Etude des biais de composition en acides aminés des protéines microbiennes

Géraldine Pascal

Université d'Evry Val d'Essonne

Ecole doctorale « des génomes aux organismes »

Thèse

Pour obtenir le grade de
Docteur de l'université d'Evry val d'Essonne
En Bioinformatique, Biologie structurale et Génomique
Présentée et soutenue publiquement par

Géraldine Pascal

Le 25 octobre 2005

Etude des biais de composition en acides aminés des protéines microbiennes

Jury :

M. Francis Quétier : président du jury
M. Jacques Batut : rapporteur
M. Manolo Gouy : rapporteur
M. Gunnar Von Heijne : examinateur
Mme Hélène Chiapello : examinatrice
M. Antoine de Daruvar : examinateur
Mme Claudine Médigue : directrice de thèse
M. Antoine Danchin : directeur de thèse

Durant ces dernières années, j'ai pu vivre (tout au moins devant ma télévision) une coupe du monde et un championnat d'Europe de football, les jeux olympiques d'hiver et d'été, deux championnats du monde d'athlétisme, quatre Rolland Garros et quatre tours de France.

En observant tous ces champions, je pensais que le dépassement de soi, la persévérance, la patience, le stress des grands rendez-vous, étaient des valeurs que partageait également un étudiant en thèse, notamment moi.

Mais le sportif ne gagne jamais seul !

Et je tiens aujourd'hui à remercier toutes les personnes qui ont fait que les prochaines pages que vous lirez ont pu s'écrire malgré les embûches et les découragements.

Je tiens à remercier tout d'abord les membres du jury, pour continuer la métaphore, ils sont les juges arbitres de ce challenge. Votre temps est précieux, merci d'en avoir consacré une partie à juger mon travail.

Je remercie les présidents de Clubs, M. Jean Weissenbach et M. Philippe Kourilsky, qui ont bien voulu m'accueillir et m'ont donné les moyens matériels et financiers de travailler sereinement au Genoscope et à l'Institut Pasteur respectivement.

Viennent ensuite les directeurs sportifs, Claudine et Antoine. Je leur dois beaucoup : encouragements, aboutissements (les papiers), entraînements (corrections, directions), apprentissages (transmission du savoir), déplacements extérieurs (à Hong Kong, notamment), rencontres (conférences, collaborations).

Et puis, il y a le staff, c'est-à-dire toutes les personnes sans qui rien ne fonctionne. Il vous épaule aussi bien moralement que techniquement. Je remercie tous les membres de l'AGC et de l'unité GGB. En particulier, je tiens à remercier de tout cœur Zoé, Cathy, Aurélie et David pour leurs précieuses aides entre autres lors de la manipulation des banques et des bases de données. Je remercie également les Stefs, Laurent, Eduardo et Philippe pour les nombreuses discussions qui, à plusieurs reprises, ont orienté mon travail. Et puis de grands mercis à Lili, Hi, Mike, Gang, Evelyne T., Evelyne K., Isabelle et Susan qui ont toujours été à mes cotés lorsque j'avais besoin d'eux. J'ai une pensée particulière pour Agnieszka grâce à qui le dépaysement à Hong Kong fut moins difficile. Merci à Laurent, le root, qui a toujours fait en sorte que les problèmes informatiques ne soient qu'illusion. Je n'oublie pas Annie, Catherine et Corinne, nos merveilleuses secrétaires sans qui les petits tracas administratifs deviendraient de véritables tourments.

Tout ce travail n'aurait pas vu le jour sans mon coach, celui qui vit nuit et jour avec son « poulain », ressent les mêmes angoisses et remporte les mêmes victoires. Pour tout, Ced', merci.

N'oublions pas la famille, papa, maman, soeurette, Yo et les Toulousains. Je vous remercie tendrement de votre présence et de votre chaleureux soutien durant toutes ces années.

Je terminerai en remerciant mes supporters, Phine, Tonio, Oliv', Carole et Sev', pour leur précieuse amitié toujours restée sans faille malgré mon emploi du temps contraignant et la rare disponibilité que j'ai pu leur offrir ces temps derniers.

A ma grand-mère, Lucie.

« Le hasard ne favorise que les esprits préparés »

Louis Pasteur (1822 – 1895)

Avant Propos

La rédaction de cette thèse s'est déroulée parallèlement à celle d'un article scientifique qui va être soumis à publication très prochainement au journal Bioessays. Ce nouveau papier mêle à la fois une brève revue sur l'analyse compositionnelle des protéines procaryotes et des résultats originaux. De nombreux extraits de cet article (en langue anglaise) ont été repris dans l'introduction de ce manuscrit, et l'un des chapitres de cette thèse (C.II) est consacré à l'ensemble de ces nouveaux résultats.

Introduction

A.I Contexte.

L'ensemble des organismes vivants est soumis à une variété de pressions de sélection qui agit non seulement au niveau du phénotype global, mais également à chaque niveau de l'organisation de la cellule. Il est généralement admis que les protéines seraient, entre autres, soumises à une pression de sélection associée à leur fonction. C'est en effet le cas. Cependant, la comparaison de protéines aux fonctions identiques d'organismes pris dans l'ensemble des grands groupes évolutifs, montre qu'à l'exception de quelques dizaines d'acides aminés, la presque totalité des résidus ont pu être échangés, sans évolution conséquente de la fonction des protéines. En exemple, il a été montré chez les Bactéries que des protéines aux fonctions similaires voient leur composition en acides aminés modifiée suivant que leurs gènes soient localisés sur le brin direct ou le brin indirect de l'ADN (Rocha, Danchin et al. 1999). Ceci indique non seulement que les protéines ont des structures extrêmement robustes, mais démontre également que l'identité chimique de chaque résidu, à la plupart des positions, pourrait intégrer des signaux subtils dérivés de pressions de sélection sévisant dans d'autres régions de la cellule. Un acide aminé dans une protéine intègre une préférence pour une certaine composition en bases nucléiques, une préférence pour un codon et une préférence pour cet acide aminé en particulier. Ces choix résultent de contraintes aux niveaux de la structure de l'ADN et de l'ARN, mais également des contraintes au niveau de la biosynthèse des nucléotides et/ou des acides aminés. De fait, Gojobori et ses collègues ont suggéré dans l'une de leurs dernières études, que le coût énergétique de la synthèse d'un acide aminé donné a un certain effet sur la composition générale des protéines (Akashi et Gojobori 2002). C'est donc dans la continuité de ces observations que nous avons entrepris l'analyse approfondie de facteurs variants pouvant affecter la composition des protéines. La condition préalable à cet exercice sera d'élaborer une nouvelle approche de comparaison des protéines d'organismes phylogéniquement distants, une tâche notoirement difficile.

A.II Bref état de l'art.

