologie entre les deux séquences). Aucun support théorique n'était disponible pour soutenir la pertinence de ce seuil et le point de départ de la thèse présentée dans ce mémoire a été consacré à l'examen du fondement théorique de cette propriété remarquable.

[14] www.gene-it.com

Résultats et discussion

Chapitre 1
(*Article 1*)

Résultats - Chapitre 1

Majoration de la probabilité du score d'alignement de deux séquences déterminé à l'aide de la *Z-value* : le théorème TULIP

Article 1

Olivier Bastien, Jean-Christophe Aude, Sylvaine Roy & Eric Maréchal (2004)
« Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of Z-value statistics »
Bioinformatics 20:534-537

Préambule

Quelle « confiance » accorder à l'alignement de *deux séquences particulières* ? Comment trier les « meilleurs » alignements obtenus entre *une séquence d'intérêt* et *les séquences stockées dans une banque* ? Comment détecter les alignements « significatifs » parmi les résultats de *comparaisons dites massives*, pour lesquelles un grand nombre de séquences d'un organisme (ou d'une base) est comparé à un grand nombre de séquences d'un organisme (ou d'une base) identique ou différent(e) ?

Nous avons détaillé dans le chapitre bibliographique qu'un alignement de deux séquences était le résultat d'un processus d'optimisation d'une grandeur appelée score d'alignement, somme des scores des résidus alignés. Nous avons de plus rappelé que les différences de composition et de taille des séquences comparées pouvaient peser sur l'estimation des scores d'alignements ainsi que sur l'estimation de la significativité de ces scores. L'analyse comparative de séquences de *Plasmodium falciparum*, de compositions biaisées et de longueurs atypiques, nécessite donc l'emploi des méthodes les moins sensibles aux variations de ces paramètres.

Dans des travaux pionniers, Comet et al. (1999) ont montré la robustesse des statistiques de la *Z-value* vis-à-vis des longueurs et des compositions des séquences. Un seuil empirique pour les *Z-values* a été déterminé entre 6 et 12, intervalle qualifié de *twilight-zone*, au dessous duquel un alignement est considéré comme peu fiable et au dessus duquel un alignement peut être considéré comme significatif (*i.e.* avec une forte probabilité de relation d'homologie entre les deux séquences). Aucun développement théorique n'était cependant disponible pour soutenir la pertinence de ce seuil.

L'**Article 1** présenté dans ce chapitre démontre, à l'aide du théorème de Bienaymé-Chebychev (Bienaymé, 1853; Chebyshev, 1867), comment la *Z-value* permet de déterminer un majorant de la probabilité d'un score d'alignement, et de ce fait, le risque statistique correspondant au seuil de *Z-value* défini par Comet et al. (1999).

Le théorème TULIP (Theorem of the Upper Limit of a score Probability) que nous avons démontré dans cet article, ainsi que ses corollaires, constituent un fondement important de l'analyse automatique de séquences, entre autre pour des comparaisons massives. Ce travail a ainsi été cité en tant que fondement théorique important pour la base de comparaison massive CluSTr de l'EBI, fondée sur les statistiques de *Z-value* (Petryszak et al., 2005) ainsi que pour déterminer un risque statistique dans une procédure de recherche de domaines conservés (Lefebvre et al., 2005). Enfin l'intérêt du théorème TULIP pour les méthodes de clustering de séquences protéiques a été mentionné par Arnold et al. (2005), bien qu'ils ne l'exploitent pas. La poursuite de notre étude a montré à quel point le théorème TULIP permettait d'exploiter les valeurs de *Z-value* pour l'analyse rigoureuse de comparaisons massives de séquences protéiques, et en particulier de reconstituer la phylogénie de protéines homologues.

BIOINFORMATICS

Vol. 20 no. 4 2004, pages 534–537
DOI: 10.1093/bioinformatics/btg440

Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of Z-value statistics

Olivier Bastien[1,2], Jean-Christophe Aude[3], Sylvaine Roy[4] and Eric Maréchal[1,]*

[1]Laboratoire de Physiologie Cellulaire Végétale, Département Réponse et Dynamique Cellulaire, UMR 5168 CNRS-CEA-INRA-Université J. Fourier, CEA Grenoble, 17 rue des Martyrs, F-38054, Grenoble cedex 09, France, [2]Gene-IT, 147 avenue Paul Doumer, F-92500 Rueil-Malmaison, France, [3]Laboratoire de Bioinformatique, Génomique et Modélisation, Département de Biologie Joliot Curie, CEA Saclay, F-91191 Gif sur Yvette Cedex, France and [4]Service de Développements pour la Bioinformatique Sud-Est, CEA Grenoble, 17 rue des Martyrs, F-38054, Grenoble cedex 09, France

Received on May 23, 2003; revised on July 18, 2003; accepted on August 4, 2003
Advance Access publication January 22, 2004

ABSTRACT

Motivation: Different automatic methods of sequence alignments are routinely used as a starting point for homology searches and function inference. Confidence in an alignment probability is one of the major fundamentals of massive automatic genome-scale pairwise comparisons, for clustering of putative orthologs and paralogs, sequenced genome annotation or multiple-genomic tree constructions. Extreme value distribution based on the Karlin–Altschul model, usually advised for large-scale comparisons are not always valid, particularly in the case of comparisons of non-biased with nucleotide-biased genomes (such that of *Plasmodium falciparum*). *Z*-values estimates based on Monte Carlo technics, can be calculated experimentally for any alignment output, whatever the method used. Empirically, a *Z*-value higher than ~8 is supposed reasonable to assess that an alignment score is significant, but this arbitrary figure was never theoretically justified.

Results: In this paper, we used the Bienaymé–Chebyshev inequality to demonstrate a theorem of the upper limit of an alignment score probability (or *P*-value). This theorem implies that a computed *Z*-value is a statistical test, a single-linkage clustering criterion and that $1/Z$-value^2 is an upper limit to the probability of an alignment score whatever the actual probability law is. Therefore, this study provides the missing theoretical link between a *Z*-value cut-off used for an automatic clustering of putative orthologs and/or paralogs, and the corresponding statistical risk in such genome-scale comparisons (using non-biased or biased genomes).

Contact: emarechal@cea.fr

*To whom correspondence should be addressed.

INTRODUCTION

Biological sequence analysis has become an everyday task for biologists. Different automatic methods of sequences comparison, such as Smith–Waterman (Smith and Waterman, 1981) or BLAST algorithms (Altschul *et al*., 1990), are routinely used to align sequences as a starting point for homology searches and function inference (Dardel and Képès, 2002). Biologists often re-examine alignment outputs, either to re-sort the produced alignments or even to re-adjust local residue matches, as judged from their biological expertise. Confidence in an alignment score probability is therefore one of the major fundamentals of automatic genome-scale pairwise comparisons, for clustering of putative orthologs and paralogs, sequenced genome annotation or multiple-genomic tree constructions. An important question is to know whether a computed alignment is evolutionarily relevant or whether it could have arisen simply by chance (Doolittle, 1981). The improbability of the observation called 'alignment' is a critical probability, also termed *P*-value or alignment score probability. In order to test an alignment score probability, two methods have been proposed.

The first method proposed by Karlin and Altschul (1990) is an estimate of the probability of an observed local alignment score according to an Extreme Value Distribution (EVD). The Karlin–Altschul formula is the consequence of interpreting the number of high-scoring matching regions above a threshold by a Poisson distribution. Briefly, when one consider two random sequences A and B ($A = A_1A_2 \cdots A_m$ and $B = B_1B_2 \cdots B_n$), given the distribution of individual residues (i.e. amino acids), and given a scoring matrix, the *P*-value, or probability of finding an ungapped segment pair

Published by Oxford University Press2004.

with a score greater than or equal to s is:

$$P(\tilde{S}(A,B) \geq s) = 1 - \exp(-K \cdot m \cdot n \cdot e^{-\lambda s})$$

where $\tilde{S}(A,B)$ is the random variable named 'score of two random sequences', and λ and K can be calculated from the scoring matrix and sequence compositions. The validity of this model requires two restrictive conditions: first, the individual residues distributions for the two sequences should not be 'too dissimilar' and second, the sequence lengths (m and n) should 'grow at roughly equal rates' (Karlin and Altschul, 1990). The length dependency of alignment scores has been extensively discussed in the literature (Arratia and Waterman, 1994; Waterman, 1994; Waterman and Vingron, 1994; Mott and Tribe, 1999). Arratia and Waterman (1994) demonstrated that the growth of the best matching score of gapped alignments was linear, when gap penalties were small, becoming logarithmic with increasing sequence length for larger gap penalties. In the logarithmic domain, Waterman and Vingron (1994a,b) extended the Poisson approximation techniques of the Karlin–Altschul model to gapped alignments using the Aldous clumping heuristic. Mott and Tribe (1999) further developed a specific heuristic approximation to the score distribution of gapped alignments in the logarithmic domain. The Karlin–Altschul inspired models are usually appropriate for pairwise sequence alignments, and the BLAST method (Altschul *et al*., 1990) that provides P-values accordingly, is undoubtedly the most popular for routine homology searches. Still, the Karlin–Altschul model reaches a validity limit when sequences are sampled in organisms, which residue distribution diverges spectacularly from the average. For instance, *Plasmodium falciparum* amino acid distribution is governed by a strong nucleotidic bias (82% of adenosine and thymidine), extremely divergent from that of other complete eukaryotic genomes sequenced to date (Gardner *et al*., 2002; Nishizawa and Nishizawa, 1998; Nishizawa *et al*., 1999; Pizzi and Frontali, 1999, 2001; Brocchieri, 2001; Singer and Hickey, 2000; Musto *et al*., 1995). In this given example, any pairwise comparison of sequences sampled in *P.falciparum* and another proteomes cannot be estimated theoretically solely based on P-values calculated with the Karlin–Altshul model. When the restricting conditions are not filled, computed P-values are strongly over-estimated, cannot be used to sort best alignment scores and therefore cannot be trusted for an automatic large-scale pairwise comparison without an in-depth validation of the outputs without the help of an expert curator.

An alternative method, proposed by Lipman and Pearson (1985) and described extensively by Comet *et al*. (1999) and Bacro and Comet (2001), uses the Monte Carlo technics to investigate the significance of a given score calculated from the alignment of two real sequences a and b ($a = a_1 a_2 \cdots a_m$ and $b = b_1 b_2 \cdots b_n$). The method in computing C alignments consists of a shuffled sequence from a with a shuffled sequence from b (Fitch, 1983). The variable corresponding to the shuffled sequences from a and b are termed A and B, respectively. These comparisons allow the estimate of an empirical mean score ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) from the distribution of the random variable $\tilde{S}(A,B)$. The Z-score is then defined as:

$$Z(a,b) = \frac{s(a,b) - \hat{\mu}}{\hat{\sigma}}$$

As discussed by Comet *et al*. (1999), for the computation of a Z-value between two sequences, only one of the two sequences is generally shuffled. However, Comet *et al*. (1999) report that shuffling one sequence [for instance, $Z(a,b^*)$ computed after C shuffling of sequence b] leads to Z-values that are markedly different from those obtained by shuffling the second, particularly when sequence aminoacid distributions are markedly divergent. As a consequence, Comet *et al*. (1999) reformulated the Z-value owing to a conservative principle, that is the minimum of $Z(a,b^*)$ and $Z(a^*,b)$. By choosing the minimum of $Z(a,b^*)$ and $Z(a^*,b)$, one do not generally overestimate the Z-value. In this paper, the theoretical discussion only considers the real Z-value as defined by Bacro and Comet (2001).

The computation of $Z(a,b)$ is known to be convergent and depends on the accuracy of the estimation of μ and σ, and therefore on C, ranging practically from 100 to 1000 (Comet *et al*., 1999; Aude and Louis, 2002). Bacro and Comet (2001) show that the asymptotic law of Z-value was independent of sequences length and composition. However, they proved that the estimate of the Z-value was additionally dependent on the shuffling method because the shuffling procedure respects the sequence composition but breaks down biological structures. An improved random model would take into account pattern conservation between the biological real sequence and its random counterparts (if such a definition for a 'Natural' random sequence could be given). Bacro and Comet (2001) proposed a correction factor for the estimate of the Z-value. Since Z-values are estimates based on Monte Carlo technics, they can be calculated experimentally for any pairwise sequence alignment score, whatever the method used, e.g. Smith–Waterman (Smith and Waterman, 1981), BLAST (Altschul *et al*., 1990), etc. Statistically, it was proposed that a Z-value higher than $\sim$8 was reasonable to assess a significant alignment score and a cut-off of 8 was proposed for genome-scale automatic clustering (Comet *et al*., 1999; Codani *et al*., 1999), although this arbitrary figure was never theoretically justified.

Here, we demonstrate a simple theorem assessing that Z-values can be used as a statistical test, a single-linkage clustering criterion and that $1/Z\text{-value}^2$ is an upper limit to the probability of an alignment score whatever the actual probability law is.

THEOREM OF THE UPPER LIMIT OF A SEQUENCE ALIGNMENT SCORE PROBABILITY

Mathematical statement

Given two real sequences a and b ($a = a_1a_2\cdots a_m$ and $b = b_1b_2\cdots b_n$), let s be the maximal score of a pairwise alignment obtained with any alignment method: $s = S(a,b)$. Given A and B are the variables corresponding to the shuffled sequences from a and b respectively, and $P\{\tilde{S}(A,B) \geq s\}$ is the probability that an alignment by chance has a score higher than that calculated with the real sequences, whatever the distribution of the random variable $\tilde{S}(A,B)$ is, relation (1) is true:

$$s \geq \mu + k\sigma \Rightarrow P\{\tilde{S}(A,B) \geq s\} \leq \frac{1}{k^2} \quad (1)$$

where $k > 1, \mu$ is the mean of $\tilde{S}(A,B)$ and σ its standard deviation.

PROOF. We consider a random variable $\tilde{S}(A,B)$ that has a finite mean (μ) and a finite variance (σ^2), the integral of the square of the random variable being finite. Therefore, given $k > 1$, the Bienamé–Chebyshev (Bienaymé, 1853; Chebyshev, 1867) inequality holds:

$$P\{|\tilde{S}(A,B) - \mu| \geq k\sigma\} \leq \frac{1}{k^2} \quad (2)$$

Hereafter, we only consider the right part of the tail of the distribution $[\tilde{S}(A,B) \geq \mu]$:

$$P\{\tilde{S}(A,B) - \mu \geq k\sigma\} \leq P\{|\tilde{S}(A,B) - \mu| \geq k\sigma\} \leq \frac{1}{k^2} \quad (3)$$

thus,

$$P\{\tilde{S}(A,B) \geq \mu + k\sigma\} \leq \frac{1}{k^2} \quad (4)$$

In addition:

$$s \geq \mu + k\sigma \Rightarrow P\{\tilde{S}(A,B) \geq s\} \leq P\{\tilde{S}(A,B) \geq \mu + k\sigma\} \quad (5)$$

Combining relations (4) and (5), we deduce the theorem.

COROLLARY TO THE THEOREM: *The computed Z-value is a statistical test for the probability of a sequence alignment score.*

A direct application of the theorem is that Z-values allow the determination of an upper limit for the probability that an alignment by chance has a score higher than that calculated with the real sequences. Indeed, the term $s \geq \mu + k\sigma$ is true if $(s-\mu)/\sigma \geq k$, that is to say $Z(a,b) \geq k$. Expression of the theorem becomes:

$$Z(a,b) \geq k \Rightarrow P\{\tilde{S}(A,B) \geq s\} \leq \frac{1}{k^2} \quad (6)$$

From this inequality, the computed Z-value appears as a statistical test for a sequence alignment probability.

APPLICATION: DEFINITION OF A STATISTICAL ALPHA-RISK FOR AN AUTOMATIC CLUSTERING OF PUTATIVE ORTHOLOGS AND PARALOGS IN GENOME-SCALE COMPARISONS

For any sequence comparison, whatever the pairwise alignment algorithm, a statistical alpha-risk ρ for an automatic clustering of putative homologs can be expressed as $\rho = 1/k^2$. For example $\rho = 0.015$ is given when $k \approx 8$. One simply has to estimate the Z-values corresponding to all possible alignments. The confidence in the inequality $s \geq \mu + k\sigma$ is tested by comparing the computed Z-value and k. If the inequality is true, the theorem implies that the probability to observe a random alignment with a score superior to that of (a,b) is lower than $1/k^2$, that is to say 0.015. The pragmatic use of a Z-value cut-off of 8 for pairwise alignments clustering is therefore justified theoretically as a way to cluster sequences that may share a common biological structure with a risk of 0.015. For example, when searching homologous sequences of TIC22 from *Arabidopsis thaliana* (accession number At3g23710.1 from the TAIR database, http://www.arabidopsis.org) among annotated protein sequences from the biased genome of *P.falciparum* (PlasmoDB, http://plasmodb.org), using WU-BLASTP algorithm and Blosum62 similarity matrix, no homologous sequence is returned with a significant E-value (threshold = 10), according to the Karlin–Altschul model. By contrast, a SW search returns the sequence PFE1460W from the PlasmoDB database, with a comparison score of 102. The Z-value estimated according to Comet *et al*. (1999) with 500 shuffling is 16.39, a value higher than 10. The statistical risk that PFE1460W be annotated as an TIC22 homolog is therefore lower than 1%. From the biologist expertise, TIC22 from *A.thaliana* is a chloroplast protein harboring a N-terminus plastid sequence that is also found in PFE1460W from *P.falciparum*, an additional argument to assess that PFE1460W is an TIC22 homolog.

CONCLUSION

In this paper, we demonstrate with a simple theorem (named the TULIP theorem, for Theorem of the Upper LImit of a sequence alignment score Probability) that Z-values can be used as a statistical test, a single-cluster criterion and provide an upper limits to the probability of a alignment $S(a,b)$ whatever the actual score probability distribution is. This latter feature is particularly valuable in comparative studies including singularly biased genomes, such as that of *P.falciparum*, in which the score probability distribution is outside the validity domain of the EVD proposed by Karlin and Altschul (1990). Eventually, the pragmatic setting of a cut-off of 8 in clustering sequences from pairwise alignments is theoretically supported as a statistical alpha-risk of 0.015, and is therefore confirmed as a trustworthy cut-off value for

a pairwise similarity clustering after massive genome-scale comparisons.

REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Arratia,R. and Waterman,M.S. (1994) A phase transition for the score in matching random sequences allowing deletions. *Ann. Appl. Prob.*, **4**, 200–225.

Aude,J.C. and Louis,A. (2002) An incremental algorithm for Z-value computations. *Comput. Chem.*, **26**, 403–411.

Bacro,J.-N. and Comet,J.-P. (2001) Sequence alignment: an approximation law for the Z-value with applications to databank scanning. *Comput. Chem.*, **25**, 401–410.

Bienaymé,I.J. (1853) Considérations a l'appui de la découverte de Laplace sur la loi de probabilité dans la méthode des moindres carrés. *C.R. Acad. Sci. Paris*, **37**, 309–324.

Brocchieri,L. (2001) Low-complexity regions in Plasmodium proteins: in search of a function. *Genome Res.*, **11**, 195–197.

Chebyshev,P.L. (1867) Des valeurs moyennes. *Liouville's J. Math. Pures. Appl.*, **12**, 177–184.

Codani,J.J., Comet,J.P., Aude,J.C., Glémet,E., Wozniak,A., Risler,J.L., Hénaut,A. and Slonimski,P.P. (1999) Automatic analysis of large-scale pairwise alignments of protein sequences. *Methods Microbiol.*, **28**, 229–244.

Comet,J.P., Aude,J.C., Glémet,E., Risler,J.L., Hénaut,A., Slonimski,P.P. and Codani,J.J. (1999) Significance of Z-value statistics of Smith–Waterman scores for protein alignments. *Comput. Chem.*, **23**, 317–331.

Dardel,F. and Képès,F. (2002) *Bioinformatique: Génomique et Post-Génomique*. Les éditions de l'école polytechnique edit, Paris.

Doolittle,R.F. (1981) Similar amino acid sequences: chance or common ancestry? *Science*, **214**, 149–159.

Fitch,W.M. (1983) Random sequences. *J. Mol. Biol.*, **163**, 171–176.

Gardner,M.J., Hall,N., Fung,E., White,O., Berriman,M., Hyman,R.W., Carlton,J.M., Pain,A., Nelson,K.E., Bowman,S. *et al*. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, **419**, 498–511.

Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci., USA*, **87**, 2264–2268.

Lipman,D.J. and Pearson,W.R. (1985) Rapid and sensitive protein similarity searches. *Science*, **227**, 1435–1441.

Mott,R. and Tribe,R. (1999) Approximate statistics of gapped alignment. *J. Comput. Biol.*, **6**, 91–112.

Musto,H., Rodriguez-Maseda,H. and Bernardi,G. (1995) Compositional properties of nuclear genes from *Plasmodium falciparum*. *Gene*, **152**, 127–132.

Nishizawa,M. and Nishizawa,K. (1998) Biased usages of arginines and lysines in proteins are correlated with local-scale fluctuations of the G + C content of DNA sequences. *J. Mol. Evol.*, **47**, 385–393.

Nishizawa,K., Nishizawa,M. and Kim,K.S. (1999) Tendency for local repetitiveness in amino acid usages in modern proteins. *J. Mol. Biol.*, **294**, 937–953.

Pizzi,E. and Frontali,C. (1999) Molecular evolution of coding and non-coding regions in *Plasmodium*. *Parassitologia*, **41**, 89–91.

Pizzi,E. and Frontali,C. (2001) Low-complexity regions in *Plasmodium falciparum* proteins. *Genome Res.*, **11**, 218–229.

Singer,G.A. and Hickey,D.A. (2000) Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol. Biol. Evol.*, **17**, 1581–1588.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Waterman,M.S. (1994) Estimating of statistical significance of sequence alignment. *Philos. Trans. R. Soc. Lond.*, **344**, 383–390.

Waterman,M.S. and Vingron,M. (1994a) Sequence comparison significance and Poisson approximation. *Stat. Sci.*, **9**, 367–381.

Waterman,M.S. and Vingron,M. (1994b) Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc. Natl Acad. Sci., USA*, **91**, 4625–4628.

Résultats et discussion

Chapitre 2
(*Articles 2 & 3*)

Résultats - Chapitre 2

Une représentation géométrique, topologique et probabiliste pour les séquences protéiques, inspirée de la physique lagrangienne et de la théorie synthétique de l'évolution : l'espace de configuration des protéines homologues (CSHP)

Article 2

Olivier Bastien, Philippe Ortet, Sylvaine Roy & Eric Maréchal (2005)
« A configuration space of homologous proteins conserving mutual information and allo-wing a phylogeny inference based on pair-wise Z-score probabilities »
BMC Bioinformatics 6(1):49

Article 3

Olivier Bastien, Philippe Ortet, Sylvaine Roy & Eric Maréchal
« The configuration space of homologous proteins: a theoretical and practical framework to reduce the diversity of the protein sequence space after massive all-by-all sequence comparisons **»**
Futur Generation Comput. Syst., in press

Préambule

La recherche d'une homologie entre deux séquences repose sur une *mesure de leur ressemblance* et sur la *fiabilité statistique* de cette mesure (voir **Article 1**). Une fois établi que deux séquences se ressemblent de façon significative, et qu'elles ont *probablement* un ancêtre commun, peut-on corréler le niveau mesuré de leur ressemblance (et dissemblance) avec une grandeur temporelle qui les sépare dans l'histoire évolutive ?

La séparation temporelle de séquences homologues issues d'un ancêtre commun peut se reconstruire selon une approche phylogénétique de type « méthode de distance » reposant sur les principes suivants. Des protéines homologues sont placées dans un espace abstrait, muni d'une fonction de proximité telle que les séquences qui se ressemblent sont proches et celles qui ne se ressemblent pas sont éloignées. Ces protéines sont supposées évoluer selon un ensemble de mutations et recombinaisons génétiques au hasard (une sorte de relaxation), soumises à une pression évolutive de conservation d'un certain niveau fonctionnel. A partir d'hypothèses sur v, la vitesse d'évolution de ces protéines dans cet espace, la distance géométrique d qui les sépare est corrélée avec le temps de divergence t par la formule $v = d/t$. Deux problèmes sont à traiter pour cette reconstruction. Le premier est la mise en espace des

protéines. Le deuxième problème est celui de la fonction de proximité dans cet espace qui soit 1) mesurable dans le sens métrologique du terme (Perdijon, 2004) et 2) qui ait un sens biologique pertinent.

La notion d'espace de configuration appliqué à la représentation des protéines homologues

La mise en espace des protéines présentée dans l'**Article 2** repose sur l'idée d'*espace de configuration* (pour une introduction, voir Arnold, 1989, Hladik et Chrysos, 2000). Dans un espace de configuration, le nombre de grandeurs indépendantes déterminant de façon univoque la position d'un élément est appelé *nombre de degrés de liberté du système*. Chaque degré de liberté peut donc définir une coordonnée du système et de façon générale, ces coordonnées ne seront pas nécessairement les coordonnées cartésiennes dans l'espace usuel (à 3 dimensions). Pour un système à n degrés de liberté, les n grandeurs quelconques $g_1, g_2, \ldots, g_n$ qui caractérisent la position du système sont appelées ses *coordonnées généralisées*. L'espace mathématique de dimension n, défini par les g_i est appelé *espace de configuration*.

La construction d'un espace de configuration permet par exemple de simplifier le problème posé par un pendule pesant illustré sur la **Figure 25A**. Une masse M est suspendue à l'extrémité d'un fil de masse négligeable, attaché à un point P. On se donne pour étude le mouvement du pendule. Une première approche, appelée *mécanique newtonienne*, consiste à étudier ce mouvement en fonction du temps dans l'espace physique usuel, de dimension 3 et muni d'un repère cartésien. L'étude de ce mouvement prend en compte les deux forces extérieures au système : le poids et la force de tension exercée sur la masse par le fil. Il faut donc 3 coordonnées, x, y et z pour décrire la position du pendule à un moment donné. Les *contraintes* imposées par le fil sur le système impliquent que le pendule ne possède qu'un seul degré de liberté, l'angle de rotation θ du pendule avec la verticale V. Il est par conséquent possible de construire une autre représentation spatiale à 1 dimension en définissant la coordonnée généralisée $g_1 = \theta$ et d'étudier le mouvement du pendule dans cet espace de configuration (**Figure 25B**). Cette approche est appelée *mécanique lagrangienne* (Arnold, 1989).

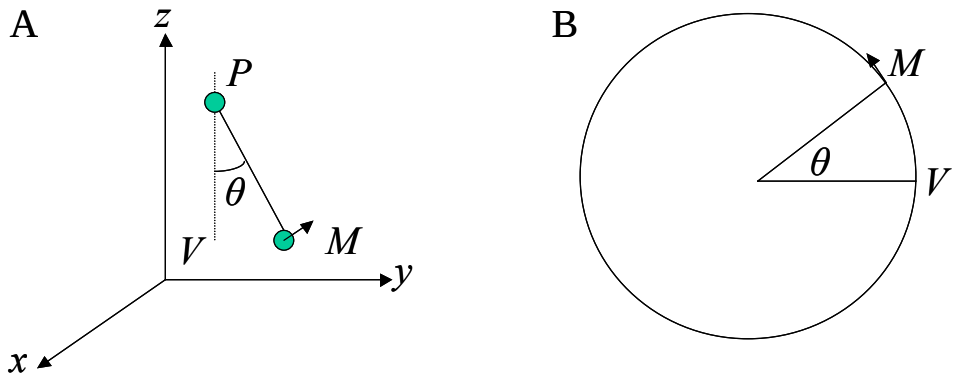


Figure 25: Un pendule pesant dans une représentation spatiale euclidienne (mécanique newtonienne) et dans un espace de configuration (mécanique lagrangienne). A. Description de la cinématique du pendule dans l'espace physique usuel. **B.** Description de la même cinématique dans l'espace de configuration où la coordonnée généralisée est l'angle θ.

Il est intéressant de remarquer que la diminution du nombre de dimensions simplifie le problème et est en meilleure adéquation avec la complexité réelle ; les contraintes ne sont plus considérées comme des forces extérieures aux systèmes et sont intégrées dans la définition du nouvel espace. Pour le pendule par exemple, l'interaction avec le fil était considérée dans l'approche newtonienne comme une force extérieure. Dans l'espace de configuration, il n'y a plus qu'une variable, l'angle. Cet espace de configuration a la topologie d'un cercle (**Figure xxB**) : lorsqu'on on se déplace sur le cercle dans une direction, on revient au point de départ. La force de liaison avec le fil ne se retrouve plus dans l'expression de la cinématique du système, celle-ci étant intégrée dans la conception de l'espace.

L'**Article 2** montre comment la prise en compte des contraintes 1) des propriétés des scores d'alignements déduites de la théorie de l'information (voir ci-dessous), 2) des méthodes de calcul des alignements, et 3) des probabilités d'alignements, en particulier en conformité avec le théorème TULIP (**Article 1**), permet d'aboutir à la construction d'un *Espace de Configuration des Protéines Homologues* (CSHP). Un CSHP n'est pas un espace absolu dans lequel il serait possible de projeter toute protéine, mais un espace indissociable de ses éléments, construit par ses éléments. Dans un CSHP, chaque protéine est positionnée relativement à toutes les autres protéines qui lui sont homologues.

La théorie de l'information et la définition de la proximité des protéines homologues dans le CSHP

Comment établir la ressemblance de deux protéines dans un CSHP ? Nous avons cherché à définir une fonction de proximité sur cet espace reposant sur des hypothèses cohérentes avec la biologie, en particulier en considérant que les séquences ne sont comparables qu'en regard de l'information qu'elles partagent (information génétique codée au niveau nucléotidique, soumise aux variations par mutations/recombinaisons, et information physico-chimique et structurale au niveau protéique, soumises aux contraintes de pérennité fonctionnelle ; Wu et al., 1974). Nous avons établi que le score d'alignement permettait une estimation de l'information partagée par les séquences, interprétée dans le cadre de la théorie de l'information (Shannon, 1948). Nous avons déduit par ce moyen une méthode de reconstruction phylogénétique dans le CSHP.

La théorie de l'information a été formalisée par Claude Shannon aux Bell laboratories (Shannon, 1948), reprenant certains concepts énoncés vingt ans plus tôt par Hartley (Hartley, 1928) et répondant à certains besoins en télécommunications concernant la fiabilité de la transmission des messages, les critères de codes correcteurs d'erreur, etc. Cette théorie a eu des retombées dans différents domaines de la science. Partant de ces concepts, Kolmogorov a en particulier montré qu'il était possible de définir rigoureusement la notion de *complexité* d'une suite aléatoire (Kolmogorov, 1968). Il a été également possible d'établir que l'entropie thermodynamique et celle définie par Shannon étaient deux concepts équivalents. Ces mêmes concepts ont permis d'aboutir à une autre interprétation des notions d'information en statistiques mathématiques et à une nouvelle formulation constructiviste des lois de probabilités (Cover et Thomas, 1991; Féménias, 2003).

Parmi les concepts majeurs de la théorie de l'information, nous avons repris ceux d'*incertitude*, d'*information relative* et d'*information mutuelle*.

Soit un ensemble fini d'éléments sur lesquels on dispose d'une mesure de probabilité Pr définie sur l'ensemble des évènements W. L'***incertitude*** h d'un événement i, au sens de Hartley, est défini par :

$$h(i) = -\log(\Pr(i))$$

Elle est souvent interprétée comme la quantité d'information que l'on peut obtenir sur le système si i se produit. Une fois la notion d'incertitude définie, on peut approcher la notion d'*information mutuelle entre deux événements* définie comme la réduction de l'incertitude de l'un des événements dû à la survenue de l'autre.

Notant incertitude de i/j l'incertitude liée à l'événement i sachant que l'événement j s'est produit, on appelle ***information*** sur i ***relative*** à la survenue de j, noté $I_{j\to i}$, la réduction de l'incertitude de i par celle de i/j :

$$I_{j\to i} = h(i) - h(i/j)$$

En utilisant le théorème de Bayes, on obtient la relation $I_{j\to i} = I_{i\to j}$. *La* quantité $I_{j\to i} = I_{i\to j}$, noté $I(i;j)$ est appelée ***information mutuelle*** entre i et j. Il s'agit d'une fonction du produit de l'ensemble des évènements W par lui-même ($W^2=W\times W$). Dans l'ensemble des réels $I(i;j)$ peut être négative. Le théorème Bayes implique que :

$$h(i \cap j) = h(i) + h(j) - I(i;j)$$

Si l'occurrence d'un évènement tend à rendre l'autre impossible, alors l'information mutuelle sera fortement négative. Si les événements sont complètement indépendants, alors l'information mutuelle est nulle. On notera ici qu'il est également possible de définir une information mutuelle entre variables aléatoires, celle-ci exprimant alors leur liaison statistique. L'avantage par rapport au coefficient de corrélation est qu'elle prend en compte également les liaisons non linéaires entre ces variables aléatoires (Steuer et al., 2002).

Considérant que le score d'alignement de résidus i et j de deux séquences alignées permet d'évaluer la réduction de l'incertitude de l'occurrence de i sachant qu'il est aligné avec j nous avons établi que les ***scores***, calculés à l'aide des matrices de similarité, étaient des expressions de l'***information mutuelle*** à l'échelle des amino acides. L'information mutuelle étant sommable, le score d'un alignement est une expression de l'information mutuelle des séquences alignées (**Article 2**).

Problème de la proximité des séquences dans le CSHP et reconstructions phylogéniques

La projection des protéines homologues dans un CSHP est stable. Tout ajout ou retrait d'une séquence n'altère que sa position relativement aux autres. Les degrés de libertés étant définis par les alignements de chaque résidu, ils sont directement définis par l'information mutuelle entre chaque séquence.

Nous avons recherché une notion de proximité qui soit compatible avec les contraintes de l'information mutuelle. Dans ce contexte, deux séquences identiques, « jumelles », ne sont pas confondues. Une protéine de séquence identique à celle d'une protéine prise comme référence devrait pouvoir être positionnée de façon probabiliste, comme « probablement » la plus proche, mais pas nécessairement confondue. Nous avons introduit une fonction de proximité, appelée *q*-dissimilarité (pour quasi-dissimilarité, voir les rappels bibliographiques) et calculée à partir du score d'alignement par la formule $q=\exp(-s)$, comme un moyen de respecter ces contraintes, et de disposer d'une géométrie permettant d'exprimer le théorème TULIP.

A l'aide de cette fonction de proximité, nous avons proposé une méthode de reconstruction phylogénétique, prenant comme modèle évolutif initial l'horloge moléculaire, modèle qui peut être raffiné, et obtenu des arbres, appelés TULIP trees (en français, tulipiers), qui se sont avérés cohérents par comparaison avec les reconstructions phylogénétiques basées sur des alignements multiples. La projection dans un CSHP et la reconstruction phylogénétique sous forme d'arbres TULIP permet par ailleurs de résoudre une inconsistance de la phylogénie des énolases, incluant les séquences de *Plasmodium falciparum*, rapportée par Keeling et Palmer (2001), les ayant conduit à des conclusions erronées.

L'**Article 3** illustre l'intérêt de la projection dans des CSHP et de la reconstruction d'arbres TULIP, dans le contexte général 1) de la comparaison massive de séquences génomiques et 2) de l'intégration des résultats de ces comparaisons dans des bases de données dédiées. Les aspects pratiques de la méthode que nous avons développée, en particulier pour les mises à jour des bases de comparaison, sont soulignés.

BMC Bioinformatics

Research article

Open Access

A configuration space of homologous proteins conserving mutual information and allowing a phylogeny inference based on pair-wise Z-score probabilities

Olivier Bastien[1,2], Philippe Ortet[3], Sylvaine Roy[4] and Eric Maréchal*[1]

Address: [1]UMR 5019 CNRS-CEA-INRA-Université Joseph Fourier, Laboratoire de Physiologie Cellulaire Végétale; Département Réponse et Dynamique Cellulaire; CEA Grenoble, 17 rue des Martyrs, F-38054, Grenoble cedex 09, France, [2]Gene-IT, 147 avenue Paul Doumer, F-92500 Rueil-Malmaison, France, [3]Département d'Ecophysiologie Végétale et de Microbiologie; CEA Cadarache, F-13108 Saint Paul-lez-Durance, France and [4]Laboratoire de Biologie, Informatique et Mathématiques; Département Réponse et Dynamique Cellulaire, CEA Grenoble, 17 rue des Martyrs, F-38054, Grenoble cedex 09, France

Email: Olivier Bastien - obastien@cea.fr; Philippe Ortet - portet@cea.fr; Sylvaine Roy - sroy@cea.fr; Eric Maréchal* - emarechal@cea.fr

* Corresponding author

Published: 10 March 2005

BMC Bioinformatics 2005, **6**:49 doi:10.1186/1471-2105-6-49

Received: 25 November 2004

Accepted: 10 March 2005

This article is available from: http://www.biomedcentral.com/1471-2105/6/49

© 2005 Bastien et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Popular methods to reconstruct molecular phylogenies are based on multiple sequence alignments, in which addition or removal of data may change the resulting tree topology. We have sought a representation of homologous proteins that would conserve the information of pair-wise sequence alignments, respect probabilistic properties of Z-scores (Monte Carlo methods applied to pair-wise comparisons) and be the basis for a novel method of consistent and stable phylogenetic reconstruction.

Results: We have built up a spatial representation of protein sequences using concepts from particle physics (configuration space) and respecting a frame of constraints deduced from pair-wise alignment score properties in information theory. The obtained configuration space of homologous proteins (CSHP) allows the representation of real and shuffled sequences, and thereupon an expression of the TULIP theorem for Z-score probabilities. Based on the CSHP, we propose a phylogeny reconstruction using Z-scores. Deduced trees, called TULIP trees, are consistent with multiple-alignment based trees. Furthermore, the TULIP tree reconstruction method provides a solution for some previously reported incongruent results, such as the apicomplexan enolase phylogeny.

Conclusion: The CSHP is a unified model that conserves mutual information between proteins in the way physical models conserve energy. Applications include the reconstruction of evolutionary consistent and robust trees, the topology of which is based on a spatial representation that is not reordered after addition or removal of sequences. The CSHP and its assigned phylogenetic topology, provide a powerful and easily updated representation for massive pair-wise genome comparisons based on Z-score computations.

Background

Past events that gave birth to biological entities can be tentatively reconstructed based on collections of descriptors traced in ancient or present-day creatures. Using genomic

sequences, an estimate of the relative time separating branching events, previously supported by geological records, could be formalized using mathematical models. The use of proteins for evolutionary reconstructions was vastly explored as soon as the first amino acid sequences were made available [1-9]. The rich biological information contained in protein sequences stems from their being, on the one hand, translation of genes that reflect the history of genetic events to which the species has been subjected, and on the other hand, effectors of the functions constituting a living creature [10] Since protein sequences are encoded in a 20-amino acid alphabet, they are also considered to embody more *information-per-site* than DNA or RNA [11]; they also exhibit smaller compositional trends [12,13]. When compared, sequences that share substantial features are considered as possible homologues [14], based on the fundamental postulate that can be simply stated as "the closer in the evolution, the more alike and conversely, the more alike, *probably* the closer in the evolution".

As summarized by Otu and Sayood [15], the techniques of molecular phylogenetic analyses can be divided into two groups. In the first case, a matrix representing the distance between each pair of sequences is calculated and then transformed into a tree. In the second case, a tree is found that can best explain the observed sequences under evolutionary assumptions, after evaluation of the fitness of different topologies. Some of the approaches in the first category utilize distance measures [16-19] with different models of nucleotide substitution or amino acid replacement. The second category can further be divided into two groups based on the optimality criterion used in tree evaluation: parsimony [20,21] and maximum likelihood methods [22,23]. For a detailed comparison of these methods see [24].

In phylogeny inference based on distance methods, features separating related proteins are used to estimate an observed distance, also called the p-distance, the simplest measure of which is just the number of different sites between proteins. Divergence time (t), also called genetic distance or evolutionary time, is calculated from the p-distance, depending on assumptions derived from evolutionary models [11,24]. For example, the assumption that mutational events happen with equal probability at each site of any sequence leads to the molecular clock model [2]. Although widely used, it is well-known to be unrealistic and numerous corrections have been proposed to refine it [19,25,26]. By definition, the distance matrix is given as $T = (t_{ab})$ where a and b represent the homologous sequences from the analyzed dataset. Tree reconstruction algorithms are then applied to these matrices [11,24]. Eventually, phylogenetic trees corresponding to the classified sequences are statistically evaluated with bootstrap methods and, when available, calibrated using dated fossils [25,26].

Doolittle et al. [27] have proposed methods for converting amino acid alignment scores into measures of evolutionary time. Similarity between amino acids [28-30] provides a way to weight and score alignments [31]. In practice the optimal alignment of two sequences (a and b) is determined from the optimal score $s(a,b)$ [25,27], computed with a dynamic programming procedure [32,33]. In aligned sequences, conservation is measured at identical sites, whereas variation is scaled at substituted sites. To estimate the variation/conservation balance, the *p-distance* can be given as a function of f_{id}, the fraction of identical residues: *p-distance* = 1 - f_{id}. To take into account that multiple mutations can happen at the same site, an expression of f_{id} was proposed by Doolittle et al. [27] using $s(a^*,b^*)$, the score obtained from randomized a and b sequences [34] and s_{id}, the average score of the sequences compared with themselves [19,25,27]:

$$f_{id}(a,b) = \frac{s(a,b) - s(a^*,b^*)}{s_{id} - s(a^*,b^*)} \qquad (1)$$

To connect pair-wise alignments and phylogeny, divergence time has been approximated:

$$t(a,b) = -\lambda \log[f_{id}(a,b)] \qquad (2)$$

introducing a Poisson correction [2] as a reasonable stochastic law relating amino acid changes and elapsed time. As mentioned earlier, adjustments and corrections of equation (2) were proposed to fit more realistically the complexity of evolution [11,25,35]. This attempt of unification helped reconstructing phylogeny of major lineages [27]. However, detailed phylogenic trees obtained from evolutionary close sequences are not satisfactory. In practice, phylogenies are reconstructed based on multiple alignments. Multiple alignment based (MAB) trees are recalculated when incremented with additional sequences; although MAB methods are usually considered accurate, numerous cases of inconsistencies (incongruence) between observed data and deduced MAB trees are recorded (see [15,36]).

Here, we re-examine the estimate of the *p-distance* between two homologous sequences, based on f_{id}, as a source for geometric positioning, divergence time calculations and evolutionary reconstruction. We based our model on mathematical properties that alignment scores should respect; i) information theory [37,38] applied to sequence similarity, ii) algorithmic theory applied to alignment optimization [28] and iii) alignment probability, particularly in conformity with the TULIP theorem [39]. We used these properties as a framework of constraints to build a

geometric representation of a space of probably homologous proteins and define a theoretically explicit measure of protein proximity. This unified model conserves information in the way physical models conserve matter or energy. The obtained representation of protein sequences is unaltered by adding or removing sequences. Applications include therefore the reconstruction of evolutionary consistent and robust trees, the topology of which is based on a spatial representation that is not reordered after addition or removal of sequences.

Results and discussion

Pair-wise sequence alignment scores in information theory

Criteria to measure the variation/conservation balance between proteins should embody as much as possible the structural and functional potentiality within sequences of amino acids. In the absence of explicit physical criteria, amino acid similarity was solved empirically by measuring amino acid substitution frequencies in alignments of homologous sequences [30,40]. Given two amino acids i and j, the similarity function $s(i,j)$ was set as:

$$s(i,j) = \log \frac{\varpi_{ij}}{\pi_i \pi_j} \qquad (3)$$

where ϖ_{ij} is the observed frequency of substitution of i by j or j by i, and π_i and π_j are the frequencies of i and j in the two aligned sequences. The ϖ_{ij} frequency is the estimate of the probability of substitution of i by j in real alignments; whereas $\pi_i\pi_j$ is the estimate of the probability of substitution under the independency hypothesis. The similarity function gives a 20 × 20 similarity matrix usable to score protein sequence alignments, that can be interpreted in the information theory [37,38] according to the following proposition.

Proposition 1

Amino acid substitution matrix values are estimates of the mutual information between amino acids in the sense of Hartley [37,38]. Consequently, the optimal alignment score computed between two biological sequences is an estimate of the optimal mutual information between these sequences.

Proof

Given a probability law P that characterizes a random variable, the Hartley self-information h is defined as the amount of information one gains when an event i occurred, or equivalently the amount of uncertainty one loses after learning that i happened:

$$h(i) = -\log(P(i)) \qquad (4)$$

The less likely an event i, the more we learn about the system when i happens. The mutual information I between two events, is the reduction of the uncertainty of one event i due to the knowledge of the other j:

$$I_{j \to i} = h(i) - h(i/j) \qquad (5)$$

Mutual information is symmetrical, *i.e.* $I_{j \to i} = I_{i \to j}$, and in the following will be expressed by $I(i;j)$. The self and mutual information of two events i and j are related:

$$h(i \cap j) = h(i) + h(j) - I(i;j) \qquad (6)$$

If the occurrence of one of the two events makes the second impossible, then the mutual information is equal to $-\infty$. If the two events are fully independent, mutual information is null. The empirical measure of the similarity between two amino acids described in equation (3) can therefore be expressed in probabilistic terms:

$$s(i,j) = \log\left(\frac{\varpi_{ij}}{\pi_i \pi_j}\right) = \log(P_{\varpi}(i \cap j)) - \log(P_{\pi}(i)) - \log(P_{\pi}(j)) \qquad (7)$$

where P_{ϖ} is the joint probability to have i and j aligned in a given alignment and P_{π} the measure of probability that amino acids occur in a given sequence. From equations (4) and (6), equation (7) becomes:

$$s(i,j) = h(i) + h(j) - h(i \cap j) \qquad (8)$$

that is

$$s(i,j) = I(i;j) \qquad (9)$$

As a consequence, the similarity function (or score) is the mutual information between two amino acids. Additionally, the score between sequences (the sum of elementary scores between amino acids, [32,33,41,42]) is, according to the hypothesis of independence of amino acid positions, the estimated mutual information between the two given biological sequences.

Once two sequences are aligned, we pose the question whether the alignment score is sufficient to assess that the proteins are conceivably alike and thus evolutionarily related? The theorem of the upper limit of a sequence alignment score probability (TULIP theorem, [39]), sets the upper bound of an alignment score probability, under a hypothesis less restrictive than the Karlin-Altschul model [43]. Given two real sequences a and b ($a = a_1a_2...a_m$ and $b = b_1b_2...b_n$), where $s = s(a,b)$ the maximal score of a pair-wise alignment obtained with any alignment method, b^* the variable corresponding to the shuffled sequences from b, and given $P\{S(a,b^*) \geq s\}$ the probability that an alignment by chance between a and b^* has a higher score than s, then whatever the distribution of the random variable $S(a,b^*)$ the TULIP theorem states:

$$s \geq \mu + k\sigma \Rightarrow P\{S(a,b^*) \geq s\} \leq \frac{1}{k^2}$$

with $k > 1$, μ the mean of $\tilde{S}(a,b^*)$ and σ its standard deviation. The unique restriction on $S(a,b^*)$ is that it has a finite mean and a finite variance. A first corollary of the TULIP theorem is that the Z-score is a statistical test for the probability of a sequence alignment score. We additionally state the following new corollary.

TULIP corollary 2

Given the TULIP theorem conditions, let $z(a,b^*) = \frac{s(a,b) - \mu}{\sigma}$ be the Z-score [44]. Then, $z(a,b^*)$ is the greatest possible value for k ($k \in]1,+\infty[$), which holds relation (10) true. In consequence, with $k = z(a,b^*)$, then

$$P\{S(a,b^*) \geq s\} \leq \frac{1}{z(a,b^*)^2}$$

The best upper bound value for $P\{S(a,b^*) \geq s\}$ is termed $l_U(a,b^*) = \frac{1}{z(a,b^*)^2}$. From the TULIP theorem and corollaries, the comparison of a protein to a given reference a, weighed by an alignment score, is characterized by a bounded probability that the alignment is fortuitous.