A.II.1 L'analyse du code, une quête d'hier et d'aujourd'hui.

Extrait de l'article « Persistent biases in the amino acid composition of prokaryotic proteins », issu du paragraphe « Introduction » et de l'introduction du paragraphe « Results and Discussion »..

« Genome studies analyse the bulk of the proteins of an organism, its proteome, as the result of the conceptual translation of protein-coding DNA sequences (Grantham, Gautier et al. 1980; Grantham, Gautier et al. 1980; Grantham, Gautier et al. 1981; Gouy et Gautier 1982; Blake et Hinds 1984; Medigue, Rouxel et al. 1991; Lobry et Gautier 1994; Karlin, Mrazek et al. 1997). With the development of new rapid sequencing methods concentrated in specialised centres, the genome sequence of prokaryotic organisms, Bacteria and Archaea, have become widely available. Several analyses of the codon usage bias or the amino acid usage in reference proteomes have been published, such as

those of *Thermotoga maritima* (Zavala, Naya et al. 2002), *Pseudomonas aeruginosa* (Gupta et Ghosh 2001), Buchnera species (Palacios et Wernegreen 2002), or of the three omnipresent model prokaryotes *Escherichia coli*, *Bacillus subtilis* and *Methanococcus jannaschii* (Pascal, Medigue et al. 2005). »

« The choice of each amino acid residue in a protein results from superposition of a wide range of selection pressures, some indirect (such as the metabolic cost to obtain each residue (Akashi et Gojobori 2002), or the availability of a given pathway for the synthesis of nucleotides (Rocha et Danchin 2002), with a limited contribution of the nature of the protein function (on average, less than ten percent of the residues in an enzyme are directly involved in its catalytic properties, for example). Global features, such as the genomic G+C content (from 28.6% to 72.1%), the optimal growth temperature (from 10°C to 103°C), the ecological niche (living in animal or plant tissues, in soil, in marine, alkaline or acidic environments…) and other constraints due to differences in growth rate or pathogenicity, all contribute to the final outcome. »

A.II.2 *La composition en acides aminés des protéines membranaires.*

Extrait de l'article « Persistent biases in the amino acid composition of prokaryotic proteins », issu du paragraphe « A universal rule: integral inner membrane proteins cluster together ».

« The amino acids which constitute proteins are roughly split into two major classes, depending on their interaction with water. In the cytoplasm, they contribute to protein folding, with the hydrophobic amino acids usually clustered within the inside of the protein. In contrast, membrane-associated proteins have to interact with a highly hydrophobic lipid bilayer, which they usually perform through sequences of exposed hydrophobic residues. In particular, hydrophobic alpha helices made of 19-22 residues span the lipid bilayer (Tie, Nicchitta et al. 2005). It has been shown that proteins that are imbedded in the membrane, with limited outside stretches, are rich in hydrophobic residues (van Geest et Lolkema 2000; Ulmschneider et Sansom 2001). Integral inner membrane proteins, melted in the phospholid bilayer of membrane, include a significant amount of hydrophobic residues such as Phe, Leu and Ile while they have only a few charged residues: Asp, Arg, Glu, and Lys, mostly located outside of the lipid core or the membrane (Wallin et von Heijne 1998; Ulmschneider et Sansom 2001; Pascal, Medigue et al. 2005). These proteins, atypical in their amino acid composition, differ from the proteins from other cellular compartments, including those of the outer membrane, when it exists (Perriere, Lobry et al. 1996; Guerdoux-Jamet, Henaut et al. 1997). »

A.II.3 *Les protéines hautement exprimées, une composition en acides aminés optimale.*

Extrait de l'article « Persistent biases in the amino acid composition of prokaryotic proteins », issu du paragraphe « Highly expressed ancestral proteins display common biases ».

« The bulk of the expressed proteins in fast growing organisms consists of those which constitute the translation machinery. They are generally expressed at a high level under exponential growth conditions. These proteins are considered as ancestral, and are generally used as relevant markers for phylogenetic analyses, in parallel with studies involving ribosomal RNA (Woese et Fox 1977; Woese, Kandler et al. 1990). Because they constitute the core of all cells, they have been used as a reference to compare the proteomes of all organisms, including those which grow poorly or slowly. Generally, highly expressed proteins have a biased amino acid composition, frequently linked with a bias in the way they use the genetic code, their Codon Adaptation Index (CAI). The highly expressed genes use a subset of optimal codons due to selection for efficient translation of their mRNAs (Gouy et Gautier 1982; Sharp et Li 1987; Medigue, Rouxel et al. 1991; Lobry et Gautier 1994; Pan, Dutta et al. 1998; Karlin et Mrazek 2000). In particular, ribosomal proteins are characterised by an enrichment in basic amino acids (Lys and Arg) and small hydrophobic residues (Ala, Val and Gly) (Karlin, Mrazek et al. 1998; Lin, Kuang et al. 2002). »

A.II.4 *L'influence de la cystéine et du tryptophane dans les protéines.*

Extrait de l'article « Persistent biases in the amino acid composition of prokaryotic proteins », issu du paragraphe « Rare amino acids create specific clusters ».

« Not all amino acids are equal in terms of frequency in proteins. Some are systematically rare, while others are frequent. Among the former, cysteine (Cys) and tryptophan (Trp) are particularly important as their rarity often results in a strong bias in the proteome cloud's shape. Cys and Trp each represent less than 1% of the amino acids of the proteomes. Cys contains the highly reactive sulphhydryl group, and takes part in numerous active sites, for example as a catalytic centre, zinc ligand or core component of iron-sulphur clusters. It also plays an important role in structural stabilisation of exported proteins in forming disulphide bridges. Trp is an aromatic amino acid with a voluminous side chain. Its role is not well established: besides its aromaticity, it is mildly hydrophobic, because of the nitrogen in the indole component, and often contributes to the stabilisation of protein structures, having a positive impact on the folding of proteins, because indole can also accept hydrogen bonds under certain circumstances (Zhu, Jutila et al. 2001; Clark, East et al. 2003). Trp is however not used very frequently (≤1%), perhaps because it is very costly in terms of metabolism and quite reactive towards reactive oxygen species (also, it is usually coded by only one codon TGG or two in many mycoplasma, including TGA). These two residues, Cys and Trp, are so rare in proteomes that proteins having several of these amino acids are atypical and therefore worth investigation. »

A.II.5 *L'impact du contenu en G+C des gènes sur la composition en acides aminés des protéines.*

Extrait de l'article « Persistent biases in the amino acid composition of prokaryotic proteins », issu du paragraphe « The G+C-content of the Coding DNA Sequences creates an unexpected bias in amino acid composition ».

« The amino acids present in proteins are constrained by the nucleotide composition of the corresponding genes. It has thus been demonstrated that, because there is often a strong bias in the composition of the leading and lagging strands of chromosomes, which are distinguished from each other by an enrichment in G+T and C+A, respectively, proteins coded from the leading strand are enriched in valine as compared to those coded from the lagging strand, which are enriched in isoleucine and threonine (Lobry 1996; Francino et Ochman 1997; Frank et Lobry 1999; Rocha, Danchin et al. 1999; Lobry et Louarn 2003). Moreover, there is a general bias in GNN codons in CDSs, possibly acting as a ratchet-like mechanism during the translation elongation process that could influence the overall amino acid composition of the proteomes (Brooks et Fresco 2003). In the same way, the overall G+C base composition of genomes influences strongly the choice of amino acids that constitute the corresponding proteomes (Lobry 1997; Wilquet et Van de Casteele 1999; Pascal, Medigue et al. 2005). The G+C content will drive the codon usage bias and not the reverse (Knight, Freeland et al. 2001). As a consequence, because it is adapted to the temperature which is optimal for growth of the organism (Lynn, Singer et al. 2002; Musto, Naya et al. 2004; Naya, Zavala et al. 2004), the resulting constraints on the proteome create the first discriminant factor for the thermophilic organisms (Kreil et Ouzounis 2001; Tekaia, Yeramian et al. 2002). »

A.II.6 *Usage du code et biotope.*

Extrait de l'article « Persistent biases in the amino acid composition of prokaryotic proteins », issu du paragraphe « Specificities linked to the biotope ».

« Many studies have been devoted to the physiological consequences of extreme temperatures, trying to correlate them with the gene and protein composition allowing an optimal adaptation under these extreme conditions (Carbone, Kepes et al. 2005; Liang, Huang et al. 2005). Membrane fluidity is an important parameter associated with survival, as leakage of components from the cytoplasm or influx of reactive molecules from the outside is deleterious to the cell. Permeability strongly depends on temperature, usually increasing and decreasing in parallel with temperature. As a consequence, the cell must constantly tune its membrane fluidity to compensate for temperature variations. This is made possible by modulating the lipid composition of the membrane. Long chain and saturated fatty acids decrease membrane fluidity. At low temperature, fatty acids become unsaturated and are often branched-chain. To compensate for temperature increase, the cell tends to increase the length of the fatty acids, which constitute the phospholipid bilayer of the membrane. Archaea have membranes characterised by lipids of unique structure and chemical composition. To form phospholipids their fatty acids are linked to glycerol by ether bonds instead of ester bonds. In addition to affecting membranes in a dramatic manner, high temperature causes proteins to denature, causing irreparable damage to the cell. As a consequence, not only must heat-adapted proteins have specific features, but protein synthesis has to be adapted to allow efficient folding, presumably requiring a particular codon usage bias, superimposed on constraints in the nucleotide composition of the genome as a whole (Singer et Hickey 2003). Temperature has been described as an external selective force that was correlated with variation at the nucleic acid level (Lobry et Chessel 2003). This was consistent with a report showing

that genomes of thermophiles had more AGR than CGN codons for arginine, the latter being commonplace in mesophiles (Farias et Bonato 2003). Others studies showed that the surfaces of thermophilic proteins are generally richer in the charged residues Glu, Arg and Lys (Fukuchi et Nishikawa 2001; Nakashima, Fukuchi et al. 2003). Added to these biases, we know also that asparagine distinguishes proteins of psychrophilic bacteria (Medigue, Krin et al. in press). It is likely that the physico-chemical constraints linked to extreme temperatures are managed by the cell both at the nucleic acid and protein level, resulting in a complex outcome in terms of amino acid distribution in proteins. »

A.II.7 *Composition en acides aminés selon la phylogénie des espèces.*

Extrait de l'article « Persistent biases in the amino acid composition of prokaryotic proteins », issu du paragraphe « The prokaryotic domain split ».

« Most known genomes of thermophilic organisms are from Archaea. Phylogenetic studies of thermophiles were often developed to support theories proposing that the first living cells were thermophiles (Kreil et Ouzounis 2001; Tekaia, Yeramian et al. 2002; La, Silver et al. 2003). Since the seventies, the study of the evolution of organisms has been mainly based on the molecular comparison of ribosomal RNAs (Woese et Fox 1977; Woese, Kandler et al. 1990). More recent studies have attempted to expand phylogenetic analyses, first to protein evolution, then to amino acid composition and codon usage biases, to establish independently whether the similarities uncovered using rRNAs phylogenies between domains, species and organisms were conserved in all situations (Sorimachi 1999; Tekaia, Lazcano et al. 1999; Pupo, Lan et al. 2000). Proteomic signatures were found to distribute organisms into domains according to their protein amino acid and peptide composition as broad clusters (Pe'er, Felder et al. 2004). Likewise, enhancement of the content in some residues such as Cys, Met, Phe, His and Ser, with concomitant loss of others such as Pro, Ala, Glu and Gly in the proteins of most organisms was demonstrated when comparing proteomes with a variety of statistical approaches (Brooks et Fresco 2002; Jordan, Kondrashov et al. 2005). All these studies concur to show that the amino acid composition of proteomes can provide clues to enrich our knowledge about processes underlying the evolution of organisms. It should be stressed however that only coarse-grained relationships could thus be uncovered, as most of these methods consider the overall content of the proteome rather than the fine distribution of the amino acids within the proteins, which constitute the proteomes. Furthermore, the statistical approaches used lack reference comparisons, while methods such as Principal Component Analysis are prone to create large distortions when the objects analysed form clusters of widely different size or are very distant from one another (Hill 1974). Moreover, there is certainly a contribution of the level of protein expression (Pal, Papp et al. 2001; Papp, Pal et al. 2003; Rocha et Danchin 2004) and horizontal gene transfer to the final proteome composition (Lawrence et Ochman 2002), while biases induced by the location of a gene in the leading or the lagging strand of a genome will influence the amino acid and peptide composition of the corresponding protein (Rocha et Danchin 2001). »

A.III Lorsque les voyages forment la recherche.

Ce travail de recherche a été réalisé sur plusieurs années, dans différents laboratoires et pays. Les 18 premiers mois se sont écoulés au HKU-Pasteur Research Centre de Hong Kong co-dirigé par Antoine Dandin et K.Y. Yuen. Venue dans ce pays lointain afin d'analyser de plus près les codes nucléiques et protéiques du champignon *Penicillium marneffei*, endémique à l'Asie du Sud-est, mes horizons se sont en fait étendus à un bien plus large spectre d'organismes, non pas eucaryotes mais procaryotes. De retour en France et rejoignant les équipes d'Antoine Danchin à l'unité de Génétique des Génomes Bactériens (GGB) de l'Institut Pasteur de Paris et l'Atelier de Génomique Comparative (AGC) dirigé par Claudine Médigue au Genoscope d'Evry, j'ai poursuivi pleinement mes investigations par l'analyse compositionnelle des protéines procaryotes. J'ai également participé à de nombreuses collaborations notamment sur l'analyse du protéome de *Photorhabdus luminescens* avec Evelyne Turlin et Sylvianne Denzelle de l'Institut Pasteur de Paris ainsi que sur les stratégies de méthodes de partitionnement avec le Pr. Ngaiming Mok de l'université de Hong Kong. J'ai constamment évolué dans un environnement mêlant bioinformatique et biologie aussi bien à Hong Kong qu'en France et ceci a été un fait essentiel à la réussite de nombre de mes entreprises.