Question of the proximity between protein sequences in the light of information theory

Since the optimized alignment score of two protein sequences allows an access to both the mutual information between proteins and an upper bound that the alignment is not fortuitous, one would expect that it is an accurate way to spatially organize proteins sets. A simple relation would be "the higher the mutual information, the nearest". There are three ways to assess the proximity between two objects a and b in a given space E [41]. The first is dissimilarity, a function $f(a,b)$: $E \times E \rightarrow \Re^+$ such that $f(a,b) = 0 \Leftrightarrow a = b$ and $f(a,b) = f(b,a)$; the second is the distance *per se*, that is a dissimilarity such that the triangle inequality is respected: $\forall\ a,b,c \in E$, $f(a,c) \leq f(a,b) + f(b,c)$; and the third is the similarity defined as a function $f(a,b)$: $E \times E \rightarrow \Re$ such that $f(a,a) = \max_b f(a,b)$ and $f(a,b) = f(b,a)$. Representing objects in a space is convenient using the notion of distance. When the optimal alignment is global, *i.e.* requiring that it extends from the beginning to the end of each sequence [32], it is theoretically possible to define a distance *per se*, that is to spatially organize the compared sequences [41]. However, from a biological point of view, global alignment algorithms are not reliable to assess homology of protein domains. Local alignments are better suited, using scoring matrices to find the optimum local alignment and maximizing the sum of the scores of aligned residues [28,31]. In contrast with global alignments, local alignments do not allow any trivial definition of distances [41].

Although amino acid similarity is a function $f(i, j)$: $E \times E \rightarrow \Re$, owing to the local alignment optimization algorithms, the computed score is a function $f(a,b)$: $E \times E \rightarrow \Re^+$, requiring the existence of at least one positive score in the similarity matrices. Thus, when constructing an alignment with the Smith and Waterman [33] method, the constraint that $s(a,b)>0$ (*i.e.* $I(a;b) > 0$) is imposed. This condition is consistent with proposition 1: if two sequences are homologous, knowledge about the first has to bring information about the second, that is to say, the mutual information between the two sequences cannot decrease below zero: $I(a;b) > 0$ (*i.e.* $s(a,b) > 0$). As a consequence, in the following geometric construction we sought a refined expression for the proximity of proteins.

Geometric construction of a configuration space of homologous proteins (CSHP) conserving mutual information

In a set of homologous proteins, any sequence a can be selected as a reference, noted a_{ref}, in respect to which the others are compared. A geometric representation of objects relatively to a fixed frame is known as a configuration space (CS). In physics, a CS is a convenient way to represent systems of particles, defined by their positional vectors in some reference frame. Here, given n similar sequences, it is therefore possible to consider n references of the CSHP. In a given (CSHP, a_{ref}), each amino acid position aligned with a position in the a_{ref} sequence, corresponds to a comparison dimension (CS dimension). Proteins are simply positioned by a vector, the coordinates of which are given by the scores of aligned amino acids. Gaps are additional dimensions of the CS. When considering that local algorithms identify the space of biological interest, *i.e.* a CSHP, the gap penalty is a parameter that maximizes the shared informative dimensions. Thus, given the amino acids mutual information, alignment optimization methods define the relative positions of proteins.

At this point in our construction, a first important property of the CSHP can be deduced. Since mutual information with a_{ref} is sufficient for the full positioning, then positioning of proteins in a given (CSHP, a_{ref}) is unambiguous, unique, and is not altered when proteins are added or removed. In other words, a (CSHP, a_{ref}) is a univocal space.

Given two sequences a and b, if b occurs in (CSHP, a_{ref}), then a also occurs in (CSHP, b_{ref}). The pair-wise alignment

of a and b having no order (symmetry of the mutual information), the positions of b in (CSHP, a_{ref}) is dependent of the position of a in (CSHP, b_{ref}). Thus, once a (CSHP, a_{ref}) has been built, $\forall b \in$ (CSHP, a_{ref}), part of the geometry of (CSHP, b_{ref}) is learnt. Thus, in a CSHP, information needed for the position of n sequences is totally contained in the geometry of the n (CSHP, a_{ref}). This geometric stability is not observed with multiple alignments, which can be deeply modified by addition or removal of sequences. In the CSHP, protein position is unaltered by additions or removals of other proteins. In practice, the construction of CSHP is therefore completely deduced from any all-by-all protein sequence comparison [45,46] and can be easily updated.

The q-dissimilarity, a proximity notion for a geometric representation of the CSHP

In the CSHP, the definition of a distance *per se* based on mutual information is reduced *ad absurdum* (For demonstration, see methods). To define a proximity function i) sharing properties of distance, *i.e.* increasing when objects are further apart, ii) deriving from similarity and iii) relying on mutual information, particularly the property "$f(a,a) \neq f(b,b)$ is possible", we introduce a fourth notion of proximity. Such proximity was called *q-dissimilarity* (for quasi-dissimilarity), a function $f(a,b)$: $E \times E \rightarrow \Re^+$ is defined such that

$$\forall a \in E, f(a,a) = \min_{b \in E} f(a,b) \qquad (12)$$

$$\forall\, a \in E, \forall\, b \in E, f(a,b) = f(b,a) \qquad (13)$$

Let s be a similarity, then $q = e^{-s}$ is a q-dissimilarity, named the 'canonical q-dissimilarity' associated to s. Accordingly, the TULIP theorem allows a statistical characterization of $q(a,b)$ the canonical q-dissimilarity between two sequences a and b.

TULIP corollary 3

From the TULIP corollary 2, relation (14) is simply deduced:

$$P\{Q(a,b^*) \leq q(a,b)\} \leq \frac{1}{z(a,b^*)^2} \qquad (14)$$

with $Q(a,b^*)$ being the random q-dissimilarity variable associated with $S(a,b^*)$. Given a (CSHP, a_{ref}), each sequence b aligned with a is characterized by a q-dissimilarity $q(a,b)$. In geometric terms, b can be represented as a point contained in a hyper-sphere B of radius $q(a,b)$.

The representation of a (CSHP, a_{ref}) shown in Figure 1 is therefore in conformity with all constraints listed earlier and can also serve as a Venn diagram for the setting of events realized following a continuous random variable $Q(a,b^*)$. When a is compared to itself, it is set on a hyper-sphere A of radius $q(a,a)$, which is not reduced to one point. In the context of information theory, it is therefore possible to express that the proximity respects the property "$q(a,a) \neq q(b,b)$ is possible". Considering Figure 1, $P\{Q(a,b^*) \leq q(a,b)\} \leq \frac{1}{z(a,b^*)^2}$ is the probability for a random sequence b* to be in the hyper-sphere B. In conclusion, the *q-dissimilarity* is therefore a proximity notion that allows a rigorous geometric description of the configuration space of homologous proteins, real or simulated, (CSHP, a_{ref}, q).

Unification of pair-wise alignments theory, information theory, p-distance and q-dissimilarity in the CSHP model

A geometric space is a *topological* space when endowed with characterized *paths* that link its elements. Here, paths can be defined as the underlying evolutionary history separating sequences [11]. Given u the common unknown ancestor, then the divergence time $t(a,b)$ is theoretically the summed elapsed times separating u to a and to b. Without any empirical knowledge of u, the simplest approximation for $t(a,b)$ was sought as a function of the fraction of identical residues f_{id}, thus of the p-distance. With the hypothesis of the molecular clock, this function can be given as equation (2), where the transmutation of a and b is a consequence of a Poisson process. By using relation (9) on the equivalence between score similarity and mutual information, then the fundamental postulate "the closer in the evolution, the more alike and conversely, the more alike, *probably* the closer in the evolution" can be reformulated:

Fundamental postulate

Given two homologous proteins a and b, the closer in the evolution, the greater the mutual information between a and b (*i.e.* the optimal computed score $s(a,b)$) and conversely, the greater the mutual information between a and b, *probably* the closer in the evolution.

Whereas the first part of the postulate is a consequence of the conservational pressure on mutual information, the second assertion founds the historical reconstruction underlying a set of biological sequences on statistical concepts. A corollary is that evolution of two homologous proteins is characterized by a loss of mutual information.

In the CSHP, this formulation of the fundamental postulate allows a novel mathematical formalization of the p-distance in probabilistic terms. Basically, the p-distance is the divergence observed between two sequences *knowing* that they share some features (the observed sequences a and b) and that they were identical before the speciation event (sequence u).

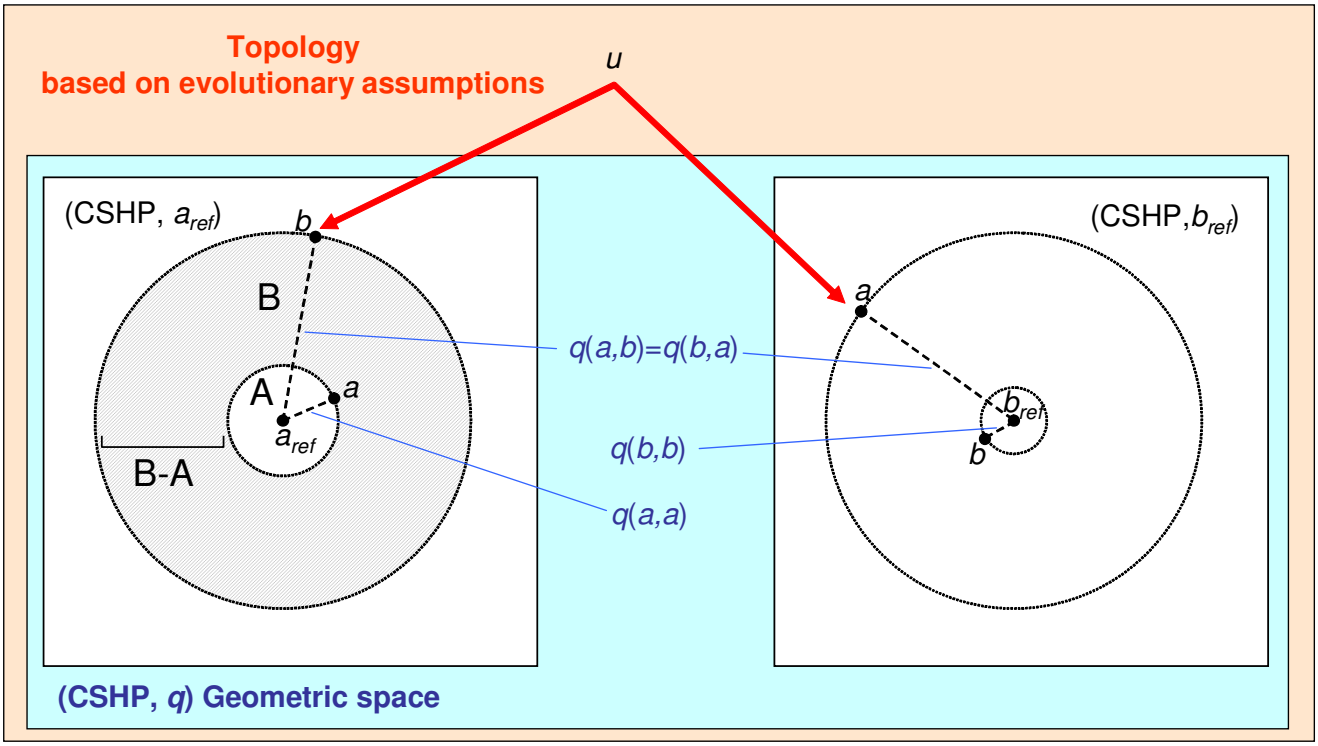


Figure 1
Geometric and probabilistic representation of the configuration space of homologous proteins (CSHP). For any sequence *a* taken as a reference (a_{ref}), one can build a configuration space (CSHP, a_{ref}) where all sequences that are homologous to *a* can be set. When two sequences *a* and *b* are aligned with a score *s*(*a*,*b*), then *b* is positioned in (CSHP, a_{ref}) and *a* in (CSHP, b_{ref}). The sequence alignment length determines the number of configuration dimensions; pair-wise amino acid scores determine the unique solution for its positioning. The q-dissimilarity ($q = e^{-s}$) defines a proximity between sequences allowing a geometric representation (CSHP, *q*). Remarkable properties are i) the conservation of mutual information, [$I(a;b) = I(b;a) \Rightarrow q(a,b) = q(b,a)$], between (CSHP, a_{ref}) and (CSHP, b_{ref}), ii) a probabilistic representation of homologies based on q-dissimilarities by Venn diagrams (A and B) and iii) the assignment of a topology relying on protein evolution assumptions. Evolutionary paths for *a* and *b* lineages, sharing an unknown ancestor *u*, have a probabilistic expression, bounded above (see text), supporting a phylogenetic topology (TULIP trees).

Looking back to equation (1), we can re-formulate f_{id} in probabilistic terms, considering the fraction of shared features (identical sites) *knowing* the observed data and the existence of a common ancestor. Given two proteins *a* and *b*, let us consider the random variable Q(a,b*), defined in TULIP corollary 3. In (CSHP, a_{ref}), shown in Figure 1, one can define the probability law $P\{Q(a,b^*) \leq \rho\}$ as the probability that the q-dissimilarity between b^* and a_{ref} is lower than ρ. The hyper-sphere of radius ρ contains therefore the b^* random sequences sharing informative features with *a* accordingly. The probability $p_{id/a}$ that b^* shares identity with *a*, *knowing* that the q-dissimilarity between b^* and *a* is lower than that between the real sequences *b* and *a*, is:

$$p_{id/a}(b^*) = P\{Q(a,b^*) \leq q(a,a) \,/\, Q(a,b^*) \leq q(a,b)\} \qquad (15)$$

which is a probabilistic expression of f_{id} in respect to the reference a_{ref}. According to the Venn diagram in Figure 1: $p_{id/a}(b^*) = P(A/B)$

Using the Bayes theorem, equation (15) can be expressed as:

$$p_{id/a}(b^*) = P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} \qquad (16)$$

In consequence:

$$p_{id/a}(b^*) = \frac{P\{Q(a,b^*) \leq q(a,a)\}}{P\{Q(a,b^*) \leq q(a,b)\}} \qquad (17)$$

which can be expressed as

$$p_{id/a}(b^*) = \frac{P\{S(a,b^*) \geq s(a,a)\}}{P\{S(a,b^*) \geq s(a,b)\}} \qquad (18)$$

Assuming that substitution rates are independent of lineages [35], then random sequence models a* and b* are equivalent, that is to say $Q(a,a^*) \approx Q(a,b^*)$ and

$$p_{id/a}(b^*) = \frac{P\{Q(a,b^*) \leq q(a,a)\}}{P\{Q(a,b^*) \leq q(a,b)\}} \approx \frac{P\{Q(a,a^*) \leq q(a,a)\}}{P\{Q(a,b^*) \leq q(a,b)\}} \qquad (19)$$

Thus $p_{id/a}$, and symmetrically $p_{id/b}$, provide a probabilistic expression of f_{id} *knowing* the data, *i.e.* the observed mutual information between *a* and *b* expressed as $Q(a,b)$.

Given two homologous sequences *a* and *b*, when their optimal score is $s(a,b) \geq \mu + \psi$ with ψ being a critical threshold value depending on the score distribution law (See Methods for the demonstration for the critical threshold), owing to the TULIP corollary 2, we can state that $p_{id/a}$ is bounded above:

$$p_{id/a}(b^*) \approx \frac{P\{S(a,a^*) \geq s(a,a)\}}{P\{S(a,b^*) \geq s(a,b)\}} \leq \frac{l_v(a,a^*)}{l_v(a,b^*)} = \frac{z(a,b^*)^2}{z(a,a^*)^2} \qquad (20)$$

This expression can also be developed as:

$$p_{id/a}(b^*) \leq \frac{\left(\frac{s(a,b)-\mu_1}{\sigma_1}\right)^2}{\left(\frac{s(a,a)-\mu_2}{\sigma_2}\right)^2} \qquad (21)$$

where μ_1, σ_1, μ_2 and σ_2 are the mean and the standard deviation of $S(a,b^*)$ and $S(a,a^*)$ respectively. The right term in relation (21) exhibits analogies with f_{id} given by equation (1), showing that the pragmatic approach by Feng and Doolittle [19] could be supported and generalized in a theoretical elaboration.

Using the Poisson correction, an expression of $t(a,b)$ is given as the linear combination of the two corrections of the p-distance deduced from $p_{id/a}$ and $p_{id/b}$:

$$t(a,b) = -[\log(p_{id/a}(b^*)) + \log(p_{id/b}(a^*))] \qquad (22)$$

with a^* and b^* the random variables corresponding to the shuffled sequences of *a* and *b* respectively. The sum of the logarithms corresponds to the product of the two probabilities, an expression of the hypothesis of independence of lineage. Interestingly, equation (22) provides an expression of the symmetric effect of time on the variations that independently affected *a* and *b*.

From relation (20), $t(a,b)$ appears as a function of Z-score ratios. For any set of homologous proteins, it is therefore possible to measure a table of pair-wise divergence times and build phylogenetic trees using distance methods.

Reconstruction of protein phylogeny: first example, case study of the glucose-6-phosphate isomerase phylogeny

We compared the trees we obtained, called TULIP trees, to phylogenetic trees built using classical methods, for instance the popular PHYLIP [47] or PUZZLE-based [48] methods, termed here MAB trees (for multiple alignment-based trees). Firstly, because MAB trees are constructed from multiple alignments, removals or additions of proteins modify the multiple alignments. Inclusion of sequences is considered as a way to improve the quality of multiple alignments and to increase the sensitivity of the comparison of distant sequences [49,50]. By contrast, the protein space used to build TULIP trees is not reordered when data sets are incremented or decremented (drawing of the TULIP tree may apparently change due to the tree graphic representation methods; nevertheless the absolute tree topology is not reordered). This remarkable property is due to both the geometrical construction by pair-wise comparison and the convergence of the distance matrix elements estimated by equation (21). Indeed, the estimate of the right-hand term of equation (21) relies on a Monte Carlo method, after randomization of the biological sequences [39,44,51] and is therefore dependent on the sequence randomization model [52] and convergent in respect to the weak law of large numbers [53]. Convergence is proportional to $1/\sqrt{numb_{rand}}$, where $numb_{rand}$ is the number of randomizations. In the case studies presented here, we set $num_{rand} = 2000$ (see Methods). By contrast, stability of MAB trees is sought by bootstrapping approaches and consensus tree reconstruction. MAB trees appear as the result of a complex learning process including possible re-adjustment of the multiple alignments after eye inspection pragmatically applied to assist the reconstruction. Alternatively, Bayesian analyses have been recently proposed for phylogenetic inference [54], estimating posterior probability of each clades to assess most likely trees. Still, in a recent comparative study, Suzuki et al. [55] and Simmons et al. [56] provided evidence supporting the use of relatively conservative bootstrap and jacknife approaches rather than the more extreme overestimates provided by the Markov Chain Monte Carlo-based Bayesian methods. In the absence of any decisive methods to assess the validity of the trees obtained after such different approaches, no absolute comparison with the TULIP classification trees can be rigorously provided.

Whenever a TULIP classification was achieved on a dataset that led to a consensual MAB tree, both were always consistent. For example, Figure 2 shows the phylogenetic PHYLIP [47] and TULIP trees obtained for glucose-6-

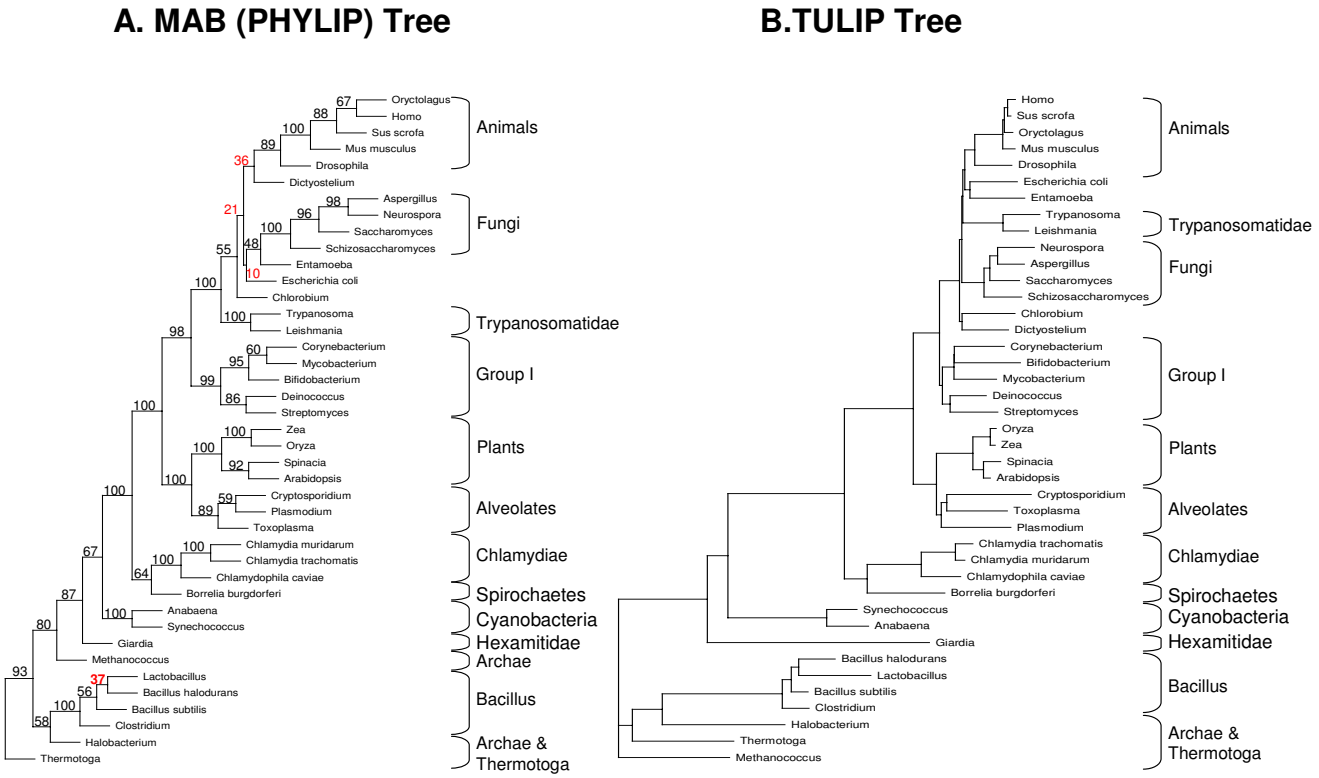


Figure 2
Glucose-6-phosphate isomerase phylogeny. **(A)** Multiple alignment based (MAB) tree. **(B)** TULIP tree. Both trees were constructed using the BLOSUM 62 similarity matrix. For MAB tree construction, bootstrap support was estimated using 1000 replicates. To build TULIP trees, Z-scores were estimated with 2000 sequence shuffling. Topology supported by high bootstrap results in the MAB tree (figures in black), are consistently recovered in the corresponding pair-wise alignement based TULIP tree.

phosphate isomerases (G6PI). Phylogeny of the G6PI enzyme has been studied by Huang et al. [57] in order to demonstrate the horizontal transfer of this enzyme in the apicomplexan phylum due to a past endosymbiosis [57]. Owing to the neighbor-joining analysis used by Huang et al. [57] (see methods) Figure 2A shows that apicomplexan G6PI is "plant-like". The TULIP tree shown in Figure 2B is consistent with this conclusion. Interestingly, differences between the two trees are found only when the bootstrap values on the MAB tree are not strong enough to unambiguously assess branching topology.

Reconstruction of protein phylogeny: second example, case study of the enolase phylogenic incongruence

TULIP classification tree further helps in solving apparent conflicting results obtained with MAB methods. In a comprehensive study from Keeling and Palmer [36] the PUZZLE-based reconstruction of the enolase phylogeny led to incongruent conclusions. Enolase proteins from a wide spectrum of organisms were examined to understand the evolutionary scenario that might explain that enolases from land plants and alveolates shared two short insertions. Alveolates comprise apicomplexan parasites, known to contain typical plant features as mentioned above, particularly a plastid relic. In this context, the shared insertion in apicomplexan and plant enolases (Figure 3) has been interpreted as a possible signature for

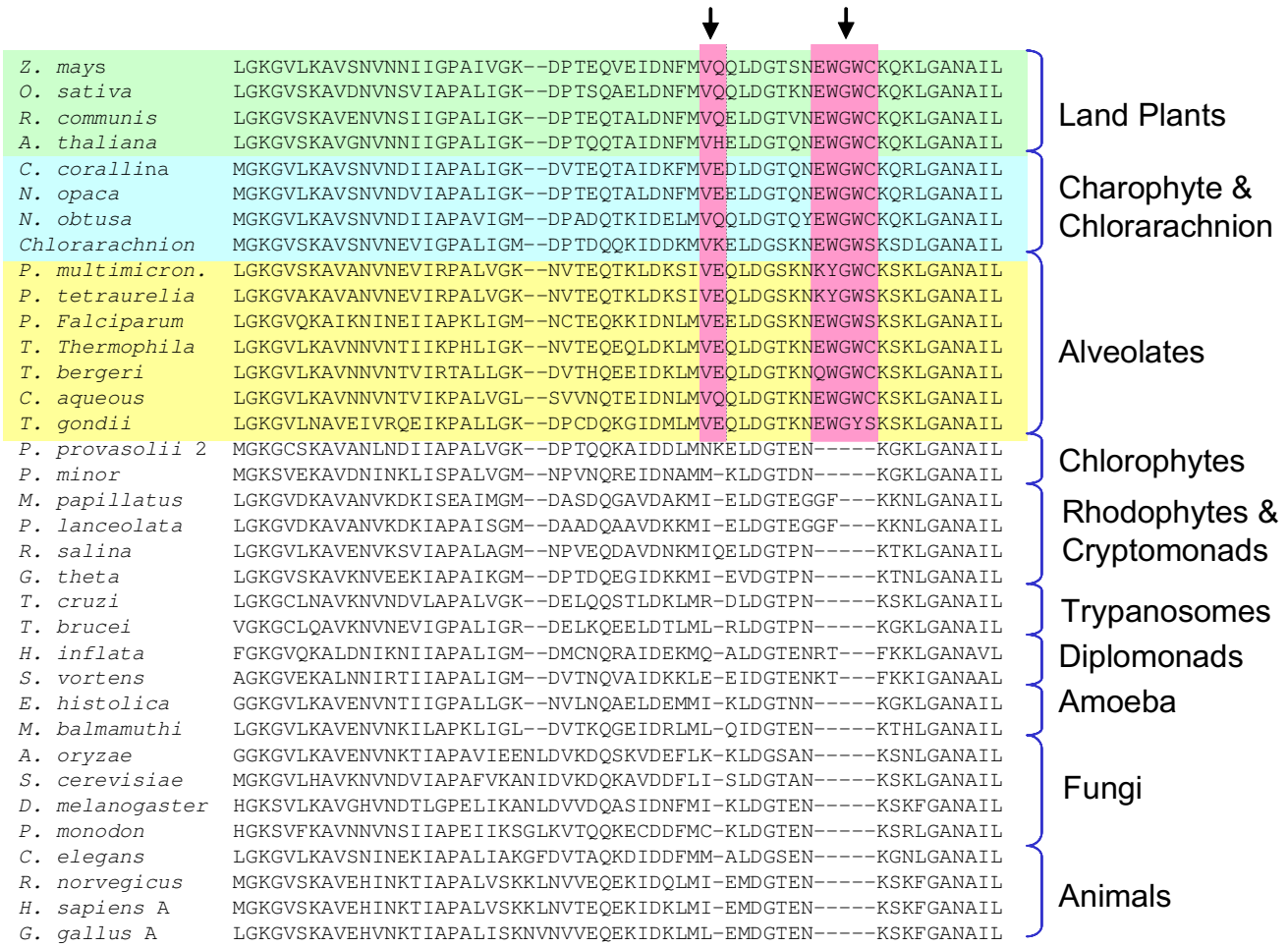


Figure 3
Enolase phylogenic incongruence. When aligned, the enolase region corresponding to amino acids 73–118 of the *Oryza sativa* gene, exhibit two insertions (red boxes) that are only present in land plants, charophytes and alveolates. In alveolates, these insertions are consistent with a horizontal gene transfer. However, to date, evolutionary reconstructions based on enolase sequences did not allow any phylogenetic branch gathering for these clades [36].

some evolutionary relationship between apicomplexans and plants [58,59] and a likely sign of a lateral transfer. From the distribution of this insertion in enolases from several key eukaryotic groups, Keeling and Palmer [36] postulated that lateral transfer had been an important force in the evolution of eukaryotic enolases, being responsible for their origin in cryptomonads, *Chlorarachnion* and *Arabidopsis*. However, they could not conclude about alveolates, finding a conflict between the distribution of the insertion and the MAB phylogenetic position (Figure 4A). The authors had to admit that lateral gene transfers failed to explain apicomplexa enolases, and were compelled to suppose that the lack of congruence between insertion and phylogeny could be because of a parallel loss of insertions in lineages, or to more complex transfers of gene portions.

Based on our theoretical model, we constructed the corresponding TULIP tree. TULIP trees given with BLOSUM 62 or PAM 250 matrices, Fitch-Margoliash or neighbor-joining methods led indistinctly to a unique tree topology (Figure 4B). Separation of great phyla (Archaebacteria, Eubacteria, Diplomonads, Trypanasomes, Animals, Fungi and Amoeba) is recovered. A plant-like cluster is additionally reconstructed, in which a distinct separation occurred between {Rhodophytes ; Cryptomonads} and {Land

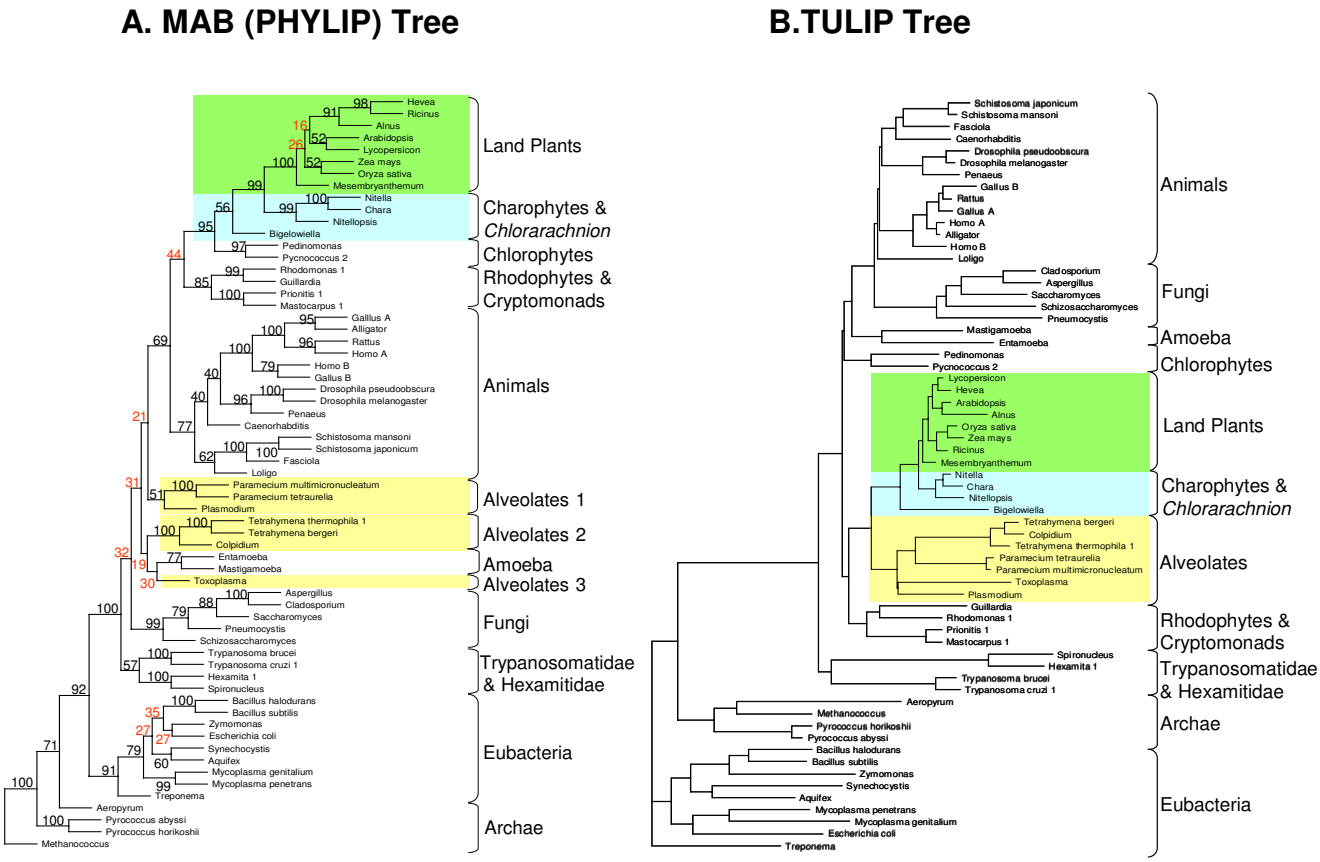


Figure 4
Solution of the enolase phylogenic incongruence. **(A)** Multiple alignment based (MAB) tree. **(B)** TULIP tree. Both trees were constructed using the BLOSUM 62 similarity matrix. For MAB tree construction, bootstrap support was estimated using 1000 replicates. To build TULIP trees, Z-scores were estimated with 2000 sequence shuffling. Clades that contain a unique insertional signature (Land plants, green box; Charophytes and Chlorarachnion, blue box; Alveolates; yellow box) are not gathered in the MAB tree, as previously reported [36]. By contrast, in the TULIP tree, the phylogeny of enolase proteins is reconciled with the insertional signature detection in Land plants, Charophytes, Chlorarachnion and Alveolates.

Plants ; Charophytes ; Chlorarachnion ; Alveolates} main clusters. It is remarkable that this latter cluster is that characterized by the enolase insertion.

This topology corresponds to the observed distribution of the enolase short insertions and provides therefore a solution to the apparent enolase phylogeny incongruence: the phylogenetic position of alveolates is not in conflict with the distribution of enolase insertion and the apicomplexa enolase is possibly a consequence of a lateral transfer, like in cryptomonads.

Large scale phylogeny based on a CSHP built from massive genomic pair-wise comparisons

A CSHP containing large sets of protein sequences can be built after any all-by-all massive comparison providing Z-score statistics. Because the space elaboration is explicit, then quality of the mutual information conservation depends on the choice of the scoring matrix, the geometric positioning depends on the local alignment method, the homology assessment depends on the alignment score and probabilistic cutoffs and the phylogenetic topology on the choice of the stochastic law correction. Eventually,

genome-scale pair-wise comparisons [39,36] find in the present CSHP a robust, evolutionary consistent and easily updatable representation.

Methods

Glucose-6-Phosphate Isomerase sequences

The 41 Glucose-6-phosphate isomerase (EC 5.3.1.9) sequences studied in the paper are taken from several representative groups, as provided from the Swiss-prot database. Group I: Archae ([Swiss-prot:G6PI_HALN1], *Halobacterium* sp.; [Swiss-prot:G6PI_METJA], *Methanococcus jannaschii*). Group II: Bacteria Actinobacteria ([Swiss-prot:G6P1_STRCO], *Streptomyces coelicolor*; [Swiss-prot:G6PI_COREF, *Corynebacterium efficiens*; [Swiss-prot:G6PI_MYCTU], *Mycobacterium tuberculosis*). Group III: Bacteria Cyanobacteria ([Swiss-prot:G6PI_ANASP], *Anabaena sp.*; [Swiss-prot:G6PI_SYNEL], *Synechococcus elongates*). Group III: Bacteria Bacillus ([Swiss-prot:G6PI_LACFE], *Lactobacillus fermentum*; [Swiss-prot:G6PI_BACHD], *Bacillus halodurans*; [Swiss-prot:G6PI_BACSU], *Bacillus subtilis*; [Swiss-prot:G6PI_CLOPE], *Clostridium perfringens*). Group IV: Bacteria Proteobacteria ([Swiss-prot:G6PI_BIFLO], *Bifidobacterium longum*; [Swiss-prot:G6PI_ECOLI], *Escherichia coli*). Group V: Bacteria Chlamydiae ([Swiss-prot:G6PI_CHLTR], *Chlamydia trachomatis*; [Swiss-prot:G6PI_CHLCV], *Chlamydophila caviae*; [Swiss-prot:G6PI_CHLMU], *Chlamydia muridarum*). Group VI: Others Bacteria ([Swiss-prot:G6PI_CHLTE], *Chlorobium tepidum*; [Swiss-prot:G6PI_DEIRA], *Deinococcus radiodurans*; [Swiss-prot:G6PI_BORBU], *Borrelia burgdorferi*; [Swiss-prot:G6PI_THEMA], *Thermotoga maritime*). Group VII: Fungi ([Swiss-prot:G6PI_SCHPO], *Schizosaccharomyces pombe*; [Swiss-prot:G6PI_YEAST], *Saccharomyces cerevisiae*; [Swiss-prot:G6PI_NEUCR], *Neurospora crassa*; [Swiss-prot:G6PI_ASPOR], *Aspergillus oryzae*). Group VII: Eukaryota Viridiplantae ([Swiss-prot:G6PI_ARATH], *Arabidopsis thaliana*; [Swiss-prot:G6PI_MAIZE], *Zea mays*; [Swiss-prot:G6PI_SPIOL, *Spinacia oleracea*; [Swiss-prot:G6PA_ORYSA], *Oryza sativa*). Group VIII: Eukaryota Alveolata Apicomplexa ([Swiss-prot:G6PI_PLAFA], *Plasmodium falciparum*; [Swiss-prot:Q9XY88], *Toxoplasma Gondii*; [Swiss-prot:269_185], *Cryptosporidium parvum*). Group IX: Animals ([Swiss-prot:G6PI_DROME, *Drosophila melanogaster*; [Swiss-prot:G6PI_MOUSE], *Mus musculus*; [Swiss-prot:G6PI_HUMAN], *Homo sapiens*; [Swiss-prot:G6PI_PIG], *Sus scrofa*; [Swiss-prot:G6PI_RABIT], *Oryctolagus cuniculus*; [Swiss-prot:G6PI_TRYBB], *Trypanosoma brucei brucei*). Group X: Other Eukaryota ([Swiss-prot:AY581147], *Entamoeba histolytica*; [Swiss-prot:G6PI_LEIME], *Leishmania mexicana*; [Swiss-prot:AY581146], *Dictyostelium discoideum*; [Swiss-prot:Q968V7], *Giardia intestinalis*).

Enolase sequences

Enolase sequences used for the case-study presented in this paper were taken from eight major groups previously studied by [36]. Group I: Land Plant, Charophytes, Chlorophytes, Rhodophytes and Cryptomonads ([Swiss-prot:CAA39454], *Zea mays*; [Swiss-prot:Q42971], *Oryza sativa*; [Swiss-prot:Q43130], *Mesembryanthemum crystallinum*; [Swiss-prot:P42896], *Ricinus communis*; [Swiss-prot:Q43321], *Alnus glutinosa*; [Swiss-prot:Q9LEJ0], *Hevea brasiliensis* 1; [Swiss-prot:P25696], *Arabidopsis thaliana*; [Swiss-prot:P26300], *Lycopersicon esculentum*; [Swiss-prot:AF348914], *Chara corallina*; [Swiss-prot:AF348915], *Nitella opaca*; [Swiss-prot:AF348916], *Nitellopsis obtusa*; [Swiss-prot:AF348918], *Pycnococcus provasolii* 2; [Swiss-prot:AF348919], *Bigelowiella natans* – Chlorarachnion -; [Swiss-prot:AF348920], *Mastocarpus papillatus* 1; [Swiss-prot:AF348923], *Prionitis lanceolata* 1; [Swiss-prot:AF348931], *Rhodomonas salina* 1; [Swiss-prot:AF348933], *Guillardia theta*; [Swiss-prot:AF348935], *Pedinomonas minor*). Group II : Animals and Fungi ([Swiss-prot:P04764], *Rattus norvegicus*; [Swiss-prot:P51913], *Gallus gallus* A; [Swiss-prot:P07322], *Gallus gallus* B; [Swiss-prot:Q9PVK2], *Alligator mississippiensis*; [Swiss-prot:P06733], *Homo sapiens* A; [Swiss-prot:P13929, *Homo sapiens* B; [Swiss-prot:P15007], *Drosophila melanogaster*; [Swiss-prot:AF025805], *Drosophila pseudoobscura*; [Swiss-prot:O02654], *Loligo pealeii*; [Swiss-prot:AF100985], *Penaeus monodon*; [Swiss-prot:Q27527], *Caenorhabditis elegans*; [Swiss-prot:Q27877], *Schistosoma mansoni*; [Swiss-prot:P33676], *Schistosoma japonicum*; [Swiss-prot:Q27655], *Fasciola hepatica*; [Swiss-prot:P00924], *Saccharomyces cerevisiae* 1; [Swiss-prot:Q12560], *Aspergillus oryzae*; [Swiss-prot:P42040], *Cladosporium herbarum*; [Swiss-prot:P40370], *Schizosaccharomyces pombe* 1; [Swiss-prot:AF063247], *Pneumocystis carinii* f.). Group III: Amoebae ([Swiss-prot:P51555], *Entamoeba histolytica*; [Swiss-prot:Q9U615], *Mastigamoeba balamuthi*). Group IV: Alveolates ([Swiss-prot:AF348926], *Paramecium multimicronucleatum*; [Swiss-prot:AF348927], *Paramecium tetraurelia*; [Swiss-prot:AF348928], *Colpidium aqueous*; [Swiss-prot:AF348929], *Tetrahymena thermophila* I; [Swiss-prot:AF348930], *Tetrahymena bergeri*; [Swiss-prot:Q27727], *Plasmodium falciparum*; [Swiss-prot:AF051910], *Toxoplasma gondii*). Group V: *Trypanosomatidae* ([Swiss-prot:AF159530], *Trypanosoma cruzi* eno1 partial; [Swiss-prot:AF152348], *Trypanosoma brucei* complete). Group VI: Hexamitidae ([Swiss-prot:AF159519], *Hexamita inflata* eno1 partial; [Swiss-prot:AF159517], *Spironucleus vortens* partial). Group VII: Archaebacteria ([Swiss-prot:Q9UXZ0], *Pyrococcus abyssi*; [Swiss-prot:O59605], *Pyrococcus horikoshi*i; [Swiss-prot:Q60173], *Methanococcus jannaschii*; [Swiss-prot:Q9Y927], *Aeropyrum pernix*). Group VII: Eubacteria ([Swiss-prot:O66778], *Aquifex aeolicus*; [Swiss-prot:P37869],*Bacillus subtilis*; [Swiss-prot:Q9K717], *Bacil-*

lus halodurans; [Swiss-prot:P77972], *Synechocystis* sp.; [Swiss-prot:P33675], *Zymomonas mobilis*; [Swiss-prot:P08324], *Escherichia coli*; [Swiss-prot:P47647], *Mycoplasma genitalium*; [Swiss-prot:Q8EW32, *Mycoplasma penetrans*; [Swiss-prot:P74934], Treponema pallidum).

Demonstration that distance of a protein to itself cannot be defined in the CSHP

In the simplest case, building a distance between amino acids (that would lead to distance between sequences) on the basis of computed similarity values would have to respect the condition:

$$\forall i \in E, \forall j \in E, d(i,j) = 0 \Rightarrow i = j \quad \text{(a)}$$

for i and j, two given amino acids and d the distance function. Using this condition in the proposition, any organization of the CSHP with a geometric distance is reduced *ad absurdum*.

Proposition

Building a distance between amino acids derived from the composed function $d(i,j) = (\phi \circ s)(i,j)$, where s is a similarity function and ϕ a bijection, is impossible without a loss of mutual information. Moreover, two proteins from distinct organisms can have the same configuration, being like "twins", and $d(i,j) = 0$ does not imply $i = j$.

Proof

Condition (a) implies that $\phi(s(i,i)) = \phi(s(j,j)) = 0$. This equality imposes that $s(i,i) = s(j,j)$ and, following equation (7) of main text, that $I(i;i) = I(j;j)$. Considering for example tryptophan (W) and glutamic acid (E), if W occurs in a sequence, the mutual information gained about the occurrence of W at the aligned position would be the same as that gained in the case of E about the occurrence of E at the aligned position in the homologous protein. This statement is easily rejected on the basis of biochemical concerns. On one hand, aspartic acid (D) shares common biochemical properties with E, particularly a carboxylic acid, and easily substitutes in homologous sequences. By contrast W, exhibiting a unique biochemical feature, is less substitutable without altering the function. Thus the mutual information $I(E;E)$ is necessarily lower than $I(W;W)$. This that can be checked in scoring matrices such as BLOSUM 62 [30] where $I(E;E) = 5$, $I(D;D) = 6$ and $I(W;W) = 11$. Condition $d(i,i) = 0$ leads to an obvious loss of information. The second assertion of the proposition is obvious.