A.IV Présentation générale de la thèse.

L'analyse des biais compositionnels des protéines procaryotes a été le fil conducteur de mon travail tout au long de ma thèse. Les questions auxquelles nous avons cherché réponse, étape par étape, touchent à la fois (i) la structure et les entités du code protéique, (ii) les relations intrinsèques entre la séquence nucléique et les protéines de la cellule ainsi que (iii) les fortes pressions de sélection attribuées à l'environnement de croissance de l'organisme et au fonctionnement cellulaire.

Ce sujet m'a amené à approfondir de nombreux mécanismes biologiques. J'ai abordé le coté fonctionnel par l'étude de certains cycles métaboliques, de leurs enchevêtrements, de leurs coûts énergétiques, et le coté structural, par l'étude des structures tridimensionnelles et des repliements des protéines notamment celles relatives aux protéines membranaires. J'ai également approfondi mes connaissances en biologie cellulaire, en liaison à la compartimentation des cellules mais aussi aux processus responsables des modifications post-traductionnelles des protéines. Enfin, et de manière plus générale, je me suis concentrée sur la compréhension des modes de vie des organismes (reproduction, résistance, pathogénicité, symbiose, survie) et de leur phylogénie, réflexions indispensables à l'élucidation de certains résultats. Du point de vu informatique, j'ai été amenée à développer un certain nombre de programmes de traitement et formatage des données en amont et en aval de l'analyse (i.e. le formatage des banques de données et des résultats). Enfin, il m'a été nécessaire d'acquérir un savoir-faire statistique (i.e. les analyses multivariées, les méthodes de partitionnement, les tests statistiques) pour le traitement des données et l'analyse des résultats.

En ce qui concerne le matériel biologique exploité, notre choix s'est porté en premier lieu sur l'étude des procaryotes modèles *Escherichia coli*, *Bacillus subtilis* et *Methanococcus jannaschii*. Ces premières analyses ont confirmé l'importance de la connaissance de la composition en acides aminés des protéines afin de comprendre la façon dont l'expression des gènes est organisée dans la cellule procaryote. Nous avons ensuite décidé de poursuivre nos investigations suivant deux voies parallèles et complémentaires. La première fut l'étude d'un ensemble de 28 organismes représentatif du monde procaryote afin d'essayer de révéler des règles universelles régissant les fonctions biologiques et l'évolution des protéines des organismes d'intérêts. La seconde fut l'étude de la bactérie Antarctique, *Pseudoalteromonas haloplanktis* TAC125 séquencée récemment au Genoscope, pour tenter d'observer la pression sélective des basses températures sur la composition en acides aminés des protéines. Conjointement, les méthodes et les développements élaborés pendant la thèse ont pu être appliqués et/ou adaptés à d'autres projets comme l'étude de la pathogénicité de *Photorhabdus luminescens* réalisée dans l'unité GGB ou dans le cadre de la mise en place de la plateforme MaGe à l'AGC. Finalement, comme une analepse, la perspective de travailler sur des eucaryotes simples est envisagée sous certaines conditions, au regard des travaux préliminaires, initiés au début de la thèse, effectués sur l'usage des codons de *Penicillium marneffei* en relation avec son dimorphisme.

B Matériels et méthodes

B.I Pourquoi étudier les procaryotes ?

Depuis les travaux entrepris par Tatum et Beadle (prix Nobel de Médecine en 1958), l'ADN est reconnu comme support de l'information génétique. Le déchiffrage d'un génome dans sa globalité devient alors une source formidable de connaissances. Cependant, pour des raisons techniques et financières, le séquençage de génomes complets est resté longtemps inaccessible. Epoque révolue, une nouvelle ère s'ouvre aujourd'hui aux biologistes puisque le séquençage à grande échelle est maintenant une réalité. Les techniques utilisées ont de plus évolué. De manière exclusive, les génomes sont décodés à l'aide de séquenceurs automatiques, ce qui réduit considérablement les coûts (~1€ par lecture, soit environ 750 nucléotides, au Genoscope), ainsi que la durée de traitement (en moyenne un an pour la production et la finition de la séquence d'un génome procaryote), et augmente la précision (limitation des erreurs humaines).

Un projet de séquençage s'établit toujours autour de l'intérêt biologique de l'organisme. Bien que de plus en plus de programmes de séquençage d'organismes eucaryotes voient le jour, l'engouement s'est toujours porté largement sur le séquençage des génomes procaryotes (simplicité et petite taille des génomes, grande diversité des organismes). Les génomes de bactéries telles que *Escherichia coli* (bactérie à Gram négatif) et *Bacillus subtilis* (bactérie à Gram positif) ont été séquencés en raison de leur importance dans le domaine de la recherche fondamentale. D'autres génomes tels que ceux d'*Agrobacterium tumefaciens* (une bactérie du sol qui peut infecter les végétaux), de *Clostridium acetobutilicum* (production de butanol et d'acétone) ou de *Lactococcus lactis* (levains utilisés dans la plupart des fabrications fromagères) l'ont été en raison de leur utilisation industrielle, en particulier dans le domaine agro-alimentaire. Les procaryotes pathogènes, de part leurs intérêts pharmacologiques et environnementaux, ont bien évidemment trouvé leur place parmi les grands projets de séquençage. Concernant les intérêts pharmacologiques, ne citons que quelques bactéries responsables d'importantes maladies humaines : *Haemophilus influenzae* (otite, bronchite), *Vibrio cholerae* (choléra), *Mycoplasma pneumoniae* (pneumonies), *Helicobacter pylori* (ulcères), *Mycobacterium tuberculosis* (tuberculose – 10 personnes touchées chaque minute), *Mycobacterium leprae* (lèpre), *Yersinia pestis* (peste). Concernant les enjeux environnementaux, nous citerons, entre autres, *Xylella fastidiosa* (phytopathogène affectant en particulier les orangers), *Xanthomonas campestris* (pathogène de divers crucifères) ou *Ralstonia solanacearum* (bactérie capable de contaminer plus de 200 espèces végétales). Enfin, les génomes des bactéries telles que *Buchnera* qui vit en symbiose avec certains insectes, ou *Deinococcus radiodurans* qui supporte de très hauts niveaux d'irradiation, ont été séquencés en raison de leurs particularités biologiques. Parmi les procaryotes, les Archaeabactéries sont elles aussi très intéressantes à séquencer. Elles vivent dans des conditions extrêmes de température, de pH, de pression et présentent souvent des métabolismes particuliers (production de méthane, réduction de sulfates, etc.) aux propriétés attractives quant aux applications industrielles.