Determination of the threshold value ψ, for topological reconstructions in the CHSP based on pair-wise alignment score probabilities

An important basis of the reconstruction of a probabilistic evolutionary topology in the CSHP is based on the demonstration that, given S the random variable corresponding to the alignment scores of pairs of shuffled sequences and μ the mean of S, given two homologous sequences a and b, when their optimal score is $s(a,b) \geq \mu + \psi$ (with ψ a critical threshold value depending on the score distribution law), owing to the TULIP corollary 2, we can state that $p_{id/a}$ is bounded above

$$p_{id/a}(b^*) \approx \frac{P\{S(a,a^*) \geq s(a,a)\}}{P\{S(a,b^*) \geq s(a,b)\}} \leq \frac{l_\upsilon(a,a^*)}{l_\upsilon(a,b^*)} = \frac{z(a,b^*)^2}{z(a,a^*)^2}$$

To the purpose of this demonstration, we considered the cumulative distribution function $F(s) = P(S \leq s)$, its derivative $f(s)$ known as the probability density function defined as $dF(s) = f(s)ds$, and the positive delta function $\delta(s) = (s - \mu)^2(1 - F(s))$. Since $\delta(s) = (s - \mu)^2(1 - F(s))$ is null for $s = \mu$ and $\lim_{s \to +\infty} \delta(s) = 0$, the Rolle's theorem implies that $\exists s_0 \in]\mu, +\infty[$ such as $\frac{\partial \delta}{\partial s}(s_0) = 0$ [60]; s_0 corresponds to a maximum of $\delta(s)$ and is therefore the solution of the equation

$$2(1 - F(s)) - (s - \mu) f(s) = 0 \quad \text{(b)}$$

one can express as

$$s - \mu = \frac{2(1 - F(s))}{f(s)} \quad \text{(c)}$$

The $\frac{2(1 - F(s))}{f(s)}$ term corresponds to a continuous function. Interestingly, $\phi(s) = \frac{f(s)}{(1 - F(s))}$ is known as the hazard function [61], that is the probability of s, per score unit (*i.e.* mutual information), conditional to the fact that the pair-wise alignment score is *at least* equal to s. The hazard function is also defined by $\phi(s) = \lim_{ds \to 0} \frac{P(s \leq S \leq s + ds \mid S \geq s)}{ds}$. A critical hypothesis is that $\phi(x)$ function is strictly increasing and conversely that $\frac{2(1 - F(s))}{f(s)} = \frac{2}{\phi(s)}$ is strictly decreasing. Considering $\psi = \lim_{s \to \mu} \frac{2}{\phi(s)} = \frac{2(1 - F(\mu))}{f(\mu)}$, equation 2 has only one solution s_0 and this solution is bounded above:

$$s_0 = \mu + \frac{2(1 - F(s_0))}{f(s_0)} \leq \mu + \psi \quad \text{(d)}$$

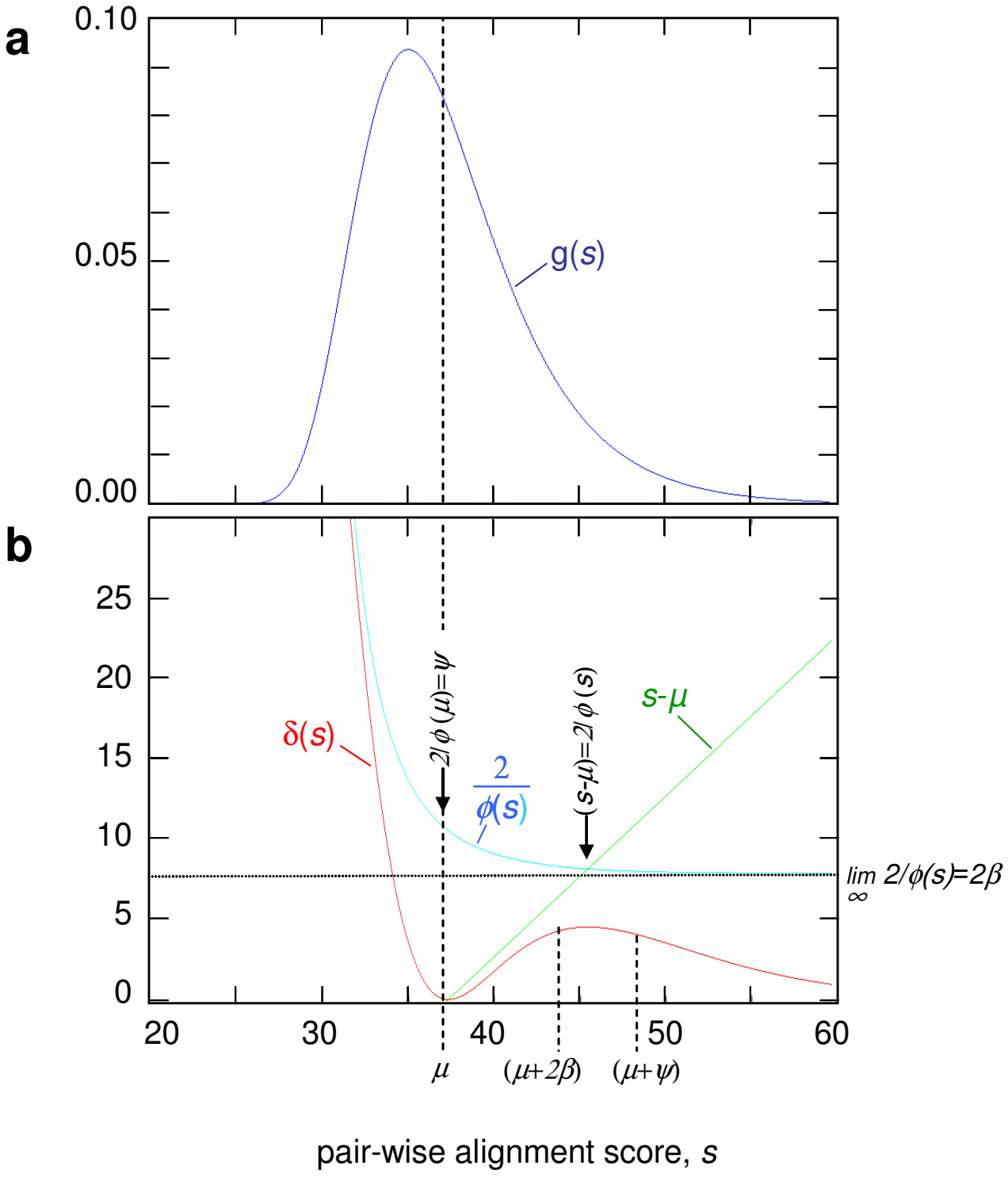


Figure 5
(A) Gumbel score distribution simulated for enolases used in the present paper **(B)** graphical determination of ψ

In consequence, $\delta(s)$ reaches its maximum for a s_0 ($s_0 \leq \mu + \psi$) and it is strictly decreasing on $]\mu + \psi, +\infty[$.

The estimation of s_0 is not trivial because it depends on the knowledge of the cumulative distribution function. Extensive studies provided experimental and theoretical supports for an extreme value distribution of alignment scores [31,43,44]. Using the extreme value distribution of type I, *i.e.* the Gumbel distribution [62], the cumulative distribution is given by

$$G(s) = P(s \leq S) = \exp\left\{-\exp(-\frac{s-\theta}{\beta})\right\} \qquad (e)$$

with θ and β ($\beta > 0$) the location and scale parameters. The probability density function $g(s)$ is defined by $dG(s) = g(s)ds$. We observe with $\varepsilon(s) = -\exp(-\frac{s-\theta}{\beta})$ that $\lim_{s\to+\infty} \varepsilon(s) = 0$. Using the Taylor's polynomial formula, *i.e.* $\exp x \underset{0}{\approx} 1 + x$:

$$\lim_{s\to+\infty} \frac{2}{\phi(s)} = \lim_{\varepsilon\to 0} \frac{2(1-\exp\varepsilon)}{-\frac{1}{\beta}\varepsilon\exp\varepsilon} = \lim_{\varepsilon\to 0} \frac{2(1-1-\varepsilon)}{-\frac{1}{\beta}\varepsilon(1+\varepsilon)} = \lim_{\varepsilon\to 0} \frac{2}{\frac{1}{\beta}+\frac{\varepsilon}{\beta}} = 2\beta \qquad (f)$$

In consequence, for a Gumbel score probability distribution:

$$\mu + 2\beta \leq \mu + \frac{2(1-G(s_0))}{g(s_0)} \leq \mu + \psi \qquad (g)$$

A graphical determination of ψ from a Gumbel distribution is illustrated in Figure 5.

If a pair-wise alignment score of two sequences *a* and *b* is relatively high, that is $s(a,b) \geq \mu + \psi$, then the trivial inequality $s(a,a) \geq s(a,b)$ implies

$$(s(a,b) - \mu)^2(1 - F(s(a,b))) \geq (s(a,a) - \mu)^2(1 - F(s(a,a))) \qquad (h)$$

that is to say

$$\frac{P\{S(a,a^*) \geq s(a,a)\}}{P\{S(a,b^*) \geq s(a,b)\}} \leq \frac{(s(a,b)-\mu)^2}{(s(a,a)-\mu)^2} = \frac{z(a,b^*)^2}{z(a,a^*)^2} \qquad (i)$$

From inequality (i), we deduce that $p_{id/a}$ is bounded above.

Construction of PHYLIP multiple alignment based trees and pair-wise alignment based TULIP trees

To build PHYLIP trees, multiple sequence alignments were created with ClustalW [63]. PHYLIP trees where constructed using the protpars and neighbor modules from the PHYLIP package [47] and the BLOSUM 62 substitution matrix. Bootstrap support was estimated using 1000 replicates. To build TULIP trees, for each couple of sequences *a* and *b*, alignment was achieved with the Smith-Waterman method and the BLOSUM 62 scoring matrices, using the BIOFACET package from Gene-IT, France [64]. We computed estimated *z*-scores $z(a,b^*)$, $z(a,a^*)$, $z(a^*,b)$, $z(b^*,b^*)$, with 2000 sequence shuffling. For all computations, an estimation of the Gumbel parameters θ and β was made using the computed μ and σ of any $S(a,b^*)$ and the formula $\beta = \frac{\sigma}{\pi}\sqrt{6}$ and $\theta = \mu - \beta\Gamma'(1)$, where $\Gamma'(1) \approx 0.577216$ is the Euler constant. In all computations, both Gumbel parameters were very close (in the case of enolases, $mean(\theta)$ = 35.04, $SD(\theta)$ = 0.12, $mean(\beta)$ = 3.92, $SD(\beta)$ = 0.08). As a consequence, the assumption $Q(a,a^*) \approx Q(a,b^*)$ was verified for any pair of sequences. We used the parameters to estimate $\mu = \theta + \beta\Gamma'(1)$ (in the case of enolases, μ = 37.33), and $\mu + \psi \approx \mu + 10.5178 \approx 47.85$. As any pairs of computed scores are higher than this critical threshold, we used relation [20]. Estimation of evolutionary time was achieved according to equations [20] and [22]. Trees were constructed using Fitch-Margoliash and Neighbor-Joining methods [47].

List of abbreviations

CSHP, configuration space of homologous proteins, TULIP, theorem of the upper limit of a score probability

Authors' contributions

OB conceived the main theoretical model, designed and developed the method to build phylogenetic trees and drafted the manuscript. SR and PO participated in the theoretical model refinement and in the design and development of computational methods to build TULIP trees. EM contributed to the conception of this study, participated in its design and coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank John Dunlop for copy editing and Joël Bleuse and Jacques Demongeot for critical reading of this manuscript. We are indebted to Jean-Paul Eynard, Jean-Michel Pabiot and Jean-Jacques Codani for supportive help. Part of this work was financed by Ministère de la Recherche et de la Technologie and by Agence Nationale de la Valorisation de la Recherche Rhône-Alpes.

References

1. Zuckerkandl E, Pauling L: **Molecules as documents of evolutionary history.** *J Theor Biol* 1965, **8:**357-366.
2. Zukerkandl E: **The evolution of hemoglobin.** *Sci Am* 1965, **212:**110-118.
3. Fitch WM, Margoliash E: **Construction of phylogenetic trees.** *Science* 1967, **155:**279-284.
4. Arnheim N, Taylor CE: **Non-Darwinian evolution: consequences for neutral allelic variation.** *Nature* 1969, **223:**900-903.
5. Dayhoff MO: **Computer analysis of protein evolution.** *Sci Am* 1969, **221:**86-95.

6. Arnheim N, Steller R: **Multiple genes for lysozyme in birds.** *Arch Biochem Biophys* 1970, **141:**656-661.
7. DeLange RJ, Smith EL: **Histones: structure and function.** *Annu Rev Biochem* 1971, **40:**279-314.
8. Zuckerkandl E: **Some aspects of protein evolution.** *Biochimie* 1972, **54:**1095-102.
9. Dayhoff MO, Barker WC, McLaughlin PJ: **Inferences from protein and nucleic acid sequences: early molecular evolution, divergence of kingdoms and rates of change.** *Orig Life* 1974, **5:**311-330.
10. Wu TT, Fitch WM, Margoliash E: **The information content of protein amino acid sequences.** *Annu Rev Biochem* 1974, **43:**539-566.
11. Brocchieri L: **Phylogenetic inferences from molecular sequences: review and critique.** *Theor Popul Biol* 2001, **59:**27-40.
12. Singer GA, Hickey DA: **Nucleotide bias causes a genomewide bias in the amino acid composition of proteins.** *Mol Biol Evol* 2000, **17:**1581-1588.
13. Bastien O, Lespinats S, Roy S, Metayer K, Fertil B, Codani JJ, Maréchal E: **Analysis of the compositional biases in Plasmodium falciparum genome and proteome using Arabidopsis thaliana as a reference.** *Gene* 2004, **336:**163-173.
14. Doolittle RF: **Similar amino acid sequences: chance or common ancestry?** *Science* 1981, **214:**149-159.
15. Otu HH, Sayood K: **A new sequence distance measure for phylogenetic tree construction.** *Bioinformatics* 2003, **19:**2122-2130.
16. Jukes TH, Cantor CR: *Mammalian Protein Metabolism* New York: Academic Press; 1969.
17. Kimura M: **A simple model for estimating evolutionary rates of base substitiutions through comparative studies of nucleotide sequences.** *J Mol Evol* 1980, **16:**111-120.
18. Lake JA: **Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances.** *Proc Natl Acad Sci USA* 1994, **91:**1455-1459.
19. Feng DF, Doolittle RF: **Converting amino acid alignment scores into measures of evolutionary time: a simulation study of various relationships.** *J Mol Evol* 1997, **44:**361-370.
20. Camin J, Sokal R: **A method for deducing branching sequences in phylogeny.** *Evolution* 1965, **19:**311-326.
21. Fitch WM: **Toward defining the course of evolution: minimum change for a specific tree topology.** *Syst Zool* 1971, **35:**406-416.
22. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum likelihood approach.** *J Mol Evol* 1981, **17:**368-376.
23. Felsenstein J, Churchill GA: **A hidden Markov model approach to variation among sites in rate of evolution.** *Mol Biol Evol* 1996, **13:**93-104.
24. Salemi M, Vandamme AM: *The Phylogenetic Handbook* Cambridge University Press; 2003.
25. Feng DF, Cho G, Doolittle RF: **Determining divergence times with a protein clock: update and reevaluation.** *Proc Natl Acad Sci USA* 1997, **94:**13028-13033.
26. Nei M, Xu P, Glazko G: **Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms.** *Proc Natl Acad Sci USA* 2001, **98:**2497-2502.
27. Doolittle RF, Feng DF, Tsang S, Cho G, Little E: **Determining divergence times of the major kingdoms of living organisms with a protein clock.** *Science* 1996, **271:**470-477.
28. Dayhoff MO, Barker WC, Hunt LT: **Establishing homologies in protein sequences.** *Methods Enzymol* 1983, **91:**524-545.
29. Risler JL, Delorme MO, Delacroix H, Henaut A: **Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix.** *J Mol Biol* 1988, **204:**1019-1029.
30. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci USA* 1992, **89:**10915-10919.
31. Waterman MS: *Introduction to computational biology* CRC Press; 1995.
32. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *J Mol Biol* 1970, **48:**443-453.
33. Smith TF, Waterman MS: **Identification of common molecular subsequences.** *J Mol Biol* 1981, **147:**195-197.
34. Fitch WM: **Random sequences.** *J Mol Biol* 1983, **163:**171-176.
35. Grishin NV: **Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites.** *J Mol Evol* 1995, **41:**675-679.
36. Keeling PJ, Palmer JD: **Lateral transfer at the gene and subgenic levels in the evolution of eukaryotic enolase.** *Proc Natl Acad Sci USA* 2001, **98:**10745-10750.
37. Hartley RVL: **Transmission of Information.** *The Bell System Technical Journal* 1928, **3:**535-564.
38. Shannon CE: **A Mathematical Theory of Communication.** *The Bell System Technical Journal* 1948, **27:**379-423.
39. Bastien O, Aude JC, Roy S, Maréchal E: **Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of Z-value statistics.** *Bioinformatics* 2004, **20:**534-537.
40. Dayhoff MO, Schwartz RM, Orcutt BC: **A Model of Evolutionary Change in Proteins.** *Atlas of Protein Sequence and Structure* 1978, **5:**345-352.
41. Setubal J, Meidanis J: *Introduction to Computational Molecular Biology* PWS Publishing Compagny; 1997.
42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
43. Karlin S, Altschul SF: **Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.** *Proc Natl Acad Sci USA* 1990, **87:**2264-2268.
44. Comet JP, Aude JC, Glemet E, Risler JL, Henaut A, Slonimski PP, Codani JJ: **Significance of Z-value statistics of Smith-Waterman scores for protein alignments.** *Comput Chem* 1999, **23:**317-331.
45. Bacro JN, Comet JP: **Sequence alignment: an approximation law for the Z-value with applications to databank scanning.** *Comput Chem* 2001, **25:**401-410.
46. Louis A, Ollivier E, Aude JC, Risler JL: **Massive sequence comparisons as a help in annotating genomic sequences.** *Genome Res* 2001, **11:**1296-1303.
47. Felsenstein J: **PHYLIP – Phylogeny Inference Package (Version 3.2).** *Cladistics* 1989, **5:**164-166.
48. Schmidt HA, Strimmer K, Vingron M, von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002, **18:**502-504.
49. Thompson JD, Plewniak F, Poch O: **A comprehensive comparison of multiple sequence alignment programs.** *Nucleic Acids Res* 1999, **27:**2682-2690.
50. Simmons MP, Freudenstein JV: **The effects of increasing genetic distance on alignment of, and tree construction from, rDNA internal transcribed spacer sequences.** *Mol Phylogenet Evol* 2003, **26:**444-451.
51. Manly BFJ: *Randomization, Bootstrap and Monte Carlo Methods in Biology* CRC Press; 1997.
52. White S: **Global statistics of protein sequences: implications for the origin, evolution, and prediction of structure.** *Annu Rev Biophys Biomol Struct* 1994, **23:**407-439.
53. Capinski M, Kopp E: *Measure, Integral and Probability* New-York: Springer-Verlag; 1999.
54. Rannala B, Yang Z: **Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference.** *J Mol Evol* 1996, **43:**304-311.
55. Suzuki Y, Glazko GV, Nei M: **Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics.** *Proc Natl Acad Sci U S A* 2002, **99:**16138-16143.
56. Simmons MP, Pickett KM, Miya M: **How meaningful are Bayesian support values?** *Mol Biol Evol* 2004, **21:**188-199.
57. Huang J, Mullapudi N, Lancto CA, Scott M, Abrahamsen MS, Kissinger JC: **Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in Cryptosporidium parvum.** *Genome Biol* 2004, **5:**R88.
58. Read M, Hicks KE, Sims PF, Hyde JE: **Molecular characterisation of the enolase gene from the human malaria parasite Plasmodium falciparum. Evidence for ancestry within a photosynthetic lineage.** *Eur J Biochem* 1994, **220:**513-520.
59. Dzierszinski F, Popescu O, Toursel C, Slomianny C, Yahiaoui B, Tomavo S: **The protozoan parasite Toxoplasma gondii expresses two functional plant-like glycolytic enzymes. Implications for evolutionary origin of apicomplexans.** *J Biol Chem* 1999, **274:**24888-24895.
60. Lang S: *Undergraduate analysis* New-York: Springer-Verlag; 1997.
61. Valleron AJ: *Introduction à la Biostatistique* Paris: Masson; 1998.
62. Coles S: *An introduction to Statistical Modeling of Extreme Values* New-York: Springer-Verlag; 2001.

63. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
64. Codani JJ, Comet JP, Aude JC, Glémet E, Wozniak A, Risler JL, Hénaut A, Slonimski PP: **Automatic analysis of large-scale pairwize alignments of protein sequences.** *Methods in Microbiology* 1999, **28:**229-244.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral

Title:

The configuration space of homologous proteins: a theoretical and practical framework to reduce the diversity of the protein sequence space after massive all-by-all sequence comparisons

Journal:

Future Generation Computer Systems, in press

Keywords:

Protein space, protein sequence comparison, TULIP, configuration space of homologous protein, CluSTr, Z-value

Authors:

Olivier Bastien (a,b), Philippe Ortet (c), Sylvaine Roy (d) and Eric Maréchal (a,*)

Affiliations :

(a) UMR 5168 CNRS-CEA-INRA-Université Joseph Fourier, Laboratoire de Physiologie Cellulaire Végétale; Département Réponse et Dynamique Cellulaires; CEA Grenoble, 17 rue des Martyrs, F-38054, Grenoble cedex 09, France

(b) Gene-IT, 147 avenue Paul Doumer, F-92500 Rueil-Malmaison, France

(c) Département d'Ecophysiologie Végétale et de Microbiologie; CEA Cadarache, F-13108 Saint Paul-lez-Durance, France

(d) Laboratoire Biologie, Informatique, Mathématiques; Département Réponse et Dynamique Cellulaires, CEA Grenoble, 17 rue des Martyrs, F-38054, Grenoble cedex 09, France

Corresponding author:

* Eric Maréchal

UMR 5019 CNRS-CEA-INRA-Université Joseph Fourier, Laboratoire de Physiologie Cellulaire Végétale, Département Réponse et Dynamique Cellulaire, CEA Grenoble, 17 rue des Martyrs, F-38054, Grenoble cedex 09, France

Tel : +33 (0)4 38 78 49 85

Fax : +33 (0)4 38 78 50 91

E-mail : emarechal@cea.fr

Abstract:

Most of the millions of virtual protein sequences deduced from genomic DNA, and the millions to come, will not be experimentally confirmed, neither their function directly analyzed. Given the cost and low throughput of experimental characterization and of manual examination of sequence annotations, our exploration of the majority of the protein space will rely on our ability to automatically extrapolate the portion of knowledge we have on characterized sequences to unknown sequences, using accurate navigation maps of the protein space. A first reduction of the diversity of proteins at the primary sequence level is based on the detection of shared features according to proximity criteria. Numerous large scale comparisons of hundreds of thousands of protein sequences have been released using the calculation power of supercomputers or grid frameworks. Following these massive comparisons, pragmatic rules have been used to reduce the protein sequence diversity, but none was based on a rigorous and robust spatial projection of protein sequences. Here, we present the configuration space of homologous proteins (CSHP) that can be constructed from the output of an all-by-all pair-wise sequence alignment in which Z-values are computed after Monte-Carlo simulations. In the CSHP, reduction of the protein diversity can be carried out according to an evolutionary model raising real phylogenetic clusters. Accuracy of the phylogenetic clusters designed from the CSHP projection depends on the method used to perform sequence comparisons, the amino acid scoring matrix used to weight alignments, the sequence randomization technique and null model used for Monte-Carlo simulations and the molecular evolution model used to reconstruct phylogenetic trees. In addition, the projection of proteins in the CSHP can be easily updated after sequence database updates, and the CSHP accuracy and phylogenetic topology can be upgraded by improving any of the implemented sub-models. Clusters of homologous proteins can be represented as phylogenetic trees, named TULIP trees, making sense to a bench biologist. We evaluate the use of the CSHP projection to map the protein space using current massive protein comparison outputs, and propose guidelines for future generations of massive protein sequence comparison projects.

1. Introduction.

In 2005, more than 3 millions protein sequences obtained from numerical translation of DNA coding sequences (the reading frames of genes) were electronically stored in public databases. Production rate of virtual proteins is exponential, fueled by international funding, laboratory networking and competition. Additional productivity gain is expected soon, following a recent technological breakthrough [1], *i.e.* a DNA sequencing method allowing a 100 fold increase in throughput, reading for instance 25 million bases of genetic code - the entire genome of some fungi - within hours. Publicity following the release of complete genomes raised higher expectations than what could be reasonably achieved in short term. This is particularly true for human, with a hope of accelerating the identification of genes involved in genetic diseases or cancer; for pathogens, with a hope of identifying new fighting targets; and for crops, with a hope of improving productivity and quality. Meanwhile, chemical technologies made progress in the throughput of small molecule syntheses. Millions of compounds have been physically stored in chemolibraries; their molecular structures were electronically stored. Having in hand 3D structures of protein active sites and of chemolibraries, methods are available to predict interactions in virtual docking experiments (this issue, [2]). PubChem, a repository for molecules acting on biological targets was recently launched [3] and the UniProt protein knowledge base was recently upgraded to report toxic doses of small molecules on proteins [4]. Access to an ocean of small molecular structures and to a deluge of biological sequences raised therefore an enthusiastic challenge: “The goal for the coming decades will be to explore the overlap between chemistry space and protein space” [5]. No doubt a gigantic computing power will be demanded to reach this chemogenomic horizon. Is this post genomic prediction exuberant? Besides upstream quality of data and downstream methods by which interactions might be screened, can we at least define what we mean by protein and small molecule spaces? This paper focuses on our recent advances toward the definition of a stable reference for a protein space and navigation map at the sequence level.

In recent years, large scale comparisons of hundreds of thousands of protein sequences have been released using the calculation power of supercomputers or grid frameworks (Tables 1 and 2). Putting some order into thousands of proteins following these massive comparisons, using representations based on “proximity” criteria, have provided pragmatic rules to reduce the initial diversity. For a long term use, updating pipelines should prevent initial results from being lost and should be easily carried out. Wealth of bioinformatic methods exists to compare and sort proteins at the level of full-length sequences or of sequence sub-domains and motifs. Combining methods is wise to circumvent limitations of each technique, however no combination is considered as a standard. Thus, although some attempts to organize the protein space were undergone by pragmatic combinations of bioinformatic methods, a representation based on such complex workflows is risky on a long term perspective. Given the number of virtual protein sequences (soon exceeding two-digit millions) and the long-term investment, organizing the protein space should start by a simple process, should be biologically sound, theoretically rather than pragmatically defined so as to be as explicit as possible, and should be statistically accurate, with the widest validity domain. Because of the CPU cost, such pre-treatment of the virtual protein sequences should also be easily incremented with newly released protein sequences.

Comparing protein primary sequences can be carried out following two main methods, first is based on multiple alignments, usually achieved on sets of sequences presumed to be homologous, second is based on naïve pairwise comparisons. Multiple alignment-based comparisons, raising phylogenetic protein clusters, are expensive, have to be completely repeated after each data update and are susceptible to raise modified phylogeny

reconstructions when proteins are added or removed from the multiple alignment sets. Thus, although protein clusters based on phylogeny is appealing to biologists, the reduction of protein space diversity based on multiple alignments demands an exorbitant computing power and does not allow the conservation of results through updating. For these reasons, multiple alignment-based comparisons do not appear as a practical way to assist the organization of the protein space.

Few massive pairwise sequence comparisons have been performed, some of which being described in Table 1 and Table 2. Output of an all-by-all comparison of *n* protein sequences is an *n* x *n* table (Figure 1). According to the output table processing, it can be either totally recomputed at each database update, *i.e.* {new+old}-by-{new+old}, or stored and updated by computing complementary {new}-by-{new} and {new}-by-{old} tables. Information is extracted from the output table to help reducing complexity and diversity at the sequence level. Sets of sequences sharing features are named "clusters" [6], that can be used to extrapolate some functional information from characterized to uncharacterized sequences. In most cases, results were not exploited by the potential end-users, *i.e.* bench biologists, at the level one would expect following such expensive experiments. Numerous projects, unlisted here, simply closed the internet portal providing access to their results two or three years after release. One reason for the poor public success of massive comparisons lies in the uneasy access to protein clusters for the bench biologists and the lack of biologically sound or explicit representation.

Here, we present the configuration space of homologous protein sequences, or CSHP [7], as a projection of the virtual protein sequence space that is theoretically supported by information theory [8,9]. It is constructed from pair-wise sequence alignment scores and Z-values obtained after Monte-Carlo simulations. In the CSHP, reduction of the protein diversity can be carried out according to evolutionary models raising phylogenetic clusters. The accuracy of the Darwinian topology underlying the CSHP projection depends on the method used to perform sequence comparisons, the amino acid scoring matrix used to weight the alignments, the sequence randomization technique and null model used for Monte-Carlo simulations and the molecular evolution model used to reconstruct phylogenetic trees. The spatial representation of proteins in the CSHP can therefore be easily updated, and the CSHP accuracy and phylogenetic topology be upgraded by improving any of the implemented sub-models. Diversity and complexity of the protein world is reduced by detecting clusters of phylogenetically related proteins, and clusters can be viewed as real phylogenetic trees, named TULIP trees, making sense to a bench biologist. The CSHP projection aims at providing an underlying order to the protein space, and to assist the definition of other sub-protein spaces at the structural (sequence similarity only) or functional (with links to annotation) levels, as well as subspaces interacting with family of drug scaffolds. It can therefore be a milestone toward a structured definition of a referential protein space for phylogenomics and chemogenomics. We evaluated the possible use of the CSHP projection to map and navigate into the protein space using current massive protein comparison outputs, and we propose guidelines for future generations of massive protein sequence comparison projects.

2. Virtual protein sequence databases

Most of the millions of virtual protein sequences deduced from DNA automatic sequencing, and the millions to come, will not be experimentally confirmed, neither their function directly analyzed. Electronic storage of genes is scattered in specific databases (*e.g.* GDB for Human [10], TAIR for *Arabidopsis* [11], PlasmoDB, for *Plasmodium* [12], etc.) and gathered in three major generalist repositories (*e.g.* EMBL, at the European Bioinformatic Institute [13],

GenBank at the National Center for Biotechnology Information [14] and DDBJ at the DNA Data Bank of Japan [15]). Institutes in charge of these databases, aware of the risk of ending with an inextricable labyrinth with redundancy and inconsistencies, unified the storage of gene entries and exchange their data daily [16]. Coexisting protein databases with differing coverage and priorities were recently unified. Briefly, the PIR-PSD [17] and Swiss-Prot databases [18] contain best characterized sequences, with detailed and manually curated annotations of protein function; they represent therefore a reference, the biologists enrich at their pace. Complementarily, TrEMBL (translation of EMBL nucleotide sequence database) consists of computer-annotated entries derived from the translation of all coding sequences in the nucleotide sequence databases, except for entries already referenced in Swiss-Prot [18]. A centralized authoritative resource, UniProt, has been formed by uniting PIR-PSD, Swiss-Prot and TrEMBL databases [4]. Proportion of protein sequences with high-quality functional annotation in the complete virtual protein sequence set was ~13 % in 2003, ~10 % in 2004 and lower than 8 % in 2005. This proportion will fall below 1 % in 2008. Given the cost and low throughput of experimental protein characterization and of manual curation of sequence annotations, our exploration of > 92 % of the protein space (> 99 % in 3 years) relies on our ability to extrapolate most accurate navigation maps in the protein space.

3. Sequence alignment score measures, optimization methods and statistics for a homology-based reduction of the protein space diversity

Given a new protein sequence, the widely used method to start with is a homology-based annotation transfer, *i.e.* the transfer of portions of knowledge from related sequences stored in databases, on the basis of a suspected common evolutionary origin. When homologues are detected in distinct organisms, they are termed orthologues and when they are suspected to derive from gene duplication inside a given organism, they are termed paralogues. Sequence homology is assessed using dynamic programming techniques similar to those developed to calculate the edit distances between strings. In all cases, underlying rationale for homology search at the sequence level relies on a fundamental postulate that can be stated as: “the closer in the evolution, the more alike and reversely, the more alike, *probably* the closer in the evolution”. Genome annotations based on homology, performed by numerous teams without any standard workflow, ended with a mild result since ~25 % of the virtual protein sequences have no known homologues [5], a figure that reaches ~60 % in some organisms, such as *Plasmodium falciparum* [19]. Paucity of detected homologues may be either due to a real lack of homologues or a technical failure (false negatives). Selection of scoring measures, sequence alignment methods and statistics for a most accurate homology-based reduction of the protein space diversity should therefore be carefully examined before one achieves a massive comparison.

Firstly, alignments are obtained by maximizing (or minimizing) a quantity named “score”, determined at the level of aligned amino acids. To that purpose, scoring matrices are used to weight and sum scores of compared amino acids and find optimal alignments, computed with a dynamic programming procedure. Scoring matrices have been found to be similarity matrices as well [20,21]. Many scoring matrices are available [22-25] and evaluation studies led to the conclusion that those based on a log-odds ratio, like BLOSUM [23], over performed the others [26]. BLOSUM was computed using blocks of aligned sequences with:

$$s(i,j) = \frac{1}{\lambda}\log\left(\frac{q_{ij}}{p_i pj}\right)$$

where i and j are aligned amino acids, $s(i, j)$ the score, q_{ij} the frequency of the observation: “i is aligned with j”, *i.e.* the target frequency, p_i and p_j the background frequencies of i and j

respectively, and λ a scaling factor. We further demonstrated [25] that the score was also related to the *mutual information* in the sense of Hartley [8,9], between the two considered amino acid:

$$s(i,j) = \frac{1}{\lambda} I(i;j)$$

where I is the mutual information between events, *i.e.* the reduction of the uncertainty of event i due to the knowledge of j. Massive comparisons using scoring matrices thus defined allow therefore the extraction of mutual information shared by sequences. The postulate for homology-based annotation transfer is therefore supported by information theory [8,9] and is re-formulated accordingly as: "Given two homologous proteins, the closer in the evolution, the greater their mutual information at the sequence level and reversely, the greater the mutual information at the sequence level, *probably* the closer in the evolution". Thus selection of the amino acid scoring matrix can influence the quality of the information one can transfer, and therefore the accuracy of the homology-based reduction.

Secondly, few algorithms exist to optimize the protein sequence alignments, being either global (at the full-length sequence level) or local. Global alignment algorithms [27] are not accurate to assess homology of domains in modular proteins [21]; local alignments are better suited [28-30]. The SW dynamic algorithm [28] is considered exact, but its computational cost is too high for traditional computing resources, even for small samples of sequences. Heuristic approaches were successful in speeding up alignment processes, most popular being the low CPU demanding BLASTP [30] and FASTA [29] algorithms. BLASTP proved to be efficient from routine comparisons of small sample of proteins directly handled by bench biologists, to batch comparisons of large numbers of sequences monitored by bioinformaticians. Rank in computation speed, *i.e.* SW < FASTA < BLASTP, is considered reverse to that of accuracy. The BLASTP, FASTA and SW methods were successfully parallelized and Grid-enabled (e.g. [31-33]).

Thirdly, confidence in alignment results is estimated according to statistical criteria. It is important to note that an alignment algorithm comes with a statistical model implemented in the code, particularly in the BLAST package, and it is therefore difficult to discuss the current view on sequence comparison methods and statistics independently. Two major statistical models are currently used to test alignment scores. Most common test is an estimate of the *E-value* (short for Expectation value), *i.e.* the number of alignments one expects to find in the database by chance, with equivalent or better scores. It can be determined from the complete distribution of scores. The BLASTP associated statistics defined by Karlin and Altschul [34] are based on the probability of an observed local alignment score according to an extreme value distribution. The number of high-scoring sequence matching regions is estimated above a threshold by a Poisson distribution and allows the computation of a *P-value*, that the score could have occurred by chance, related to the *E-value* by the formula:

$E = \log (1/1\text{-}P)$

The *E-values* can therefore be computed based on the score distribution or on the Karlin-Altschul model (see Table 1). Validity of the Karlin-Altschul *E-value* computation model (and subsequent improvements) requires two restrictive conditions: first, individual residue distributions for the two sequences should not be 'too dissimilar' and second, sequence lengths 'should grow at roughly equal rates' [34]. Validity restrictions listed here are fully acceptable when dealing with protein sequences of average lengths and amino acid distribution, and BLASTP is a good compromise when having access to limited CPU power. However, compositionally biased genomes such as that of *Plasmodium falciparum*, fall

outside of the validity domain for a BLASTP comparison with unbiased sequences [35,36]. Based on a BLASTP semi-automatic annotation procedure, around 60 % of the *Plasmodium* sequences did not have any apparent homology with sequences from other genomes [19] although such proteomic uniqueness appeared doubtful. In addition, the dependence of the *P-value* calculation on the data bank size implies that results may fluctuate through updating, which is not compatible with the construction of a stable reference.

An alternative method to assess the relevance of an alignment was introduced by Lipman and Pearson [37]. It uses the Monte Carlo techniques to investigate the significance of a given score calculated from the alignment of two real sequences *a* and *b*. It can be used to sort results obtained by any comparison methods including BLASTP, although this has not yet been achieved at a massive scale. It is currently used to estimate the probabilities of SW comparisons. The method consists in computing alignments of shuffled sequences from *a* and *b* [38]; the variables corresponding to the shuffled sequences are termed *a** and *b** respectively. These comparisons allow the estimate of an empirical mean score ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) from the distribution of the random variable $\tilde{S}(a^*,b^*)$. The *Z-value* is then defined as:

$$Z = \frac{s(a,b) - \hat{\mu}}{\hat{\sigma}}$$

For the computation of a *Z-value* between two sequences, only one of the two sequences is generally shuffled [39]. For practical reasons, the *Z-value* was reformulated according to a conservative principle, that is the minimum of Z(*a*,*b**) and Z(*a**,*b*) [40]. Alternatively, *Z-value* can be estimated by an average of Z(*a*,*b**) and Z(*a**,*b*) (see Table 2). Computation of Z(*a*,*b*) is known to be convergent and depends on the accuracy of the estimation of μ and σ, and therefore on the number of shuffling, ranging from 100 to 1000 [35,39,41]. The asymptotic law of *Z-value* was shown to be independent of sequences length and amino acid distribution [40]. The estimate of the *Z-value* was additionally dependent on the shuffling method because the shuffling procedure respects the sequence composition but breaks down biological structures [40]. We demonstrated the TULIP theorem (theorem of the upper limit of a score probability, [35]) assessing that *Z-values* can be used as a statistical test, a single-linkage clustering criterion and that $1/Z\text{-}value^2$ was an upper limit to the probability of an alignment score whatever the actual probability law was. From the TULIP theorem and corollaries [7,35], the comparison of a protein to a given reference sequence *a*, weighed by an alignment score, is characterized by a bounded probability that the alignment is obtained by chance.

In practice, a *Z-value* table can be analyzed using the TULIP theorem to detect pairs of proteins that are probably homologues following a *Z-value* confidence cutoff. For instance, a *Z-value* above 10 allows an estimate that the alignment is significant with a statistical risk of $1/Z\text{-}value^2$, *i.e.* 0,01.

The TULIP theorem provides therefore sustained arguments in favor of the Lipman-Pearson model to estimate an alignment probability for massive comparisons; in particular the two restrictions of the Karlin-Altschul model, *i.e.* that the amino acid distributions of compared sequences should not be too dissimilar and that their lengths should be relatively close, are not required. In addition, *Z-values* are completely independent of the alignment length and are therefore normalized values, and *Z-value* statistics are independent of the databank size and therefore stable after each database update. Eventually, update is made easy since the simple

collection of {new}-by-{old} and {new}-by-{new} is sufficient to complete the *Z-value* table obtained from an existing all-by-all comparison (Figures 2 and 3).

Tables 1 and 2 shows a comparative record of protein all-by-all massive comparisons that have been performed to organize, map, and reduce the diversity of the protein space. When available, information on the computing resource (mostly grids and supercomputers) was provided. Table 1 describes projects (COG [42-45], Tribe [46,47], ProtNet [48-50], ProtoMap [51-53], SIMAP [54,55], SYSTERS release 4 [56-58]) in which alignment significance was assessed from *E-value* estimates that are less CPU-demanding than *Z-value* computations in a single massive comparison experiment. The handling of the output *n* x *n* table of *E-values* requires pragmatic post-processing normalization, including asymmetric corrections of *E-values* obtained after permutation of the two aligned sequences (*e.g.* Tribe), or consensus *E-value* computation after alignment with different algorithms (*e.g.* ProtoNet). In all cases, there is no theoretical support to justify that an *E-value* table can be converted into a rigorous and stable metric. The *E-value* table can be converted into a Markov matrix (*e.g.* Tribe, SIMAP), or a close graphic equivalent, *i.e.* graphs connecting protein entries with *E-values* as weights for graph edges (*e.g.* COG, ProtoMap, SYSTERS). The protein sets are organized either by detecting graphs and sub-graphs following pragmatic rules, with granularities depending on *E-value* thresholds, or by distance clustering using *E-value* as a pseudo-metrics, or by Markov-random-field clustering. None of the obtained organization of the protein sequences can be named a spatial projection and none of the obtained clusters can be represented as a phylogenetic reconstruction. Eventually, the economy of computing *E-values* in an all-by-all comparison experiment is not gained in the updating process that requires a complex pipeline or a complete re-calculation.

Table 2 describes projects (Decrypthon [59], TeraProt [60], PhytoProt [61-63], CluSTr [64-67]) in which alignment significance was assessed from *Z-value* estimates. All computations were undergone with SW. It is possible to compute scores and *Z-values* with other alignment algorithms such as BLASTP, but it is very likely that this was not done simply because *E-value* calculation is directly implemented in the BLASTP code, and because *Z-value* computation with BLASTP is not yet coded. Handling of the output *n* x *n* table does not require in-depth processing and is compliant with a straightforward updating process, *i.e.* {new}-by-{new} and {new}-by-{old} computations. Because *Z-value* are sufficient to assess an alignment significance [7,35], results are provided either as raw lists one can sort following *Z-values* (Decrypthon, TeraProt), or as clusters using *Z-values* as single linkage criterion (ClusTr) or more sophisticate clusters obtained following pyramidal classification (PhytoProt). Although *Z-value* computation has a high initial CPU-cost, the statistic validity and easy update of the *Z-value* table are arguments supporting that future massive comparisons using grid and supercomputer power at best should get inspired of this first generation of projects listed in Table 2.

From this overview of alignment score measures, optimization methods and statistics, the methods of choice for a reduction of the virtual proteome from an all-by-all sequence comparison seem to include the use of log odds ratio-derived amino acid scoring matrices (such as BLOSUM), the exact and CPU demanding SW algorithm, and statistics based on *Z-values*. This workflow is summarized in Figure 2 and 3.

4. The configuration space of homologue proteins (CSHP) and its phylogenetic topology, based on pair-wise *Z-value* probabilities

We introduced a representation of protein sequences using pair-wise alignments in which *Z-values* are computed and stored [7]. In a set of *n* homologous proteins, any sequence *a* can be selected as a reference in respect to which the *n*-1 others are compared. Such geometric

representation of objects relatively to a fixed frame is known as a configuration space (CS), and named accordingly the configuration space of homologous proteins or CSHP. In a set of n homologous proteins, it is therefore possible to define n references for the CSHP by permuting the sequences considered as referential. An interesting property of the CSHP derives from the demonstration that sequence similarity computed for an alignment is an expression of the mutual information shared by protein sequences, as stated in the information theory [8,9]. The CSHP is conservative for mutual information, in the way physical spaces can be conservative for metrics, forces or energy. Mutual information with a referential sequence a is sufficient for the full positioning of the n-1 homologues, and the positioning of proteins that share some homology with the reference is unambiguous, unique, and unaltered when proteins are added or removed.

In the CSHP, features separating objects, *i.e.* the sequences, can be used to compute a probabilistic expression of a proximity [7], out of which one can deduce a divergence time t, following assumptions derived from an evolutionary model [68]. For instance, given a and b two homologous sequences, the simple molecular clock hypothesis [69] supposes that t is a measure of the transmutation of a and b as a consequence of a Poisson process. In the CSHP, an evolutionary distance is given by the divergence observed between two sequences *knowing* that they share some features (the observed sequences a and b) and that they were identical before the speciation event (an unknown ancestral sequence u). Considering b^* a shuffled sequence of b, then the probability $p_{id/a}$ that b^* shares identity with a, *knowing* that the proximity between b^* and a is lower than that between b and a, was shown to be bounded above according to the following formula:

$$p_{id/a}(b^*) \leq \frac{\left(\frac{s(a,b)-\mu_1}{\sigma_1}\right)^2}{\left(\frac{s(a,a)-\mu_2}{\sigma_2}\right)^2}$$

where $s(a,a)$ and $s(a,b)$ are the pair-wise optimal alignment scores of a with a and a with b, and μ_1, σ_1, μ_2 and σ_2 are the mean and the standard deviation of $S(a,b^*)$ and $S(a,a^*)$ respectively. Using the Poisson correction, an expression of $t(a,b)$ is given as the linear combination of the two corrections of the evolutionary distances deduced from $p_{id/a}$ and $p_{id/b}$:

$$t(a,b) = -[\log(p_{id/a}(b^*)) + \log(p_{id/b}(a^*))]$$

with a^* and b^* the random variables corresponding to the shuffled sequences of a and b respectively. The sum of the logarithms corresponds to the product of the two probabilities, an expression of the hypothesis of independence of lineage. Thus, $t(a,b)$ appears as a function of *Z-value* ratios.

For any set of homologous proteins, it is therefore possible to measure a table of pair-wise divergence times and build phylogenetic trees using distance methods (see Figure 2). These trees were called TULIP trees. TULIP trees were compared to phylogenetic trees built using conventional methods, for instance the popular PHYLIP [70] or PUZZLE [71] methods based of multiple sequence alignments. TULIP trees proved to perform as well in any unbiased sets of proteins. Moreover, some phylogenetic inconsistencies in trees built with multiple-alignment based methods, particularly including subsets of compositionally biased sequences, or with low boostraping values, could be spectacularly solved with the TULIP tree [7,72].

An advantage of the phylogenetic inference from the CSHP over that obtained from multiple alignments lies precisely in the TULIP tree construction from pair-wise alignments. Whereas

the addition or removal of a sequence can deeply alter the multiple alignment result, and the deduced phylogeny, the *Z-value* and divergence time tables that serve to reconstruct the phylogenetic trees from the CSHP are the result of a Monte-Carlo simulation, which is a converging process at the level of the pair-wise comparison and is not altered by database updates. As a result, whereas a phylogenetic database computed from multiple alignments would require a complete and increasing computation for any update, the TULIP tree calculation simply requires the calculation of the {new}-by-{old} and {new}-by-{new} *Z-values* and divergence times (Figure 3). The CSHP and its assigned Darwinian phylogenetic topology is therefore a theoretical frame of choice for the treatment of results of massive pairwise protein sequence comparisons.

5. CSHP projection and TULIP tree construction from existing massive comparison tables

Can the CSHP be used to obtain a spatial projection of proteins and detect clusters using existing massive comparison tables? Can the Darwinian topology assigned to the CSHP be used to return phylogenetic trees (TULIP trees) from these CSHP projections? We downloaded complete results of one of the projects listed in Table 2, *i.e.* the complete release of TeraProt, which output was made publicly available as lists, one can sort using alignment parameters including *Z-values*. Thus, given a protein entry, we can extract the complete list of sequences which alignment *Z-value* was above a defined cutoff. The subsequent projection of this set of sequences in the CSHP requires the complete table of pairwise *Z-values*. Because of SW score cutoff used in the TeraProt workflow (see Table 2), these required *Z-values* may not have been all collected, and the table may contain holes. A default *Z-value* should therefore be introduced to circumvent this missing data, or the corresponding pairs of sequences should be discarded as the missing alignment may reflect a non-transitive homology.

Figure 4a shows an example of a TULIP tree obtained with sequences that were aligned with a spinach protein (MGD1, UniProt reference Q9SM44) with a Z-value above 30, computed after 100 sequence shuffling. This set contains plant proteins that have the same function (MGD, *i.e.* catalyzing the transfer of a galactose from a UDP-galactose donor onto a lipidic diacylglycerol acceptor) and bacterial proteins having a distinct but related activity (MURG, *i.e.* catalyzing the transfer of an N-acetyl-glucosamine from a UDP-N-acetyl-glucosamine donor onto the Lipid 1 acceptor). The obtained cluster is consistent with the common origin of MGD and MURG sequences and their 3D structural similarity [72]. The TULIP tree highlights the clear evolutionary divergence between these groups of sequences. Figure 4b and c show the TULIP trees obtained after a re-computation of the *Z-value* table with 100 and 2000 sequence randomizations. Although the accuracy of the tree is improved, the reliability of the initial clusters and derived phylogeny shown in Figure 4a, as estimated from the refined tree shown Figure 4b, illustrates that the information collected in the original massive comparison could be extracted with a good level of confidence. The TULIP tree constructed from the TeraProt database was obtained after <15 sec using a bench workstation (HP ProLiant G4 Bi-Xeon - 2,8 Ghz) whereas trees reconstructed with 100 shuffling required 6 minutes computation and reconstruction with 2000 shuffling required less than 2 hours using the bench workstation. This example illustrates the power of the method and the benefit one can get by exploring massive comparison results using the CSHP projection and TULIP phylogenetic underlying topology.

6. Conclusion

The proportion of protein sequences in database that will be of high-quality annotation will fall below 1% in less than 3 years, and the total number of proteins in the virtual proteome will soon exceed 10 millions. The need for a standard theoretical and practical projection of the protein space is urgent. The reduction of the virtual proteome from an all-by-all sequence comparison seem to include the use of odd ratio-derived amino acid scoring matrices (such as BLOSUM), the exact and CPU demanding SW algorithm, and statistics based on *Z-values*. Comparison of the performance and accuracy of BLASTP and SW based on *Z-value* statistics might be determinant in selecting the alignment method for future generations of massive comparison projects. The use of *Z-values* is extremely important to embrace the diversity of the protein world, and particularly proteins that diverge from the average amino acid distribution (compositionally biased [35,36]). The CSHP is obtained by a rigorous projection of proteins, respecting biological constraints and being conservative for mutual information shared by sequences [7]. It can be assigned a topology based on a protein evolution model that allows a construction of phylogenetic trees, named TULIP trees [7]. TULIP trees have numerous advantages over current clustering methods, including their stability and explicit Darwinian construction that makes sense to a bench biologist. The CSHP projection can be used to process the outputs of current massive comparison projects (TeraProt results can be explored at the TULIP 1.1 server that will soon be launched, but the CSHP projection and TULIP tree cluster representation can also be an important improvement for other *Z-value* based systematic massive comparisons) given minor definition of uncollected low Z-values in the database. The use of massive comparison outputs are not restricted to phylogenic explorations, and are used for automatic massive annotation by homology transfers, which require other clustering methods, including species specific occurrence profiles. The CSHP projection is therefore proposed as a sequence-based underlying organization of sequences that should be linked to other protein classification (particularly based on sub-domains and available tertiary structures) and knowledge bases.