Après l'obtention de la séquence complète d'un organisme, il faut encore identifier ses gènes (annotation syntaxique). Chez les procaryotes, cette étape est plus simple de part la structure de leur(s) chromosome(s) (petite taille, peu ou pas d'introns, reconnaissance aisée des phases ouvertes de lecture et de leur terminaison de transcription, faible taille des séquences intergéniques). Enfin, reste à associer une ou des fonctions biologiques à ces gènes, opération ardue et délicate (annotation fonctionnelle), puis à analyser les processus cellulaires et physiologiques (annotation relationnelle) pour tenter de déterminer comment les gènes et les protéines interagissent entre eux pour accomplir une tâche spécifique dans la cellule.

Ainsi, ces dernières années, le séquençage des génomes procaryotes s'est révélé particulièrement productif et a crû à un rythme exponentiel. Un génome bactérien est achevé en moyenne tous les deux mois, et à ce jour, plus de 230 séquences complètes de génomes bactériens sont disponibles. L'importante diversité de ces génomes et des gènes qu'ils contiennent a ainsi été mise en évidence mais pour un large tiers d'entre eux, la fonction de la protéine encodée reste inconnue. Il est alors nécessaire d'utiliser de nouvelles stratégies, comme la génomique comparative, pour identifier de plus en plus de fonctions biologiques. Servie par les biostatistiques et la bioinformatique, elle permet d'analyser et d'exploiter ces données plus rapidement et les éventuelles interprétations dégagées seront susceptibles d'avoir des implications biologiques, agricoles et/ou médicales.

B.II *Le formatage des données.*

La mise en forme homogène des séquences protéiques étudiées (i.e. leur formatage) a pris une grande partie du temps de la thèse. Ce qui aurait pu paraître trivial s'est avéré fort complexe. En effet, les sources de données et leurs différents formats sont si hétérogènes qu'il est quasi impossible de transposer une méthode de formatage d'une source à l'autre. La vigilance est le mot d'ordre du traitement des séquences. Prenons un seul et unique cas, celui de l'organisme le mieux décrit, parce que le plus étudié, la bactérie *Escherichia coli* K-12. Malgré tous les efforts consentis à l'annotation de son génome, il reste très difficile de localiser sa source la plus fiable et la plus complète. Fort heureusement (mais ce n'est valable uniquement que pour *E. coli* K-12), la banque de données GenProtEC (Serres, Goswami et al. 2004), maintenue par Monica Riley, répond à nos attentes. L'utilisation conjointe des formats de séquences les plus usuels (i.e. EMBL, GenBank, Swiss-Prot, et leurs variants selon les mises à jour : ajout, suppression, modification de gènes, suppléments d'informations) rend extrêmement délicate (sinon impossible) la correspondance entre les différentes sources de données. Aussi, l'analyse des données d'un ensemble d'organismes est-elle le plus souvent réalisée à partir d'un format unique, ce qui limite bien évidemment les sources d'informations et leur diversité.

Dans les travaux suivants, à l'exception du protéome d'*E.coli* K-12, chaque protéome étudié provient de la banque EMBL (Kanz, Aldebert et al. 2005), ce qui a permis d'obtenir des séquences d'acides aminés (produits des formatages) homogènes d'un organisme à un autre. Il s'avère que les opérations

de formatage elles-mêmes sont plus simples à réaliser sur des données provenant de l'EMBL que provenant d'autres banques. Dans ces fichiers, de nombreuses informations ont pu être récupérées telles que la séquence protéique, son nom, sa longueur, sa fonction ou encore la position du gène correspondant sur le chromosome. Avant traitement par l'analyse factorielle des correspondances, seules les séquences de plus de cent acides aminés ont été retenues afin d'éviter des biais uniquement dus à la petite taille de ces protéines. De fait, une cystéine parmi cinquante autres acides aminés joue un rôle bien plus fort, en raison de sa rareté (moins de 1% des acides aminés utilisés), que lorsqu'elle est noyée au milieu de plusieurs centaines d'autres résidus. Successivement, de manière à masquer les contraintes liées aux processus d'initiation et de terminaison de traduction comme la sur-expression de résidus hydrophiles aux extrémités (Rocha, Danchin et al. 1999), toutes les protéines ont été tronquées de dix acides aminés en N-terminal et de cinq acides aminés en C-terminal .

B.III *L'analyse factorielle des correspondances, le principal outil.*

Les biais compositionnels des protéines ont été étudiés, entre autres, par l'Analyse Factorielle des Correspondances (AFC), une méthode statistique développée par J.P Benzécri (Benzecri 1973). Cette méthode établit une composition moyenne des entités étudiées et regroupe ou oppose ces entités suivant la composition de chacune d'entre elles. L'AFC va donc hiérarchiser l'information contenue dans un tableau de données. Elle va aussi bien s'intéresser à l'étude des colonnes (les acides aminés) qu'à l'étude des lignes (les protéines) pour confronter les différentes distributions, permettre de découvrir des irrégularités dans ces distributions et de mettre en évidence des combinaisons plus ou moins systématiques entre les variables ; en bref, de dégager des structures dans l'espace géographique étudié résultant de l'AFC, qui ne sont pas forcément linéaires. L'AFC, dans sa représentation spatiale, va offrir une simplification de l'information d'origine.

Toute AFC débute par un tableau de comptage (ou de contingence), c'est-à-dire, dans le cadre de nos travaux, un tableau où les lignes représentent les protéines du protéome étudié et les colonnes les acides aminés. Ainsi, dans chaque case du tableau se trouve le nombre d'acide aminé X de la protéine Y, et ce pour chaque acide aminé et chaque protéine. Ensuite des distances mathématiques entre chaque protéine et entre chaque acide aminé sont calculées (métrique du Chi2). Les protéines et les acides aminés sont ensuite représentés par des points dans un espace à plusieurs dimensions (nombre de colonnes ôté de 1 soit 19 axes). Une particularité intéressante de l'AFC est l'interchangeabilité entre les données des colonnes et des lignes, contrairement à l'Analyse en Composante Principale (ACP, méthode très proche de l'AFC) qui ne travaille que sur l'ensemble des colonnes. Ceci permet la représentation simultanée des deux nuages de points résultants de l'AFC, celui des protéines et celui des acides aminés, et offre ainsi la possibilité d'interpréter la localisation des éléments de l'un par rapport aux éléments de l'autre (Figure 1).