References

[1] M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, J.M. Rothberg, Genome sequencing in microfabricated high-density picolitre reactors. Nature 437 (7057) (2005) 376-380.

[2] this issue, *refererence to be included*. Please, note that reference #2 was left blank so as to be filled with a reference of this journal issue dealing with virtual pharmacological screening using Grids. If this was not the case we would provide a reference.

[3] H.J. Feldman, M. Dumontier, S. Ling, N. Haider, C.W. Hogue, CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules. FEBS Lett. 579 (21) (2005) 4685-4691.

[4] A. Bairoch, R. Apweiler, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, L.S. Yeh, The Universal Protein Resource (UniProt). Nucleic Acids Res. 33 (Database issue) (2005) D154-D159.

[5] Y. Ofran, M. Punta, R. Schneider, B. Rost, Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. Drug Discov. Today 10 (21) (2005) 1475-1482.

[6] J. Liu, B. Rost, Domains, motifs and clusters in the protein universe. Curr. Opin. Chem. Biol. 7 (1) (2003) 5-11.

[7] O. Bastien, P. Ortet, S. Roy, E. Maréchal, A configuration space of homologous proteins conserving mutual information and allowing a phylogeny inference based on pair-wise Z-score probabilities. BMC Bioinformatics 6 (1) (2005) 49.

[8] R.V.L. Hartley, Transmission of Information. The Bell System Technical Journal 3 (1928) 535-564.

[9] C.E. Shannon, A Mathematical Theory of Communication. The Bell System Technical Journal 27 (1948) 379-423.

[10] Access to the genome database for Human at http://gdbwww.gdb.org/

[11] Access to the genome database for Arabidopsis at http://www.arabidopsis.org/

[12] Access to the genome database for Plasmodium at http://www.plasmodb.org/

[13] EMBL genome sequence repository at the European Bioinformatic Institute web service: http://www.ebi.ac.uk/embl/

[14] GenBank genome sequence repository at the National Center for Biotechnology Information web service: http://www.ncbi.nlm.nih.gov/entrez/

[15] DDBJ genome sequence repository at the DNA Data Bank of Japan: http://www.ddbj.nig.ac.jp/

[16] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, D.L. Wheeler, GenBank. Nucleic Acids Res. 33 (Database issue) (2005) D34-D38.

[17] C.H. Wu, L.S. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, P. Kourtesis, R.S. Ledley, B.E. Suzek, C.R. Vinayaka, J. Zhang, W.C. Barker, The Protein Information Resource. Nucleic Acids Res. 31 (1) (2003) 345-347.

[18] B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, M. Schneider, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31 (1) (2003) 365-370.

[19] M.J. Gardner, N. Hall, E. Fung, O. White, M. Berriman, R.W. Hyman, J.M. Carlton, A. Pain, K.E. Nelson, S. Bowman, I.T. Paulsen, K. James, J.A. Eisen, K. Rutherford, S.L. Salzberg, A. Craig, S. Kyes, M.S. Chan, V. Nene, S.J. Shallom, B. Suh, J. Peterson, S. Angiuoli, M. Pertea, J. Allen, J. Selengut, D. Haft, M.W. Mather, A.B. Vaidya, D.M. Martin, A.H. Fairlamb, M.J. Fraunholz, D.S. Roos, S.A. Ralph, G.I. McFadden, L.M. Cummings, G.M. Subramanian, C. Mungall, J.C. Venter, D.J. Carucci, S.L. Hoffman, C. Newbold, R.W. Davis, C.M. Fraser, B. Barrell, Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419 (6906) (2002) 498-511.

[20] M.S. Waterman, Introduction to computational biology: Maps, Sequences, and Genomes. CRC Press, 1995.

[21] J. Setubal, J. Meidanis, Introduction to Computational Molecular Biology, PWS Publishing Compagny, Boston, 1997.

[22] M.O. Dayhoff, W.C. Barker, L.T. Hunt, Establishing homologies in protein sequences. Methods Enzymol. 91 (1983) 524-545.

[23] S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U. S. A. 89 (22) (1992) 10915-10919.

[24] J.L. Risler, M.O. Delorme, H. Delacroix, A. Henaut, Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. J. Mol. Biol. 204 (1988) 1019-1029.

[25] O. Bastien, S. Roy, E. Maréchal, Construction of non-symmetric substitution matrices derived from proteomes with biased amino acid distributions. C R Biol. 328 (5) (2005) 445-453.

[26] S. Henikoff, J.G. Henikoff, Performance evaluation of amino acid substitution matrices. Proteins 17 (1) (1993) 49-61.

[27] S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48 (3) (1970) 443-453.

[28] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences. J. Mol. Biol. 147 (1) (1981) 195-197.

[29] W.R. Pearson, D.J. Lipman, Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U. S. A. 85 (8) (1988) 2444-2448.

[30] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool. J. Mol. Biol. 215 (3) (1990) 403-410.

[31] E. Glemet, J.J. Codani, LASSAP, a LArge Scale Sequence compArison Package. Comput. Appl. Biosci. 13 (2) (1997) 137-143.

[32] T. Rognes, E. Seeberg, Six-fold speed-up of Smith-Waterman sequence database searches using parallel processing on common microprocessors. Bioinformatics 16 (8) (2000) 699-706.

[33] A. YarKhan, J. Dongarra, Biological sequence alignment on the computational grid using the GrADS framework. Future Generation Comp. Syst. 21(6) (2005) 980-986.

[34] S. Karlin, S.F. Altschul, Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. U. S. A. 87 (6) (1990) 2264-2268.

[35] O. Bastien, J.C. Aude, S. Roy, E. Maréchal, Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of Z-value statistics. Bioinformatics 20 (4) (2004) 534-537

[36] O. Bastien, S. Lespinats, S. Roy, K. Métayer, B. Fertil, J.J. Codani, E. Maréchal, Analysis of the compositional biases in Plasmodium falciparum genome and proteome using Arabidopsis thaliana as a reference. Gene 336 (2) (2004) 163-173.

[37] D.J. Lipman, W.R. Pearson, Rapid and sensitive protein similarity searches. Science 227 (4693) (1985) 1435-1441.

[38] W.M. Fitch, Random sequences. J. Mol. Biol. 163 (2) (1983) 171-176.

[39] J.P. Comet, J.C. Aude, E. Glemet, J.L. Risler, A. Henaut, P.P. Slonimski, J.J. Codani, Significance of Z-value statistics of Smith-Waterman scores for protein alignments. Comput. Chem. 23 (3-4) (1999) 317-331.

[40] J.N. Bacro, J.P. Comet, Sequence alignment: an approximation law for the Z-value with applications to databank scanning. Comput Chem. 25 (4) (2001) 401-410.

[41] J.C. Aude, A. Louis, An incremental algorithm for Z-value computations. Comput Chem. 26 (5) (2002) 403-411.

[42] R.L. Tatusov, E.V. Koonin, D.J. Lipman, A genomic perspective on protein families. Science 278 (5338) (1997) 631-637.

[43] R.L. Tatusov, D.A. Natale, I.V. Garkavtsev, T.A. Tatusova, U.T. Shankavaram, B.S. Rao, B. Kiryutin, M.Y. Galperin, N.D. Fedorova, E.V. Koonin, The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29 (1) (2001) 22-28.

[44] R.L. Tatusov, N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, S. Smirnov, A.V. Sverdlov, S. Vasudevan, Y.I. Wolf, J.J. Yin, D.A. Natale. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4 (2003) 41.

[45] Acces to COG protein clusters at http://www.ncbi.nlm.nih.gov/COG/

[46] A.J. Enright, V. Kunin, C.A. Ouzounis, Protein families and TRIBES in genome sequence space. Nucleic Acids Res. 31 (15) (2003) 4632-4638.

[47] TribeMCL downloadable at http://www.ebi.ac.uk/research/cgg/tribe/

[48] O. Sasson, A. Vaaknin, H. Fleischer, E. Portugaly, Y. Bilu, N. Linial, M. Linial, ProtoNet: hierarchical classification of the protein space. Nucleic Acids Res. 31 (1) (2003) 348-352.

[49] N. Kaplan, O. Sasson, U. Inbar, M. Friedlich, M. Fromer, H. Fleischer, E. Portugaly, N. Linial, M. Linial, ProtoNet 4.0: a hierarchical classification of one million protein sequences. Nucleic Acids Res. 33 (Database issue) (2005) D216-D218.

[50] Access to ProtoNet massive comparison output at http://www.protonet.cs.huji.ac.il/

[51] G. Yona, N. Linial, M. Linial, ProtoMap: automatic classification of protein sequences and hierarchy of protein families. Nucleic Acids Res. 28 (1) (2000) 49-55.

[52] G. Yona, N. Linial, M. Linial, ProtoMap: automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. Proteins 37 (3) (1999) 360-378.

[53] Acces to ProtoMap protein clusters at http://protomap.cornell.edu/

[54] R. Arnold, T. Rattei, P. Tischler, M.D. Truong, V. Stumpflen, W. Mewes, SIMAP--The similarity matrix of proteins. Bioinformatics 21 (Suppl. 2) (2005) ii42-ii46.

[55] Access to SIMAP massive comparison output at http://mips.gsf.de/services/analysis/simap/ and http://webclu.bio.wzw.tum.de/cgi-bin/simap/start.pl

[56] A. Krause, J. Stoye, M. Vingron, Large scale hierarchical clustering of protein sequences. BMC Bioinformatics 6 (1) (2005) 15.

[57] A. Krause, M. Vingron, A set-theoretic approach to database searching and clustering. Bioinformatics 14 (5) (1998) 430-438.

[58] Acces to SYSTERS V4 protein clusters at http://systers.molgen.mpg.de/

[59] Access to Decrypthon massive comparison output at http://www.infobiogen.fr/services/decrypthon/decrypthon_gb.htm

[60] Access to TeraProt massive comparison output at http://www.infobiogen.fr/services/Teraprot/

[61] J.C. Aude, Y. Diaz-Lazcoz, J.J. Codani, J.L. Risler, Applications of the pyramidal clustering method to biological objects. Comput Chem. 23 (3-4) (1999) 303-315.

[62] A. Louis, E. Ollivier, J.C. Aude, J.L. Risler, Massive sequence comparisons as a help in annotating genomic sequences. Genome Res. 11 (1) (2001) 1296-1303.

[63] Access to PhytoProt protein clusters at http://genoplante-info.infobiogen.fr/phytoprot/

[64] R. Apweiler, M. Biswas, W. Fleischmann, A. Kanapin, Y. Karavidopoulou, P. Kersey, E.V. Kriventseva, V. Mittard, N. Mulder, I. Phan, E. Zdobnov, Proteome Analysis Database: online application of InterPro and CluSTr for the functional classification of proteins in whole genomes. Nucleic Acids Res. 29 (1) (2001) 44-48.

[65] E.V. Kriventseva, F. Servant, R. Apweiler, Improvements to CluSTr: the database of SWISS-PROT+TrEMBL protein clusters. Nucleic Acids Res. 31 (1) (2003) 388-389.

[66] R. Petryszak, E. Kretschmann, D. Wieser, R. Apweiler, The predictive power of the CluSTr database. Bioinformatics 21 (18) (2005) 3604-3609.

[67] Access to CluSTr protein clusters at http://www.ebi.ac.uk/clustr/

[68] L. Brocchieri, Phylogenetic inferences from molecular sequences: review and critique. Theor. Popul. Biol. 59 (1) (2001) 27-40.

[69] E. Zückerkandl, The evolution of hemoglobin. Sci. Am. 212 (1965)110-118.

[70] J. Felsenstein, PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5 (1989) 164-166.

[71] H.A. Schmidt, K. Strimmer, M. Vingron, A. von Haeseler, TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18 (3) (2002) 502-504.

[72] C. Botté, C. Jeanneau, L. Snajdrova, O. Bastien, A. Imberty, C. Breton, E. Maréchal, Molecular modeling and site-directed mutagenesis of plant chloroplast monogalactosyldiacylglycerol synthase reveal critical residues for activity. J Biol Chem. 280 (41) (2005) 34691-34701.

Table 1. Massive all by all protein sequence pairwise comparisons based on alignment E-value statistics

Project name	Protein sequence DB		Pairwise comparison							Diversity reduction and clustering				Updating		Reference and internet access
	Number of non-redundant sequences (first release date)	source	Computing resources	Amino acid scoring matrix	Alignment algorithm	Alignment statistics			Output table	organization principles (pragmatic rule vs theoretical spatial projection)	Clustering method and produced hierarchy	Cluster representation	Cluster statistics	pipeline	frequency	
						Model	Statistical parameter	Cutoff								
COG	17,967 (October 1997) >170,000 (November 2005)	Initially: entries from 7 complete genomes. In 2005: entries from 73 complete genomes.	na	na	BLASTP	Karlin-Altschul model for E-value computation	E-value	na	Best hit (BeT) of each entry, compared to all entries belonging to a distinct genome	Pragmatic organization of data / no explicit spatial projection. Pragmatic rule: construction of BeT graphs including entries in at least three genomes	Detection of consistent patterns in BeT graphs including entries in at least 3 genomes + inspection for protein domains. No hierarchy produced.	Lists of entries belonging to clusters (named COG) classified by occurrence/absence in each genome (species profiling)	Initially none. In 2001, a test was introduced based on an estimate of the probability that an entry is assigned to a COG by chance	Update protein database genome by genome; compute BeT with new entries against entries from genomes previously compared; complete and increment BeT graphs; repeat clustering from new BeT graphs; apply statistical test on graphs	Last database update at time of writing: May 2003. Update frequency of the COG database depends on the release frequency of reliable complete genome. More frequent updates of the entry references are undergone.	[42-45]
Tribe	311,257 (March 2003)	COGENT (83 genomes) + SwissProt	400 node Compaq Alpha DS10 cluster (8h)	na	BLASTP	Karlin-Altschul model for E-value computation	E-value	E-value ≤ 1.10^{-10}	Alignment score and E-values processed to correct asymmetric BLAST hits and scores	Pragmatic organization of data / no explicit spatial projection. Conversion of E-value table into Markov Matrix	Markov-random-field clustering with TribeMCL algorithm. No hierarchy produced.	Clusters of multiple granularities by altering inflation value used for clustering	na	Update protein database; compute missing comparisons (old by old and old by new); apply E-value cutoff; repeat clustering from new Markov Matrix	No current public access to protein clusters	[46,47] No current public access to protein clusters
ProtoNet	94,153 (July 2002) > 1,000,000 (September 2004)	SwissProt, latter expanded to TrEMBL and eventually to UniProt	na	BLOSUM 62	BLASTP	Karlin-Altschul model for E-value computation	E-value	E-value ≤ 10	E-values	Pragmatic organization of data / no explicit spatial projection. Pseudo-metrics based on E-values.	Hierarchical detection of clusters following a non supervised procedure. Degrees of granularity are defined by merging clusters based on E-value average defined by the arithmetic, square, geometric or harmonic means computed with E-values taken as a kind of metrics.	Clusters of multiple granularities following the merging rule. Number and size of clusters vary with the method of mean computation. Default strategy is unweighted pair group with arithmetic mean (UPGMA), generating UPGM trees (which are not phylogenetic trees)	Non supervised definition of cluster hierarchies based on automatic condensation rules	Update protein database; the massive comparison update, ie all comparisons (old + new by old + new) or only marginal (old by new and new by new), is na; the update procedure for E-values is na; repeat UPGMA clustering.	Last database update: September 2004. Update frequency is currently once every 6 months.	[48-50]

na: not available

Table 1 (Continued). Massive all by all protein sequence pairwise comparisons based on alignment E-value statistics.

Project name	Protein sequence DB		Pairwise comparison							Diversity reduction and clustering				Updating		Reference and internet access
	Number of non-redundant sequences (first release date)	source	Computing resource	Amino acid scoring matrix	Alignment algorithm	Alignment statistics			Output table	organization principles (pragmatic rule vs theoretical spatial projection)	Clustering method and produced hierarchy	Cluster representation	Cluster statistics	pipeline	frequency	
ProtoMap	365,174 (May 2000)	SwissProt + TrEMBL	MOSIX Parallel system and cluster of PCs	BLOSUM 50 + BLOSUM 62	BLASTP + FASTA + SW	Expectation value for an alignment score computed with the real distribution of all scores obtained with one of the aligned sequences against all the protein database; for BLASTP, E-value is obtained from the Karlin-Altschul model	E-value	E-value ≤ 1	Consensus E-values obtained after numerical normalization of E-values obtained with all alignment algorithms	Pragmatic organization of data / no explicit spatial projection. Pragmatic rule: detection of graphs, in which the weight of an edge connecting two proteins is the E-value (Markovian representation). Pseudo-metrics based on E-values: Consensus E-values are taken as distances for tree representations.	Hierarchical detection of graphs starting with a granularity determined by most stringent E-values (< 10^{-100}), and growing by detecting relations with E-values lower than 10^{-95}, 10^{-90}, etc. 10^{0} (=1)	Graphs with E-values as edge weights and E-value based dendograms (which are not phylogenetic trees)	Statistical test applied to detect and eliminate 'problematic' and 'possibly false' connections between unrelated proteins. Quality of connections is defined as minus the log of the geometric mean of pairwise E-values between connected sequences	Update protein database; save original score distribution and scores with E-values <1; compute missing comparisons (old by old and old by new); increment score distribution; compute all E-values with the new score distribution; apply statistical test on E-value; define graphs; apply statistical test on graphs	Last database update at time of writing: May 2000.	[51-53]
SIMAP	> 3,500,000 (From 2003 to October 2005)	UniProt + mips nonredH + PFAM + PDB + entries of complete genomes	Versatile pipeline running on Sun Gridengine clusters and BOINC-based grid systems	BLOSUM 50	FASTA + SW	FASTA is used in a low cost prerun to detect best hits and SW is used in a second run on the selected pairs of sequences to comput most exact alignment scores	E-value	SW score ≥ 80	Alignment raw parameters including SW score, and E-values	Pragmatic organization of data / no explicit spatial projection. Conversion of E-value table into Markov Matrix. Detection of bidirectional best hits.	Markov-random-field clustering with Amsterdam MCL algorithm. No hierarchy produced.	Clusters of multiple granularities by altering inflation value used for clustering. Clusters are provided as lists of entries that can be sorted based on alignment parameters (e.g. length, identity, score, Z-score). List is used to generate multiple alignments and hidden Markov models.	na	Update protein database; compute missing comparisons (old by old and old by new); apply SW score cutoff; repeat clustering from new Markov Matrix	Last database update at time of writing: October 2005. Important update frequency is currently ~bimonthly with peaks every 6 months.	[54,55]
SYSTERS V4	1,168,498 (November 2003)	SwissProt + TrEMBL + entries of 11 complete eukaryotic genomes	Paracel GeneMatcher Workstation	na	SW	Expectation value for an alignment score computed with the real distribution of all scores obtained with one of the aligned sequences against all the protein database	E-value	E-value ≤ 0.05	E-values	Pragmatic organization of data / no explicit spatial projection. Pragmatic rule: detection of graphs, in which the weight of an edge connecting two proteins is the E-value (Markovian representation). Pseudo-metrics based on E-values: E-values are taken as distances for tree representations.	SYSTEmatic RE-Searching clustering: stepwise construction of of single linkage clustering trees; hierarchy based on internal structure of trees; construction of superfamily E-value graphs (E-values as distances); subcluster detection at weak connections.	Lists of entries belonging to clusters; post-computation of DIALIGN multiple alignments and of the corresponding unweighted pair group with arithmetic mean (UPGMA) trees. In previous release, multiple alignements were performed using CLUSTALW	Cluster construction is estimated self-validating	Update protein database; the massive comparison update, ie all comparisons (old + new by old + new) or only marginal (old by new and new by new), is na; the update procedure for E-values is na; repeat SYSTEmatic RE-Searching clustering.	Last database update at time of writing: November 2003. Four releases since 1999, with major method modifications (from BLASTP to SW). Public accesses to releases 3 and 4.	[56-58]

na: not available

Table 2. Massive all by all protein sequence pairwise comparisons based on alignment Z-value statistics.

Project name	Protein sequence DB		Pairwise comparison							Diversity reduction and clustering				Updating		Reference and internet access
	Number of non-redundant sequences (first release date)	source	Computing resource	Amino acid scoring matrix	Alignment algorithm	Alignment statistics			Output table	organization principles (pragmatic rule vs theoretical spatial projection)	Clustering method and produced hierarchy	Cluster representation	Cluster statistics	pipeline	frequency	
						Model	Statistical parameter	Cutoff								
Decrypthon	559,275 (January 2002)	SwissProt + TrEMBL + entries of 76 complete genomes	Grid computing model running with unused computing power of 75000 PCs (200h each), network architecture corresponds to a virtual calculation power > 40 Teraflops)	BLOSUM 62	SW	Z-score computation by 50 shuffling of each aligned sequence (100 shuffling per comparison)	Z-score	Z-score ≥ 5	Alignment raw parameters and Z-score	na	Basic definition based on Z-score thresholds. No hierarchy produced.	Clusters are provided as lists of entries that can be sorted based on alignment parameters (e.g. length, identity, score, Z-score)	na	na	Last database update at time of writing: January 2002	[59]
TeraProt	240,000 (July 2002)	entries of 67 complete genomes	645 node Compaq ES45 cluster (16 to 512 processors used depending on the jobs; 76.3h cumulated time, equivalent Tera)	PAM 10	SW	Z-score computation by 100 shuffling of each aligned sequence (200 shuffling per comparison) and conservation of the minimum Z-score obtained	Z-score	SW score ≥ 220	Alignment raw parameters and Z-score	na	Basic definition based on Z-score thresholds. No hierarchy produced.	Clusters are provided as lists of entries that can be sorted based on alignment parameters (e.g. length, identity, score, Z-score)	na	Update protein database; compute missing comparisons (old by old and old by new); apply score cutoff	Last database update at time of writing: July 2003 with entries of 22 additional genomes	[60]
PhytoProt	14,723 (January 2001)	All entries of plants in SwissProt + TrEMBL	Multiprocessor Sun Sparc server 4500	na	SW	Z-score computation by n shuffling (varying from 30 to 600 to keep control to the variance of Z) of each aligned sequence and conservation of the minimum Z-score obtained.	Z-score	Z-score > 14	Alignment raw parameters and Z-score	Pragmatic organization of data / no explicit spatial projection. Use of Z-score table for pyramidal classification. Post-analysis of sequences to detect possibly shared sub-domains within clusters.	Clusters defined using Z-score thresholds. Pyramidal classification of each cluster based on Z-scores. Some proteins may belong simultaneously to two classes.	Pyramidal classification trees (which are not phylogenetic trees) and list of proteins with PRODOM based sub-domain representation.	na	Update protein database; compute missing comparisons (old by old and old by new); apply Z-score cutoff; repeat cluster detection and pyramidal classification.	Last database update at time of writing: July 2002	[61-63]
CluSTr	462,000 (May 2004)	Initially mammalian and plant proteins in SwissProt + TrEMBL and entries of 3 complete eukaryote proteins, latter expanded to UniProt (including 195 complete genomes)	SARA supercomputer, 1024-CPU system consisting of two 512-CPU SGI Origin 3800 (one with 32 to 256 CPU partition, the other with a single 512 CPU partition, number of processors was adjusted depending on jobs, total run: 2 months)	na	SW	Z-score computation by 100 shuffling of each aligned sequence (200 shuffling per comparison) and conservation of the minimum Z-score obtained	Z-score	na	Alignment raw parameters and Z-score	Pragmatic organization of data / no explicit spatial projection. Use of Z-score table for single-linkage clustering. Links of results with Knowledge bases, through the Integr8 platform.	Hierarchical clusters defined using Z-score. Clusters are determined by single-linkage method. Hierarchy is built incrementally.	Binary forest representing the hierarchy of clusters, in which each parent cluster has only two clusters.	Pruning of clusters following a set of rules (e.g. exclusion of a cluster when its members form ≥90% of its parent set, singletons). Obtained clusters are collected in CluSTr Slim.	Update protein database; compute missing comparisons (old by old and old by new); apply Z-score cutoff; repeat cluster detection and hierarchy classification.	Last update at time of writing: May 2004. Marginal update of database with UniProt entries is announced to be bi-weekly at CluSTr internet portal.	[64-67]

na: not available

Figures

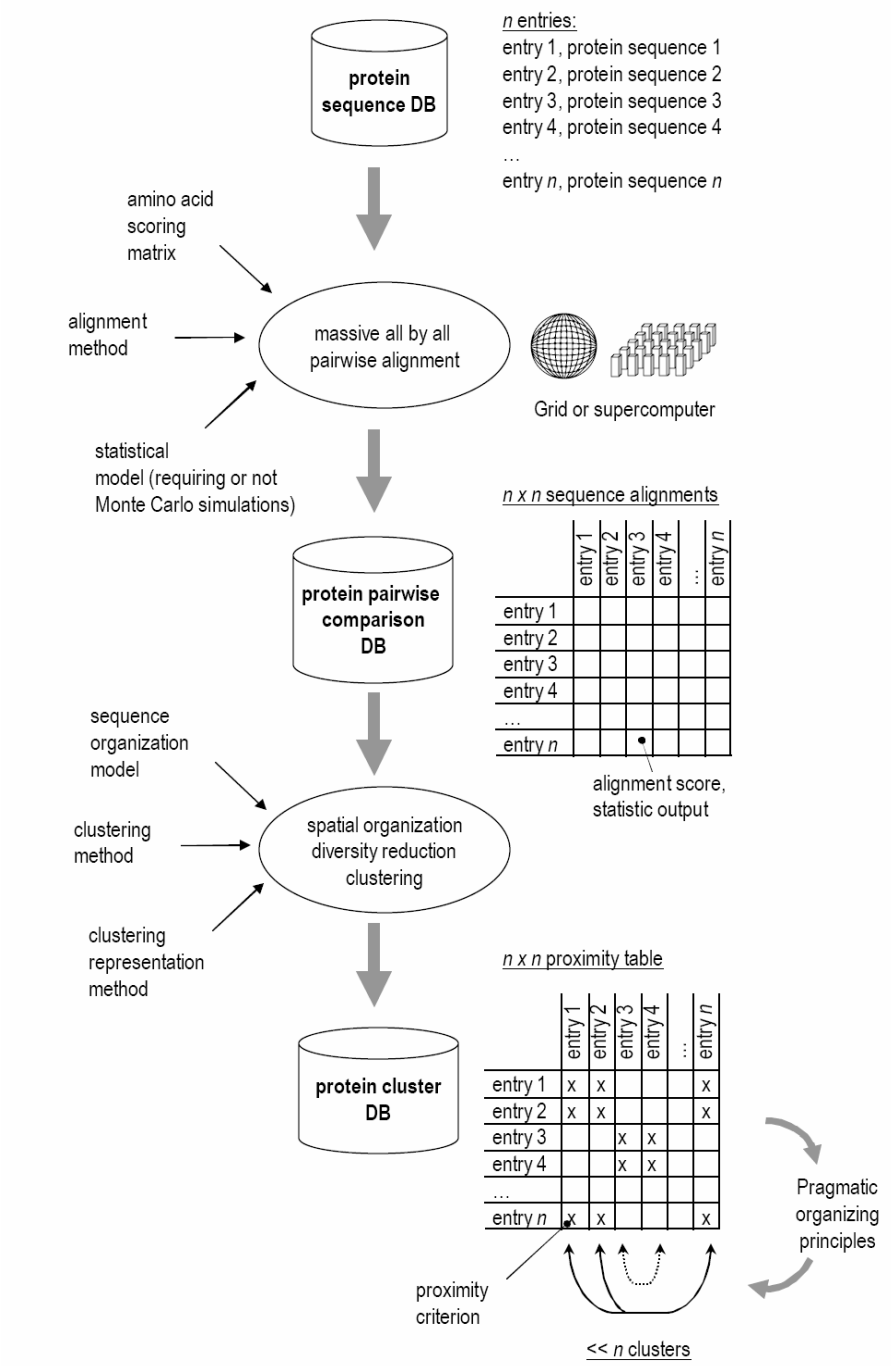


Fig. 1. Massive all-by-all protein sequence comparison and protein space reduction workflow. From a protein sequence database, a massive comparison is carried out using an amino acid scoring matrix, an alignment method and a corresponding statistical model. The output of the comparison is a table that is used for the reduction of the initial protein diversity at the sequence level, following pragmatic rules. Obtained clusters can be viewed as lists, graphs, or classification trees which are not phylogenetic trees. According to the statistical model used to estimate the accuracy of alignment scores, the update of the initial protein sequence database can imply a complete re-computation of the massive all by all pairwise alignment in order to update the clusters.

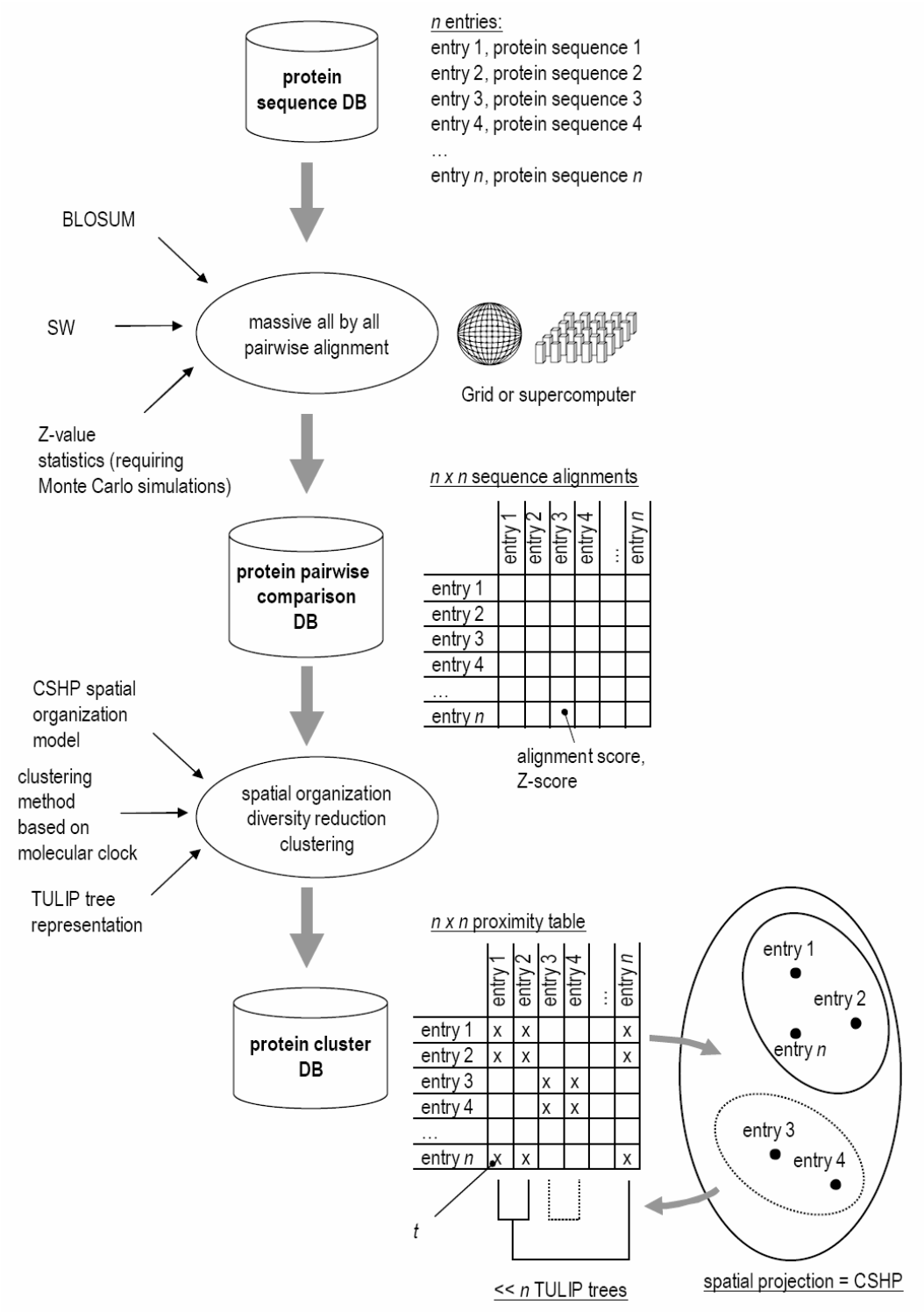


Fig. 2. Massive all-by-all protein sequence comparison and protein space reduction workflow, based on Z-value statistics and protein projection in the CSHP. From a protein sequence database, a massive comparison is carried out using an amino acid scoring matrix such as BLOSUM, the SW alignment method and the TULIP statistical model based on Z-values (requiring Monte Carlo simulation). The output of the comparison is a table containing SW scores and Z-values that are used for the reduction of the initial protein diversity at the sequence level, following the rigorous CSHP spatial projection. Obtained clusters can be viewed as phylogenetic TULIP trees.

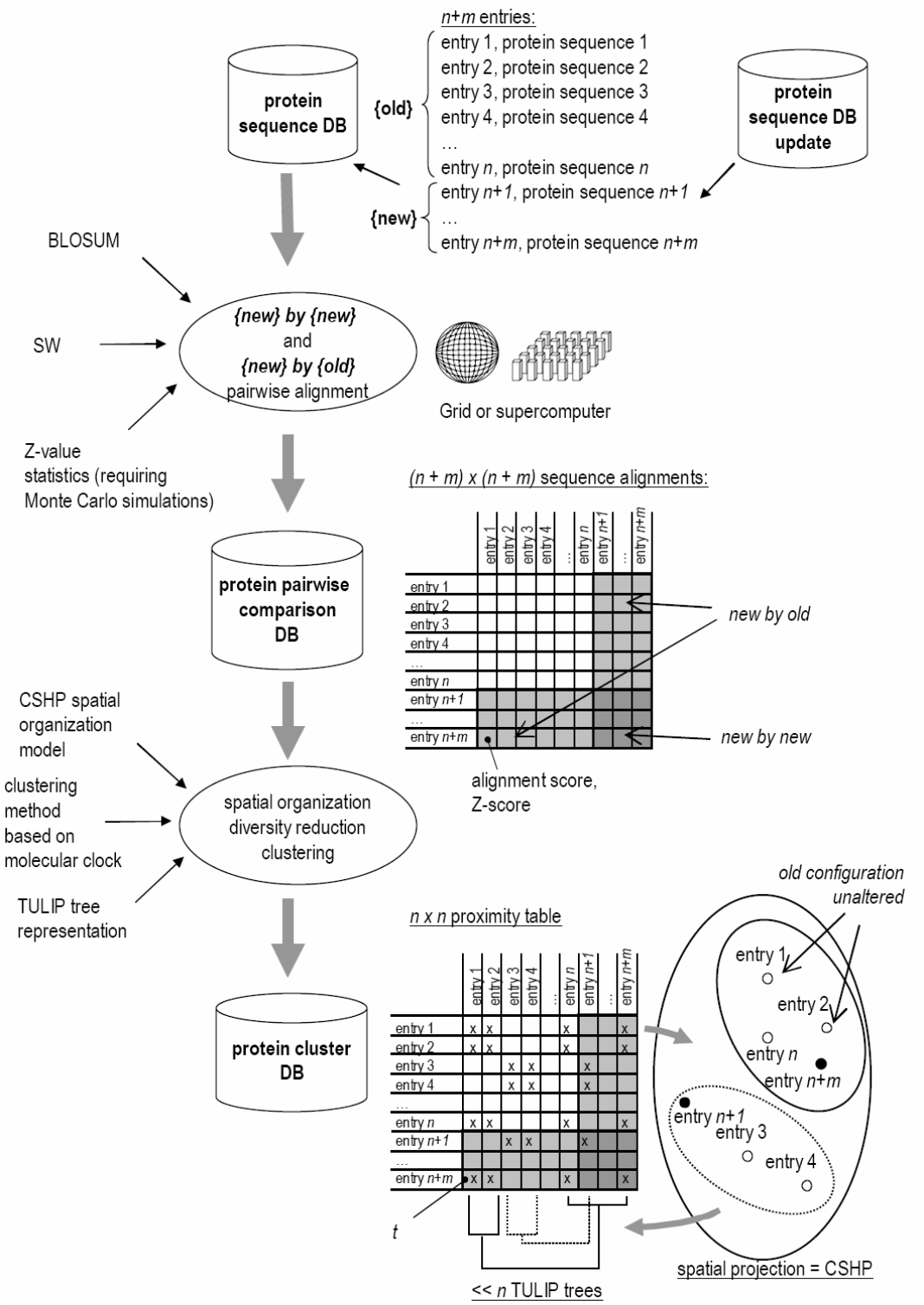


Fig. 3. Updating process for a massive all-by-all protein sequence comparison based on Z-value statistics and protein projection in the CSHP. According to the statistical model used to estimate the accuracy of alignment scores, the update of the initial protein sequence database {new} implies that a simple {new}-by-{new} and {new}-by-{old} computation is sufficient in order to update the clusters.

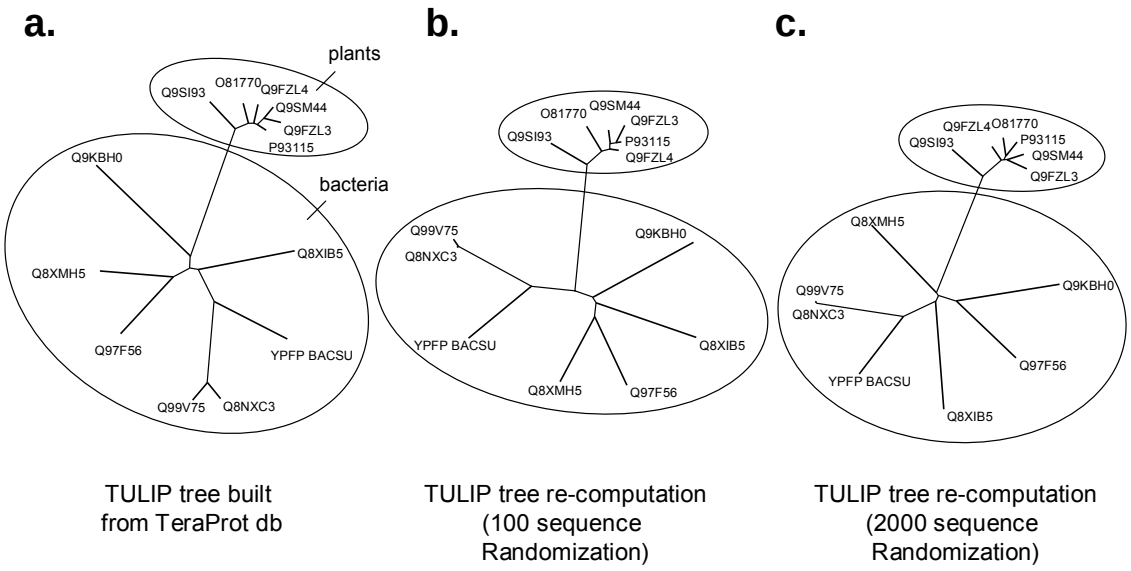


Fig. 4. Exploration of the protein space based on the TeraProt sequence massive comparison and protein projection in the CSHP. Given a UniProt protein entry (here, Q9SM44) we extracted sequences from the TeraProt output table that could be aligned with the query with a *Z-value* > 30. This set contains plant proteins (Q9FZL4, O81770, Q9S193, Q9FZL3, P93115) that have the same function as the query (MGD, *i.e.* catalyzing the transfer of a galactose from a UDP-galactose donor onto a lipidic diacylglycerol acceptor) and bacterial proteins (Q8XMH5, Q99V75, Q8NXC3, YPFP BACSU, Q8XIB5, Q97F56, Q9KBH0) having a distinct but related activity (MURG, *i.e.* catalyzing the transfer of an N-acetyl-glucosamine from a UDP-N-acetyl-glucosamine donor onto the Lipid 1 acceptor). A. TULIP tree computed from the Z-value table obtained from the TeraProt project (Monte Carlo simulation with 100 sequence shuffling; amino acid scoring matrix: PAM 10). This tree highlights the clear evolutionary divergence between plant and bacterial sequences. B. TULIP tree obtained after a re-computation of the *Z-value* table with 100 sequence shuffling and the BLOSUM 62 amino acid scoring matrix. C. TULIP tree obtained after a re-computation of the *Z-value* table with 2000 sequence shuffling and the BLOSUM 62 amino acid scoring matrix. The TULIP tree constructed from the TeraProt database was obtained after <15 sec using a bench workstation (HP ProLiant G4 Bi-Xeon - 2,8 Ghz) whereas trees reconstructed with 100 shuffling required 6 minutes computation and reconstruction with 2000 shuffling required less than 2 hours. This example illustrates the power of the method and the benefit one can get by exploring massive comparison results using the CSHP projection and TULIP phylogenetic underlying topology.

Résultats et discussion

Chapitre 3
(*Article 4*)

Résultats - Chapitre 3

La théorie de la fiabilité appliquée aux systèmes biologiques : déterminisme des propriétés remarquables de la *Z-value* dans un modèle de vieillissement et de longévité des séquences

Article 4

Olivier Bastien & Eric Maréchal
« System aging and longevity as a model of protein sequence evolution »
Soumis

Préambule

La *théorie de la fiabilité* est une théorie générale sur la défaillance des systèmes, de leur longévité et de leur vieillissement (pour revue: Gavrilov and Gavrilova, 2001). Dans cette théorie, un système est constitué de composants qui assurent des fonctions élémentaires, en interrelation au sein d'une architecture fonctionnelle. Les composants peuvent contenir un certain niveau de redondance, et se relayer si l'un d'eux venait à défaillir. Le niveau de redondance intrinsèque au système est donc un facteur essentiel de sa fiabilité au cours du temps. Lorsqu'un ensemble de composants redondants est défectueux au point de ne plus pouvoir assurer une fonction dans l'architecture d'ordre supérieure, le système peut devenir défaillant. Un composant d'un système peut être lui même un sous-système, constitué de sous-composants. Une entité biologique telle qu'une cellule, un organisme ou une population est évidemment un système de la sorte, caractérisé par une forte redondance.

La théorie de la défaillance se donne pour objet la distribution des temps de vie des systèmes, pour lesquels on suppose connue l'architecture ainsi que les distributions des temps de vie de ses composants. La fonction de fiabilité, R pour *reliability* (appelée aussi fonction de survie) est la probabilité qu'un système dépasse une certaine durée, c'est à dire $R(x) = P(X > x)$ où x est fonction du temps. On définit le taux de défaillance $h(x)$, appelé aussi fonction de risque, comme le taux relatif de la décroissance de la fonction de défaillance. C'est à dire :

$$h(x) = -\frac{dR(x)}{R(x)dx} = -\frac{d(\log R(x))}{dx}$$

Le taux de défaillance est équivalent à la force de mortalité en démographie (Valleron, 1998; Shkovskii, 2005). Par exemple, si $h(x)$ est constante, une simple intégration conduit à $R(x) = R(0)\exp(-\lambda x)$ qui est la distribution exponentielle, caractéristique des systèmes non

vieillissants. De façon intéressante, un système constitué de composants non vieillissants peut être lui même vieillissant (avec un taux de défaillance qui croît avec le temps).

Dans l'**Article 4**, nous avons posé l'hypothèse qu'un ensemble de *protéines homologues* était un système soumis à une pression évolutive visant à conserver un certain niveau commun de fonctionnalité. Au sein du monde vivant, la présence de gènes homologues dans des espèces distinctes (et la redondance qui en découle) est liée d'une part à des héritages verticaux après divergence à partir d'une espèce ancestrale commune, et d'autre part à des transferts horizontaux (ou latéraux) de fragments d'ADN entre certaines espèces. En considérant les protéines homologues au sein d'une même espèce (des allèles, des paralogues par exemple), il est possible de considérer un sous-système pour lequel l'invalidation d'un gène (devenant non opérationnel, ou bien divergeant vers une nouvelle fonctionnalité) peut être compensée par la présence d'un gène sain. Le gène sain qui relaye le gène invalidé peut être un allèle porté par un chromosome apparié (apporté par exemple au cours de la fécondation) ou encore un paralogue distant dans le même génome. Ce phénomène est connu en génétique dans le cas des mutations récessives et nécessite pour certaines analyses phénotypiques, d'invalider l'ensemble des gènes d'une même famille multigénique et d'étudier les mutants homozygotes, afin d'éviter tout problème de complémentation. Un individu est lui même un sous-système dans lequel les populations de protéines homologues assurent une redondance fonctionnelle. Enfin, chaque protéine peut être prise comme un sous-système dont les composants sont les acides aminés. Une redondance peut-être assurée à l'échelle d'une protéine par des amino acides voisins (si un amino acide hydrophobe venait à être substitué par un amino acide hydrophile, suite à une mutation génétique, au sein d'une hélice transmembranaire, il est possible que les amino acides voisins compensent cette perte par leur propre hydrophobicité). Il est intéressant de noter que dans cette architecture, les acides aminés sont aussi le support d'un niveau de redondance entre séquences distinctes, du fait que la pression de conservation fonctionnelle s'exerce au niveau des acides aminés. On peut mesurer cette redondance grâce à la conservation d'acides aminés (ou la substitution par des acides aminés équivalents) aux sites critiques pour la fonction.

Un alignement de deux séquences permet une mesure du niveau de redondance assuré par conservation ou substitution de certains résidus. Nous sommes partis du constat que l'information mutuelle entre deux résidus de deux séquences biologiques homologues décroissait avec le temps, et ceci dès l'instant où les séquences ont commencé à diverger (soit par un événement de duplication, soit par un événement de spéciation). En effet, la mutation d'un acide aminé en un autre conduit nécessairement à une diminution de cette information mutuelle (aucune substitution ne possède de score supérieur à l'identité dans une matrice de score). Nous avons supposé que l'évolution de l'information mutuelle était de ce fait orientée vis-à-vis de la flèche du temps. Partant de ce constat, et en suivant le cadre conceptuel de la théorie de la fiabilité, nous avons déduit que

- les amino acides sont des composants non vieillissants
- les séquences protéiques sont des systèmes vieillissants
- la distribution des scores d'alignement suit une loi de Gumbel

- la distribution des *Z-value* suit une loi de Gumbel.

Cet article permet donc de proposer un déterminisme pour la forme de la loi de distribution des *Z-values* suivant quelques hypothèses simples sur l'évolution des protéines, et d'autre part fournit une base unificatrice entre la biologie des systèmes et l'analyse des séquences.

TITLE:

System aging and longevity as a model of protein sequence evolution: derivation of remarkable properties of sequence alignment statistics

Authors:

Olivier Bastien (a,b) and Eric Maréchal (a,*)

Adresses:

(a) *UMR 5168 CNRS-CEA-INRA-Université J. Fourier, Laboratoire de Physiologie Cellulaire Végétale; Département Réponse et Dynamique Cellulaire; CEA Grenoble, 17 rue des Martyrs, F-38054, Grenoble cedex 09, France*

(b) *Gene-IT, 147 avenue Paul Doumer, F-92500 Rueil-Malmaison, France*

Corresponding author:

* Eric Maréchal

UMR 5168 CNRS-CEA-INRA-Université J. Fourier, Laboratoire de Physiologie Cellulaire Végétale, Département Réponse et Dynamique Cellulaire, CEA Grenoble, 17 rue des Martyrs, F-38054, Grenoble cedex 09, France

Tel : +33 (0)4 38 78 49 85

Fax : +33 (0)4 38 78 50 91

E-mail : emarechal@cea.fr

Keywords

Aging, reliability theory, Z-value, pairwise alignment, Smith-Waterman, hazard function

ABSTRACT

Confidence in pairwise alignments of biological sequences, obtained by various methods such as Blast or Smith-Waterman, is critical for automatic analyses of genomic data. In the asymptotic limit of long sequences, two statistical models have been proposed: 1) the Karlin-Altschul model that computes a *P-value* assuming that the number of high scoring matching regions above a threshold is Poisson distributed; the Lipman-Pearson model that computes a *Z-value* from a random score distribution obtained by a Monte-Carlo simulation. *Z-values* allow the deduction of an upper bound of the *P-value* ($1/Z\text{-value}^2$) following the TULIP theorem. Simulations of *Z-value* distribution is known to fit with a Gumbel law. This remarkable property was not demonstrated and had no obvious biological support. Here, we built a model of aging and longevity for homologous molecular sequences. We used principles of the "reliability theory" about systems aging, in which the failure rate (the rate by which systems deteriorate) is related to the systems longevity. The system can be a machine with structured components, or a living entity or population. "Reliability" refers to the ability to operate properly according to a standard. We considered a set of homologous sequences as a hierarchical system, which components are the sequences *per se*, having a high redundancy reflected by their alignment scores. Sequences were considered as systems of lower rank, which components are the amino acids that can independently be damaged by random DNA mutations. From these assumptions, we deduced that the amino acids were non-aging components (constant hazard rate) and that pairwise sequence alignment score should follow a Gumbel distribution, which parameters could find some theoretical rationale. In particular one parameter is the constant hazard rate of non-aging amino acids. A *P-value* can be derived from this remarkable property of *Z-value* distribution. This model also provides a link between concepts of biological sequence analysis and of systems biology.

MAIN TEXT

1. Introduction

Automatic analysis of biological sequences is crucial for the treatment of massive genomic outputs. Our understanding of more than 90 % of protein sequences stored in public databases, deduced from automatic translation of gene sequences, will not result from direct experimentation, but from our ability to predict informative features using *in silico* workflows (Ofran et al., 2005 ; Bastien et al., subm.). An underlying postulate is that the molecular sequences determined in biological individuals or species, which have evolved from a common ancestor sequence and are therefore homologous, have conserved enough of the original features to be similar. Popular sequence alignment methods, such as Blast (Altschul et al., 1990) or Smith-Waterman (Smith and Waterman, 1981) algorithms are used as a starting point for homology searches. Because re-examination of alignments obtained after massive comparisons is not manageable, confidence in alignment score probabilities is critical for automatic sequence comparisons, clustering of orthologs and paralogs, homology-based annotations or phylogeny reconstructions based on pairwise alignments (for review: Bastien et al., subm.). Assessing whether a computed alignment is evolutionarily relevant or whether it could have arisen simply by chance is therefore a question that has been extensively studied (for review: Ouzonis and Valencia, 2003). Two major methods have been proposed.

The first method proposed by Karlin and Altschul (1990) is an estimate of the probability of an observed local ungapped alignment score according to an Extreme Value Distribution, (or EVD; for review: Coles, 2001) in the asymptotic limit of long sequences. The Karlin-Altschul formula is the consequence of interpreting the number of high scoring matching regions above a threshold by a Poisson distribution. Briefly, considering *A* and *B* two random sequences, *m* and *n* their lengths, given the distribution of individual residues (*i.e.*

amino acids), and given a scoring matrix, the number of distinct local alignments with score values of at least *s* is approximately Poisson distributed with mean

$$E(s) \approx K.m.n.\exp(-\lambda.s) \qquad [1]$$

where λ and K can be calculated from the scoring matrix and average sequence compositions based on the Poisson distribution hypothesis. $E(s)$ is known as the *E-value*. As a consequence, if *s* is the score obtained after aligning two real sequences *a* and *b* (with *m* and *n* their respective lengths), the probability of finding an ungapped segment pair with a score lower than or equal to *s*, follows a Gumbel distribution (EVD type I as defined by Coles, 2001):

$$P(S(A,B) \leq s) \approx \exp(-K.m.n.\exp(-\lambda.s)) \qquad [2]$$

where $S(A,B)$ is the random variable corresponding to the score of two random sequences. The *P-value*, defined as the probability of finding an ungapped segment pair with a score higher than *s*, is simply given by $1\text{-}P(S(A,B) \leq s)$. Pragmatically, the *P-value* is approximated by the *E-value* when $E(s) < 0.01$. The validity of the Karlin-Altschul model depends on restrictive conditions: first, the residue distributions for the two sequences should not be "too dissimilar" and second, the sequence lengths (*m* an *n*) should "grow at roughly equal rates" (Karlin and Altschul, 1990). The length dependency of alignment scores has been discussed (Karlin and Altschul, 1990; Vingron and Waterman, 1994). In particular, it has been demonstrated that the growth of the best matching score of gapped alignments was linear when gap penalties were small, becoming logarithmic when increasing sequence length and for larger gap penalties (Vingron and Waterman, 1994). Although the Karlin-Altschul formula given by equation [2] is not valid for gapped alignments and although no asymptotic score distribution has been analytically established for local alignments allowing gaps, simulations (Pearson, 1998; Comet et al., 1999; Altschul et al., 2001; Webber and Barton, 2001) showed

that, for both local and global alignments, the Gumbel law was well-suited to the distribution of scores after pragmatic estimation of the λ and K parameters.

An alternative method, proposed by Lipman and Pearson (1985) and described extensively by Comet et al. (1999), Bacro and Comet (2001) and Bastien et al. (2004a, 2005a), uses Monte Carlo simulations to investigate the significance of a score s calculated from the alignment of two real sequences a and b. This method consists in computing C alignments of a with sequences obtained after shuffling b (Fitch, 1983). The variable corresponding to the shuffled sequence b is termed B. The C alignments allow an estimate of an empirical mean score ($\hat{\mu}$) and standard deviation ($\hat{\sigma}$) from the distribution of the random variable $S(a,B)$. A *Z-value* is then defined as:

$$Z(a,b^*) = \frac{s - \hat{\mu}}{\hat{\sigma}} \qquad [3]$$

where * indicates the sequence that was submitted to randomization.

In practice, the computation of $Z(a,b^*)$ is known to be convergent and depends on the accuracy of the estimation of μ and σ, and therefore on C, ranging usually from 100 to 1000 (Comet et al., 1999; Aude and Louis, 2002). Bacro and Comet (2001) showed that the asymptotic law of the Z-value was independent of the length and composition of sequences. Bastien et al. (2004a, 2005a) further demonstrated that whatever the distribution of the random variable $S(a,B)$ was, the relation

$$P(S(a,B)) \leq \frac{1}{z(a,b^*)^2} \qquad [4]$$

is true. This relation, known as the TULIP theorem, shows that the *Z-value* computed for pairwise sequence alignments 1) provides an upper bound of alignment score probability (Bastien et al., 2004a), 2) can be used to reconstruct molecular phylogenies (Bastien et al.,

2005a) and is an accurate clustering criterion to reduce the diversity of protein sequence databases (Petryszak et al., 2005). Here we term the *T-value* the upper bound deduced from the TULIP theorem, *i.e.* $1/Z(a,b^*)^2$.

Simulations of *Z-value* distribution (Pearson, 1998; Comet et al., 1999) show that it fits with a Gumbel distribution, which suggests that the distribution of alignment scores might follow a Gumbel distribution as well (Coles, 2001).

In this paper, we deduced biological rationale for the remarkable Gumbel-like distribution of sequence alignment scores and *Z-values*, based on a limited number of assumptions on sequences evolution. We considered that a set of homologous sequences was a population of entities that shared structural features (in particular some important conserved or functionally similar amino acids detected by alignment methods) that should not be modified, or deteriorated, beyond a certain point without loosing the original function. Over time, genetic mutants occurring in the population of homologous sequences are subjected to this conservative pressure. Some substitutions of amino acids by others having redundant properties may be permitted without functional break down, but some mutations may lead to a functional defect (or a shift to a novel function). In other words, a set of homologous proteins may be considered as a system, with a high level of structural redundancy, which components may age and die. We introduced principles of the *reliability theory of aging and longevity* (Gavrilov and Gravrilova, 2001), that apply to a wide range of systems, from artificial machines to biological population or organisms, applied here to molecular sequences. Based on the deduced model, we could provide biological basis for the *Z-value* Gumbel distribution, and significance for the corresponding Gumbel parameters (termed *K'* and λ').

2. Can the reliability theory of aging and longevity be applied to biological sequences?

The "reliability theory" is a general theory about systems aging, in which the failure rate (the rate by which systems deteriorate) is related to the systems longevity (Gavrilov and Gravrilova, 2001). The system can be a machine with structured components, or a living entity or population. "Reliability" of a system (or of one of its components) refers to its ability to operate properly according to a standard (Crowder et al., 1991). The relation between the age of a system and its failure rate shows that aging is a direct consequence of redundancies within the system. For instance, when applied to a biological system in which redundant vital structures ensure a function, damage of a component that is compensated by another redundant intact one, does not lead to a complete impairment of the system. Defects do accumulate, resulting in redundancy exhaustion and giving rise to the phenomenon of aging. As the system (or one of its components) degenerates into a system with no redundancy, new defects can eventually lead to death. Reliability of the system (or component) is described by the "reliability function" $R(x)$, also named "survival function", which is the probability that the system (or component) will carry out its mission through time x (Rigdon and Basu, 2000), expressed as the probability that the failure time X is beyond time x:

$$R(x) = P(X > x) = 1 - P(X \leq x) = 1 - F(x) \qquad [5]$$

where $F(x) = P(X \leq x)$ is a cumulative distribution function (Gavrilov and Gravrilova, 2001) reflecting the resistance of the system to failures (at time x, distribution of the probability that the system could have failed previously). $R(x)$ evaluates therefore the probability that the systems becomes completely defective after a time x (x can be a direct measure of time t or an increasing function of time).

The "hazard rate" $h(x)$, also called "failure rate", is defined as the relative rate for reliability function decline:

$$h(x) = -\frac{dR(x)}{R(x).dx} = -\frac{d(\log R(x))}{dx} \qquad [6]$$

Hazard rate is equivalent to mortality force in demography (Valleron, 1998; Shkovskii, 2005). When $h(x)$ is a constant h, the system does not deteriorate more often with age, and is therefore a *non-aging* system. A simple integration of equation [6] leads to

$$R(x) = R(0)\exp(-h.x) \qquad [7]$$

which is the exponential distribution that characterizes non-aging systems. Interestingly, a system with redundant *non-aging components* can be an *aging system*.

As discussed by Gavrilov and Gravrilova (2001), the "reliability theory" provided explanations for some fundamental problems regarding aging, longevity, death of organisms within populations. Organisms or populations are considered as systems in which categories of components (molecules, biological processes, cells, individuals, etc.) can be highly redundant, and be key elements for the system longevity.

Here, we propose to consider the particular case of *a set of homologous protein sequences as a system*, in which redundancy is ensured at various levels:

- at the residue level by the occurrence of functionally redundant amino acids (*e.g.* after a DNA damage that leads to a genetic mutation, an aspartic acid may be substituted by a functionally redundant glutamic acid),

- at the sequence level by the occurrence of functionally redundant amino acids, regions or domains (*e.g.* in a transmembrane helix, the spanning through a biological membrane is ensured by a series of hydrophobic amino acid; damage a some of these residues by mutations into polar amino acids may be compensated by remaining hydrophobic residues in the vicinity of the mutated residues),

- at the level of the set of sequences within an individual by the occurrence of functionally redundant homologues inside the same genomic background (a knocked-out mutant can be complemented by an intact allele or gene duplicate),

- at the level of the homologous sequences in species populations by the occurrence of functionally redundant homologues that may be transmitted vertically (from parents to progenies) following mitosis or meiosis, or transmitted horizontally (or laterally) by transfers of DNA segments.

The structural redundancy within the system, which is a key criterion for mortality, is conversely a key criterion for the resistance of the system to structural modifications and therefore its longevity. A sequence would "die" (*i.e.* would no more ensure the initial function) if a mutation would substitute an amino acid by a non redundant residue, at a position where no other amino acids of the sequence would compensate the defective mutation introduced. Such biological system belongs to the category of repairable systems as defined generally by Rigdon and Basu (2000). In a given species, if this event occurred, previous duplications of the gene, or presence of a healthy allele, may ensure a redundancy of the subset of homologous sequences and ensure the organism survival. If all redundant genes failed in a given organism, due to mutations, then the organism would reach another level of redundancy exhaustion, and maybe die. Considering DNA lateral transfers and chromosome exchanges, intact homologous sequences from other organisms may be introduced and compensate the defect. A population (or sub-population) of homologous sequences would theoretically extinct, if all homologous sequences became defective. In other words, the redundancy within a system of homologous proteins, basically ensured by the conservation of identical or functionally equivalent amino acids, is related to the "longevity" of an effective biological function through evolution.

To measure the rate of conservation of a shared structure/function relationship at time x within a system of homologous proteins (*i.e.* the time of observation) we defined a *homology longevity rate* Ψ from the cumulative distribution function $F(x) = P(X \leq x)$ (at time x, probability that the system could have previously failed), supposed continuously differentiable, as:

$$\psi(x)=\lim_{dx\to 0}\frac{P(x-dx<X\leq x/X\leq x)}{dx} \qquad [8]$$

The homology longevity rate has the following properties.

Theorem. Given $f(x) = F(x)dx$ the density function of x, we have the equality

$$\psi(x) = \frac{f(x)}{F(x)} = \frac{f(x)}{P(X \leq x)} \qquad [9]$$

Proof: Using the Bayes theorem, we have

$$\psi(x) = \lim_{dx\to 0}\frac{P(x-dx < X \leq x / X \leq x)}{dx} = \lim_{dx\to 0}\frac{P(x-dx < X \leq x \cap X \leq x)}{P(X \leq x).dx} \qquad [10]$$

Since $\{x-dx<X\leq x\}\subset\{X\leq x\}$, then

$$\psi(x) = \lim_{dx\to 0}\frac{P(x-dx < X \leq x)}{P(X \leq x).dx} = \lim_{dx\to 0}\frac{f(x)dx}{P(X \leq x).dx} = \frac{f(x)}{P(X \leq x)} \qquad [11]$$

***Corollary*:** From this theorem, we can state the following three equalities:

$$\psi(x)=\frac{d(\log F(x))}{dx} \qquad [12]$$

$$P(X \leq x) = \exp(-\int_{x}^{+\infty}\psi(u)du) \qquad [13]$$

$$f(x) = \psi(x)\exp(-\int_{x}^{+\infty}\psi(u)du) \qquad [14]$$

3. Mutual information of aligned sequences reflects system redundancy at the amino acid and sequence levels and allows a derivation of the distribution of alignment scores based on reliability theory.

Dobzhansky (1974) and Wu et al. (1974) established that *information* harbored by a protein 1) emerged from the three-dimensional self organization of its residues (*i.e.* the sequence of amino acids) and had to do with information harbored by amino acids, and 2) was submitted through time to evolutionary pressure (achievement of a minimal functional level fitting environmental and species survival conditions). Using previous empirical results (Dayhoff et al., 1978; Henikoff and Henikoff, 1992; Risler et al., 1988), Bastien et al. (2005a) have shown that the alignment score of two homologous sequences *a* and *b* was proportional to the estimate of the information they share due to their common origin and parallel evolution under similar conservative pressure, *i.e.* the *mutual information* $I(a;b)$ in the sense of Hartley (Hartley, 1928; Shannon, 1948):

$$s(a,b) = \Gamma . I(a;b) \qquad [15]$$

with Γ a constant. The mutual information being additive, $I(a;b)$ is the sum of the mutual information of aligned residues, *reflecting the magnitude of the redundancy between the sequences at the amino acid level*. Mutual information between residues is therefore simply deduced from the 20x20 amino acid substitution matrix (*e.g.* Dayhoff et al., 1978; Henikoff and Henikoff, 1992; Risler et al., 1988) used to compute the alignment.

Within a given sequence, mutual information was also shown to *reflect the dependency of close or remote amino acids*, a phenomenon known as the residue co-evolution, due to their co-contribution to the sequence function (Aynechi and Kuntz, 2005a, 2005b).

Considering a *protein* as a *system* of lower rank, which *components* are *amino acids*, we examined how these components should age, and how amino acid aging affected protein sequence aging inside the system of homologous sequences. We simply hypothesized that an amino acid may deteriorate over time following random DNA mutations. As stated earlier, reliability of the component refers to its ability to operate properly according to a standard (Crowder et al., 1991), which can be measured here by the mutual information between the original residue and the new one, *i.e.* the corresponding substitution score in a 20x20 substitution matrix. If the score is lower than a threshold characterizing a functional conservation, the component is considered as damaged.

Over time, an amino acid i is either conserved or substituted by another one. The similarity of i compared with its descent is therefore either that of identity (the diagonal term in the scoring matrix) or a lower value (no score is higher than that of identity). The magnitude of the similarity of i compared with its descent is therefore a decreasing function of elapsed time. The probability that i was mutated into a defective residue, *i.e.* with a score S_i lower than a threshold s_i defined to allow the component to operate like i, can be deduced from the distribution of substitution scores. For most amino acids (F, P, W, Y, V, E, G, H, I, L, K, R, N, D and C), the distribution of scores deduced from BLOSUM 62 fits an exponential distribution (see the case of valine in Figure 1A) in a way one would expect for a *non-aging* component. For five amino acids (M, S, T, A and Q), the distribution of scores does not fit an exponential distribution (see the case of Threonine in Figure 1B). The distribution of scores deduced from the BLOSUM 62 matrix is exponential (Figure 1C) supporting a general model of amino acids as *non-aging* components. Thus, taking the average situation, *i.e.* an exponential score distribution, the probability P_i that a residue i is mutated into a residue with insufficient functional redundancy, *i.e.* below s_i is:

$$P_i(S_i \leq s_i) = 1 - \exp(-\lambda_i . s_i) \qquad [16]$$

where λ_i is the *constant hazard rate, or failure rate, for reliability function decline of the amino acid i*.

Given a sequence *a*, what is the cumulative probability that any of its *m* residues (termed *i*) had previously mutated into the *n* residues (termed *j*) of a sequence *b*, so that sequence *b* could no more operate like *a* (with exhaustion of functional redundancy)? We can consider $m \neq n$ due to insertion or deletion events. If *m* is high, we can state that this probability corresponds to the failure of the alignment, *i.e.* with a score *S* lower than a threshold *s*, with the following approximations: $S \approx m\langle S_i \rangle$, $\langle S_i \rangle = \lim_{m \to +\infty} S_i$ and $s \approx m\langle s_i \rangle$, $\langle s_i \rangle = \lim_{m \to +\infty} s_i$ (Waterman, 1995). Asymptotically, hazard rate for reliability function decline of amino acids (non-aging components, see Figure 1) is considered as a constant λ'.

Considering that residues are independent, the cumulative probability that the *n* components of sequence *b* are defective, *i.e.* that they do not operate like the *m* components of sequence *a* is given by:

$$P(S \leq s) = \left(P_i(\langle S_i \rangle \leq \langle s_i \rangle)\right)^{K'.m.n} \qquad [17]$$

with *K'*<1 a correcting factor, for pairwise alignments likelihood and edge effects. Considering the approximation of $\langle S_i \rangle$ and $\langle s_i \rangle$ respectively by *S*/*m* and *s*/*m*

$$P(S \leq s) = \left(P_i(S \leq s)\right)^{K'.m.n} \qquad [18]$$

The density function $f(s)$ is therefore given by:

$$f(s) = \frac{\partial P}{\partial s} = K'.m.n.f_i(s).\left(P_i(S \leq s)\right)^{K'.m.n-1} \qquad [19]$$

We can deduce the *homology longevity rate* Ψ, defined earlier as a function of the pairwise alignment score:

$$\psi(s) = \frac{f(s)}{P(S \le s)} = \frac{K'.m.n.f_i(s).\left(P_i(S \le s)\right)^{K'.m.n-1}}{\left(P_i(S \le s)\right)^{K'.m.n}} \quad [20]$$

Equation [16] implies that:

$$\psi(s) = \frac{K'.m.n.\lambda'.\exp(-\lambda'.s).\left(1-\exp(-\lambda'.s)\right)^{K'.m.n-1}}{\left(1-\exp(-\lambda'.s)\right)^{K'.m.n}} = \frac{K'.m.n.\lambda'.\exp(-\lambda'.s)}{1-\exp(-\lambda'.s)} \quad [21]$$

Asymptotically, the homology longevity rate is therefore given by

$$\psi(s) = K'.m.n.\lambda'.\exp(-\lambda'.s) \quad [22]$$

Using equation [13], we deduce that the distribution of alignment scores should respect the following formula

$$P(S \le s) \approx \exp(-K'.m.n.\exp(-\lambda'.s)) \quad [23]$$

From this expression, one can simply deduce the *P-value* as 1- $P(S \le s)$.

4. Discussion and Conclusion

We built a model of aging and longevity for homologous molecular sequences, in which a set of homologous sequences is a hierarchical system, which components are the sequences *per se*, having a high functional redundancy, reflected by the sequence alignment scores. Sequences were also considered as systems of lower rank, which components are the amino acids that can independently be damaged by random DNA mutations. Residues harbor a functional redundancy reflected by the amino acid substitution scores.

From these assumptions, we deduced that the pairwise sequence alignment score should follow a Gumbel distribution (equation [23]). The λ' parameter is the constant failure rate for the reliability function decline of amino acids, taken as *non-aging* components. The λ' parameter depends 1) on the distribution of the amino acids and 2) on the distribution of amino acid similarities deduced from a substitution matrix. The *K*' parameter has a more

complex meaning, because it depends on likelihood of an alignment of two sequences, with edge effects, gaps, length difference and repartition of the information (the local score) in the alignment.

The Gumbel parameters for score alignments can be estimated by two kinds of simulations. First is by adjusting EVD to the simulated distribution of scores (Coles, 2001). In that case, it is simpler to express the Gumbel law as

$$P(S \leq s) \approx \exp(-\exp(-\frac{s-\theta}{\beta})) \qquad [24]$$

with $\beta = \frac{1}{\lambda'}$ and $\theta = \frac{1}{\lambda'}\log(\lambda'.K'.m.n)$. The estimate of Gumbel parameters is achieved by determining β and θ, allowing an easy estimate of the λ' and K' parameters of equation [23].

Second estimation of the Gumbel parameters is by computing the *Z-value* corresponding to the simulation of score distribution. Using the fact that for a Gumbel distribution, $\mu = \theta + \gamma\beta$ and $\sigma^2 = \frac{\pi^2}{6}\beta^2$, then the *Z-value* allows a computation of the β and θ constants.

Simulations of *Z-value* distribution (Pearson, 1998; Comet et al., 1999) showed that it fitted with a Gumbel law. Based on the Gumbel distribution of scores (equations [23] and [24]), then the distribution of *Z-values* should respect the following equality:

$$P(Z \leq z) = \exp(-.\exp(-z\frac{\pi}{\sqrt{6}} - \gamma)) \qquad [25]$$

with γ the Euler-Mascheroni constant ($\gamma \approx 0.5772$). Equation [25] is the precise expression of the distribution of *Z-values* deduced by Pearson (1998) from simulations. It is important to note that this expression of the *Z-value* distribution is independent of sequence lengths and amino acid distributions.

This consideration has practical implications, since it allows a refined estimate of the *P-value* based on *Z-value* computation, and a real gain over available methods, particularly in some documented cases where the Karlin-Alstchul formula failed to assess the significance of an alignment. Table 1 shows for instance the different statistical estimates for the alignment of two homologous TFIIA gamma sequences from *Plasmodium falciparum* and *Arabidopsis thaliana*. The compositional bias in the proteome of *Plasmodium falciparum*, the malarial parasite, is known to limit the use of Karlin-Altschul statistics for pairwise comparisons with unbiased proteins such as those of *Arabidopsis thaliana* (Bastien et al., 2004b). The TFIIA gamma subunit sequence of *Plasmodium* could not be deduced from BLASTP-based homology searches (Callebaut et al., 2005). The Blastp apparent search failure was due to the overestimate of the *P-value* following the Karlin-Altschul formula (0.008, using unfiltered BLASTP, see Table 1). Alignment score *Z-value*, computed with either Blastp (P. Ortet, unpublished algorithm) or Smith-Waterman was above 10. The upper bound for the *P-value* based on the TULIP theorem, given by the formula $T\text{-value} = 1/Z\text{-value}^2$ (Bastien et al., 2004a), was therefore below 10^{-2}. Eventually, the *P-value* deduced from the *Z-value* Gumbel distribution was below 10^{-6} (see Table 1) indicating that, for both the Blastp and Smith-Waterman methods, the homology could be statistically assessed, even in the limit case of unbiased vs biased sequence comparisons. We noticed that the asymmetric DirAtPf100 matrix specified for *Plasmodium* vs. *Arabidopsis* comparisons that we developed earlier (Bastien et al., 2005b) allowed an additional gain in estimating this missed homology.

Eventually, considering that the current molecular biodiversity finds its origin in a common ancestral pool of living entities, then all proteins are related in some way. We can introduce an additional hierarchical level in the system proposed here, embracing all known sequences, in which the system architecture is determined by families of homologous sequences of various granularities. Distribution of the alignment scores of a sequence against

all the others (as stored in a sufficiently big and diverse database) should therefore exhibit a Gumbel-like distribution. In support to this prediction, Comet et al. (1999) reported that the comparison of a sequence with a large database led to a Gumbel distribution of *Z-values*.

Besides a theoretical support for pragmatic observations, this report shows therefore that the alignment score Gumbel distribution is a particular and general evolutionary law for molecular sequences taken as aging systems. This model provides additionally a link between concepts of biological sequence analysis and the emerging field of systems biology.

Acknowledgements.

Authors whish to thank Philippe Ortet for computing facilities, particularly for the use of an unpublished implementation of the *Z-value* statistical model using Blastp.

References.

Altschul, S. F., Bundschuh, R., Olsen, R. & Hwa, T. (2001). The estimation of statistical parameters for local alignment score distributions. Nucleic Acids Res 29, 351-61.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. J Mol Biol 215, 403-10.

Aynechi, T. & Kuntz, I.D. (2005a) An information theoretic approach to macromolecular modeling: I. Sequence alignments. Biophys J. 89:2998-3007.

Aynechi, T. & Kuntz, I.D. (2005b) An information theoretic approach to macromolecular modeling: II. Force fields. Biophys J. 289:3008-3016.

Aude, J. C. & Louis, A. (2002). An incremental algorithm for Z-value computations. Comput Chem 26, 403-11.

Bacro, J. N. & Comet, J. P. (2001). Sequence alignment: an approximation law for the Z-value with applications to databank scanning. Comput. Chem .25, 401-10.

Bastien, O., Aude, J. C., Roy, S. & Marechal, E. (2004a). Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of Z-value statistics. Bioinformatics 20, 534-7.

Bastien, O., Lespinats, S., Roy, S., Metayer, K., Fertil, B., Codani, J. J. & Marechal, E. (2004b). Analysis of the compositional biases in Plasmodium falciparum genome and proteome using Arabidopsis thaliana as a reference. Gene 336, 163-73.

Bastien, O., Ortet, P., Roy, S. & Marechal, E. (2005a). A configuration space of homologous proteins conserving mutual information and allowing a phylogeny inference based on pair-wise Z-score probabilities. BMC Bioinformatics 6, 49.

Bastien, O., Roy, S. & Marechal, E. (2005b). Construction of non-symmetric substitution matrices derived from proteomes with biased amino acid distributions. C R Biol 328, 445-53.

Botte, C., Jeanneau, C., Snajdrova, L., Bastien, O., Imberty, A., Breton, C. & Marechal, E. (2005). Molecular modeling and site-directed mutagenesis of plant chloroplast monogalactosyldiacylglycerol synthase reveal critical residues for activity. J Biol Chem 280, 34691-701.

Callebaut, I., Prat, K., Meurice, E., Mornon, J.P. & Tomavo, S. (2005). Prediction of the general transcription factors associated with RNA polymerase II in Plasmodium falciparum: conserved features and differences relative to other eukaryotes. BMC Genomics. 23;6:100.

Coles, S. (2001). An introduction to Statistical Modeling of Extreme Values, Springer-Verlag, New-York.

Comet, J. P., Aude, J. C., Glemet, E., Risler, J. L., Henaut, A., Slonimski, P. P. & Codani, J. J. (1999). Significance of Z-value statistics of Smith-Waterman scores for protein alignments. Comput Chem 23, 317-31.

Crowder, M.J., Kimber, A.C., Smith, R.L. & Sweeting, T.J. (1991) Statistical analysis of reliability data. London, Chapman and Hall

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A Model of Evolutionary Change in Proteins. Atlas of Protein Sequence and Structure 5, 345-352.

Dobzhansky, T. (1974). Studies in the Philosophy of Biology: Reduction and Related Problems. Francisco J. Ayala and Theodosius Dobzhansky edit, University of California Press.

Fitch, W. M. (1983). Random sequences. J Mol Biol 163, 171-6.

Gavrilov, L. A. & Gavrilova, N. S. (2001). The reliability theory of aging and longevity. J Theor Biol 213, 527-45.

Hartley, R.V.L. (1928). transmission of Information. The Bell System Technical Journal 3, 535-564.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89, 10915-9.

Karlin, S. & Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci U S A 87, 2264-8.

Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. Science 227, 1435-41.

Ofran, Y., Punta, M., Schneider, R. & Rost, B. (2005) Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. Drug Discov Today. 10, 1475-1482.

Ouzonis, C.A. & Valencia, A. (2003) Early bioinformatics: the birth of a discipline – a personal view. Bioinformatics 19, 2176-2190

Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. J Mol Biol 276, 71-84.

Petryszak, R., Kretschmann, E., Wieser, D. & Apweiler, R. (2005). The predictive power of the CluSTr database. Bioinformatics 21, 3604-9.

Rigdon, S.E. & Basu, A.P. (2000) Statistical methods for the reliability of repairable systems. New York, Wiley and Son, Inc.

Risler, J. L., Delorme, M. O., Delacroix, H. & Henaut, A. (1988). Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. J Mol Biol 204, 1019-29.

Shannon, C.E. (1948). A Mathematical Theory of Communication. The Bell System Technical Journal 27, 379-423,623-656.

Shkovskii, B. I. (2005). A simple derivation of the Gompertz law for human mortality. Theory in Biosciences 123, 431-433.

Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. J Mol Biol 147, 195-7.

Tekaia, F., Lazcano, A. & Dujon, B. (1999). The genomic tree as revealed from whole proteome comparisons. Genome Res 9, 550-7.

Valleron, A. J. (1998). Introduction à la Biostatistique, Masson, Paris.

Vingron, M. & Waterman, M. S. (1994). Sequence alignment and penalty choice. Review of concepts, case studies and implications. J Mol Biol 235, 1-12.

Waterman, M. S. (1995). Introduction to computational biology (Hall, C., Ed.), CRC Press.

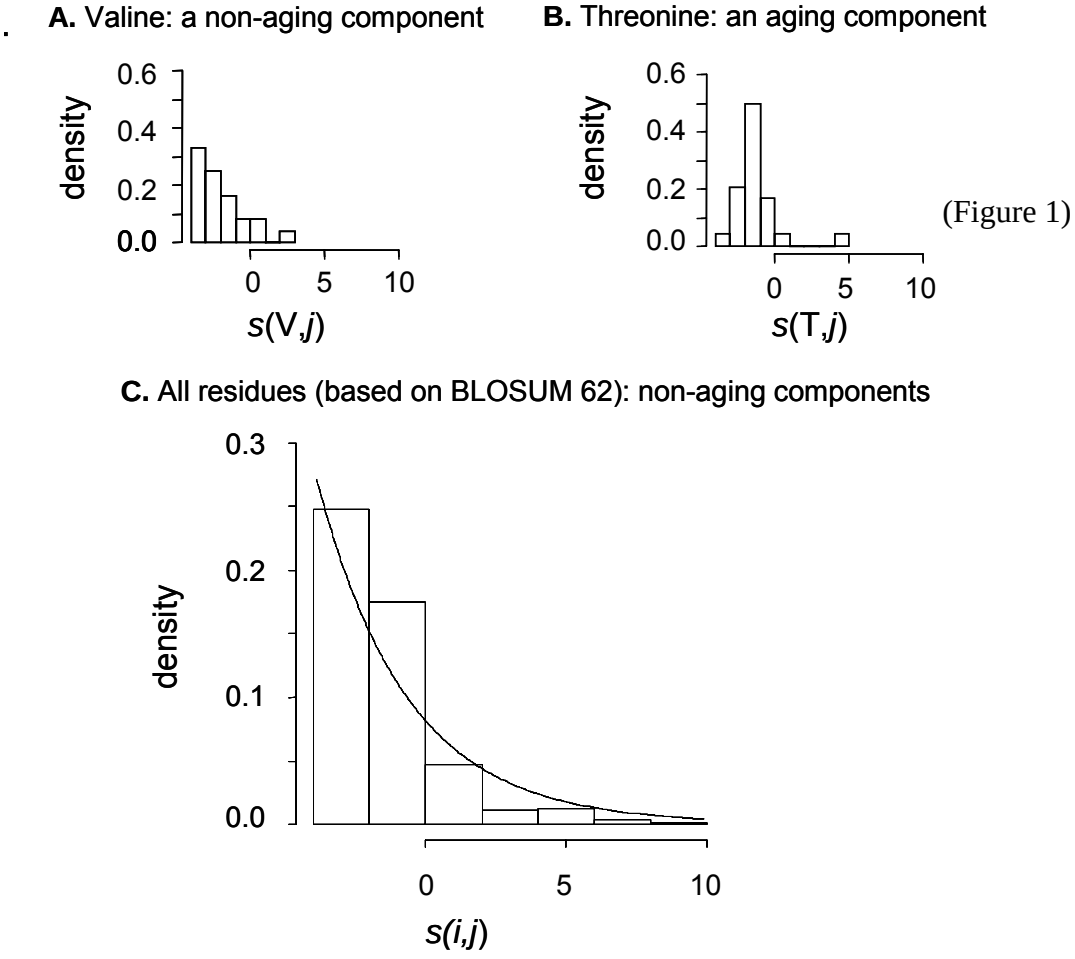
Webber, C. & Barton, G. J. (2001). Estimation of P-values for global alignments of protein sequences. Bioinformatics 17, 1158-67

Wu, T.T., Fitch, W. M. & Margoliash, E. (1974) The information content of protein amino acid sequences. Annu Rev Biochem 43, 539-566

Table 1: Alignment statistics of the homologous Transcription initiation factor IIA (TFIIA) gamma chain sequences from *Plasmodium falciparum* and *Arabidopsis thaliana*. TFIIA gamma sequences from *Plasmodium* (UniProtKB Q8I4S7_PLAF7) and *Arabidopsis* (UniProtKB T2AG_ARATH) were aligned with Blastp and Smith-Waterman methods. Statistics were computed following the Karlin-Altschul model (as implemented in the Blastp algorithm) or the Lipman-Pearson *Z-value* model. The upper bound for the *P-value* based on the TULIP theorem is given following the formula: $T\text{-value} = 1/Z\text{-value}^2$. The *P-value* deduced from the *Z-value* Gumbel distribution was computed following the model presented here. Substitution matrices were either BLOSUM62, or the asymmetric DirAtPf100 matrix specified for *Plasmodium vs. Arabidopsis* comparisons. NA: not applicable.

Alignment method	Blastp	Smith-Waterman	
Substitution matrix	BLOSUM62	BLOSUM62	DirAtPf100
Statistics			
P-value (Karlin-Altchul)	0.008	NA	NA
Z-value (Pearson-Lipman)	10	11	12
T-value (TULIP theorem)	0.01	8.10^{-3}	7.10^{-3}
P-value (this work)	$1.5.10^{-6}$	$3.7.10^{-7}$	1.10^{-7}

Figure 1: Aging properties of amino acids. Protein sequences are considered as systems, which components are amino acids. Over time, either amino acids are conserved (similarity of a residue with its descent is that of identity, diagonal term of a substitution matrix) or modified due to random DNA mutations. Similarity decreases therefore with time, since no similarity is higher than that of identity. When the similarity falls below a threshold that is necessary for the residue to operate according to a standard (functional conservation), the component is damaged. **(A) Score distribution corresponding to valine substitution.** In this case, the score distribution is exponential, suggesting that valine (V) is a non-aging component. Based on BLOSUM62, residues of this type are V, F, P, W, Y, E, G, H, I, L, K, R, N, D and C **(B) Score distribution corresponding to threonine substitution.** The score distribution shows a peak, indicating a probable accelerated process of aging (functional damage) when the residue is substituted by random mutation in some other amino acids. Based on BLOSUM62, residues of this type are T, S, M, A and Q. **(C) Score distribution in the BLOSUM62 similarity matrix.** The complete distribution in the BLOSUM62 matrix is exponential (0.287.exp(-0.287.(*s*+4))), supporting a general model of amino acids as non-aging components. The exponential law for positive scores is characterized by the same parameter (λ'=0.287). The original residue is termed *i*; its descent is termed *j*.



(Figure 1)

Résultats et discussion

Chapitre 4
(*Articles 5 & 6*)

Résultats - Chapitre 4

Divergence compositionnelles des génomes et protéomes de *Plasmodium falciparum* et d'*Arabidopsis thaliana* : étude du déterminisme du biais compositionnel malarial et proposition de correction des matrices de substitution mesurant la similarité des amino acides des deux protéomes

Article 5

Olivier Bastien, Sylvain Lespinats, Sylvaine Roy, Karine Métayer, Bernard Fertil, Jean-Jacques Codani & Eric Maréchal (2004)
« Analysis of the compositional biases in Plasmodium falciparum genome and proteome using Arabidopsis thaliana as a reference »
Gene 336:163-173

Article 6

Olivier Bastien, Sylvaine Roy & Eric Maréchal (2005)
« Construction of non-symmetric substitution matrices derived from proteomes with biased amino acid distributions »
C. R. Biol. 328:445-53.

Préambule

La composition extrême en A+T (%A+T > 80 %) des gènes de *Plasmodium falciparum* a été soulignée pour la première fois par Musto et al. (1995) et confirmée à l'échelle du génome entier par le séquençage du parasite (Gardner et al., 2002) (voir le chapitre Introduction). Le biais en %A+T est variable selon les positions de codons et la classe des protéines traduites (Musto et al., 1995). Sur divers modèles procaryotes et eucaryotes, dont quelques données génomiques partielles de *Plasmodium falciparum*, Singer et Hickey (2000) ont mis en évidence que la composition en nucléotides affectait la composition en acides aminés des protéines encodées tout le long du génome. Ces auteurs ont montré que les patrons de biais (*i.e.* les ensembles d'acides aminés sur- ou sous-représentés) étaient semblables chez les Archea et les Eubacteria et que le biais nucléotidique au sein de familles de séquences homologues augmentait avec la divergence des séquences. Dans l'**Article 5**, nous avons repris les conclusions de Singer and Hickey (2000) et examiné s'il existait aussi chez *Plasmodium falciparum* une corrélation entre les biais des séquences de gènes codant pour des protéines et leur divergence évolutive estimée en prenant comme références « stables » les protéines

homologues d'*Arabidopsis thaliana*. Pour cette étude, nous avons sélectionné 71 couples de protéines homologues, à la fois retenues pour leur haut niveau de similitude structurale et pour la concordance de leurs annotations. Nous avons montré que la composition en nucléotides de *Plasmodium falciparum* affectait les patrons de compositions des protéines, suivant la flèche du temps. Par ailleurs, en collaboration avec Sylvain Lespinats et Bernard Fertil (Unité Inserm 494, Hôpital Pitié-Salpêtrière, Paris), nous avons caractérisé finement le biais nucléotidique de *Plasmodium falciparum* sous forme de signatures trinucléotidiques (à l'aide de représentation de type *Chaos Game*, voir Lespinats, 2006) à l'échelle du génome entier et au niveau de chaque gène comparé. Lorsqu'un gène possède une signature proche de la signature génomique moyenne, l'homologie est plus difficile à détecter par les méthodes d'alignement de séquences paramétrées par défaut. L'**Article 5** été cité par Chanda et al. (2005), pour une analyse d'un autre aspect du déterminisme du biais compositionnel malarial. Ces auteurs ont en effet montré un usage plus important de codons non synonymes riches en G+C, dans les séquences les plus exprimées de *Plasmodium falciparum*, montrant ainsi qu'outre la divergence évolutive, le niveau d'expression des gènes est un facteur déterminant du biais à l'échelle nucléotidique. Par ailleurs, l'**Article 5** a été mentionné par Callebaut et al. (2005), rapportant la recherche de facteurs de transcriptions généraux, associés à l'ARN polymérase II, non détectés lors de l'annotation automatique initiale du génome malarial, et retrouvés par ces auteurs en recherchant dans le protéome non annoté du parasite, les séquences ayant conservé certains patrons HCA (hydrophobic cluster analysis).

Notre étude a mis en évidence que dans des séquences d'*Arabidopsis thaliana* et de *Plasmodium falciparum* ayant une similarité très forte, aisément détectable par les méthodes conventionnelles d'alignement, pour une paire de résidus alignés donnée, l'information (information mutuelle, voir le Chapitre 2 des Résultats) apportée par la présence d'un résidu comme la lysine, très abondant dans la séquence de *Plasmodium falciparum*, est plus pauvre que celle apportée par une lysine dans la séquence d'*Arabidopsis thaliana*. On peut tenir compte de cette différence en revenant à la définition de la similarité établie par Henikoff et Henikoff (1992) (voir Rappels Bibliographiques) et en considérant que les matrices de similarités sont un moyen de mesurer l'information mutuelle moyenne entre deux acides aminés.

Nous avons par conséquent examiné la possibilité de corriger les différences d'information apportées par les acides aminés dans les deux protéomes de *Plasmodium falciparum* et d'*Arabidopsis thaliana* au niveau des matrices de substitution utilisées pour aligner des séquences de ces deux organismes. A partir d'un ensemble de séquences homologues données (comme celles étudiées dans l'**Article 5**) il est possible de déduire les termes de matrices de substitution soit par extrapolation d'un modèle de Markov (Dayhoff et al., 1978), soit par estimation directe sur les données (les alignements) (Henikoff and Henikoff, 1992). L'approche par chaîne de Markov à été utilisée pour construire des matrices asymétriques visant à corriger les mesures d'alignements de séquences trans-membranaires biaisées par l'abondance de résidus hydrophobes (Muller et al., 2001). Nous avons adopté une approche alternative, et développé un processus itératif de calcul de matrice de substitution asymétrique, présenté dans l'**Article 6**. Nous avons pris comme ancrage initial les alignements des séquences générés à l'aide de matrices usuelles ou de la matrice identité (le résultat final ne dépend pas de la matrice initiale), puis en utilisant les alignements obtenus, nous avons mesuré les termes d'une nouvelle matrice tenant compte des différences de compositions aux positions alignées. Ce calcul a été itéré jusqu'à convergence des valeurs de la matrice. Une famille de matrices dirigées, appelées

DirAtPf (pour dirigée *Arabidopsis thaliana* → *Plasmodium falciparum*) a ainsi été générée, indiciée par le pourcentage de clustering des blocks ayant servi au calcul (voir Rappels Bibliographiques). Se posant la même question de l'alignement des séquences de *Plasmodium falciparum* et d'*Arabidopsis thaliana*, Yu et Altschul (2005) se sont engagés dans une étude semblable. Ces auteurs ont proposé de dériver numériquement les valeurs des matrices de substitution usuelles à partir de la composition des séquences comparées. Les matrices asymétriques que nous avons rapportées dans l'**Article 6** ont été mentionnées par Bulka et al. (in press), parmi les matrices à façon qui sont actuellement développées par différents groupes. Que ce soient les matrices DirAtPf (**Article 6**) ou les matrices dédiées à la comparaison de séquences de *Plasmodium falciparum* et d'*Arabidopsis thaliana* proposées par Yu et al. (2005), aucune n'a encore été exploitée à grande échelle pour en évaluer la performance. Dans l'**Article 6**, nous montrons que l'entropie relative de ces matrices augmente par rapport aux matrices usuelles, indicateur théorique du gain de sensibilité.

Available online at www.sciencedirect.com

SCIENCE DIRECT

Gene 336 (2004) 163–173

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Analysis of the compositional biases in *Plasmodium falciparum* genome and proteome using *Arabidopsis thaliana* as a reference

Olivier Bastien[a,b], Sylvain Lespinats[c], Sylvaine Roy[d], Karine Métayer[b], Bernard Fertil[c], Jean-Jacques Codani[b], Eric Maréchal[a,*]

[a] *Laboratoire de Physiologie Cellulaire Végétale, Département Réponse et Dynamique Cellulaire, UMR 5168 CNRS-CEA-INRA-Université Joseph Fourier, CEA Grenoble, 17 rue des Martyrs, F-38054, Grenoble cedex 09, France*

[b] *Gene-IT, 147 Avenue Paul Doumer, F-92500, Rueil-Malmaison, France*

[c] *Centre Hospitalier Universitaire Pitié-Salpêtrière, INSERM U 494, 91 Boulevard de l'Hôpital, F-75634, Paris cedex 13, France*

[d] *Service de Développements pour la Bioinformatique Sud-Est, CEA Grenoble, 17 rue des Martyrs, F-38054, Grenoble cedex 09, France*

Received 21 January 2004; received in revised form 14 April 2004; accepted 26 April 2004
Available online 17 June 2004
Received by W. Martin

Abstract

Comparative genomic analysis of the malaria causative agent, *Plasmodium falciparum*, with other eukaryotes for which the complete genome is available, revealed that the genome from *P. falciparum* was more similar to the genome of a plant, *Arabidopsis thaliana*, than to other non-apicomplexan taxa. Plant-like sequences are thought to result from horizontal gene transfers after a secondary endosymbiosis involving an algal ancestor. The use of the *A. thaliana* genome and proteome as a reference gives an opportunity to refine our understanding of the extreme compositional bias in the *P. falciparum* genome that leads to a proteome-wide amino acid bias. A set of pairs of non-redundant protein homologues was selected owing to rigorous genome-wide sequence comparison methods. The introduction of *A. thaliana* as a reference was a mean to weight the magnitude of the protein evolutionary divergence in *P. falciparum*. The correlation of the amino acid proportions with evolutionary time supports the hypothesis that amino acids encoded by GC-rich codons are directionally substituted into amino acids encoded by AT-rich codons in the *P. falciparum* proteome. The long-term deviation of codons in malarial sequences appears as a possible consequence of a genome-wide tri-nucleotidic signature imprinting. Additionally, this study suggests possible working guidelines to improve the accuracy of *P. falciparum* sequence comparisons, for homology searches and phylogenetic studies.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Nucleotide bias; Amino acid bias; Genomic signatures; Malaria; Apicomplexa

1. Introduction

The causative agent of malaria, *Plasmodium falciparum*, is a eukaryotic unicellular organism of the Apicomplexa phylum. Apicomplexans are thought to derive from the engulfing of a red or green alga (here termed primary endosymbiont) in an ancestral eukaryotic cell (here termed secondary endosymbiont) (Archibald and Keeling, 2002). Consistently with this multiple endosymbiosis involving an algal ancestor, the comparison with other eukaryotes for which the complete genome was available, revealed that the genome from *P. falciparum*, was more similar to that of a plant, i.e., *Arabidopsis thaliana*, than to other non-apicomplexan taxa (Gardner et al., 2002). Apicomplexans exhibit plant features such as a chloroplast-related organelle, named the apicoplast (McFadden et al., 1996; Köhler et al., 1997; Maréchal and Cesbron-Delauw, 2001). Numerous metabolites and biosynthetic processes that are unique to red algae, green algae and land plants have also been reported, being related to the plastid or not. In the carbohydrates metabolism, glycolytic enzymes are of plant type (Dzierszinski et al., 1999), and an algal storage form, the amylopectin, is synthesized within the cytosol like in read algae and dino-

* Corresponding author. Laboratoire de Physiologie Cellulaire Végétale, Département Réponse et Dynamique Cellulaire, UMR 5019 CNRS-CEA-INRA-Université Joseph Fourier, CEA Grenoble, 17 rue des Martyrs, F-38054, Grenoble cedex 09, France. Tel.: +33-4-3878-4985; fax: +33-4-3878-5091.
E-mail address: emarechal@cea.fr (E. Maréchal).