La superposition du nuage de protéines et du nuage d'acides aminés nous permet de mieux comprendre les attractions et les répulsions entre acides aminés et protéines. Deux protéines dont la composition en acides aminés est similaire seront donc très proches dans le nuage de points. Les éléments proches du centre de l'espace de l'AFC sont peu informatifs car ils correspondent à la moyenne, c'est-à-dire (i) des protéines dont la composition en acides aminés est proche de la composition moyenne en acides aminés du protéome et (ii) des acides aminés dont l'utilisation dans les protéines est proche de l'utilisation moyenne des acides aminés dans le protéome. Par opposition, les éléments éloignés du centre sont les plus atypiques donc les plus intéressants à étudier (Figure 1).

L'information (ou inertie) portée par les axes, formant le multi-espace, est dégressive d'axe en axe, le premier portant le plus d'information. L'inertie est exprimée en pourcentage sur chaque axe. On recherche bien évidemment à traiter le maximum d'information. Dans la plupart des travaux présentés dans ce manuscrit, l'analyse n'a pas été poursuivie au-delà du quatrième axe, les suivants ne portant pas d'information suffisamment significative. En moyenne, les quatre premiers axes représentent 50% de l'information totale.

La lecture des ouvrages suivants sera très utile à l'approfondissement et à une meilleure connaissance de la méthode de l'AFC :

- Thèse de doctorat d'Hélène Chiapello, université Paris VI, 2 juillet 1999,
- Thèse de doctorat de Stéphanie Bocs, université Paris VI, le 19 mai 2004,
- Livre « Statistique exploratoire multidimensionnelle ». L. Lebart, A. Morineau, M. Piron, 3^{e} édition, 2002, Dunod.

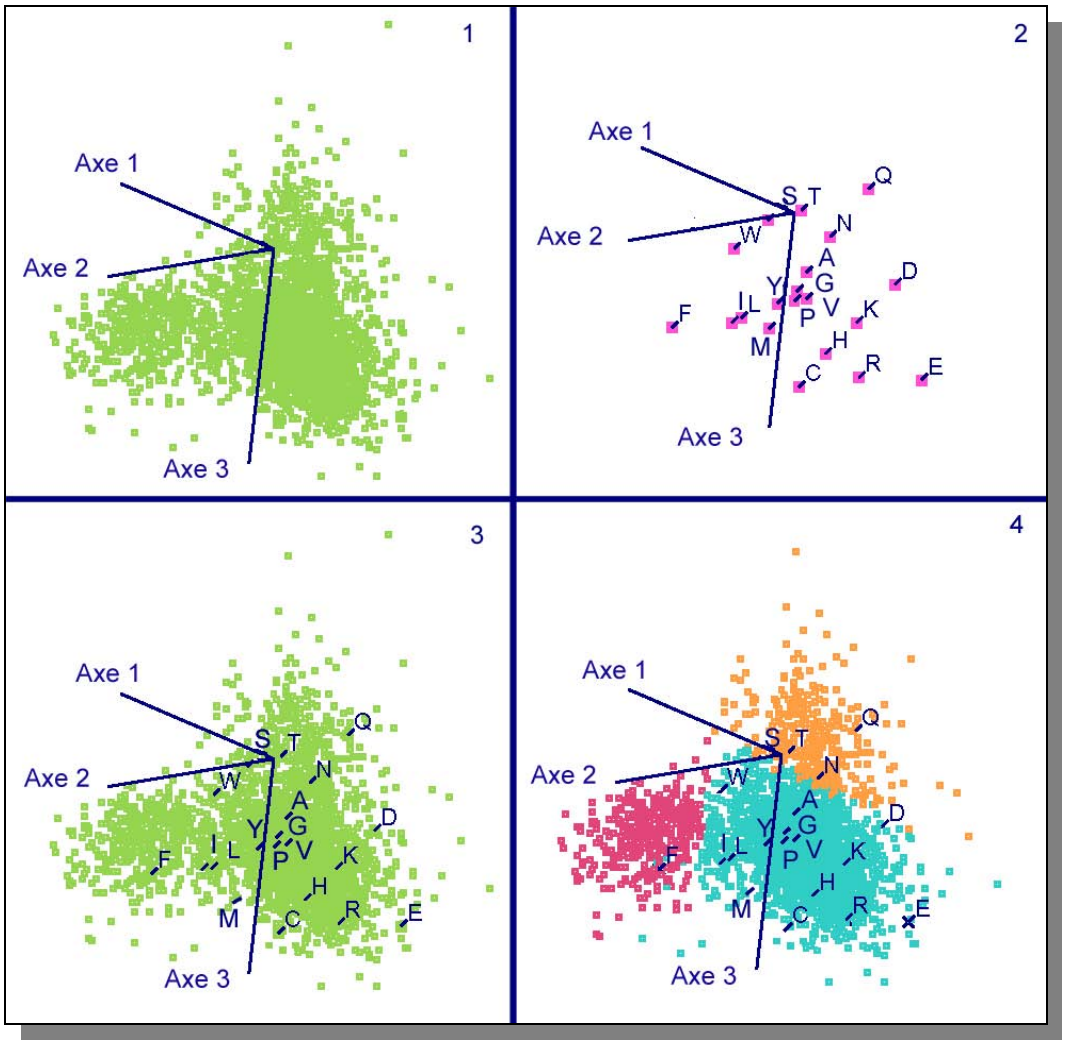


Figure 1 : Représentation graphique de l'AFC

1- **Chaque point représente une protéine (les lignes du tableau), le nuage de points se place dans un espace à 19 dimensions (ici visualisé en 3D) : deux points voisins ont une composition en acides aminés similaire.**

2- **Chaque point représente un acide aminé, le nuage de points ainsi formé se place dans un espace à 19 dimensions (ici visualisé en 3D).**

3- **Superposition de l'espace des protéines et des acides aminés.**

4- **Résultat de l'application de la méthode de partitionnement « les nuées dynamiques » qui aide à l'interprétation du résultat de l'AFC (Diday 1971).**

C Composition en acides aminés des protéines procaryotes : mise en scène, enquête et dénouement

C.I L'étude des procaryotes modèles, aromaticité et évolution.

C.I.1 Préambule.