0378-1119/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2004.04.029

flagellates (Coppin et al., 2003). In the C1 metabolism, the folate biosynthetic pathway is functionally and structurally related to that of plants (for a review, see Ivanetich and Santi, 1990). In the lipid metabolism, the fatty acid biosynthesis is achieved by a chloroplastic-like multi-enzymatic complex (for a review, see Gornicki, 2003); the synthesis of glycerolipids that are unique to chloroplasts was further reported (Maréchal et al., 2002) and lipoic acid and isoprenoid plastid pathways were additionally demonstrated (Thomsen-Zieger et al., 2003; Jomaa et al., 1999). In signal transduction processes, plant-like calcium dependent protein kinases were found in *P. falciparum* (Kappes et al., 1999). As a complex heritage of the secondary endosymbiosis, one would expect that the genome of apicomplexans be a mosaic of genes deriving from the cyanobacterial plastid ancestor, the bacterial mitochondrion ancestor (an alpha-proteobacteria), the primary algal and the secondary ancestral endosymbiotionts. Like for other complex eukaryotes, the definition of a unique *P. falciparum* 'lineage' is not absolute and the contribution of the different endosymbiotic partners to the final parasite genome is still unknown (Huang et al., 2004). After horizontal transfers, from the ancestral algal genome (including the "eukaryotic" nuclear chromosomes and the "prokaryotic" organellar non-chromosomal DNAs) to the nucleus of the secondary endosymbiotic eukaryotic cell, plant-like sequences can be viewed as recent elements in the nuclear genome of the parasite. A genome-wide comparison of *P. falciparum* and *A. thaliana* proteomes appears therefore as a first possible step toward the characterization of the plant-side of the parasite. Additionally, the comparative analysis of homologues in the two organisms is an opportunity to refine our understanding of the extreme compositional bias in the *P. falciparum* genome (18 % G+C) as compared to *A. thaliana* (45% G+C).

Until recently, genome-wide studies of nucleotidic biases could only be achieved on bacterial models (Lobry, 1997). To analyze the possible correlation between the nucleotide bias and the amino acid composition correlated, Foster et al. (1997) proposed to classify the universal codon table into GC-rich, AT-rich and neutral codons: AT-rich sequences are expected to encode proteins enriched in F, Y, M, I, N and K (collectively termed FYMINK) and GC-rich sequences to encode proteins enriched in G, A, R and P (collectively termed "GARP"). Singer and Hickey (2000) investigated the accuracy of this prediction in twenty-one prokaryotes, yeast and the two *P. falciparum* chromosomes that were sequenced at that date (chromosomes 2 and 3). This study suggested that the mutational pressure, driven by the G+C bias on DNA composition, superimposed to the environmental pressure, driven by the amino acids functional conservation, in the long-term evolution of the protein composition. In *P. falciparum*, the amino acid bias is so strong that six amino acids account for more than half of the protein content (the summed frequencies of N, K, I, L, E and D in *P. falciparum* is 0.54, Fig. 1). In a pioneering work, Musto et al. (1995) showed that the amino acids from *P. falciparum* proteins were encoded by codons particularly impoverished in GC in the third position, suggesting that the GC bias might be mostly effective on silent codon positions. Re-examining the third codon bias on an updated set of malarial proteins, Musto et al. (1999) noticed that some sequences diverged from the average rule, i.e., an increase in C-ending codons in some subsets of sequences. The magnitude of the nucleotide bias impact at the amino acid level is therefore scaled, but to date, this phenomenon is unquantified.

In the present paper, we introduced the *A. thaliana* genome as a referential genome and proteome to refine the analysis of the influence of the nucleotidic bias in *P. falciparum* on the protein amino acid composition. Our study focused therefore on evolutionary conserved portions of proteins. A set of pairs of non-redundant protein homologues was selected owing to rigorous genome-wide sequence comparison methods. Using this set of pairs of homologues, the genomic compositional bias was analyzed at the tri-nucleotidic level. The aminoacid bias was investigated according to the FYMINK/GARP classification. By taking *A. thaliana* as a reference, we could investigate a correlation of the impact of the nucleotidic bias at the amino

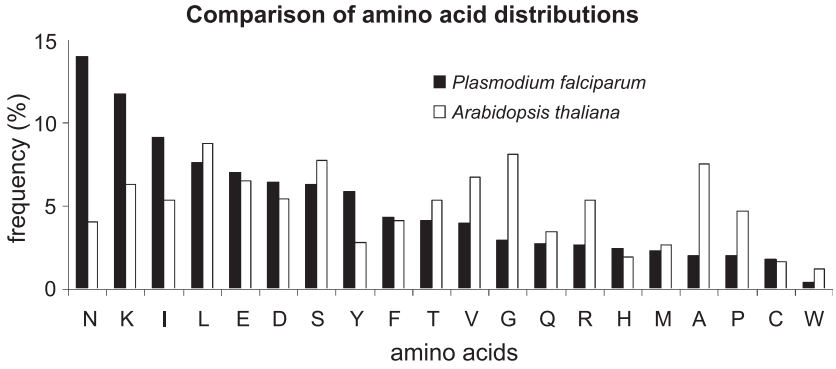


Fig. 1. Comparison of the amino acid distribution in the *P. falciparum* and *A. thaliana* proteomes. Frequencies were calculated from automatically annotated sequences at the downloading date, i.e., 5334 from *P. falciparum* and 25,545 sequences from *A. thaliana* as of December 2002. Amino acids were ranked owing to their frequencies in *P. falciparum*.

acid level, with the evolutionary divergence that separate protein homologues. Additionally, this work suggests possible working guidelines to improve the accuracy of *P. falciparum* sequence comparisons, for homology searches and phylogenetic studies.

2. Materials and methods

2.1. Annotated genomic and protein sequences from A. thaliana and P. falciparum

The genomic sequence materials were collected from *A. thaliana* and *P. falciparum* Internet genome resources, i.e., the National Center for Biotechnology Information (ftp://ftp.ncbi.nih.gov) and the PlasmoDB (http://www.plasmodb.org/) databases. We selected the automatically annotated sequences at the downloading date, i.e., 25,545 sequences from *A. thaliana* as of December 2002, 5334 from *P. falciparum* as of December 2002.

2.2. Alignment algorithms

Local sequence alignments were sought using either the exact SW (Smith and Waterman, 1981) or the heuristic BLASTP (Altschul et al., 1990) algorithms provided in the multipurpose BIOFACET software package (Gene-IT, Rueil-Malmaison, France) (Codani et al., 1999). Computations were carried out on a 1-GHz four-processor Sparc-Sun-Solaris platform hosted by CEA-Saclay, France. Default similarity matrix was Blosum 62, with a 12 gap-open cost and a 2 gap-extend cost. The BLASTP algorithm was used for intra-genomic comparisons. BLAST statistics are not theoretically valid to compare strongly divergent sequences (Altschul et al., 1990). According to the TULIP theorem (Bastien et al., 2004), non-redundant *P. falciparum* and *A. thaliana* proteome comparisons could be validly carried out with the SW algorithm, which results were sorted according to computed *Z* values.

2.3. Removal of similarity redundancies from P. falciparum and A. thaliana protein sequence databases

Redundancies were removed by stepwise computations, following the partitioning strategy from Tekaia et al. (1999). Briefly, for each proteome, we built up a random proteome database containing an identical number of protein sequences, of identical size and amino acid distribution, in which each sequence was an obligate shuffling of a corresponding sequence from the original database. Each real protein of a given organism was compared to all the sequences of the random database, using BLASTP algorithm and the best alignment *P* value was collected. From the distribution of the self × random *P* values, a 5-percentile was set to define a cutoff. Then, for each species, the calculated cutoff was used as a criterion to partition the proteome owing to the single-linkage clustering method. Eventually, the longest sequence was drawn from each similarity cluster to build up non-redundant proteomes.

2.4. Determining pairs of non-redundant homologous sequences from P. falciparum and A. thaliana

All proteins from the *A. thaliana* non-redundant proteome were compared to the *P. falciparum* non-redundant proteome, using the SW algorithm. For each alignment, a *Z* value was computed (Comet et al., 1999). A *Z* value-cutoff of 8 was used to create clusters of aligned *P. falciparum* × *A. thaliana* sequences (Bastien et al., 2004), owing to the single-clustering method (Codani et al., 1999). Clusters were examined manually to selects pairs of sequences which functional annotation appeared analogous.

2.5. Partition of the codon table

The codon table was partitioned according to Foster et al. (1997) and Singer and Hickey (2000). Amino acid encoded by codons that are GC-rich at the two first positions are termed GARP (glycine, alanine, arginine and proline) and those encoded by codons that are AT-rich at the two first

Table 1
Selection of pairs of *P. falciparum* and *A. thaliana* homologuous sequences

	P. falciparum	*A. thaliana*
Initial genomic sequences		
Number of chromosomes	14	5
Base pairs	22,853,754	115,409,949
Number of annotated sequences	5334	25,545
Genome partition (BLASTP)		
P value cutoff	1e − 17	1e − 5
Number of redundancy clusters	3344	3185
Average number of sequence per cluster	160	8.02
Non-redundant protein sequences databases		
Number of non-redundant sequences	3344	3185
Non-redundant protein sequences databases comparison (SW)	(*P. falciparum* × *A. thaliana*)	
Number of comparisons	10,637,264	
Z-value cutoff	8	
Single linkage clusters containing:		
2 sequences	287	
3 sequences	92	
4 sequences	43	
5 sequences	19	
>5 sequences	25	
Total number of clusters	466	
Average number of sequences per cluster[a]	2.79	
Set of pairs of similar sequences		
After examination of the sequence unambiguous functional annotations	71	

[a] After removal of the largest cluster containing 2764 lowest complexity sequences (the average with this cluster is 7.8 sequences per cluster).

Table 2
Pairs of selected homologues ranked by evolutionary divergence

Evolutionary divergence (rank)[a]	Pairs of homologous sequences (access numbers)	
	A. thaliana	*P. falciparum*
1	GI|15234001|REF|NP_193608.1	PFA3D7|CHR7|PF07_0079
2	GI|15242241|REF|NP_197024.1	PFA3D7|PFAL_CHR5|PFE1005W
3	GI|15220698|REF|NP_173743.1	PFA3D7|CHR6|MAL6P1.244
4	GI|15228111|REF|NP_181264.1	PFA3D7|CHR7|PF07_0088
5	GI|15223382|REF|NP_174013.1	PFA3D7|CHR14|PF14_0141
6	GI|15228152|REF|NP_178531.1	PFA3D7|CHR11|PF11_0447
7	GI|15222320|REF|NP_177693.1	PFA3D7|PFAL_CHR5|PFE0965C
8	GI|15241768|REF|NP_201036.1	PFA3D7|CHR10|PF10_0038
9	GI|15240695|REF|NP_196326.1	PFA3D7|CHR11|PF11_0065
10	GI|15221107|REF|NP_177543.1	PFA3D7|CHR10|PF10_0155
11	GI|15222741|REF|NP_173985.1	PFA3D7|CHR10|PF10_0154
12	GI|15230011|REF|NP_187210.1	PFA3D7|CHR13_1|MAL13P1.20
13	GI|15236757|REF|NP_193544.1	PFA3D7|PFAL_CHR9|PFI0190W
14	GI|15229845|REF|NP_187143.1	PFA3D7|PFAL_CHR5|PFE0975C
15	GI|15218602|REF|NP_171777.1	PFA3D7|CHR6|MAL6P1.281
16	GI|15232356|REF|NP_190957.1	PFA3D7|CHR11|PF11_0454
17	GI|15221783|REF|NP_175831.1	PFA3D7|CHR12|PFL2095W
18	GI|15218473|REF|NP_174665.1	PFA3D7|CHR11|PF11_0272
19	GI|15219362|REF|NP_173122.1	PFA3D7|CHR12|PFL1700C
20	GI|15241190|REF|NP_200446.1	PFA3D7|CHR6|MAL6P1.160
21	GI|15236042|REF|NP_194898.1	PFA3D7|CHR13_1|PF13_0228
22	GI|15239135|REF|NP_196166.1	PFA3D7|CHR7|PF07_0117
23	GI|15219044|REF|NP_175669.1	PFA3D7|CHR13_1|MAL13P1.16
24	GI|15237632|REF|NP_198953.1	PFA3D7|CHR10|PF10_0369
25	GI|15225470|REF|NP_181478.1	PFA3D7|CHR13_1|PF13_0132
26	GI|15230258|REF|NP_191281.1	PFA3D7|PFAL_CHR3|PFC0935C
27	GI|15233198|REF|NP_191077.1	PFA3D7|CHR11|PF11_0260
28	GI|15223800|REF|NP_173447.1	PFA3D7|CHR13_1|MAL13P1.19
29	GI|15224284|REF|NP_181874.1	PFA3D7|CHR11|PF11_0312
30	GI|15239406|REF|NP_198495.1	PFA3D7|CHR7|PF07_0059
31	GI|15223082|REF|NP_177781.1	PFA3D7|PFAL_CHR9|PFI0755C
32	GI|15217996|REF|NP_175576.1	PFA3D7|CHR7|MAL7P1.75
33	GI|15237847|REF|NP_197778.1	PFA3D7|CHR8|PF08_0075
34	GI|15242100|REF|NP_197592.1	PFA3D7|CHR10|PF10_0103
35	GI|15224678|REF|NP_180080.1	PFA3D7|CHR12|PFL0595C
36	GI|15231112|REF|NP_188668.1	PFA3D7|CHR13_1|PF13_0240
37	GI|15240842|REF|NP_198627.1	PFA3D7|CHR12|PFL1180W
38	GI|15227321|REF|NP_179285.1	PFA3D7|CHR10|PF10_0332
39	GI|15231796|REF|NP_190902.1	PFA3D7|CHR11|PF11_0477
40	GI|15229476|REF|NP_189000.1	PFA3D7|CHR12|PFL0210C
41	GI|15240250|REF|NP_200949.1	PFA3D7|CHR12|PFL0960W
42	GI|15219956|REF|NP_173697.1	PFA3D7|CHR13_1|MAL13P1.55
43	GI|15218176|REF|NP_177917.1	PFA3D7|CHR10|PF10_0187
44	GI|15226690|REF|NP_181581.1	PFA3D7|CHR6|MAL6P1.76
45	GI|15233740|REF|NP_194150.1	PFA3D7|CHR11|PF11_0188
46	GI|15217575|REF|NP_177324.1	PFA3D7|PFAL_CHR5|PFE0880C
47	GI|15227556|REF|NP_180514.1	PFA3D7|CHR13_1|PF13_0023
48	GI|15232448|REF|NP_190989.1	PFA3D7|PFAL_CHR5|PFE1125W
49	GI|15232110|REF|NP_186788.1	PFA3D7|CHR13_1|PF13_0130
50	GI|15223471|REF|NP_171680.1	PFA3D7|PFAL_CHR5|PFE0165W
51	GI|15240461|REF|NP_198072.1	PFA3D7|CHR12|PFL0415W
52	GI|15218281|REF|NP_173022.1	PFA3D7|CHR13_1|PF13_0061
53	GI|15226240|REF|NP_180344.1	PFA3D7|CHR12|PFL0380C
54	GI|15224925|REF|NP_181401.1	PFA3D7|CHR13_1|PF13_0092
55	GI|15238408|REF|NP_198367.1	PFA3D7|CHR10|PF10_0086
56	GI|15226877|REF|NP_181047.1	PFA3D7|CHR13_1|MAL13P1.21
57	GI|15222852|REF|NP_175419.1	PFA3D7|PFAL_CHR3|PFC0310C
58	GI|15237018|REF|NP_192839.1	PFA3D7|CHR6|MAL6P1.59|
59	GI|15239319|REF|NP_196217.1	PFA3D7|CHR13_1|PF13_0140
60	GI|15237750|REF|NP_200686.1	PFA3D7|CHR8|MAL8P1.22
61	GI|15229443|REF|NP_191908.1	PFA3D7|PFAL_CHR9|PFI0385C

Table 2 (*continued*)

Evolutionary divergence (rank)[a]	Pairs of homologous sequences (access numbers)	
	A. thaliana	*P. falciparum*
62	GI|15221343|REF|NP_176996.1	PFA3D7|CHR7|PF07_0046
63	GI|15238395|REF|NP_201331.1	PFA3D7|CHR10|PF10_0275
64	GI|15225372|REF|NP_179641.1	PFA3D7|PFAL_CHR5|PFE1265W
65	GI|15234278|REF|NP_192078.1	PFA3D7|CHR11|PF11_0265
66	GI|15218837|REF|NP_176170.1	PFA3D7|CHR10|PF10_0368
67	GI|15222767|REF|NP_175376.1	PFA3D7|CHR11|PF11_0295
68	GI|15218306|REF|NP_175009.1	PFA3D7|PFAL_CHR9|PFI0890C
69	GI|15223252|REF|NP_177239.1	PFA3D7|CHR10|PF10_0122
70	GI|15238559|REF|NP_198413.1	PFA3D7|PFAL_CHR9|PFI1110W
71	GI|15229914|REF|NP_187167.1	PFA3D7|PFAL_CHR2|PFB0210C

[a] Pairs are ranked according to the evolutionary divergence shown in Fig. 3.

positions are termed FYMINK (phenylalanine, tyrosine, methionine, isoleucine, asparagine and lysine).

2.6. Tri-nucleotide signatures according to the chaos game representation

Tri-nucleotides are three-letter words with 64 (4^3) possible combinations. Given a nucleotidic sequence, chaos game representation allows the depiction of tri-nucleotide frequencies as an image made up of 64 squares (8 lines, 8 columns) as described in Deschavanne et al. (1999). Frequencies determined from counting in 800-nucleotide portions, are indicated using a color scale. Signatures of complete genomes, individual chromosomes or sample sequences from *P. falciparum* and *A. thaliana*, were computed, on a 1-GHz bi-processor power Mac G4 platform, using the Matlab 6.5 software (The Mathworks, Natick, Massachusetts, USA). Obtained images being points defined by 64 coordinates, signatures were compared using a two-dimensional projection obtained by a principal component analysis, using the JMP 5.0.1a statistical software (SAS, Cary, North Carolina, USA).

2.7. Conversion of the amino acid alignment scores into measure of evolutionary divergence

Theoretical measures of the evolutionary divergence, based on the distance (D) that separates two similar sequences can be obtained after conversion of amino acid alignment scores (Feng and Doolittle, 1997). In the simplest pragmatic approach, all residues in a sequence are supposed equally likely to change and every amino acid is supposed to have the same likelihood for changing into any of the other amino acids. Owing to these hypotheses, D is estimated according to Eq. (1):

$$D = -\log(S) \times 100, \text{ with } S = \frac{S_{obs} - S_{rand}}{S_{ident} - S_{rand}} \quad (1)$$

In these equations, S_{obs} is the observed similarity score between a pair of sequences, S_{rand} the score obtained from randomized sequences of the same lengths and compositions and S_{ident} the average score of the sequences compared with themselves (Feng and Doolittle, 1997).

2.8. Statistical tests and graphical representations

All statistical tests and graphical representations were produced using the R 1.7 statistical software (Free Software Foundation, Boston, USA). Linear models were calculated using the $Y = aX + y_0$ estimate (with a, the slope of the curve, y_0, the intercept at the ordinate) and P_a and P_{y0}, the probabilities that a and y_0 values are not significantly distinct from zero.

3. Results

3.1. Construction of a set of non-redundant pairs of homologous sequences from P. falciparum and A. thaliana

Proteomes from *P. falciparum* and *A. thaliana* were deduced from automatically annotated protein coding sequences (5334 and 25,545 sequences, respectively) (Table 1). Similarity redundancies were removed based on the partitioning method from Tekaia et al. (1999). A self-× self proteome comparison was carried out using the BLASTP algorithm. Then, a self × random proteome comparison provided a distribution of best P value per protein compared, out of which a 5-percentile cutoff was determined, i.e., 1e − 17 for *P. falciparum* and 1e − 05 for *A. thaliana* (Table 1). Taking these P value cutoffs as single-linkage clustering criteria, *P. falciparum* and *A. thaliana* proteomes were partitioned. Longest sequences were extracted from each similarity cluster to build up non-redundant proteomes, i.e., 3344 sequences from *P. falciparum* and 3185 sequences from *A. thaliana* (Table 1). A total of 466 clusters of similar sequences from *P. falciparum* × *A. thaliana* non-redundant proteome comparison was deduced, owing to a Z value cutoff of 8 as a single-linkage clustering criterion (TULIP theorem, Bastien et al., 2004). Cluster comprising more than two sequences indicates than multiple sequences of a given organism happen to be similar to distinct portions of a sequence of the other organism. After

re-examination of the clusters by a biologist curator, 71 clusters of pairs of (*P. falciparum*, *A. thaliana*) sequences were selected based on analogous functional annotations (see Table 1). The complete list of accession numbers of the identified (*P. falciparum*, *A. thaliana*) homologues is given in Table 2. The set of 71 pairs of homologues selected was used to examine the effects of the nucleotide bias in the *P. falciparum* genome, using the homologues in the non-biased *A. thaliana* genome as references.

3.2. Comparative analysis of amino acids encoded by AT-rich and GC-rich codons in homologous sequences from P. falciparum and A. thaliana

In the following analyses, the G+C proportions in coding sequences were measured separately, in the first codon di-nucleotides (collectively termed non-synonymous sites) and third codon nucleotides. Musto et al. (1995, 1999) reported that in *P. falciparum*, the mutational pressure was seemingly stronger at the third codon site. Consistently, we recorded that the average G+C content in *A. thaliana* sequences was 45% at both non-synonymous and third codon sites, whereas the average G+C content in *P. falciparum* was greater at non-synonymous sites, i.e., 28 %, than at third codon site, i.e., 19%. We measured the proportions of amino acids coded by AT-rich codons, i.e., FYMINK, and by GC-rich codons, i.e., GARP, in each of the 142 sequences of the 71 pairs of homologues. The FYMINK and GARP proportions of *P. falciparum* (Fig. 2, triangles) and *A. thaliana* (Fig. 2, circles) were plotted as a function of the average G+C proportion in the respective coding sequences, at the codon third sites (Fig. 2A) and non-synonymous sites (Fig. 2B). In all cases, no overlap is observed in the G+C proportions of the selected set of *P. falciparum* and *A. thaliana* homologues, whatever the codon sites (Fig. 2). The global overlap in the amino acid

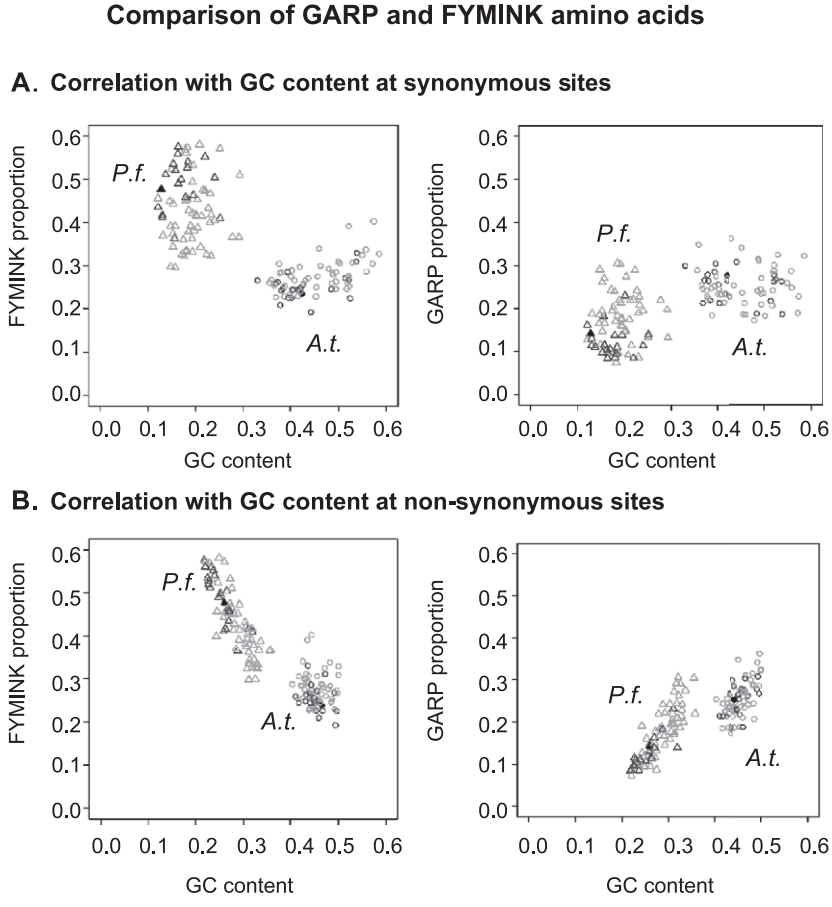


Fig. 2. Comparison of the GARP and FYMINK amino acid contents in homologous sequences from *P. falciparum* and *A. thaliana*. Both nucleotide and amino acid sequences for all pairs of homologous sequences selected in *P. falciparum* and *A. thaliana* (71 pairs) were extracted from GenBank flat files. The amino acid encoded by GC-rich codons are termed 'GARP' (glycine, alanine, arginine and proline) and those encoded by AT-rich codons are termed 'FYMINK' (phenylalanine, tyrosine, methionine, isoleucine, asparagines and lysine). Each point represents the data of a single sequence. The graphs show the relationship between the amino acid content of each sequence, i.e., proportions of FYMINK (left panels) and of GARP (right panels), and the G+C content at synonymous (A) and non-synonymous (B) codon positions of the complete given sequence. Data for *P. falciparum* sequences are shown as triangles and data for *A. thaliana* sequences are shown as circles. The evolutionary divergence between *P. falciparum* and *A. thaliana* sequences was estimated (see Fig. 3): three-color codes were set to scale the lowest (0–200, light-grey), average (200–400, dark-grey) and highest (>400, black) evolutionary divergence (see Discussion).

content (Fig. 2, projections on FYMINK and GARP proportion axes) is apparent since the GARP/FYMINK ratio in *A. thaliana* homologue was always higher than in the corresponding sequence of *P. falciparum*. Together these results show that, although the selected sequences are similar, all sequences from *P. falciparum* exhibit strict nucleotidic and amino-acid biases as compared to their *A. thaliana* counterparts, and suggest that no codon usage balance was efficiently mobilized to "erase" a possible effect of the nucleotide bias at the amino acid level.

3.3. Correlation of the compositional biases in P. falciparum sequences with the evolutionary divergence from the A. thaliana reference

To compare the *P. falciparum* compositional biases and the evolutionary divergence between the homologous sequences, we plotted the difference of GARP ($\Delta GARP = GARP_{Arabidopsis} - GARP_{Plasmodium}$) and FYMINK ($\Delta FYMINK = FYMINK_{Arabidopsis} - FYMINK_{Plasmodium}$) in the pairs of homologous sequences from *A. thaliana* and *P. falciparum*, as a function of the evolutionary divergence estimated by Eq. (1) (Fig. 3). Fig. 3 shows that the amino

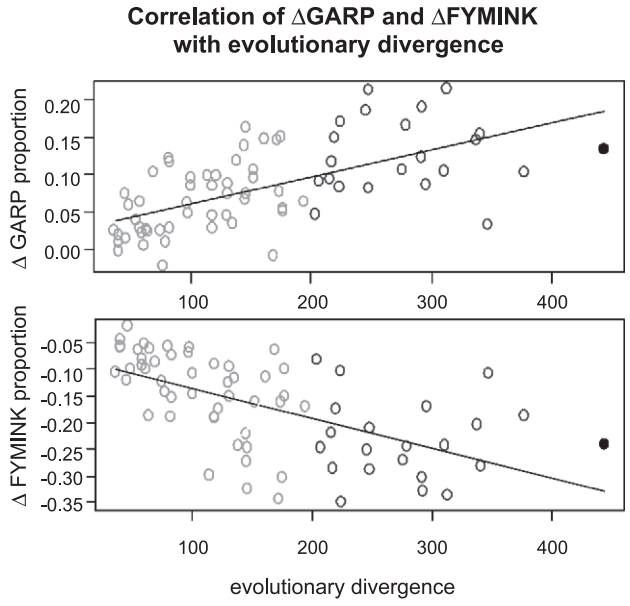


Fig. 3. Correlation between the degree of amino acid bias in pairs of *P. falciparum* and *A. thaliana* homologues and the sequence divergence. The difference of proportions of GARP ($\Delta GARP = GARP_{Arabidopsis} - GARP_{Plasmodium}$, upper panel) and of and FYMINK ($\Delta FYMINK = FYMINK_{Arabidopsis} - FYMINK_{Plasmodium}$, lower panel) in the pairs of homologous sequences from *A. thaliana* and *P. falciparum*, were plotted as a function of the evolutionary divergence estimated by Eq. (1). ΔGARP increase and Δ FYMINK decrease fit linear models, $Y = 3.50e-04X + 2.72e-02$ ($P_a < 2.78e-08$ and $P_{y0} < 0.012$) and $Y = -5.47e-04X - 8.04e-2$ ($P_a < 4.54e-08$ and $P_{y0} < 7.30e-06$), respectively. A three-color code set to scale the lowest (0–200, light-grey), average (200–400, dark-grey) and highest (>400, black) evolutionary divergence, and a ranking based on the increasing divergence shown here were used to examine results from Figs. 2 and 4.

acid composition correlates with the divergence, as judged by an increase of ΔGARP and a decrease of ΔFYMINK. Interestingly, when the GARP and FYMINK absolute proportions in individual species are plotted with the evolutionary divergence, linear deviations are measured in *P. falciparum*, whereas in *A. thaliana* sequences, no correlation with evolutionary time could be measured (not shown). These results indicate that the two proteomes have distinct amino acid directional evolutions and that the apparent orientated divergence in GARP and FYMINK proportions in *P. falciparum* and *A. thaliana* is primarily dictated by the extreme compositional biases in the *P. falciparum* genome and proteome. In that respect, *A. thaliana* is a neutral reference.

3.4. Comparison of the tri-nucleotide signatures in homologous sequences from P. falciparum and A. thaliana

In the coding regions of the genome, the codons are peculiar tri-nucleotides which mutation is trivially limited by the possible deleterious functional impact such modification would infer at the protein level. It is also known that coding sequences are enriched in GC as compared to non-coding regions. To determine if the magnitude of the evolutionary divergence between *P. falciparum* and *A. thaliana* homologues was mostly related to a genome-wide or local nucleotidic deviation in the *P. falciparum* genome, we compared the tri-nucleotidic signatures of each sequence used in this study (local signatures), to the genome-wide signatures. Independently of the possible codons they might constitute, all tri-nucleotide frequencies were displayed as images, where each pixel represents the frequency of one of the 64 possible tri-nucleotides (Deschavanne et al., 1999). AT-rich tri-nucleotides appear in the lower part of images, whereas GC-rich tri-nucleotides appear in the upper part. In Fig. 4A and B, the average signatures of the *A. thaliana* and *P. falciparum* genomes are shown. The homologous sequences were ranked according to the magnitude of their evolutionary divergence like in Fig. 3. Signatures of six of the 71 pairs of homologues (ranks 4, 5, 28, 29, 40 and 44) are shown as representative examples. Signatures were additionally plotted according to a principal component analysis (Fig. 4C). Dark-red and dark-green spots correspond to the average tri-nucleotidic signatures of the 14 chromosomes of *P. falciparum* and the 5 chromosomes of *A. thaliana*, respectively. Red ellipses correspond to the local variability of the tri-nucleotidic signature in the *P. falciparum* genome as judged from genomic signature computation in 800-nucleotide segments. The remote ellipse in the upper right corner correspond to a specific telomeric local signature in the *P. falciparum* genome and is not observed when the telomers are excluded from this computation. Green ellipses correspond to the local variability of the tri-nucleotidic signature in the *A. thaliana* genome. Eventually, light-red and light-green spots correspond to the local tri-nucleotidic signatures of the 71 selected sequences from *P. falciparum* and *A. thaliana*, respectively.

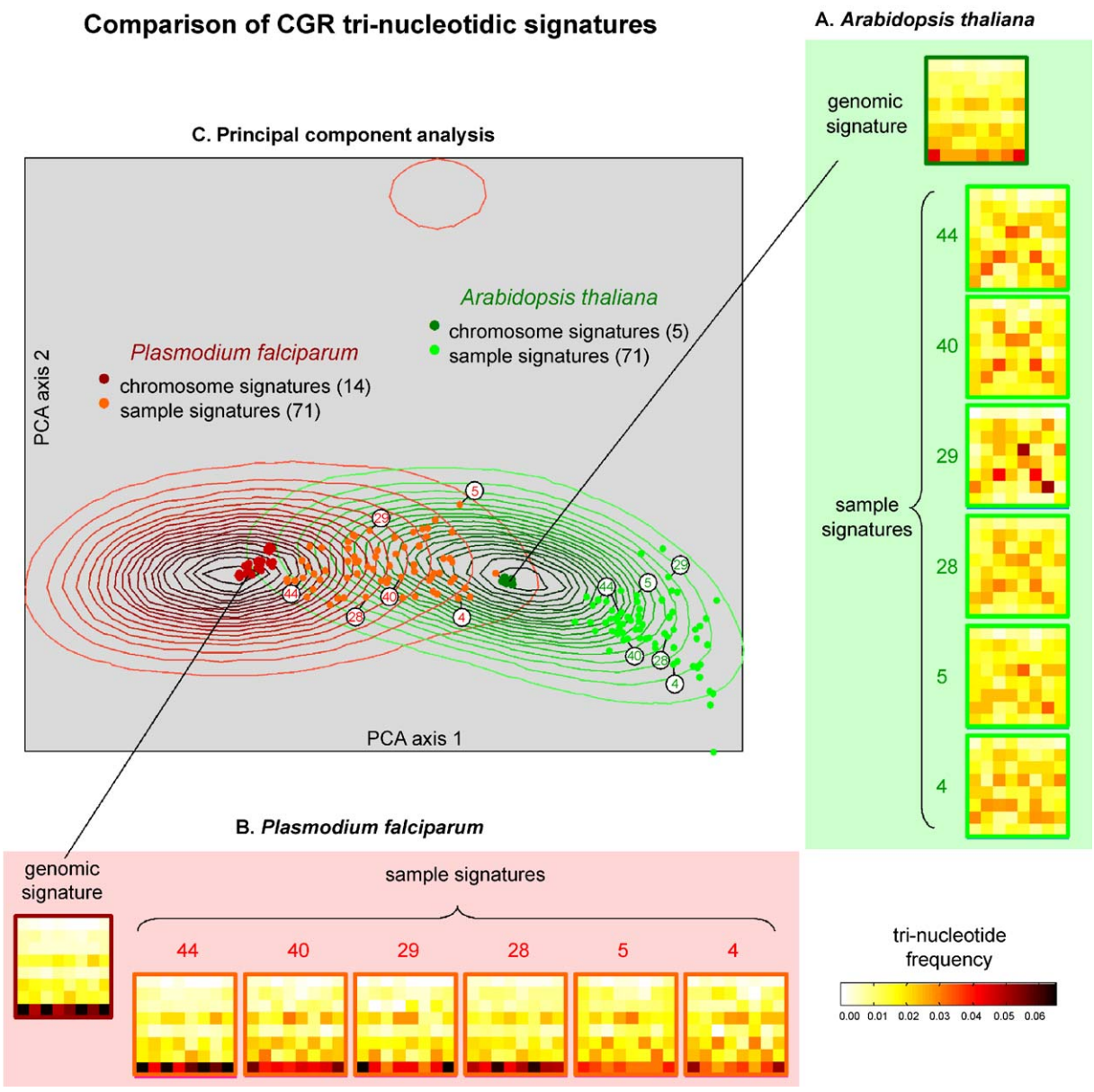


Fig. 4. Comparison of the tri-nucleotidic signatures of *P. falciparum* and *A. thaliana* homologous coding sequences. Tri-nucleotides of the coding sequences of each of the *P. falciparum* and *A. thaliana* sequences selected in this study were counted and their frequencies were displayed as images, where each square represents the frequency of one of the 64 possible tri-nucleotides. The homologous sequences were ranked according to the magnitude of their evolutionary divergence shown in Fig. 3. The signatures of six of the 71 pairs of homologues (ranks 4, 5, 28, 29, 40 and 44) are shown as representative examples. Tri-nucleotide frequencies are indicated using a color scale shown in the lower right corner. (A) Tri-nucleotidic signatures of the *A. thaliana* genome and samples. (B) Tri-nucleotidic signatures of the *P. falciparum* genome and samples. (C) Principal component analysis of the genomic tri-nucleotidic signatures. Dark-red and dark-green spots correspond to the tri-nucleotidic signatures of the 14 chromosomes of *P. falciparum* and the 5 chromosomes of *A. thaliana*, respectively. Red and green ellipses correspond to the genome-wide distribution of the tri-nucleotidic signatures in the *P. falciparum* and *A. thaliana*, respectively, as judged from signature computation in 800-nucleotide segments. Light-red and light-green spots correspond to the tri-nucleotidic signatures of the 71 selected sequences from *P. falciparum* and *A. thaliana*, respectively (average length ~ 800 nucleotides).

The GC-balance roughly determines the horizontal axis. The selected 142 sequences exhibit signatures that are distinct from the average or chromosomic genomic signatures. Our analysis further shows that coding sequences from *P. falciparum* and *A. thaliana* (mean length = 800 nucleotides) harbor signatures expectedly found in the GC-rich portions of the *P. falciparum* and *A. thaliana* distributions of 800-nucleotide genomic signatures respectively. Very interestingly, taking the evolutionary divergence ranking defined in Fig. 3, the ranks of the *P. falciparum* selected sequences globally increased when their local signature were closer to the average *P. falciparum* signature. As an illustration, signatures of the 4th, 5th, 28th, 29th, 40th and 44th sequences (Fig. 4B) show a directional orientation toward the average genomic signature (Fig. 4B). By contrast, the ranks of the *A. thaliana* selected sequences did not correlate with

the position of their local signature. As an illustration, signatures of the 4th, 5th, 28th, 29th, 40th and 44th sequences shown in Fig. 4A do not exhibit any obvious trend. This refined analysis suggests that the evolutionary divergence of the *P. falciparum* sequences with their *A. thaliana* counterparts increases when the nucleotidic signature is closer to the average genomic signature of the *P. falciparum* genome. The closest from the *P. falciparum* average genomic signature, the most divergent from *A. thaliana* and the most biased amino acid content. From this analysis, the compositional bias in *P. falciparum* sequences appears as a possible consequence of a genomic signature bias imprinting.

3.5. Correlation of GARP decrease and FYMINK increase in P. falciparum sequences with the evolutionary divergence from the A. thaliana reference

The change of a lysine, encoded by AT-rich codons, into an arginine, encoded by GC rich-codons, in sequences enriched in G + C, was extensively described in a comparative study of homologous sequences from *Caenorhabditis elegans* and *Saccharomyces cerevisiae* (Nishizawa and Nishizawa, 1998). Accordingly, in the set of 71 pairs of homologous sequences identified in *P. falciparum* and *A. thaliana*, the substitution profiles of lysine showed that a conservation is favored in the *A. thaliana* → *P. falciparum* direction as compared to an increased substitution into arginine in the *P. falciparum* → *A. thaliana* direction (not shown). The non-symmetrical substitution pattern observed for lysine might reflect a general rule that orientates the substitution of AT-rich encoded amino acids into GC-rich amino acids. In Fig. 5, we plotted GARP vs. FYMINK proportions in the homologous sequences from *P. falciparum* and *A. thaliana*. In *P. falciparum*, the increase in GARP amino acids is correlated with a decrease of FYMINK amino acids (Fig. 5A). By contrast, there is no correlation between GARP and FYMINK amino acid proportions in *A. thaliana* (Fig. 5B). In addition to the long-term increase of FYMINK and decrease of GARP proportions with evolutionary time, this study supports the hypothesis that GARP amino acids are directionally substituted into FYMINK amino acids in the *P. falciparum* proteome.

4. Discussion

Sequences of *P. falciparum* proteins are difficult to analyze using conventional bio-analytical tools, particularly because of the striking amino acid bias shown in Fig. 1, which reflects the extreme nucleotide compositional bias of the malarial genome. Thank to the recent complete genome sequence release, the genome-wide distribution of the compositional bias is now fully recorded (Gardner et al., 2002). In this study, we used *A. thaliana* as a referential genome, to refine the analysis of the *P. falciparum* nucleotide and amino acid biases. Non-redundant pairs of homologues were rigorously selected (Tables 1 and 2). The comparison of the proportion of G + C in third and non-synonymous codon positions shows that although the codon usage does not mute the nucleotidic bias at the amino acid level, an environmental pressure seems to partly counteract the nucleotidic mutational pressure at non-synonymous codon positions, consistently with hypotheses suggested from previous reports (Musto et al., 1995, 1999). In spite of the amino acid conservational pressure, we further noticed that the amino acids encoded by GC-rich codons (GARP) and amino acids encoded by AT-rich codons (FYMINK) were strictly distinct in *P. falciparum* and *A. thaliana* pairs of homologues (Fig. 2). The introduction of *A. thaliana* as a referential genome allowed an estimate of the evolutionary

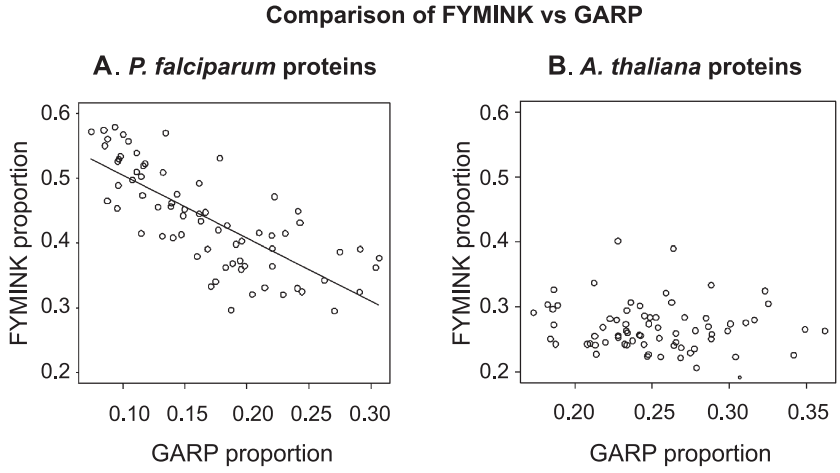


Fig. 5. Relationship between GARP and FYMINK proportions in *P. falciparum* proteins. Proportions of FYMINK were plotted as a function of GARP proportions in the set of homologous sequences from *P. falciparum* (A) and *A. thaliana* (B), selected in the present study. In *P. falciparum*, the increase in GARP proportion is correlated with a decrease in FYMINK proportion (linear model: $Y = -0.97338X + 0.601$, $P_a < 2.36e-14$ and $P_{y0} < 2e-16$). In *A. thaliana*, no correlation between GARP and FYMINK proportions is detected (linear model: $Y = -0.11X + 0.298$, $P_a > 0.3$).

divergence of malarial sequences from the plant homologues. In that respect, *A. thaliana* is not considered as the closest homologues of the plant-side of the parasite, but rather as the best characterized plant model studied to date. Thus, a relative measure of the magnitude of the amino acid bias could be related to the evolution that affected malarial proteins. The biggest the evolutionary divergence, the strongest the relative enrichment in FYMINK and the impoverishment in GARP in the *P. falciparum* sequence (Fig. 3). From the analysis of individual *P. falciparum* and *A. thaliana* sequences, ranked by the increasing divergence indicated in Fig 3, we further showed that the relative enrichment in FYMINK and impoverishment in GARP in *P. falciparum* sequences was related to an absolute amino acid orientated compositional change in the malarial proteins. Thus, *A. thaliana* happens to be a neutral referential proteome. In Fig. 3, three-color codes were set to scale the lowest (0–200, light-grey), average (200–400, dark-grey) and highest (>400, black) evolutionary divergence. When re-examining the contents in GARP and FYMINK and the GC content in *P. falciparum* and *A. thaliana* pairs of homologues (Fig. 2), we observe that the distribution of divergent sequences does not exhibit any clear correlation in *A thaliana* (Fig. 2, light-grey, dark-grey and black circles). By contrast, a spectacular correlation of the GARP decrease and the FYMINK increase in *P. falciparum* sequences, and of the GC impoverishment is observed (Fig. 2, light-grey, dark-grey and black triangles). In a long-term evolution (Fig. 2, dark-grey and black triangles), sequences are further characterized by the strongest GC-impoverishment, at both third and non-synonymous sites of codons (Fig. 2A and B). Together, these analyses show that the introduction of *A. thaliana* as a reference was a mean to weight the magnitude of the protein evolutionary divergence in *P. falciparum*. The refined analysis of the nucleotidic pattern using the chaos game representation signatures further showed that the selected *P. falciparum* and *A. thaliana* homologous sequences harbored specific tri-nucleotidic signatures that were expectedly distinct from the species average genomic signature (Fig. 4A and B). The principal component analysis of the homologue pairs indicated that the sequence evolutionary divergence between *P. falciparum* and the *A. thaliana* reference increases with the convergence of the *P. falciparum* local genic signatures toward the average malarial genomic signature (Fig. 4C). The bigger the evolutionary divergence, the closer the genic local signature is to the average genomic signature. Eventually, the relation between FYMINK and GARP proportions in *P. falciparum* and *A. thaliana* sequences (Fig. 5) showed that the amino acid bias in *P. falciparum*, which increases with the evolutionary divergence, is likely a consequence of the substitution of amino acids encoded by GC-rich codons (GARP) by amino acids encoded by AT-rich codons (FYMINK).

From the data presented in this paper, one can expect that if a given *P. falciparum* sequence shared ancestry with a plant sequence (due to an horizontal gene transfer from the primary endosymbitoic algal genome) and strongly diverged, its nucleotidic composition should be strongly biased, with a signature seemingly attracted toward the average *P. falciparum* genomic signature. As a result, the homology of a *P. falciparum* sequence with a related sequence from another organism should be 'blurred' by the magnitude of the amino acid bias and no longer detected by conventional automatic similarity search procedures. In that respect, conventional alignments using symmetric substitution matrices are pragmatically valid when homologous sequences are evolutionary close (the smallest the evolutionary divergence, the less the compositional bias). However, they are not theoretically accurate to compare *P. falciparum* proteins with evolutionary distant proteins of other species. Ideally, prior to any genome-scale comparisons including highly biased genomes such as that of *P. falciparum*, dynamic correcting methods should be developed to take into account the directional amino acid substitutions reported here, as a function of the evolutionary divergence of the aligned sequences. Our study shows that the evolutionary divergence can be primarily estimated from the *P. falciparum* genic GC content or local genomic signature. Substitution matrices should therefore be refined to take into account orientated amino acid substitutions and might prove helpful to investigate the possible homology matches that are still missing for 60% of the annotated *P. falciparum* genes and to initiate the comprehensive and massive dissection of the plant-side of the parasite, at the whole genome level.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Archibald, J.M., Keeling, P.J., 2002. Recycled plastids: a 'green movement' in eukaryotic evolution. Trends Genet. 18, 577–584.

Bastien, O., Aude, J.C., Roy, S., Maréchal, E., 2004. Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of *Z*-value statistics. Bioinformatics 20, 534–537.

Codani, J.J., Comet, J.P., Aude, J.C., Glémet, E., Wozniak, A., Risler, J.L., Hénaut, A., Slonimski, P.P., 1999. Automatic analysis of large-scale pairwise alignments of protein sequences. In Methods in Microbiology, vol. 28. Automation, Genomic and Functional Analysis. Academic Press, Elsevier, Amsterdam, pp. 229-244.

Comet, J.P., Aude, J.C., Glemet, E., Risler, J.L., Henaut, A., Slonimski, P.P., Codani, J.J., 1999. Significance of *Z*-value statistics of Smith–Waterman scores for protein alignments. Comput. Chem. 23, 317–331.

Coppin, A., Dzierszinski, F., Legrand, S., Mortuaire, M., Ferguson, D., Tomavo, S., 2003. Developmentally regulated biosynthesis of carbohydrate and storage polysaccharide during differentiation and tissue cyst formation in *Toxoplasma gondii*. Biochimie 85, 353–361.

Deschavanne, P.J., Giron, A., Vilain, J., Fagot, G., Fertil, B., 1999. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Mol. Biol. Evol. 16, 1391–1399.

Dzierszinski, F., Popescu, O., Toursel, C., Slomianny, C., Yahiaoui, B., Tomavo, S., 1999. The protozoan parasite *Toxoplasma gondii* expresses two functional plant-like glycolytic enzymes. Implications for evolutionary origin of apicomplexans. J. Biol. Chem. 274, 24888–24895.

Feng, D.F., Doolittle, R.F., 1997. Converting amino acid alignment scores into measures of evolutionary time: a simulation study of various relationships. J. Mol. Evol. 44, 361–370.

Foster, P.G., Jermiin, L.S., Hickey, D.A., 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J. Mol. Evol. 44, 282–288.

Gardner, M.J., et al., 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. Nature 419, 498–511.

Gornicki, P., 2003. Apicoplast fatty acid biosynthesis as a target for medical intervention in apicomplexan parasites. Int. J. Parasitol. 33, 885–896.

Huang, J., Mullapudi, N., Sicheritz-Ponten, T., Kissinger, J.C., 2004. A first glimpse into the pattern and scale of gene transfer in the Apicomplexa. Int. J. Parasitol. 34, 265–274.

Ivanetich, K.M., Santi, D.V., 1990. Bifunctional thymidylate synthase-dihydrofolate reductase in protozoa. FASEB J. 4, 1591–1597.

Jomaa, H., Wiesner, J., Sanderbrand, S., Altincicek, B., Weidemeyer, C., Hintz, M., Turbachova, I., Eberl, M., Zeidler, J., Lichtenthaler, H.K., Soldati, D., Beck, E., 1999. Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. Science 285, 1573–1576.

Kappes, B., Doerig, C.D., Graeser, R., 1999. An overview of Plasmodium protein kinases. Parasitol. Today 15, 449–454.

Köhler, S., Delwiche, C.F., Denny, P.W., Tilney, L.G., Webster, P., Wilson, R.J., Palmer, J.D., 1997. A plastid of probable green algal origin in Apicomplexan parasites. Science 275, 1485–1489.

Lobry, J.R., 1997. Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. Gene 205, 309–316.

Maréchal, E., Cesbron-Delauw, M.F., 2001. The apicoplast: a new member of the plastid family. Trends Plant Sci. 6, 200–205.

Maréchal, E., Azzouz, N., de Macedo, C.S., Block, M.A., Feagin, J.E., Schwarz, R.T., Joyard, J., 2002. Synthesis of chloroplast galactolipids in apicomplexan parasites. Eukaryot. Cell 1, 653–656.

McFadden, G.I., Reith, M.E., Munholland, J., Lang-Unnasch, N., 1996. Plastid in human parasites. Nature 381, 482.

Musto, H., Rodriguez-Maseda, H., Bernardi, G., 1995. Compositional properties of nuclear genes from *Plasmodium falciparum*. Gene 152, 127–132.

Musto, H., Romero, H., Zavala, A., Jabbari, K., Bernardi, G., 1999. Synonymous codon choices in the extremely GC-poor genome of *Plasmodium falciparum*: compositional constraints and translational selection. J. Mol. Evol. 49, 27–35.

Nishizawa, M., Nishizawa, K., 1998. Biased usages of arginines and lysines in proteins are correlated with local-scale fluctuations of the G+C content of DNA sequences. J. Mol. Evol. 47, 385–393.

Singer, G.A., Hickey, D.A., 2000. Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol. Biol. Evol. 17, 1581–1588.

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197.

Tekaia, F., Lazcano, A., Dujon, B., 1999. The genomic tree as revealed from whole proteome comparisons. Genome Res. 9, 550–557.

Thomsen-Zieger, N., Schachtner, J., Seeber, F., 2003. Apicomplexan parasites contain a single lipoic acid synthase located in the plastid. FEBS Lett. 547, 80–86.

Available online at www.sciencedirect.com
SCIENCE DIRECT
C. R. Biologies 328 (2005) 445–453
COMPTES RENDUS BIOLOGIES
http://france.elsevier.com/direct/CRASS3/

Genetics / Génétique

Construction of non-symmetric substitution matrices derived from proteomes with biased amino acid distributions

Olivier Bastien [a,b], Sylvaine Roy [c], Éric Maréchal [a,*]

[a] *Laboratoire de physiologie cellulaire végétale, département « Réponse et Dynamique cellulaire », UMR 5019, CNRS–CEA–INRA–université Joseph-Fourier, CEA Grenoble, 17, rue des Martyrs, 38054 Grenoble cedex 09, France*
[b] *Gene-IT, 147, av. Paul-Doumer, 92500 Rueil-Malmaison, France*
[c] *Laboratoire de biologie, informatique et mathématiques, département « Réponse et Dynamique cellulaire », CEA Grenoble, 17, rue des Martyrs, 38054 Grenoble cedex 09, France*

Received 6 December 2004; accepted after revision 1 February 2005

Available online 25 February 2005

Presented by Roland Douce

Abstract

Automatic comparison of compositionally biased genomes, such as that of the malarial causative agent *Plasmodium falciparum* (82% adenosine + thymidine), with genomes of average composition, is currently limited. Indeed, popular tools such as BLAST require that amino acid distributions be similar in aligned sequences. However, the *P. falciparum* genome is so biased that six amino acids account for more than 50% of the protein composition. One reason for the comparison methods failure lies in the compositional difference between the query and the subject proteomes, which is not taken into account in the amino acid substitution matrices. This paper introduces a method to derive substitution matrices, in particular BLOSUM 62, in the frame of the information theory. It allows the construction of non-symmetrical matrices, taking into account the non-symmetric amino acid distributions. The dirAtPf family of matrices allowing the comparison of *P. falciparum* and *A. thaliana* is given as an example. This paper further provides an analysis of the obtained matrices in the frame of the information theory, supporting the discrimination advantage they bring. ***To cite this article: O. Bastien et al., C. R. Biologies 328 (2005).***
© 2005 Académie des sciences. Published by Elsevier SAS. All rights reserved.

Résumé

La comparaison automatique de génomes biaisés, tel que celui de l'agent du paludisme *Plasmodium falciparum* (82 % adénosine + thymidine), avec des génomes de composition moyenne, est limitée. En effet, les outils populaires, tels que BLAST, imposent que les distributions en amino acides des séquences comparées soient proches. Or le génome de *P. falciparum* est tellement biaisé que six aminoacides constituent plus de 50 % de la composition protéique. Une cause de l'échec des méthodes de comparaison est de ne pas tenir compte de ces différences de distributions entre protéomes « requête » et « sujet », en parti-

* Corresponding author.
E-mail address: emarechal@cea.fr (É. Maréchal).

1631-0691/$ – see front matter © 2005 Académie des sciences. Published by Elsevier SAS. All rights reserved.
doi:10.1016/j.crvi.2005.02.002

culier au niveau de la matrice de substitution des aminoacides. Cette note présente une méthode pour dériver les matrices de substitution, en particulier BLOSUM 62, dans le cadre de la théorie de l'information. Il est ainsi possible de construire des matrices non symétriques, tenant compte de la non-symétrie des distributions en amino acides. La famille dirAtPf de matrices permettant de comparer *Arabidopsis thaliana* et *Plasmodium falciparum* est proposée comme exemple. Cette note présente, de plus, une analyse de ces matrices dans le cadre de la théorie de l'information, soutenant théoriquement le gain de discrimination qu'elles peuvent apporter. ***Pour citer cet article : O. Bastien et al., C. R. Biologies 328 (2005).***
© 2005 Académie des sciences. Published by Elsevier SAS. All rights reserved.

Keywords: Substitution matrix; BLOSUM; Biased genome; *Plasmodium falciparum*; Information theory; Mutual information

Mots-clés : Matrice de substitution ; BLOSUM ; Génome biaisé ; *Plasmodium falciparum* ; Théorie de l'information ; Information mutuelle

1. Introduction

Comparison of biological macromolecules has become an everyday task for biologists, for extremely diverse purposes such as genomic sequencing, structural modelling, functional inference, phylogenetic reconstruction, allelic or mutational analyses, etc. In all cases, comparison methods rely on a fundamental postulate that one can simply state as: "the closer in the evolution, the more alike and reversely, the more alike, *probably* the closer in the evolution" [1]. Numerous computer-based tools are used to estimate the proximity of protein sequences [2]. Alignment of two sequences is typically done by maximizing (or minimizing) a given quantity, named score, which reflects the shared features of the two biological entities. Global alignment algorithms [3] are not accurate to assess homology of domains in modular proteins [4]; local alignments are better suited [5,6]. They use scoring matrices to maximize the summed scores of compared residues and find optimal local alignments, computed with a dynamic programming procedure [2–6]. Scoring matrices have been found to be similarity matrices as well [2,4,6]. Many similarity matrices are available [7–10] and evaluation studies led to the conclusion that those based on a log-odds ratio, like BLOSUM 62 [8], over performed the others [9]. BLOSUM 62 was computed using blocks of aligned sequences with

$$s_{ij} = \frac{1}{\lambda}\log\left(\frac{q_{ij}}{p_i p_j}\right) \tag{1}$$

where i and j are aligned amino acids, q_{ij} the frequency of the observation: "i is aligned with j", i.e. the target frequency, p_i and p_j are respectively the i and j frequency, i.e. the background frequency and λ is a scaling factor.

Karlin and Altschul [11] have shown that substitution matrices depend on a particular set of data in which amino acids are paired with frequencies that correspond to the matrices' target frequencies. If the set of data is made up of aligned homologous sequences, then the matrix is usable to distinguish distant local homologies, from similarities due to chance [12]. Using information theory, Altschul [12] have further reported that substitution matrices can be evaluated using the average information they contained. This average information, known as the *relative entropy* of Shannon (1948) [13], was computed as

$$H = \sum_{i,j} q_{ij} s_{ij} \tag{2}$$

This formula can be trivially applied to similarity matrices, in order to estimate their 'sensitivity'. H computation is therefore a popular parameter when new matrices are proposed. In a recent report [14], it has further been used as a computation constraint to derive substitution matrices, i.e. with constant entropy. Information theory is much more than a practical frame for matrix computation; it allows essentially the translation of biological properties into mathematical models, particularly in probabilistic terms.

Given a probability law P that characterizes a random variable, the Hartley self-information h [15] is defined as the amount of information one gains when an event i occurs:

$$h(i) = -\log\big(P(i)\big) \tag{3}$$

The less likely an event i, the more we learn about the system when i happens. The mutual-information I between two events, is the reduction of the uncertainty of one event i due to the knowledge of the other j:

$$I_{j\to i} = h(i) - h(i/j) \tag{4}$$

Mutual information being symmetrical, $I_{j\to i} = I_{i\to j}$ and is noted $I(i;j)$. Self and mutual information of two events i and j are related:

$$h(i \cap j) = h(i) + h(j) - I(i;j) \tag{5}$$

If the occurrence of one of the two events makes the second impossible, mutual information is equal to $-\infty$. If the two events are fully independent, mutual information is null.

Recently, we rigorously demonstrated that the score $s(i,j)$ between two amino acids was the *mutual information*, in the sense of Hartley, between the two considered amino acid (Bastien et al., submitted), that is to say:

$$s(i,j) = \frac{1}{\lambda} I(i;j) \tag{6}$$

This assertion was implicit in Eq. (2). It first implies that it is impossible to build separable and metric sequence spaces that conserve the mutual information between compared sequences. Second, the fundamental postulate can be reformulated in the information theory framework: "Given two homologous proteins a and b, the closer in the evolution, the greater the mutual information between a and b and reversely, the greater the mutual information between a and b, *probably* the closer in the evolution."

Whereas the BLOSUM model is efficient in most cases, it fails to estimate satisfactorily the alignment between two proteins of very different amino acid composition [16–18]. A major reason lies in Eq. (1) that does not account for the distinct sequences where the amino acids i and j are sampled. This can be of importance when compared proteins are from very different cell environments (like soluble or membrane-bound proteins) or of strongly different amino acids composition [18–21]. In a pioneering study addressing this problem, Müller et al. [22] introduced a non-symmetric substitution matrices model for the comparison of homologous trans-membrane proteins and showed that this kind of matrices had a larger discrimination power, i.e. specificity. We reformulated this model for the comparison of biased and non-biased genomes ([16] and the present work).

Considering the general case of genome comparison with distinct global amino acid compositions, we used mutual information theory to construct non-symmetric substitution matrices dedicated to the detection of homologous sequences. An important application is the computing of non-symmetric matrices for the comparison between complete proteomes with deep differences in their composition in amino acids. Comparison of *Arabidopsis thaliana* and *Plasmodium falciparum* proteomes is given as a case study.

2. Methods

2.1. Nomenclature for sequences databases (query, subject, Species 1 and 2) and for the non-symmetric $DirSp_1Sp_2$ matrices

In molecular alignments, we used the standard nomenclature for homology searches methods, i.e. *query* for a known probing sequence (or database of sequences) that is compared to another sequence (or database of sequences), termed *subject*. The family of non-symmetric matrices computed here were dedicated to the comparison of a query sequence from a first species (termed Species1) with a subject database, or a single sequence, from a second species (termed Species2). The family of amino acid substitution matrices referred to as $\mathrm{DirSp_1Sp_2}$ (Species1 → Species2) were designed to be implemented in the conventional BLASTP alignment algorithm: columns correspond to the query (or Species1) and rows to the subject (or Species2) entries.

2.2. Identification of a non-redundant set of homologous proteins between Arabidopsis thaliana *and* Plasmodium falciparum

As a source of genomic sequence material, we selected the annotated sequences from *Arabidopsis thaliana* and *Plasmodium falciparum* from Internet databases, respectively the National Center for Biotechnology Information server (ftp://ftp.ncbi.nih.gov) and the Plasmodb genome resource (http://plasmodb.org/). The massive annotation resulting from collaborative efforts, genomic annotations contain errors. We selected therefore annotated sequences that were judged trustworthy at the downloading date in the respective Internet-available databases, i.e. 25 545 sequences from *A. thaliana* as of December 2002, 5334 from *P. falciparum* as of December 2002. According to a method describe previously [18], the two proteomes

were used to identify a non-redundant set of homologous proteins using the BLASTP program and the Smith–Waterman algorithm [5,6], implemented in the Biofacet software package (Gene-IT, France, [23]). To remove the similarity redundancies from *P. falciparum* and *A. thaliana* protein-sequence databases, for each proteome, we built up a random proteome database containing an identical number of protein sequences, of identical size and amino acid distribution, in which each sequence was an obligate shuffling of a corresponding sequence from the original database. Each real protein of a given organism was compared to all the sequences of the random database using BLASTP algorithm; the best alignment P-value was collected. From the distribution of the self x random P-values, a 5-percentile was set to define a cut-off. Then, for each species, the calculated cut-off was used as a criterion to partition the proteome owing to the single-linkage clustering method. Eventually, the longest sequence was drawn from each similarity cluster to build up non-redundant proteomes. All proteins from the *A. thaliana* non-redundant proteome were compared to the *P. falciparum* non-redundant proteome, using the SW algorithm. For each alignment, a Z-value was computed and a Z-value-cut-off of 8 was used to create clusters of aligned *P. falciparum* $\times$ *A. thaliana* sequences, owing to the single-clustering method. In this paper, this set was termed 'automatic training set'. Clusters were examined manually to select pairs of sequences whose functional annotation appeared analogous. This set was termed 'manual training set' (see table in [18]).

2.3. Initial construction of a database of protein blocks, from pairs of aligned sequences from Species1 and Species2 (or query and subject)

As described by [24], local alignments can be represented as ungapped blocks with each row a different protein segment and each column an aligned residue position. In the particular case described here, blocks can be simply derived as ungapped segments in pairs of aligned sequences. These 2-line blocks were ordered so that the first sequence always belongs to the same genome. Substitutions of a given amino acid from a first sequence (Species1 or query) with another amino acid from a second (Species2 or subject) were counted in all pairs of matching amino acids in each blocks in the database and then summed. The substitution frequency table is used to calculate matrices representing the odd ratio between these observed frequencies and those expected by chance.

2.4. Families of similarity matrices computed from blocks filtered by segment clustering

Closely related blocks in blocks databases exhibit a high percentage of identity, up to 100% when no amino acid substitution is observed in aligned sequences. Evolutionary divergence is marked by a decrease in the identity percentage. Thus, the distribution of the identity percentages within a block database can lead to a biased calculation of substitution matrices, that over- or underscores alignments of close or distant sequences. Therefore, for each matrix computing, the training set of blocks was filtered using a clustering percentage, so that sequence segments that were identical for at least that percentage of amino acid were kept for the substitution frequency counting. This filtering is an alternative definition of the clustering percentage described by [8], in which the multiple contribution of segments that were identical for at least that percentage, were averaged in calculating pair frequencies. In both cases, the decrease in the clustering percentage implies a decrease in the contribution of the blocks which percentage of identity is higher than the clustering percentage. Like for the BLOSUM family of matrices, varying the clustering percentage leads to a family of matrices.

2.5. Iterative process in the computation of non-symmetric matrices

Construction of $\mathrm{DirSp_1Sp_2}$ matrices was stepwise. Frequency tables, matrices, and programs for UNIX machines were primarily designed using the Biofacet multipurpose package (Gene-IT, Rueil-Malmaison, [23]. The initial training set of pairs of sequences derived from alignments using the Smith–Waterman algorithm implemented with BLOSUM 62. For a given clustering percentage, an initial non-symmetric matrix was computed, and indexed *1*: $\mathrm{DirSp_1Sp_{21}}$. The initial training set was then re-aligned using the Smith–Waterman algorithm implemented with this first non-symmetric matrix. From these alignments, ungapped segments were selected and filtered owing

to the defined clustering percentage, and used as a new database of Blocks to compute a new non-symmetric matrix indexed *2*: $\mathrm{DirSp_1Sp_{22}}$. The process was iterated, outputting a convergent family of matrices $\mathrm{DirSp_1Sp_{2,3}}$, $\mathrm{DirSp_1Sp_{2,4}}$, ... $\mathrm{DirSp_1Sp_{2,n}}$. The stable matrices where referred to as un-indexed $\mathrm{DirSp_1Sp_2}$.

3. Results and discussion

3.1. Question of the mutual information between two homologous sequences of distinct amino acid distributions

The physical environment of proteins, (pH, water solubility or association to membranes), the codon use that is required for their synthesis or nucleotidic compositional trends are constraints that can lead to very uncommon amino acid distributions in some families of protein or even in complete proteomes [16–18,22]. Although biased amino acid distributions affect the performance of protein comparison tools built for 'average' amino acid distributions, they can bring useful information to discriminate homologous proteins. To that purpose, we considered two kinds of sequences, or set of sequences, the first named *query*, the second *subject*. For example, *query* sequence can be from a first species, called Species1 (such as *Arabidopsis thaliana*; with an average nucleotidic −55% A+T− and amino acid distribution) and the *subject* sequence from a second species, called Species2 (such as *Plasmodium falciparum*; which nucleotidic bias −82% A+T− leads to a biased compositional proteome).

We can consider two kinds of events: $\{X = i\}$ given the amino acids i in the *subject* sequence and $\{Y = j\}$ given the amino acids j in the *query* sequence (as an application, X can be defined in a given species such as *Arabidopsis thaliana* and Y in another such as *Plasmodium falciparum*). The self-information h, for the occurrence of a given amino acid does not have the same signification in the context of each sequence:

$$h_X(i) = -\log\big(P_X(i)\big) \neq -\log\big(P_Y(i)\big) = h_Y(i) \qquad (7)$$

with P_X and P_Y the probability laws assigned to the random variables X and Y, respectively. From inequality (7), knowing an amino acid in one of the aligned sequences does not bring the same quantity of information concerning an amino acid occurrence at the aligned site of the other sequence. This can be easily verified for Asparagines in the case of *Arabidopsis thaliana* and *Plasmodium falciparum*. In *P. falciparum*, this amino acid is over represented and leads to well-known low-complexity regions, whereas it does not in *Arabidopsis thaliana*. Still, Eqs. (4) and (7) allow the definition of the mutual-information I between two amino acids in two different set of sequences, defined as the reduction of the uncertainty on event j in the query sequence, gained by the knowledge of the occurrence of i in the subject sequence:

$$I_{X=i\to Y=j} = h_Y(j) - h_{Y/X}(j/i) \qquad (8)$$

Using the conditional probability theorem [25], which states that:

$$P_{X/Y}(X = i/Y = j)P_Y(j) = P_{Y/X}(Y = j/X = i)P_X(i) \qquad (9)$$

we can state that the mutual information is symmetric in respect to the amino acid occurrence event *and* to the sequence were this event occurs: $I_{X=i\to Y=j} = I_{Y=j\to X=i}$. This last expression is defined as:

$$I_{XY}(i;j) = I_{YX}(j;i) \qquad (10)$$

It is important to notice that, in general:

$$I_{XY}(i;j) \neq I_{XY}(j;i) \qquad (11)$$

and therefore that the mutual information between two amino acids *is not symmetric* when just permuting the amino acids and not sequences in the two terms of Eq. (11). Using Eqs. (5), (8), (9) and (10), we can now state that:

$$I_{XY}(i;j) = \log\left\{\frac{P_{XY}((X=i)\cap(Y=j))}{P_X(i)P_Y(j)}\right\} \qquad (12)$$

Eq. (12) can be estimated from observed homologous aligned sequences and allows computation of a substitution amino acid sequence with Eq. (6). This matrix is non-symmetric (Eq. (12)) and implementation in optimization alignment algorithm should therefore be carried out paying attention to the *query* and the *subject* sequence order. An important property of this matrix is that the application inverse (transformation of the *query* into a *subject*, and vice-versa) is done by the transposition of the scoring matrix. The matrix $S_{XY}(i,j)$ in this order (i is taken from the Species2 and is reading on the row) is called dirSp1Sp2.

3.2. Training sets to compute non-symmetrical substitution matrices

Determining sets of homologous sequences is a difficult question. Here, we examined the possibilities of an automatic or manual selection of a training set of pairs of similar sequences, in the given example of *Arabidopsis thaliana* and *Plasmodium falciparum* proteomes, obtained after an all-against-all comparison of non-redundant protein sequence databases. We selected genomic sequences from *A. thaliana* (25 545 annotated sequences as of December 2002) and *P. falciparum* (5334 annotated sequences as of December 2002). The strong nucleotidic bias of the *P. falciparum* genome (82% AT) strikingly affects the amino acid distribution within encoded proteins (Fig. 1). Six amino acids (N, K, I, L, E, and S) account for 51% of the total amino acid content in proteins. Fig. 1 shows that the amino acid distribution in *A. thaliana* is strongly divergent, with a more balanced contribution of individual amino acids to the overall composition of proteins. On top of the very strong amino acid bias found in *P. falciparum*, the protein sequences exhibit a very low complexity, marked by long stretches of repeated amino acids. It is still not known whether the very low complexity is solely due to the amino acid bias or if a generic mechanism dedicated to the insertions of amino acid repeats would contribute to this striking occurrence of repeated amino acid portions in proteins. For an in-depth discussion of the biological rational of *P. falciparum* bias as compared to *A. thaliana*, see [18]. We selected a training set of

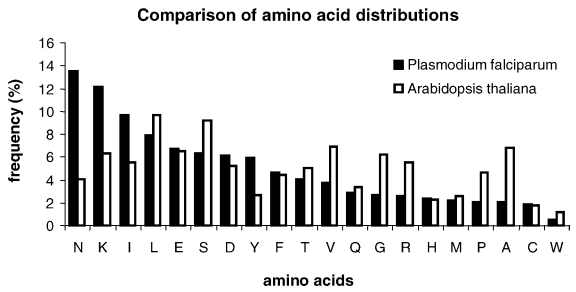


Fig. 1. Comparison of the amino acid distribution in the *Plasmodium falciparum* and *Arabidopsis thaliana* proteomes. Frequencies were calculated using the set of 71 pairs of homologous sequences selected with the method described in [17] from the two complete proteomes (see material and methods). Amino acids were ranked owing to their frequencies in *P. falciparum*.

similar sequences, which was either directly used for matrices' calculations (automatic sampling) or with the restriction that the analogy of the protein function could be assessed by inspection of an expert curator (manual sampling). The advantage of the manual training set, described in [18], lies in its quality, but it is costly. Although one might question the quality of the automatic sample, the case study in the present paper proved that no major difference between the converging matrices calculated from the manual and the automatic training sets could be noticed. Because a model generalization and a pragmatic computation of matrices would benefit mostly from a fully automated method, we therefore detailed results obtained in that context.

3.3. Convergence of the non-symmetric matrices obtained after iterative computation

Blocks used to calculate matrices were filtered as described in the Methods section, according to a clustering percentage. Matrices were termed dirSp1Sp2, with Sp1 being the query species, and Sp2 the subject species. The matrices devoted to the comparison of *A. thaliana* and *P. falciparum* are therefore called dirAtPf matrices. We generated matrices corresponding to eleven block clustering percentages (dirAtPf100_n for a clustering percentage = 100% and n iterations, dirAtPf90_n, …, dirAtPf50_n). We analysed the evolution of the matrices obtained after each iterative round (summarized in Fig. 2). Fig. 3 shows the number of amino acids that are initially aligned by the generalist matrix BLOSUM 62 in the training set and after alignments using matrices computed after 1, 4, 7 and 9 iterations. From the very first matrix computation, one notice a decrease in the number of aligned amino acid, with a very rapid convergence, as early as $\sim$ 10 iterations. Interestingly, in the present result, convergence appears as a decrease in the number of amino acids 'detected' by the non-symmetrical matrices. That decrease would suggest that, in the case of well-assessed alignments, the non-symmetrical matrices are more specific. By contrast, although they converge to a close result, matrices computed with a manual training set exhibit an increase in the number of aligned amino acid along the training process. That increase would conversely suggest that in the case of hypothetical alignments,

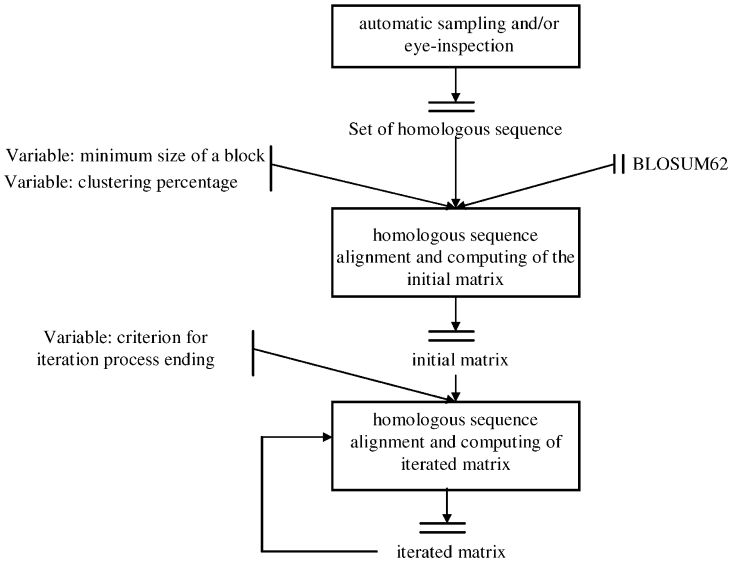


Fig. 2. dirSp1Sp2 matrices iterative computing Workflow. After sampling a set of pairs of homologous sequences between Species1 and Species2, these sequences were primarily aligned using BLOSUM 62 in order to determine conserved block. BLOSUM 62 is therefore used as a way to initially 'anchor' homologous regions. These blocks are considered only if they have both the required minimum size and a maximum given percentage of identity, named clustering percentage. The first deduced matrix, called *initial matrix*, is then used to iterate the process and lead to a sequence of dirSp1Sp2 matrices. A convergence criterion is applied on the sequence dirSp1Sp2$_n$ so as to end the process.

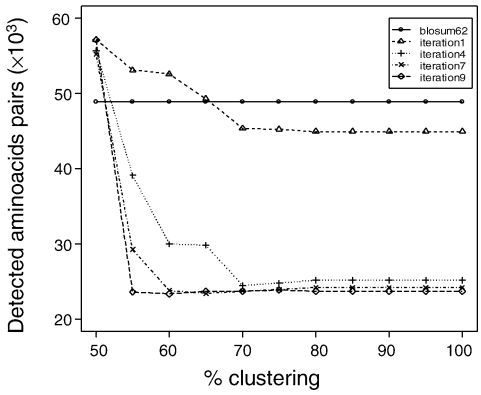


Fig. 3. Convergence of the matrices computing iterating process. Convergence of the iterating process was studied according to the number of detected aligned amino acids pairs. For all clustering percentage, convergence was also observed for the number of detected blocks and the value of the matrices (data not shown). Except for the 50% clustering percentage, convergence leads to a number of aligned amino acids pairs lower than that obtained with the BLOSUM 62 matrix. As described by Müller et al. [22], non-symmetric matrices lead to an increase of the discrimination power of the matrices, an expression of a more accurate mutual information values.

the non-symmetrical matrices would lead to a gain in sensitivity. Thus it appears that an apparent gain in selectivity and specificity would be obtained. To that extent, and as mentioned by Müller et al. [22], definition of a specificity/sensitivity gain when deriving substitution matrices is difficult to rigorously assess. The matrices obtained from automatic or manual samples exhibited identical trends for each s_{ij} terms: no opposite deviations were observed. For all clustering percentages, convergence was also observed for the number of detected blocks and the value of the matrices (data not shown).

3.4. *Asymmetry of the DirSp$_1$Sp$_2$ matrices: case study of DirAtPf*

We generated all the convergent dirAtPf matrices after a 10-iteration computation process dirAtPf100$_{10}$, dirAtPf90$_{10}$, …, dirAtPf50$_{10}$, indistinctly called dirAtPf100, dirAtPf90, …, dirAtPf50. All the matrices we obtained were non-symmetric, as shown in Fig. 4 for dirAtPf100. In detail, the sub-matrices $(\mathrm{N},\mathrm{R}) \times (\mathrm{N},\mathrm{R})$ giving the mutual information between all substitution available between Asparagines (N) and Arginines (R) stresses the different roles played by these two amino acids in the two proteomes,

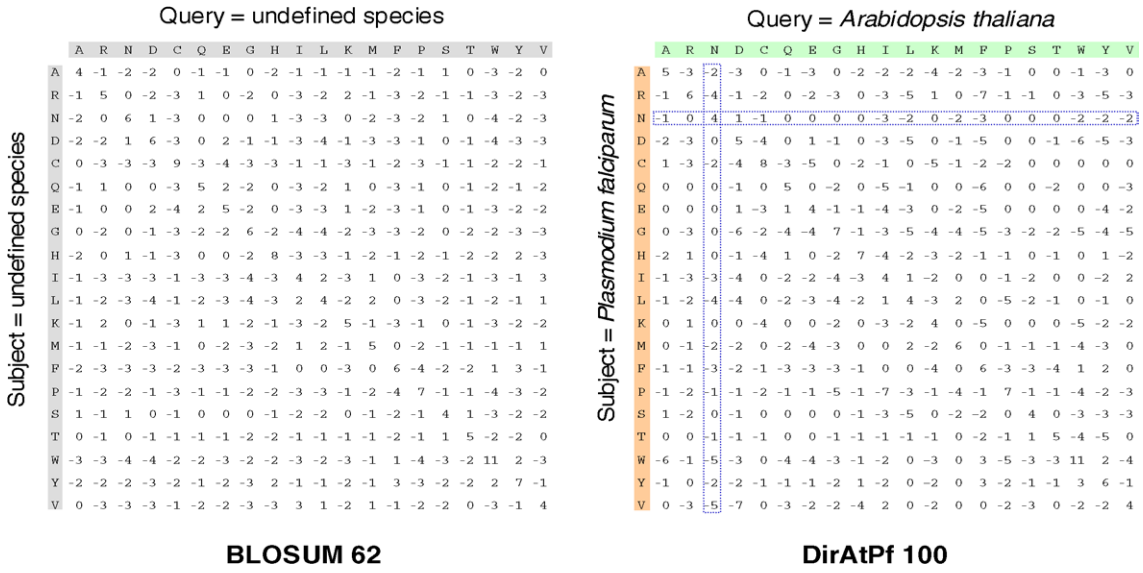


Fig. 4. BLOSUM 62 and dirAtPf100 substitution matrices.

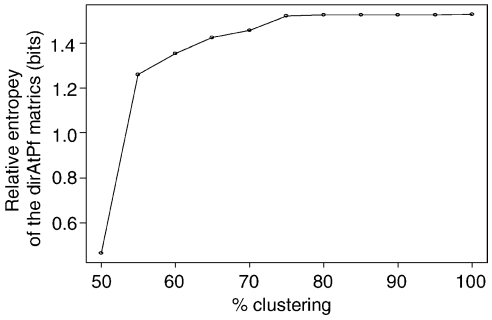


Fig. 5. Evolution of the relative Entropy as a function of the clustering percentage. As intuitively predicted by Müller et al. [22], the better definition of the mutual information between aligned amino acids leads to a higher relative entropy than this of BLOSUM 62 ($H \cong 0.69$).

as recorded by Singer and Hickey [20] and Bastien et al. [17].

3.5. Analysis of the relative entropy H of the family of DirAtPf matrices

As described earlier for the family of BLOSUM matrices [8], the relative entropy H derived from Eq. (2) decreases with the blocks clustering percentage (Fig. 5). Interestingly, relative entropy values in dirAtPf matrices (0.5–1.5 bits) are slightly higher than those of the BLOSUM or PAM matrices (0.2–1.2 bits) [7,8]. Following the theory of information, this higher relative entropy would suggest a higher sensitivity of dirAtPf matrices as compared to symmetrical matrices.

4. Conclusion

This article describes a method to compute a novel family of substitution matrices that are dedicated to the comparison of proteins, which amino acid composition deviates from the average distribution. They exhibit remarkable features such as (*i*) the possibility of computing reliable matrices from automatically selected pairs of similar sequences (automatic training sets), (*ii*) a rapidly convergent iterative process, and (*iii*) an increase in relative entropy. Still the selectivity/sensitivity gain is difficult to assess besides pragmatic use. Families of matrices for pairwise proteome comparisons including biased genomes such as that of *Plasmodium flaciparum* (AT rich) or *Chlamydomonas reinhardtii* (GC rich) are expected to be improved.

References

[1] F. Dardel, F. Képès, Bioinformatique: Génomique et postgénomique, Les Éditions de l'École polytechnique, Paris, 2002.

[2] M.S. Waterman, Introduction to Computational Biology, CRC Press, Boca Raton, FL, USA, 1995.

[3] S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol. 48 (1970) 443–453.
[4] J. Setubal, J. Meidanis, Introduction to Computational Molecular Biology, PWS Publishing Compagny, New York, 1997.
[5] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[6] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1981) 195–197.
[7] M.O. Dayhoff, R.M. Schwartz, B.C. Orcutt, A model of evolutionary change in proteins, Atlas of protein sequence and structure 5 (1978) 345–352.
[8] S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, Proc. Natl Acad. Sci. USA 89 (1992) 10915–10919.
[9] S. Henikoff, J.G. Henikoff, Performance evaluation of amino acid substitution matrices, Proteins 17 (1993) 49–61.
[10] J.L. Risler, M.O. Delorme, H. Delacroix, A. Henaut, Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix, J. Mol. Biol. 204 (1988) 1019–1029.
[11] S. Karlin, S.F. Altschul, Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, Proc. Natl Acad. Sci. USA 87 (1990) 2264–2268.
[12] S.F. Altschul, Amino acid substitution matrices from an information theoretic perspective, J. Mol. Biol. 219 (1991) 555–565.
[13] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423, 623–656.
[14] Y.K. Yu, J.C. Wootton, S.F. Altschul, The compositional adjustment of amino acid substitution matrices, Proc. Natl Acad. Sci. USA (2003) 15688–15693.
[15] R.V.L. Hartley, Transmission of Information, Bell Syst. Tech. J. 3 (1928) 535–564.
[16] O. Bastien, J.-C. Aude, K. Métayer, S. Roy, J.-J. Codani, É. Maréchal, Method for automatic pairwise alignment of protein sequences from biased and non-biased genomes: generalized model for substitution matrices and theoretical significance of Z-value statistics, in: Proc. Eur. Conf. on Computational Biology, Paris, France, 2003, pp. 525–526.
[17] O. Bastien, J.-C. Aude, S. Roy, É. Maréchal, Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of Z-value statistics, Bioinformatics 20 (2004) 534–537.
[18] O. Bastien, S. Lespinats, S. Roy, K. Metayer, B. Fertil, J.-J. Codani, É. Maréchal, Analysis of the compositional biases in *Plasmodium falciparum* genome and proteome using *Arabidopsis thaliana* as a reference, Gene 336 (2004) 163–173.
[19] A.K. Chamberlain, J.U. Bowie, Asymmetric amino acid compositions of transmembrane beta-strands, Protein Sci. 13 (2004) 2270–2274.
[20] G.A. Singer, D.A. Hickey, Nucleotide bias causes a genomewide bias in the amino acid composition of proteins, Mol. Biol. Evol. 17 (2000) 1581–1588.
[21] G.E. Tusnady, I. Simon, Principles governing amino acid composition of integral membrane proteins: application to topology prediction, J. Mol. Biol. 283 (1998) 489–506.
[22] T. Müller, S. Rahmann, M. Rehmsmeier, Non-symmetric score matrices and the detection of homologous transmembrane proteins, Bioinformatics 17 (Suppl. 1) (2001) S182–S189.
[23] J.-J. Codani, J.-P. Comet, J.-C. Aude, E. Glémet, A. Wozniak, J.-L. Risler, A. Hénaut, P.P. Slonimski, Automatic analysis of large-scale pairwize alignments of protein sequences, Methods Microbiol. 28 (1999) 229–244.
[24] S. Henikoff, J.G. Henikoff, Automated assembly of protein blocks for database searching, Nucleic Acids Res. 19 (1991) 6565–6572.
[25] A.J. Valleron, Introduction à la biostatistique, Masson, Paris, 1998.

Conclusions et perspectives

Conclusions et perspectives

Le travail présenté dans ce mémoire est né des interrogations suscitées par 1) la difficulté, voire l'incapacité, d'identifier chez *Plasmodium falciparum* certains gènes, responsables de fonctions mesurées biochimiquement chez le parasite malarial, par similarité avec des séquences homologues caractérisées dans d'autres organismes (Maréchal et al., 2001; McConkey et al., 2004; Callebaut et al., 2005) et 2) la relative faiblesse de l'annotation du génome de *Plasmodium falciparum*, avec seulement ~60 % de gènes sans fonction attribuée (Gardner et al., 2002). Cette difficulté rencontrée lors des recherches automatiques d'homologie est une limite à tout projet exploratoire du génome malarial fondé sur la phylogénie moléculaire. En particulier, l'étude des séquences héritées de l'algue ancestrale, qui a réalisé l'endosymbiose secondaire à l'origine du phylum des Apicomplexes (voir Introduction), peut être rendue incomplète. Les caractéristiques atypiques du génome et du protéome de *Plasmodium falciparum*, résumées sous le terme de biais compositionnel, ont été soupçonnées d'être un cas limite pour les outils d'analyse de séquence existants. L'objet de cette thèse a donc été d'examiner l'influence possible de ce type de biais sur les méthodologies de comparaisons de séquences et de façon plus approfondie sur leurs statistiques.

Nous avons proposé des développements théoriques associés à la statistique de la *Z-value* introduite par Lipman et Pearson (1985), et notamment

- le théorème TULIP (Article 1) permettant de déduire un majorant de la probabilité d'un score d'alignement de séquences (*i.e.* la *P-value*) par la valeur $1/Z\text{-}value^2$;
- la projection des protéines dans un espace de configuration des protéines homologues, permettant d'une part une expression du théorème TULIP et ayant d'autre part une cohérence avec la phylogénie moléculaire (Articles 2 et 3);
- la déduction des propriétés remarquables de la distribution des *Z-values*, et un raffinement du calcul de la *P-value*, à partir de quelques hypothèses sur l'évolution des protéines selon un modèle de vieillissement, dans le contexte de la théorie de la fiabilité des systèmes (Article 4).

Ces travaux ont permis d'étayer la relation entre les scores d'alignements et l'information mutuelle, au sens de la théorie de l'information, partagée par des séquences ayant un ancêtre commun et étant soumises à une pression évolutive similaire.

Ces développements théoriques ont de plus permis certaines avancées sur le plan pratique de l'identification de séquences homologues initialement non détectées par le théorème de Karlin-Altchul (avec les seuils de détection couramment admis, Karlin et Alstchul, 1990). Par exemple il a été possible d'assigner une fiabilité statistique à l'alignement du facteur de transcription TFII gamma de *Plasmodium falciparum* avec la séquence homologue d'*Arabidopsis thaliana* (voir Tableau de l'Article 4) alors que le modèle de Karlin-Altschul ne soutenait pas cette homologie (Callebaut et al., 2005). Nous pouvons aussi reprendre l'exemple abordé dans l'Article 1 visant l'identification d'une séquence homologue

de celle de TIC22 d'*Arabidopsis thaliana*, une protéine typiquement végétale impliquée dans les processus d'import de protéines vers le chloroplaste (voir Introduction), chez *Plasmodium falciparum*. Une protéine malariale avait été identifiée comme ayant une forte *Z-value* (Tableau 4) avec celle d'*Arabidopsis* dans un cluster provenant d'une comparaison Smith-Waterman tout-contre-tout en utilisant la matrice BLOSUM62. La valeur de la *Z-value* était de 15.6, ce qui correspond à un majorant de la *P-value* de $4.1.10^{-3}$ selon le théorème TULIP, une valeur qui est de l'ordre de grandeur de la *P-value* retournée selon le modèle de Karlin-Altschul après une recherche prenant en requête TIC22 d'*Arabidopsis*, dans la base de données PlasmoDB (The Plasmodium Genome Database Collaborative, 2001) selon la méthode Blastp. Toutefois, une *Z-value* de l'ordre de 15 est considérée comme significative, alors qu'une *P-value* de 1.10^{-3} (même de l'ordre de 2.10^{-5}) retournée après une comparaison suivant la méthode Blastp et en utilisant le théorème de Karlin-Altchul est souvent en dessous des seuils de significativité retenus pour les comparaisons de génome (Tekaia et al., 1999; Bastien et al., 2004). Cette ambiguïté apparente entre les statistiques selon les modèles de Karlin-Altschul et de Lipman-Perason peut être expliquée par les résultats de l'Article 4 qui énonce comment la *P-value* varie avec la *Z-value*. Ces résultats permettent de calculer une valeur plus fine de la *P-value*, de l'ordre de 10^{-9} avec BLOSUM62, hautement significative (Tableau 4).

Méthode d'alignement	**Blastp**		**Smith-Waterman**
Matrice de substitution	BLOSUM62		BLOSUM62
	avec filtre	sans filtre	
Statistiques			
P-value (Karlin-Altchul)	1.10^{-3}	2.10^{-5}	na
Z-value (Lipman-Pearson)	-	-	15.6
T-value (théorème TULIP)	-	-	$4.1.10^{-3}$
P-value (ce travail)	-	-	1.210^{-9}

Tableau 4: Statistiques des scores d'alignement des séquences de TIC22 d'*Arabidopsis thaliana* et de son homologue probable chez *Plasmodium falciparum*. Les séquences de *Plasmodium* (PlasmoDB PFE1460W) et d'*Arabidopsis* (Tair At3g23710.1) ont été alignées selon les méthodes Blastp et Smith-Waterman. Les statistiques de scores d'alignement ont été calculées selon le modèle de Karlin-Altschul, comme implémenté dans l'algorithme Blastp, ou selon le modèle de la *Z-value* de Lipman-Pearson. Le majorant de la *P-value* basé sur le théorème TULIP est donné selon la formule: $T\text{-value} = 1/Z\text{-value}^2$. La *P-value* est déduite de la distribution des *Z-values* selon une loi de Gumbel décrite dans l'Article 4. na: ne s'applique pas.

Ces résultats sont confirmés par l'alignement multiple des TIC22 d'*Arabidopsis thaliana*, *Oryza sativa*, *Pisum sativum* avec la séquence identifiée chez *Plasmodium falciparum* (Figure 26). En effet, les domaines et acides aminés conservés sur cet alignement correspondent à ceux que l'on peut observer dans le famille de domaine PRODOM PD212293 (Corpet et al., 2000), construite à partir de 11 domaine de TIC22, de 11 séquences différentes appartenant à 7 espèces différentes, dont aucune n'est issue de *Plasmodium falciparum*.

```
Arabidopsis     MNSNIFPPSKQQNELNNIQQSFSNLQSQCSNLLLNVSQTLNPLFNANTNNNKPNIFSALN
ORYZA           ----------------------MPFHFQFPWLPNNPTSS----SSSPTKPPSPAIPNPFL
PISUM           -----------------------MESQGQW---NPLLS----FSRFINHHSNHLATRLE
Plasmodium      ------------MCLLLFICLYFA----RGIYCLKTLNG----LSRDINN-SIYLRNNVH
                                         :         .   :  .  : . .

Arabidopsis     SFRDQAKQALDSRISRFNSGKAPVWARISDDGGGARAQVTVPIRGSGKGLSADAIEERLA
ORYZA           PIQAGLASFLSSLPLPRAAFPPPPWARISSASASAASASALPV---------AEIEERLA
PISUM           ETKRLA-----GTLIQSHTRTKPAF------------AATLTP---------NHVAKSLA
Plasmodium      KKKRLV---DCNLCMLKHKFRLSFWKK------------RYDE---------RPIEEKLE
                  :        .          . :                              : : *

Arabidopsis     GVPVYALSNSNEEFVLVSGTSSG-------------KSLGLLFCKEEDAETLLKEMKSMD
ORYZA           GVPVYALANSSQEFVLVSSARGGGGGGGGARAAVPPPALGLLCFRREDADALLAQMDG--
PISUM           GTSVYTVSNSDNEFVLMSDAEGA-------------KSIGLLCFRQEDAEAFLAQVRSRK
Plasmodium      VIPVFLITNYNSSPYIFQENE---------------KQVCYMFLCPYDAENMLNDMIKYN
                  .*: ::* ...  :..                    :  :     **: :* ::

Arabidopsis     PRMRKEGSKVVALALSKVFQL------------------KVNGVAFRLIPESTQVKNALK
ORYZA           -DM-AAGSTVVPVALNKVIQL------------------KSDGVAFRFVPDSSQVANAMK
PISUM           KEF-RGGAKVVPITLDQVYML------------------KVEGIAFRFLPDPVQIKNALE
Plasmodium      GMKYNGNIKIHNITMKKAYELMKEFLQLEKMEVNKEDSKKKQNIYWKLISSKRQLQNALY
                      . .:  :::.:.  *                  * :.: :::::..  *: **:

Arabidopsis     ERKTAG-IDDDDFHGVPVFQSKSLILRSENMSYRPVFFRKEDLEKSLIRASSQQNRLNPA
ORYZA           LMENEGQYVNDGFPGVPVFQSRSLVLMSDNKRYRPVFFRKEDLDNSLHRASRDQQKPNPA
PISUM           LRAA----NRGSFDGVPVFQSDLLVVKKKNKRYCPVYFSKEDLEYELSKVSRSSKGVGVS
Plasmodium      YLSFT----KKSELMYPVFYAENLYIQKDGSNIIPLFFDLEDLKEAI------EEQKNKA
                           .    *** :  * : ...    *::*  ***.  :      .:  . :

Arabidopsis     LKPGDIQVAVFEDIVKGMREST-TSNWDDIVFIPPGFE----VSTEQTQE----------
ORYZA           VKMGDIQVSSLENIIKSMKDSS-SSKWDDAVFIPPGFD----LATSSKQS-----NHDN-
PISUM           ---QHIMVGSFEDVLKKMELSEKSSGWEDLVFIPPGKK----HSQHMQEV-----IA---
Plasmodium      LSKVDYKIKVL--------------NMVDLIFTEDHKKFGFVPSTQSVKYLDKLNIGTKK
                    .  :  :                * :*     .     :    :

Arabidopsis     ---
ORYZA           ---
PISUM           ---
Plasmodium      TYF
```

Figure 26: Alignement multiple des TIC22 d'*Arabidopsis thaliana*, *Oryza Sativa*, *Pisum sativum* avec la séquence de *Plasmodium falciparum* identifiée comme potentiellement homologue. L'alignement des TIC22 d'*Arabidopsis thaliana* (AT3G23710.1), *Oryza sativa* (GI|50936021), *Pisum sativum* (GI|3769671) et de la séquence de *Plasmodium falciparum* identifiée dans l'Article 1 (PFE1460w) a été réalisé avec T-COFFEE (Notredame et al., 2000). L'alignement est affiché au format CLUSTAL.