Ce chapitre présente la première partie de l'analyse compositionnelle des protéines procaryotes ciblée tout particulièrement sur l'étude des trois organismes modèles microbiens *Escherichia coli* (bactérie à Gram négatif), *Bacillus subtilis* (bactérie à Gram positif) et *Methanococcus jannaschii* (archaeabactérie). Un modèle est un organisme dont les propriétés particulières facilitent son étude expérimentale. Son génome doit être de petite taille, son temps de génération court, sa culture (ou son élevage) doit être aisée et il doit assurer une descendance nombreuse. Ainsi, ces organismes ont toujours eu les plus grands égards et les quantités d'informations disponibles les concernant sont les plus riches. Se restreindre, dans un premier temps, à l'étude d'un si petit échantillon a permis de mettre en place de manière optimale les outils de calcul et d'analyse, les scripts informatiques et de déterminer le meilleur choix des banques, des bases de données et des sources d'informations.

C.I.2 L'essentiel de l'article.

Le nuage des points formés par les protéines d'un organisme, répartis dans un espace construit sur la composition en acides aminés, traduit une séparation selon la charge électrique des acides aminés, opposée à leur caractère hydrophobe. Cela crée une classe homogène bien identifiée, formée des protéines intégrales de la membrane interne. Un second biais est imposé par le contenu en G+C du génome, et sépare les protéines en fonction du contenu en G+C de la première position des codons correspondants aux acides aminés. Enfin, un rôle remarquable des acides aminés aromatiques crée un troisième biais universel. Les protéines "orphelines" sont enrichies en ces derniers, suggérant qu'ils ont un rôle privilégié dans la création de fonctions nouvelles au cours de l'évolution. Nous postulons que la majorité d'entre elles — les *gluons* — sont impliquées dans la stabilisation de complexes multimériques, et participent ainsi à la définition du "soi" de l'espèce.

C.I.3 Article 1.

Article paru dans le journal Proteins : Structure, Function, and Bioinformatics, le 1er juillet 2005.

PROTEINS: Structure, Function, and Bioinformatics 60:27–35 (2005)

Universal Biases in Protein Composition of Model Prokaryotes

Géraldine Pascal,[1,2]* Claudine Médigue,[1] and Antoine Danchin[2]
[1]*Genoscope/ CNRS UMR 8030, Atelier de Génomique Comparative, Evry, France*
[2]*Genetics of Bacterial Genomes, CNRS URA2171, Institut Pasteur, Paris, France*

ABSTRACT **The levels of cellular organization in living organisms are the results of a variety of selection pressures. We have investigated here the final outcome of this integrated selective process in proteins of the best known microbial models *Escherichia coli*, *Bacillus subtilis*, and *Methanococcus jannaschii*, supposed to have undergone separate evolution for more than 1 billion years. Using multivariate analysis methods, including correspondence analysis, we studied the overall amino acid composition of all proteins making a proteome. Starting from and further developing previous results that had pointed out some general forces driving the amino acid composition of the proteomes of these model bacteria, we explored the correlations existing between the structure and functions of the proteins forming a proteome and their amino acid composition. The electric charge of amino acids measured against hydrophobicity creates a highly homogeneous cluster, made exclusively of proteins that are core components of the cytoplasmic membrane of the cell (integral inner membrane proteins). A second bias is imposed by the G+C content of the genome, indicating that protein functions are so robust with respect to amino acid changes that they can accommodate a large shift in the nucleotide content of the genome. A remarkable role of aromatic amino acids was uncovered. Expressed orphan proteins are enriched in these residues, suggesting that they might participate in a process of gain of function during evolution. Proteins 2005; 60:27–35.** © **2005 Wiley-Liss, Inc.**

Key words: amino acids; hydrophobicity; GC content; aromaticity; orphans; multivariate analysis

INTRODUCTION

Living organisms are subjected to a variety of selection pressures that act not only at the level of the global phenotype but at each level of the cell's organization. It is usually assumed that proteins, for example, would be mainly subjected to selection pressures associated with their function. This is indeed the case, but if one compares proteins with identical function in organisms widely dispersed throughout evolution, one discovers that apart from a few 10's of amino acid residues, almost every residue of the protein can be replaced by any other amino acid, without dramatic change in the protein's function.[1,2] As a case in point in bacteria, proteins with similar functions will change in amino acid composition when their gene moves from the leading replicated strand to the lagging strand.[3] This not only indicates that proteins are extremely robust structures but also demonstrates that the chemical identity of each residue at most sites might integrate subtle cues derived from selection pressure operating at other places in the cell.[4] An amino acid in a protein integrates a preference for a certain DNA base composition, a preference for a codon, and a preference for that particular amino acid. This results from constraints at the level of DNA and RNA structure, as well as constraints at the level of nucleotide biosynthesis, or amino acid biosynthesis. Akashi and Gojobori have investigated the latter in a study suggesting that, indeed, the energy cost required to synthesize a given amino acid has some bearing on the overall composition of proteins.[5] Furthermore, Lobry has shown that differences in mutational bias can explain some variations in amino acid composition in bacterial species.[6] Starting from these studies that predated genome studies, we undertook a thorough analysis of the various factors that may affect amino acid composition in reference proteomes as a prerequisite to creating new ways to compare proteins from widely distant organisms, a notoriously difficult task. As a first step we analyzed the distribution of amino acids in proteins of model bacteria, where their function is best known, in contrast to that of the vast majority of proteins predicted from genome sequences.

An average protein of *Escherichia coli* K-12 is 300 residues long. If all things were kept equal, it should therefore have about 15 residues of each type of amino acid

Grant sponsor: Innovation and Technology Fund of the government of the SAR Hong Kong, China (program BIOSUPPORT), granted to A. Danchin for the creation of the HKU-Pasteur Research Centre. Grant sponsor: BioSapiens EU; Grant number: LSHG-CT-2003-503265. Grant sponsor: French Ministry of Foreign Affairs. Grant sponsor: French Centre National de la Recherche Scientifique; Grant numbers: CNRS-UMR 8030 and URA 2171. Grant sponsor: Institut Pasteur (Paris, France).

*Correspondence to: Géraldine Pascal, Genoscope/UMR 8030, Atelier de Genomique Comparative, 2 rue Gaston Cremieux, 91006 Evry Cedex, France, or to Antoine Danchin, Genetics of Bacterial Genomes, CNRS URA2171, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France. E-mail: gpascal@genoscope.cns.fr or adanchin@pasteur.fr

Received 4 August 2004; Accepted 14 January 2005

Published online 22 April 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20475

(of course, this also depends on the number of available codons, the metabolic cost of amino acid synthesis, chemical reactivity, etc.). Furthermore, if the amino acid content is normalized to the length of the protein and the number of proteins with a given density of a given residue is plotted, one would expect, all things being equal, to witness a normal distribution, with the exception of a small number of proteins with no residues of a given amino acid (the numbers considered are finite, and small). Whereas this is true for some amino acids (A, G, L, P, R, V), this is not true for others. In particular, the behavior of glutamate and aspartate is surprising: One observes a clear biphasic distribution, suggesting that proteins of *E. coli* K-12 form at least 2 classes. In addition to the presence of shoulders in some distributions, others are hyper-Gaussian (H, N, S, T, Y) (data not shown). This prompted us to study the amino acid composition in proteins using multivariate analysis methods.