Les outils de prédiction d'adressage chez *Plasmodium falciparum* tel que PATS (Zuegge et al., 2001) et PlasmoAP (Foth et al., 2003) prédisent un adressage de la séquence homologue de TIC22 vers l'apicoplaste. Ce type d'exemple illustre le gain apporté par les méthodes statistiques dérivées de la *Z-value*, décrites dans ce manuscrit, dans le cas limite de l'analyse comparative de génomes *a priori* biaisés en composition et en taille.

En construisant un espace de configuration des protéines homologues, permettant une expression du théorème TULIP et ayant une cohérence avec la théorie synthétique de l'évolution (Articles 2 et 3), nous avons déduit une méthode de reconstruction de phylogénies de séquences protéiques à l'aide des *Z-values*. Les phylogénies moléculaires reconstruites par cette méthode sont concordantes avec celles obtenues à partir d'alignements multiples (voir Articles 2). Il est par ailleurs possible de résoudre dans ce mode de représentation certaines incohérences phylogénétiques rapportées précédemment. Par exemple Keeling et Palmer (2001) ont montré que la reconstruction de la phylogénie des énolases à l'aide des méthodes conventionnelles aboutissait à la séparation des énolases des apicomplexes et de leurs homologues de plantes, malgré la présence d'insertions, conservées uniquement pour les énolases de ces groupes. La reconstruction phylogénétique que nous avons obtenue résout

cette apparente contradiction et réconcilie la conservation des insertions avec la proximité phylogénétique (voir Articles 2).

En prenant en compte le modèle statistique que nous avons élaboré (en particulier le théorème TULIP), nous avons entrepris une première analyse de l'évolution du biais en acides aminés chez *Plasmodium falciparum* 1) corrélativement à l'évolution du biais en acides nucléiques dans le génome malarial et 2) en fonction de la divergence évolutive, établie en prenant le génome non biaisé d'*Arabidopsis thaliana* comme référence (Article 5). Pour cette étude, nous avons sélectionné 71 couples de protéines homologues, retenues pour leur haut niveau de similitude structurale et pour la concordance de leurs annotations. Nous avons observé que le biais des séquences malariales était corrélé au pourcentage de divergence avec leurs homologues végétaux. Nos analyses suggèrent de plus que le biais est vraisemblablement la conséquence d'une évolution au niveau nucléique. Cette conclusion a depuis été confirmée par Chanda et al. (2005). Ces auteurs ont par ailleurs montré que le niveau d'expression des gènes était inversement corrélé au pourcentage en A+T. Les gènes hautement exprimés sont donc enrichis en G, A, R et P (classe de résidus appelée "GARP") et appauvris en F, Y, M, I, N et K (classe de résidus appelée "FYMINK") relativement aux autres gènes.

Une conséquence du biais protéique, caractérisé par le rapport GARP/FYMINK, est la nécessité de prendre en compte les dissymétries de composition lors de la comparaison de séquences de *Plasmodium falciparum* avec celles d'organismes différents. Nous avons examiné la possibilité de construire une famille de matrices tenant compte de cette dissymétrie dans le cas de la comparaison de *Plasmodium falciparum* et d'*Arabidopsis thaliana* (Article 6). Ces matrices appelées DirAtPf, possèdent une sensibilité théorique supérieure aux autres familles de matrices existantes (selon l'évaluation de l'entropie relative *HR*, voir Rappels Bibliographiques).

Est-ce que le gain de sensibilité des matrices DirAtPf s'accompagne d'une perte de spécificité ? Pour aborder cette question difficile, nous avons effectué une comparaison des *Z-values* calculées avec une matrice DirAtPf en fonction de celles calculées à l'aide de BLOSUM62, pour un ensemble de couples de séquences homologues d'*Arabidopsis thaliana* et de *Plasmodium falciparum*. Nous avons sélectionné pour cette étude les 71 couples de protéines homologues validées pour l'Article 5, dont l'homologie était détectée par les méthodes conventionnelles, ainsi que les facteurs de transcriptions non détectés par recherche d'homologie classique (Callebaut et al., 2005, et Article 4). Une première série de calculs présentée sur la Figure 27 montre une corrélation entre les *Z-values* correspondant aux scores d'alignements réalisés avec DirAtPf100 et BLOSUM62. Les valeurs obtenues avec DirAtPf100 sont plus faibles, avec des alignements de meilleure qualité, et reflètent la non prise en compte d'appariements de résidus non pertinents, observés lorsque la comparaison est réalisée à l'aide de BLOSUM62. Ce résultat suggère un gain de spécificité.

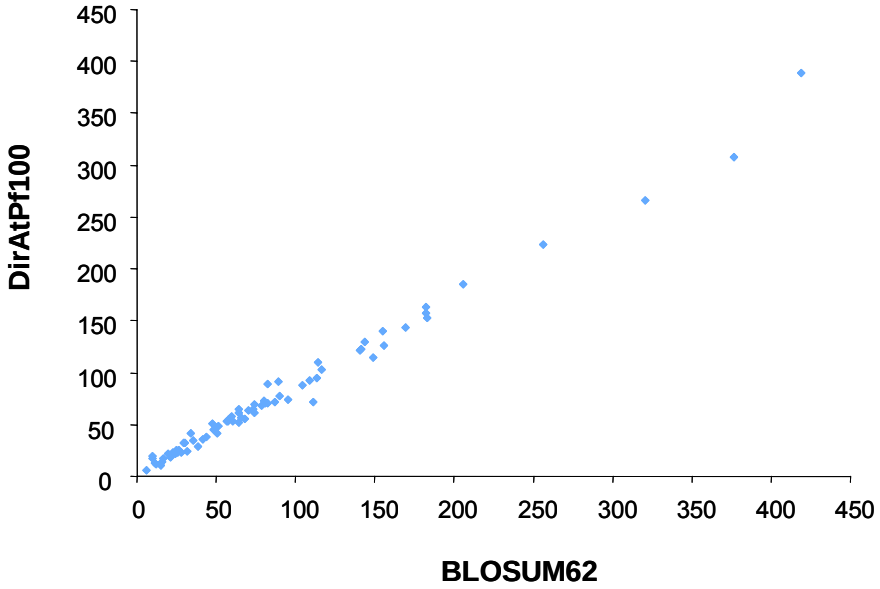


Figure 27: Comparaison des *Z-values* calculées pour les alignements d'un jeu de couples de séquences homologues de *Plasmodium falciparum* et d'*Arabidopsis thaliana*, réalisés avec les matrices DirAtPf100 et BLOSUM62 et selon la méthode Smith-Waterman.

L'obstacle majeur à l'exploitation du modèle statistique généralisé que nous avons développé, reposant sur l'usage de l'algorithme exact Smith-Waterman et sur des simulations de Monte-Carlo, est le temps de calcul imposé par ces méthodes (voir Article 3). Pour pallier cette difficulté, nous avons développé en collaboration avec Philippe Ortet (CEA-Cadarache) une méthode de calcul de la *Z-value* utilisant comme algorithme d'alignement de séquences la méthode Blastp (Figure 28)

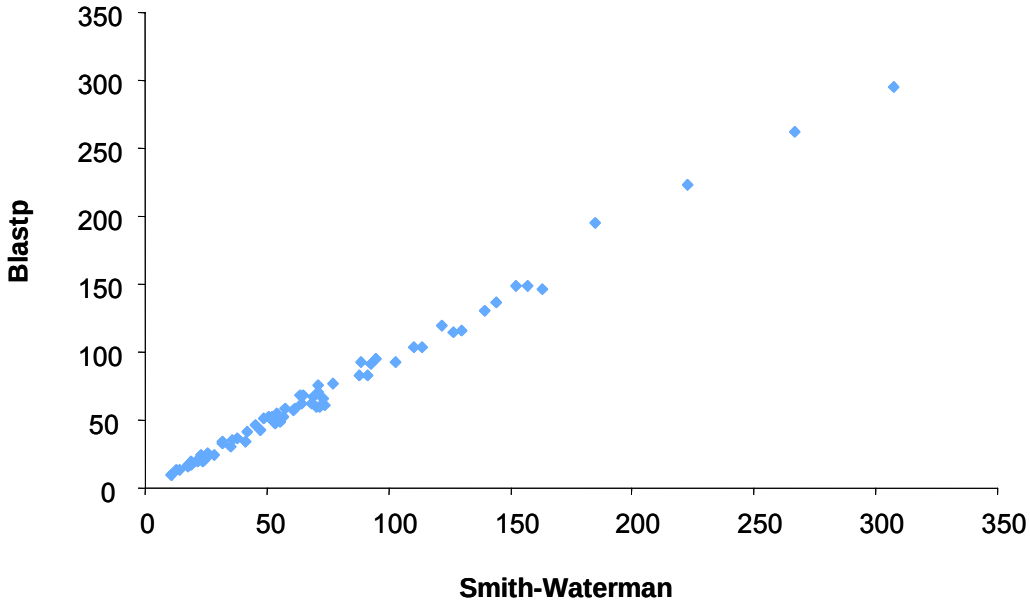


Figure 28: Comparaison des *Z-values* calculées pour les alignements d'un jeu de couples de séquences homologues de *Plasmodium falciparum* et d'*Arabidopsis thaliana*, réalisés avec la matrice BLOSUM62 et selon les méthodes d'alignement Blastp et Smith-Waterman.

Les perspectives du travail de thèse présenté dans ce mémoire incluent tout d'abord un volet théorique qui est l'optimisation du modèle de génération des séquences aléatoires sous-jacent à la simulation de la distribution des scores. Un point faible de la méthode de calcul des

Z-values à partir de simulations de Monte-Carlo est en effet, d'une part, la non prise en compte dans le modèle de génération du jeu de séquences aléatoires (par permutation des résidus) de la co-évolution possible de certains résidus à l'intérieur d'une séquence. Une perspective importante est donc de développer un modèle de randomisation reposant sur l'idée qu'il est possible d'accéder à l'information mutuelle entre résidus co-évoluant, à l'aide d'alignements multiples de séquences. D'autre part, pour l'usage d'algorithmes tels que Blastp, il sera important d'examiner l'effet du modèle de permutation sur la capacité de Blastp à identifier de courts segments d'ancrage d'alignement, et d'évaluer si la méthode de permutation doit être corrigée en conséquence.

Un des objectifs de ce travail était de développer des méthodes pour l'analyse du génome malarial en vue d'exploiter sa phylogénie singulière, en particulier la présence d'un sous génome hérité d'une algue suite à un épisode endosymbiotique secondaire, dans lequel rechercher des cibles pour des médicaments de type herbicide. Il sera nécessaire pour aboutir à cet inventaire des séquences de type végétal, d'établir la phylogénie de l'ensemble des gènes par comparaison avec plusieurs autres organismes, dont des plantes (*Arabidopsis thaliana*), mais aussi des algues dont le génome a été récemment séquencé (en particulier l'algue rouge *Cyanidioschyzon merolae*), ainsi que l'espèce humaine. La méthode de calcul de la *Z-value* utilisant comme algorithme d'alignement de séquences la méthode Blastp, rend possible une telle analyse comparative massive de *Plasmodium falciparum*, fondée sur les méthodes statistiques robustes décrites dans ce mémoire.

Publications

Publications

Publications réalisées dans le cadre de cette thèse

2006 **Bastien, O. et Maréchal, E.:** "System aging and longevity as a model of protein sequence evolution: derivation of remarkable properties of sequence alignment statistics". **Soumis**

2006 **Bastien, O., Ortet, P., Roy, S. et Maréchal, E**.: "The configuration space of homologous proteins: a theoretical and practical framework to reduce the diversity of the protein sequence space after massive all-by-all sequence comparisons". ***Future Generation Comput. Syst.*, In press**

2005 **Bastien O., Roy S. et Maréchal E.: "**Construction of non-symmetric substitution matrices derived from proteomes with biased amino acid distributions". ***C.R. Biol.*** 328 (5): 445-453.

2005 **Bastien O., Ortet P., Roy, S. et Maréchal E**.: "A configuration space of homologous proteins, conserving mutual information and allowing a phylogeny inference based on pair-wise Z-score probabilities". ***BMC Bioinformatics*** 6(1): 49.

2004 **Bastien O., Aude J.C., Roy S. et Maréchal E. :** "Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of Z-value statistics″. ***Bioinformatics*** 20(4): 534-7.

2004 **Bastien O., Lespinats S., Roy S., Métayer K., Fertil B., Codani J.J. et Maréchal E.**: ″Analysis of the compositional biases in *Plasmodium falciparum* genome and proteome using *Arabidopsis thaliana* as a reference″. ***Gene*** 336(2): 163-73.

Publications réalisées dans le cadre de collaborations

2006 **Barakat, M., Ortet, P., De Luca, G., Jourlin, C., Ansaldi, M., Py, B., Fichant, G., Coutinho, P.M., Voulhoux, R., Bastien, O., Roy, S., Maréchal, E., Henrissat, B., Quentin, Y., Noirot, P., Filloux, A., Méjean, V., DuBow, M., Barras, F. et Heulin., T. :** "Genome of the cyst-dividing bacterium Ramlibacter tataouinensis". **Soumis**.

2006 **Le Lay, P., Isaure, M.P., Sarry, J.E., Khun, L., Fayard, B., Le Bail, J.L., Bastien, O., Garin, J., Roby, C., et Bourguignon, J. :** "Metabolomic and proteomic analyses of Arabidopsis thaliana cells exposed to a caesium stress. Influence of potassium supply." ***Biochimie*, In press**.

2006 **Rivasseau, C., Couram, G., Boisson, A.M., Bastien, O. et Bligny, R.** "Capillary electrophoretic analysis of organic acids and phosphorated compounds (nucleotides) in plants and study of these metabolites during metal stress". **Soumis.**

2006 **Sarry J.E., Kuhn L., Ducruix C., Lafaye A., Junot C., Hugouvieux V., Jourdain A., Bastien O., Vailhen D.,**

Amekraz B., Moulin C., Ezan E., Garin J. et Bourguignon J. : "The early responses of Arabidopsis thaliana cells to Cd exposure explored by protein and metabolite profiling analyses***". Proteomics***. 6:2180-2198

2006 **Bisanz C., Bastien O., Grando D., Jouhet J. , Maréchal E. et Cesbron-Delauw M.F. :** "*Toxoplasma gondii* acyl-lipid metabolism: de novo synthesis from apicoplast generated fatty acids versus scavenging of host cell". ***Biochem. J.*** 394 (1) 197-206***.***

2005 **Botté, C., Jeanneau, C., Snajdrova, L., Bastien, O., Imberty, A., Breton, C. et Maréchal, E.:** "Molecular modelling and site directed mutagenesis of plant chloroplast MGDG synthase reveal critical residues for activity". ***J. Biol. Chem.*** 280 (41): 34691-34701.

Chapitres d'ouvrages réalisées en collaborations

2006 **Block, M.A., Jouhet, J., Marechal, E. Bastien, O. et Joyard, J.:** "Role of the envelope membranes in chloroplast glycerolipid biosynthesis**"**. ***Photosynthesis: A Comprehensive Treatise Biochemistry, Biophysics and Molecular Biology.*** **In press**.

2006 **Bisanz, C., Botté, C., Saïdani, N, Bastien, O., Cesbron-Delauw, M.F., et Maréchal.:** "Structure, function and biogenesis of the secondary plastid of apicomplexan parasites**"**. ***Current Research in Plant Cell Compartments.*** **In press**.

Bibliographie

Bibliographie

A

Adl, S.M., Simpson, A.G., Farmer, M.A., Andersen, R.A., Anderson, O.R., Barta, J.R., Bowser, S.S., Brugerolle, G., Fensome, R.A., Fredericq, S., James, T.Y., Karpov, S., Kugrens, P., Krug, J., Lane, C.E., Lewis, L.A., Lodge, J., Lynn, D.H., Mann, D.G., McCourt, R.M., Mendoza, L., Moestrup, O., Mozley-Standridge, S.E., Nerad, T.A., Shearer, C.A., Smirnov, A.V., Spiegel, F.W. et Taylor, M.F. (2005). The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J Eukaryot Microbiol* 52:399-451.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., et Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol* 215: 403-410.

Altschul, S.F. (1993). A protein alignment scoring system sensitive at all evolutionary distances. *J Mol Evol* 36: 290-300.

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

Altschul, S.F., Bundschuh, R., Olsen, R., et Hwa, T. (2001). The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res* 29: 351-361.

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. *Science* 181:223-230.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408: 796-815.

Archibald, J.M. et Keeling, P.J. (2002). Recycled plastids: a ‘green movement’ in eukaryotic evolution. *Trends Genet* 18: 577-584.

Arnold, R., Rattei, T., Tischler, P., Truong, M.D., Stumpflen, V. et Mewes, W. (2005). SIMAP - The similarity matrix of proteins. *Bioinformatics* 21(Suppl. 2):42-46.

Arnold, V.I. (1989). Mathematical methods of classical mechanics. (New York: Springer).

Aude, J.C. (1999). Analyse de génomes microbiens, apports de la classification pyramidale. Thèse Université de Paris-Dauphine.

Aude, J.C., et Louis, A. (2002). An incremental algorithm for Z-value computations. *Comput Chem* 26: 403-411.

B

Bacro, J.N., et Comet, J.P. (2001). Sequence alignment: an approximation law for the Z-value with applications to databank scanning. *Comput Chem* 25: 401-410.

Baird, J.K. (2005). Effectiveness of antimalarial drugs. *N Engl J Med* 352:1565-1577.

Bastien, O., Lespinats, S., Roy, S., Metayer, K., Fertil, B., Codani, J.J., et Marechal, E. (2004). Analysis of the compositional biases in Plasmodium falciparum genome and proteome using Arabidopsis thaliana as a reference. *Gene* 336: 163-173.

Bastien, O., Roy, S., et Maréchal, E. (2005a). Construction of non-symmetric substitution matrices derived from proteomes with biased amino acid distributions. *C R Biol* 328: 445-453.

Bastien, O., Ortet, P., Roy, S., et Maréchal, E. (2005b). A configuration space of homologous proteins conserving mutual information and allowing a phylogeny inference based on pair-wise Z-score probabilities. *BMC Bioinformatics* 6: 49.

Bedhomme, M., Hoffmann, M., McCarthy, E.A., Gambonnet, B., Moran, R.G., Rebeille, F., et Ravanel, S. (2005). Folate metabolism in plants: an Arabidopsis homolog of the mammalian mitochondrial folate transporter mediates folate import into chloroplasts. *J Biol Chem* 280: 34823-34831.

Benner, S.A., Cohen, M.A., et Gonnet, G.H. (1994). Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng* 7 : 1323-1332.

Bienaymé, I.J. (1853). Considérations à l'appui de la découverte de Laplace sur la loi de probabilité dans la méthode des moindres carrés. *C R Acad Sci Paris* 37: 309-324.

Botte, C., Jeanneau, C., Snajdrova, L., Bastien, O., Imberty, A., Breton, C., et Maréchal, E. (2005). Molecular modeling and site-directed mutagenesis of plant chloroplast monogalactosyldiacylglycerol synthase reveal critical residues for activity. *J Biol Chem* 280: 34691-34701.

Bisanz C., Bastien O., Grando D., Jouhet J. , Maréchal E. et Cesbron-Delauw M.F. (2006) *Toxoplasma gondii* acyl-lipid metabolism: de novo synthesis from apicoplast generated fatty acids versus scavenging of host cell. *Biochem J* 394 (1): 197-206**.**

Bray, P.G., Ward, S.A. et O'Neill, P.M. (2005). Quinolines and artemisinin: chemistry, biology and history. *Curr Top Microbiol Immunol* 295:3-38.

Brierley, R. (2005). Roll Back Malaria issue first global report. *Lancet Infect Dis* 5:332-3.

Brocchieri, L. (2001). Low-complexity regions in Plasmodium proteins: in search of a function. *Genome Res* **11:** 195-197.

Brocchieri, L. (2001). Phylogenetic inferences from molecular sequences: review and critique. *Theor Popul Biol* 59 : 27-40.

Bütschli, 0. (1880-1889). Protozoa. *In* Klassen und Ordnung des Thier-Reichs. Vol. 1, C.F. Winter, Leibzig. pp. 1-2036.

C

Callebaut, I., Prat, K., Meurice, E., Mornon, J.P. et Tomavo S. (2005). Prediction of the general transcription factors associated with RNA polymerase II in Plasmodium falciparum: conserved features and differences relative to other eukaryotes. *BMC Genomics* 6:100.

Cavalier-Smith, T. (1998). A revised six-kingdom of life. *Biol Rev Can Philos Soc* 73:203-266.

Chance, M., Warhurst, D., Baggaley V. et Peters W. (1972). Preparation and characterisation of DNA from rodent malarias. *Trans R Soc Trop Med Hyg* 66: 3-4.

Chanda, I., Pan, A. et Dutta, C. (2005). Proteome composition in Plasmodium falciparum: Higher usage of GC-rich nonsynonymous codons in highly expressed genes. *J Mol Evol* 61: 513-523.

Chebyshev,P.L. (1867) Des valeurs moyennes. *Liouville's J. Math Pures Appl* 12: 177-184.

Cheng, Z., Ventura, M., She, X., Khaitovich, P., Graves, T., Osoegawa, K., Church, D., DeJong, P., Wilson, R.K., Paabo, S., Rocchi, M., et Eichler, E.E. (2005). A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* 437: 88-93.

Ciarlet, P.G. (1998). Introduction à l'analyse numérique matricielle and à l'optimisation. (Paris: Dunod).

Coles, S. (2001). An introduction to Statistical Modeling of Extreme Values. (New-York: Springer-Verlag).

Comet, J.P., Aude, J.C., Glemet, E., Risler, J.L., Henaut, A., Slonimski, P.P., et Codani, J.J. (1999). Significance of Z-value statistics of Smith-Waterman scores for protein alignments. Comput Chem 23: 317-331.

Copeland, H.F. (1947). Progress report on basic classification. *Amer Nat* 81:340-361.

Corpet, F., Servant, F., Gouzy, J. et Kahn, D. (2000). ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res* 28:267-269

Cover, T.M., et Thomas, J.A. (1991). Elements of information theory. (New York ; Chichester: Wiley).

D

Dayhoff, M.O., Schwartz, R.M., et Orcutt, B.C. (1978). A Model of Evolutionary Change in Proteins. *Atlas of Protein Sequence and Structure* 5: 345-352.

Desowitz, R. (1991). The Malaria Capers (More Tales of Parasites and People, Research and Reality). W.W. Norton & Company eds., New York.

Dobzhansky, T. (1974). Studies in the Philosophy of Biology: Reduction and Related Problems. (University of California Press).

E

F

Farooq, U. et Mahajan, R.C. (2004). Drug resistance in malaria. *J Vector Borne Dis* 41:45-53.

Foata, D., et Fuchs, A. (1998). Calcul des probabilités. (Paris: Dunod).

Felsenstein, J. (1996). Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol* 266: 418-427.

Féménias, J.L. (2003). Probabilités and Statistiques pour les Sciences Physiques. (Paris: Dunod).

Fichera, M.E. et Roos, D.S. (1997). A plastid organelle as a drug target in apicomplexan parasites. *Nature* 390:407-409.

Fitch, W.M. (1983). Random sequences. *J Mol Biol* 163: 171-176.

Foerstner, K.U., von Mering, C., Hooper, S.D., et Bork, P. (2005). Environments shape the nucleotide composition of genomes. *EMBO Rep* 6: 1208-1213.

Foster, P.G., Jermiin, L.S., et Hickey, D.A. (1997). Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J Mol Evol* 44: 282-288.

Foth, B.J., Ralph, S.A., Tonkin, C.J., Struck, N.S., Fraunholz, M., Roos, D.S., Cowman, A.F. et McFadden G.I. (2003). Dissecting apicoplast targeting in the malaria parasite Plasmodium falciparum. *Science* 299:705-708.

Gardner, M.J., Feagin, J.E., Moore, D.J., Spencer, D.F., Gray, M.W., Williamson, D.H. et Wilson R.J. (1991). Organization and expression, of small subunit ribosomal RNA genes encoded by a 35-kilobase circular DNA in *Plasmodium falciparum*. *Mol biochem Parasitol* 48:77-88.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., Paulsen, I.T., James, K., Eisen, J.A., Rutherford, K., Salzberg, S.L., Craig, A., Kyes, S., Chan, M.S., Nene, V., Shallom, S.J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M.W., Vaidya, A.B., Martin, D.M., Fairlamb, A.H., Fraunholz, M.J., Roos, D.S., Ralph, S.A., McFadden, G.I., Cummings, L.M., Subramanian, G.M., Mungall, C., Venter, J.C., Carucci, D.J., Hoffman, S.L., Newbold, C., Davis, R.W., Fraser, C.M., et Barrell, B. (2002). Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* 419: 498-511.

Gardner, M.J., Shallom, S.J., Carlton, J.M., Salzberg, S.L., Nene, V., Shoaibi, A., Ciecko, A., Lynn, J., Rizzo, M., Weaver, B., Jarrahi, B., Brenner, M., Parvizi, B., Tallon, L., Moazzez, A., Granger, D., Fujii, C., Hansen, C., Pederson, J., Feldblyum, T., Peterson, J., Suh, B., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., White, O., Cummings, L.M., Smith, H.O., Adams, M.D., Venter, J.C., Carucci, D.J., Hoffman, S.L., et Fraser, C.M. (2002). Sequence of Plasmodium falciparum chromosomes 2, 10, 11 and 14. *Nature* 419: 531-534.

Gavrilov, L.A., et Gavrilova, N.S. (2001). The reliability theory of aging and longevity. *J Theor Biol* 213: 527-545.

Glemet, E., et Codani, J.J. (1997). LASSAP, a LArge Scale Sequence compArison Package. *Comput Appl Biosci* 13: 137-143.

Glockner, G., Eichinger, L., Szafranski, K., Pachebat, J.A., Bankier, A.T., Dear, P.H., Lehmann, R., Baumgart, C., Parra, G., Abril, J.F., Guigo, R., Kumpf, K., Tunggal, B.,

Cox, E., Quail, M.A., Platzer, M., Rosenthal, A., et Noegel, A.A. (2002). Sequence and analysis of chromosome 2 of Dictyostelium discoideum. *Nature* 418: 79-85.

Gonnet, G.H., Cohen, M.A., et Benner, S.A. (1992). Exhaustive matching of the entire protein sequence database. *Science* 256: 1443-1445.

Gribskov, M., McLachlan, A.D., et Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A* 84: 4355-4358.

Gumbel, E.J. (1958). Statistics of Extremes, Columbia University Press.

Gutteridge, W., Trigg P. et Williamson, D. (1971). Properties of DNA from some malarial parasites. *Parasitology* 62: 209-219.

H

Haeckel, E. (1866). Generelle Morphologie der Organismen. Vol II. Berlin, Georg Reimer.

Hartley, R.V.L. (1928). transmission of Information. *The Bell System Technical Journal* 3: 535-564.

Henikoff, S., et Henikoff, J.G. (1991). Automated assembly of protein blocks for database searching. *Nucleic Acids Res* 19: 6565-6572.

Henikoff, S., et Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 89: 10915-10919.

Henikoff, S., et Henikoff, J.G. (1993). Performance evaluation of amino acid substitution matrices. *Proteins* 17: 49-61.

Hladik, J., et Chrysos, M. (2000). Introduction à la mécanique quantique. (Paris: Dunod).

Hoggs, J. (1860). On the distinctions of a plant and an animal and a fourth kingdom of Nature. *Edimburgh New Phil J., NS* 12:216-225.

Huang, X. (1994). On global sequence alignment. *Comput Appl Biosci* 10: 227-235.

I

International Human Genome Sequencing Consortium (2001). Finishing the euchromatic sequence of the human genome. *Nature* 431:931-45.

J

Jones, D.T., Taylor, W.R., et Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 8: 275-282.

K

Karlin, S., et Altschul, S.F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A* 87: 2264-2268.

Karlin, S., Bucher, P., Brendel, V., et Altschul, S.F. (1991). Statistical methods and insights for protein and DNA sequences. *Annu Rev Biophys Biophys Chem* 20: 175-203.

Kawashima, S., Ogata, H., et Kanehisa, M. (1999). AAindex: amino acid index database. *Nucleic Acids Res* 27: 368-369.

Keeling, P.J., Palmer, J.D. (2001). Lateral transfer at the gene and subgenic levels in the evolution of eukaryotic enolase. *Proc Natl Acad Sci U. S. A.* 98:10745-50.

Kilejan, A. (1975). Circular mitochondrial DNA from the avian parasite *Plasmodium lophurae*. *Biophys Biochim Acta* 390:276-284.

Kohler, S., Delwiche, C.F., Denny, P.W., Tilney, L.G., Webster, P., Wilson, R.J., Palmer, J.D., et Roos, D.S. (1997). A plastid of probable green algal origin in Apicomplexan parasites. *Science* 275 : 1485-1489.

Kolmogorov, A. (1968). Logical basis for information theory and probability theory. *Information Theory, IEEE Transactions* 14 : 662-664.

L

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R.,

McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., et Chen, Y.J. (2001). Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.

Lecointre, G., et Le Guyader, H. (2002). Classification phylogénétique du vivant. (Paris: Belin).

Lefebvre, C., Aude, J.C., Glemet, E. et Neri C. (2005). Balancing protein similarity and gene co-expression reveals new links between genetic conservation and developmental diversity in invertebrates. *Bioinformatics* 21:1550-1558.

Lespinats, S. (2006). Style du génome exploré par analyse textuelle de l'ADN. *Thèse* de l'Université Pierre et Marie Curie, Paris VI.

Lipman, D.J., et Pearson, W.R. (1985). Rapid and sensitive protein similarity searches. *Science* 227: 1435-1441.

Lipman, D.J., Altschul, S.F., et Kececioglu, J.D. (1989). A tool for multiple sequence alignment. *Proc Natl Acad Sci U S A* 86: 4412-4415.

Lobry, J.R. (1997). Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. *Gene* 205: 309-316.

Louis, A., Ollivier, E., Aude, J.C., et Risler, J.L. (2001). Massive sequence comparisons as a help in annotating genomic sequences. *Genome Res* 11: 1296-1303.

Luyet, B.J. (1950). A comparative study of the life cycles of lower protozoa and protophyta. *Biodynamica* 1950 6:265-364.

M

Maréchal, E., Block, M.A., Douce, R. et Joyard, J. (1999). Procédé de criblage et de sélection d'antiparasitaires et/ou d'herbicides et ses applications. Dépôt de Brevet-France numéro 99 03434, (2000: extension Europe/Canada/USA/Japon).

Maréchal, E., Miras, S. et Joyard, J. (2000). Fractions membranaires de cellules enrichies en 1,2-sn-diacylglycérol, procédé de préparation et utilisation. Dépôt de Brevet-France numéro 0013976 (2001: extension Canada/USA/Japon/Australie).

Maréchal, E. et Cesbron-Delauw, M.F. (2001). The apicoplast: a new member of the plastid family. *Trends Plant Sci* 6: 200-205.

Maréchal, E., Azzouz, N., de Macedo, C.S., Block, M.A., Feagin, J.E., Schwarz, R.T., et Joyard, J. (2002). Synthesis of chloroplast galactolipids in apicomplexan parasites. *Eukaryot Cell* 1: 653-656.

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M. et Penny, D. (2002). Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci U. S. A.* 99:12246-12251.

Martin, W. (2003). Gene transfer from organelles to the nucleus: frequent and in big chunks. *Proc Natl Acad Sci U. S. A.* 100:8612-8614 .

Mayr, E. (1964). The Evolution of living systems. *Proc. natl. Acad. Sci. U.S.A.* 51: 934-941.

Mayr, E. (1997). The objects of selection. *Proc. Natl. Acad. Sci. U.S.A.* 94: 2091-2094.

McConkey, G.A., Pinney, J.W., Westhead, D.R., Plueckhahn, K., Fitzpatrick, T.B., Macheroux, P., et Kappes, B. (2004). Annotating the Plasmodium genome and the enigma of the shikimate pathway. *Trends Parasitol* 20: 60-65.

McFadden, G.I., Reith, M.E., Munholland, J. et Lang-Unnasch, N. (1996). Plastid in human parasites. *Nature* 381:482.

McFadden, G.I. et Roos, D.S. (1999). Apicomplexan plastids as drug targets. *Trends Microbiol* 7:328-333.

Muller, T., Rahmann, S., et Rehmsmeier, M. (2001). Non-symmetric score matrices and the detection of homologous transmembrane proteins. *Bioinformatics* 17(Suppl 1): S182-189.

Musto, H., Rodriguez-Maseda, H., et Bernardi, G. (1995). Compositional properties of nuclear genes from Plasmodium falciparum. *Gene* 152: 127-132.

Muller, T., et Vingron, M. (2000). Modeling amino acid replacement. *J Comput Biol* 7: 761-776.

Muller, S., Gilberger, T.W., Krnajski, Z., Luersen, K., Meierjohann, S., et Walter, R.D. (2001). Thioredoxin and glutathione system of malaria parasite Plasmodium falciparum. *Protoplasma* 217: 43-49.

Muller, T., Spang, R., et Vingron, M. (2002). Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Mol Biol Evol* 19: 8-13.

Musto, H., Rodriguez-Maseda, H., et Bernardi, G. (1995). Compositional properties of nuclear genes from Plasmodium falciparum. *Gene* 152: 127-132.

N

Needleman, S.B., et Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 48: 443-453.

Nishizawa, M., et Nishizawa, K. (1998). Biased usages of arginines and lysines in proteins are correlated with local-scale fluctuations of the G + C content of DNA sequences. *J Mol Evol* 47: 385-393.

Noé, L. (2005). Recherche de similarités dans les séquences d'ADN: modèles et algorithmes pour la conception de graines efficaces. Thèse Université Henri Poincaré; Nancy 1.

Notredame, C., Higgins, D.G., et Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302: 205-217.

Nwaka, S. (2005). Drug discovery and beyond: the role of public-private partnerships in improving access to new malaria medicines. *Trans R Soc Trop Med Hyg* 99:S20-S29.

O

P

Pearson, W.R., et Lipman, D.J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* 85: 2444-2448.

Pearson, W.R. (1998). Empirical statistical estimates for sequence similarity searches. *J Mol Biol* 276: 71-84.

Perdijon, J. (2004). La Mesure. Histoire, Science et Philosophie. (Paris: Dunod).

Petryszak, R., Kretschmann, E., Wieser, D., et Apweiler, R. (2005). The predictive power of the CluSTr database. *Bioinformatics* 21: 3604-3609.

Pizzi, E., et Frontali, C. (2001). Low-complexity regions in Plasmodium falciparum proteins. *Genome Res* 11: 218-229.

Q

R

Ridley, R.G. (2002). Medical need, scientific opportunity and the drive for antimalarial drugs. *Nature* 415:686-693.

Risler, J.L., Delorme, M.O., Delacroix, H., et Henaut, A. (1988). Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. *J Mol Biol* 204: 1019-1029.

Ruhla, C. (1989). La physique du hasard. De blaise Pascal à Niels Bohr. (Paris, Hachette).

Rujan, T. et Martin, W. (2001). How many genes in Arabidopsis come from cyanobacteria? An estimate from 386 protein phylogenies. *Trends Genet* 17:113-120.

S

Salzberg, S.L., Pertea, M., Delcher, A.L., Gardner, M.J. et Tettelin, (1999) Interpolated Markov models for eukaryotic gene finding. *Genomics.* 59(1):24-31.

Scholtyseck, E. et Piekarski, G. (1965). Elektronenmikroskopische Untersuchungen an Merozoiten von Eimerien (*Eimeria perforans* und *E. steidae*) und *Toxoplasma gondii* zur systematischen Stellung von *T. gondii*. *Z Parasitenkd* 26: 93-115.

Setubal, J., et Meidanis, J. (1997). Introduction to Computational Molecular Biology. (PWS Publishing Compagny).

Singer, G.A., et Hickey, D.A. (2000). Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol Biol Evol* 17: 1581-1588.

Shannon, C.E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal* 27: 379-423,623-656.

Shkovskii, B.I. (2005). A simple derivation of the Gompertz law for human mortality. *Theor Biosci* 123: 431-433.

Simonite, T. (2005). Protists push animal aside in rule revamp. *Nature* 438:8-9.

Simpson, AG. et Roger A.J. (2004). The real 'kingdoms' of eukaryotes. *Curr Biol* 14:R693-696.

Skandalis, G. (2001). Topologie et Analyse. (Paris: Dunod).

Skorokhod, A.V. (2005). Basic principles and applications of probability theory. (Berlin: Springer).

Smith, T.F., et Waterman, M.S. (1981). Identification of common molecular subsequences. *J Mol Biol* 147: 195-197.

Steuer, R., Kurths, J., Daub, C.O., Weise, J., et Selbig, J. (2002). The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics* 18(Suppl 2):S231-240.

T

Tekaia, F., Lazcano, A., et Dujon, B. (1999). The genomic tree as revealed from whole proteome comparisons. *Genome Res* 9 : 550-557.

Tekaia, F., Yeramian, E., et Dujon, B. (2002). Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. *Gene* 297: 51-60.

The Plasmodium Genome Database Collaborative (2001). PlasmoDB: An integrative database of the Plasmodium falciparum genome. Tools for accessing and analyzing finished and unfinished sequence data. *Nucleic Acids Res* 29: 66-69.

Thompson, J.D., Higgins, D.G., et Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22: 4673-4680.

U

V

Vaidya, A.B. et Arasu, P. (1987). Tandemly arranged gene clusters of malaria parasites that are highly conserved and transcribed. *Mol Biochem Parasitol* 22: 249-257.

Valleron, A.J. (1998). Introduction à la Biostatistique. (Paris: Masson).

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B, . Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X, . Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen., E, Wolfe., K, Zaveri., J, Zaveri., K., Abril, J.F., Guigo, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel., J, Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A. et Zhu, X. (2001). The sequence of the human genome. *Science* 291:1304-1351.

Verra, F., et Hughes, A.L. (1999). Biased amino acid composition in repeat regions of Plasmodium antigens. *Mol Biol Evol* 16: 627-633.

Vingron, M., et Waterman, M.S. (1994). Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J Mol Biol* 235: 1-12.

von Linné, C. (1735). Systema Naturæ, sive, Regna tria Naturæ systematice proposita per classes, ordines, genera & species, Lugduni Batavorum.

von Siebold, C.T. (1848). Anatomy of the Invertebrata. Boston, Gould et Lincoln.

W

Waller, R.F. et McFadden G.I. (2005). The Apicoplast: A review of the derived plastid of Apicomplexan parasites. *Curr Issues Mol Biol* 7: 57-79.

Waterman, M.S. et Vingron, M. (1994). Sequence comparison signifiance and Poisson approximation. *Stat Sci* 9(3): 367-381.

Waterman, M.S. (1995). Introduction to computational biology. (CRC Press).

Webber, C., et Barton, G.J. (2003). Increased coverage obtained by combination of methods for protein sequence database searching. *Bioinformatics* 19: 1397-1403.

Whittaker, R.H. (1959). On the broad classification of organisms. *Quart Rev. Biol* 34:210-226.

Wiesner, J. et Seeber, F. (2005). The plastid-derived organelle of protozoan human parasites as a target of established and emerging drugs. *Expert Opin Ther Targets* 9(1):23-44.

Wilson, R.J.M., Gardner M.J., Feagin J.E. et Williamson D.H. (1991). Have malaria parasites three genomes? *Parasitol Today* 7:134-136.

Wilson, R.J., Denny, P.W., Preiser, P.R., Rangachari, K., Roberts, K., Roy, A., Whyte, A., Strath, M., Moore, D.J., Moore, P.W. et Williamson, D.H. (1996). Complete gene map of the plastid-like DNA of the malaria parasite Plasmodium falciparum. *J Mol Biol* 261:155-72.

Wilson, R.J.M. et Williamson D.H. (1997). Extrachromosomal DNA in the Apicomplexa. Microbiol. *Mol Biol Rev* 61: 1-16.

Woese, C.R., Kandler, O. et Wheelis, M.L. (1990). Towards a natural system of organisms: proposal for the domains Archae, Bacteria and Eucarya. *Proc Natl Acad Sci U.S.A.* 87:4576-4579.

Wood, V., Rutherford, K. M., Ivens, A., Rajandream, M-A et Barrell, B. (2001). A Re-annotation of the Saccharomyces cerevisiae Genome. *Comparative and Functional Genomics* 2:143-154.

Woodrow, C.J., Haynes, R.K. et Krishna, S. (2005). Artemisinins. *Postgrad Med J* 81:71-78.

World Malaria Report (2005). Roll Back Malaria-World Health Organization-Unicef, Geneva.

Wright, C.W. (2005). Traditional antimalarials and the development of novel antimalarial drugs. *J Ethnopharmacol* 100:67-71.

Wu, T.T., Fitch, W. M. et Margoliash, E. (1974). The information content of protein amino acid sequences. *Annu Rev Biochem* 43:539-566.

Xu, P., Widmer, G., Wang, Y., Ozaki, L.S., Alves, J.M., Serrano, M.G., Puiu, D., Manque, P., Akiyoshi, D., Mackey, A.J., Pearson, W.R., Dear, P.H., Bankier, A.T., Peterson, D.L., Abrahamsen, M.S., Kapur, V., Tzipori, S., et Buck, G.A. (2004). The genome of Cryptosporidium hominis. *Nature* 431: 1107-1112.

Xue, H.Y., et Forsdyke, D.R. (2003). Low-complexity segments in Plasmodium falciparum proteins are primarily nucleic acid level adaptations. *Mol Biochem Parasitol* 128: 21-32.

Xuong, N.H. (1991). Mathématiques Discrètes et Informatique. (Paris: Dunod).

Yu, Y.K. et Altschul, S.F. (2005). The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. *Bioinformatics* 21:902-11.

Zuegge, J., Ralph, S., Schmuker, M., McFadden, G.I. et Schneider, G. (2001). Deciphering apicoplast targeting signals - feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins. *Gene* 280 : 19-26

DÉVELOPPEMENTS THÉORIQUES ET MÉTHODES NUMÉRIQUES POUR LES ANALYSES COMPARATIVES DE GÉNOMES ET PROTÉOMES BIAISÉS

Application à la comparaison des génomes et protéomes de *Plasmodium falciparum* et d'*Arabidopsis thaliana*

Le paludisme, ou malaria, est une maladie infectieuse qui touche plus de 350 millions d'êtres humains et qui tue chaque année 2,5 millions de personnes à travers le monde. Les parasites responsables de la malaria sont des apicomplexes du genre *Plasmodium*, essentiellement *P. falciparum*. Le génome de *P. falciparum*, est séquencé depuis octobre 2002, et présente un des taux les plus faibles de gènes annotés, avec ~60 % de gènes sans fonction attribuée. Il est difficile, voire impossible, d'identifier dans le génome de *P. falciparumi*, certains gènes, responsables de fonctions mesurées biochimiquement chez le parasite, par similarité avec des séquences homologues caractérisées dans d'autres organismes. Cette difficulté rencontrée lors des recherches automatiques d'homologie est une limite à tout projet exploratoire du génome malarial fondé sur la phylogénie moléculaire. En particulier, l'inventaire des séquences héritées de l'algue ancestrale, qui a réalisé l'endosymbiose secondaire qui caractérise le phylum des Apicomplexa (sous génome d'origine algale dans lequel il est possible de rechercher des cibles pour des médicaments herbicides), peut être rendu incomplet. Les caractéristiques atypiques du génome et du protéome de *Plasmodium*, résumées sous le terme de biais compositionnel (en particulier un pourcentage en adénosine+thymidine supérieur à 80%), ont été soupçonnées d'être un cas limite pour les outils d'analyse de séquence existants. L'objet de cette thèse a donc été d'examiner l'influence possible de ce type de biais sur les méthodologies de comparaisons de séquences et de façon plus approfondie sur leurs statistiques.

Nous avons proposé des développements théoriques nouveaux, associés à la statistique de la *Z-value* introduite par Lipman et Pearson pour évaluer la significativité d'un score d'alignement de deux séquences protéiques: (1) le théorème TULIP permettant de déduire un majorant de la probabilité d'un score d'alignement de séquences (*i.e.* la *P-value*) par la valeur $1/Z\text{-}value^2$ et (2) la déduction des propriétés remarquables de la distribution des *Z-values* à partir de quelques hypothèses sur l'évolution des protéines dans le contexte de la théorie de la fiabilité des systèmes. Ces développements théoriques ont permis certaines avancées sur le plan pratique de l'identification de séquences homologues initialement non détectées par le théorème de Karlin-Altschul et d'étayer la relation entre les scores d'alignements et l'information mutuelle, au sens de la théorie de l'information.

En construisant un espace de configuration des protéines homologues, permettant une expression du théorème TULIP et ayant une cohérence avec la théorie synthétique de l'évolution, nous avons déduit une méthode de reconstruction de phylogénies de séquences protéiques à l'aide des *Z-values*. Les phylogénies moléculaires reconstruites par cette méthode sont concordantes avec celles obtenues à partir d'alignements multiples et permettent par ailleurs de résoudre certaines incohérences rapportées avec les méthodes de reconstruction phylogéniques classiques.

En prenant en compte le modèle statistique que nous avons élaboré, nous avons entrepris une première analyse de l'évolution du biais en acides aminés chez *Plasmodium* corrélativement à l'évolution du biais en acides nucléiques dans le génome malarial et en fonction de la divergence évolutive, établie en prenant le génome non biaisé d'*Arabidopsis thaliana* comme référence. Nous avons observé que le biais des séquences malariales était corrélé au pourcentage de divergence avec leurs homologues végétaux. Nos analyses suggèrent de plus que le biais est vraisemblablement la conséquence d'une évolution au niveau nucléique. Nous avons examiné la possibilité de construire une famille de matrices tenant compte de cette dissymétrie dans le cas de la comparaison de *Plasmodium* et d'*Arabidopsis*. Ces matrices appelées DirAtPf, possèdent (1) une sensibilité théorique et (2) une spécificité supérieure aux familles de matrices existantes.

Les perspectives des travaux présentés dans ce mémoire incluent une progression de l'annotation automatique de *Plasmodium falciparum* et la mise en place d'une procédure statistiquement robuste et phylogénétiquement consistante pour caractériser le sous-génome algal du parasite malarial.

MOTS CLES : alignements de séquences, comparaisons de génomes, malaria, *Plasmodium falciparum*

Laboratoire de Physiologie Cellulaire Végétale
UMR 5168 CNRS-CEA-INRA-Université Joseph Fourier
Département Réponse et Dynamique Cellulaire;
CEA-Grenoble; 17, rue des Martyrs,
38054 Grenoble cedex 9

Gene-IT SA
147, Avenue Paul Doumer
92500 Rueil Malmaison France