Correspondence analysis (CA) has long been used to analyze the codon usage and amino acids composition bias of sequences of many organisms, in particular models of procaryotes.[7,8] As early as 1994, Lobry and Gautier explored the *E. coli* K-12 available data using this method, well before the complete sequence and other essential information on this organism was available.[9] Further studies substantiated that during replication, complementary DNA strands were subjected to different mutation constraints, which might be reflected at the level of amino acid composition in proteins.[10] More recently, global studies used this approach to extend investigation to a number of organisms without entering the details of the proteomes.[11–13] Taken together, these studies suggested that exploration of the detailed knowledge of genomes might allow significant integration of the statistical data with functional information about genes and genomes. This prompted us to analyze the *E. coli* K-12, *Bacillus subtilis*, and *Methanococcus jannaschii* proteomes to establish whether rules of amino acid composition might exist while placing them in a functional genomics perspective. In the present work, we correlate explicitly, using knowledge derived from genome programs, the rules that can be extracted from the statistical analysis with the function of gene and gene products as they are integrated in a coherent view of the cell. Rather than draw conclusions from uncertain annotation data, we have chosen to restrict our analysis to these 3 proteomes, because they represent well-known bacterial models of Gram-negative organisms, Gram-positive organisms, and the best known archaebacterium. Moreover, these 3 organisms are very distant in the phylogenetic tree; *E. coli* K-12 and *B. subtilis* are separated by more than 1.5 billion years. These 3 proteomes have been studied extensively for many years; consequently, the annotations in data banks are the most complete and the most correct. In choosing these 3 proteomes, we are avoiding the pitfall of basing our analysis on subcellular functions and localizations that are erroneous due to the increasingly rapid sequencing of genomes and the lack of experimental verification of information from these sequences.[14]

MATERIALS AND METHODS

Correspondence Analysis, Statistics, and Data Clustering

We used correspondence analysis[15] to identify the major factors that shape variation in amino acid usage among proteins of the organism of interest. These analyses were based on absolute frequencies in order to avoid introducing other biases.[16] Correspondence analysis was applied on the data table, including all proteins of an organism as described by their amino acid usage, to determine an orthogonal space, or factorial space, with dimension 19. The axes (called factors) are constructed according to the information they represent. They are presented in a decreasing order of importance as quantified by their corresponding "inertia."[17] Proteins and amino acids can be represented jointly in the obtained factorial space. Sequences that have a similar amino acid composition appear as neighbors.

As an additional tool, the CODONW software (http://www.molbiol.ox.ac.uk/cu/) was used for each of the 3 complete sequences, to help interpret the results. It computes the hydrophobicity levels (GRAVY score),[18] the G+C content of genes for specific proteins, and the aromaticity level (the relative frequency of aromatic amino acids) of each sequence, which are correlated with position on the main discriminating axis.

Data Sets

The complete proteome of *E. coli* K-12 is from the EcoGene17 database (http://bmb.med.miami.edu/ecogene/ecoweb/), which is known to be the most complete and reliable database for this organism. The data for *B. subtilis* and *M. jannaschii* used in this study are from the latest versions of the complete genomes in EMBL format, release 76 of 7 July 2003 and release 72 of 18 July 2002 respectively. They are available on the FTP site of the International Nucleotide Sequence Database (GenBank/EMBL/DDBJ) at http://www.ebi.ac.uk/genomes/.

In order to avoid constraints linked to the molecular processes of initiation and termination of translation, all proteins used in our study were truncated by 10 amino acids from their N-terminal end, and 5 amino acids from their C-terminal end (there is an over-representation of hydrophilic residues near both termini of proteins[19]). In order to reduce influence of stochastic variations that may occur in small proteins, only proteins longer than 100 residues (after truncating) were retained. After formatting, 3652 proteins from *E. coli* K-12, 3465 proteins from *B. subtilis*, and 1460 proteins from *M. jannaschii* were analyzed.

RESULTS

In this study, we used CA to explore the nature of the links that exist between the proteins forming the proteome of prokaryotes and the amino acid residues they contain, starting with the suggested biases created by hydrophobicity and by the local G+C content, as described in the literature.[9,11,12] New relations between amino acid composition and features of proteins were uncovered, such as a

bias created by the aromaticity of proteins. The proteomes of *E. coli* K-12, *B. subtilis*, and *M. jannaschii* have been analyzed in order to identify the common rules that drive amino acid composition of Gram-negative and Gram-positive prokaryotes and Archaea, and the differences that are typical of their metabolism or their structures. The data presented are analyzed using the best annotated models. Their generality has been substantiated by a similar analysis of the proteomes of a variety of other organisms (data not shown).

Inertia of CA

Inertia of CA can serve as a guide in determining the relative importance of a given factor. Regarding the *E. coli* K-12 data, about 41% of the inertia was distributed into the 3 first factors, and 43% for the *B. subtilis* data and for the *M. jannaschii* data. To measure their importance, these figures should be compared with an expected average inertia per axis of approximately 5% (100% of information would share equivalently in 19 axes). For this set of organisms, more than 75% of the information was present in the first 10 factors. CA was used to summarize and simplify the data. Analysis of the information carried in the CA was limited in the present study to the first 3 axes, which represent the most significant part of the whole information.

Hydrophobicity, a Discriminant Factor of Proteins

The distribution of proteins on the factorial plane made of the 2 first axes displayed 2 well-separated groups of proteins that stood out prominently (groups A and B), shown as individual clusters in all 3 model prokaryote proteomes. In *M. jannaschii*, a third small group was further revealed [Fig. 1(a–c)]. The contraposition of charged residues (E, D, and K) and of large hydrophobic amino acids (F and L) determined the clear separation of group A from group B. This discriminating factor strongly correlates with GRAVY score (*M. jannaschii*, $r = 0.87$, $p < 10^{-4}$; *E. coli* K-12, $r = 0.95$, $p < 10^{-4}$; *B. subtilis*, $r = 0.96$, $p < 10^{-4}$). Following the approach of Lobry and Gautier,[9] we identified the subcellular location of each known protein of *E. coli* K-12 and *B. subtilis* group A in Swiss-Prot databank (http://www.expasy.org) and GenProtEC database (http://genprotec.mbl.edu/). Thus, all group A proteins have an integral inner membrane location. Moreover, these integral inner membrane proteins (IIMPs) possess a transmembrane portion that makes at least 30% of the total length of the protein. Remarkably, all outer membrane proteins were found in group B. This indicated that proteins in group A are selected out of a very stringent property of their amino acid composition. After pooling the 3 proteomes together (Fig. 2), we could observe the strong conservation of group A in spite of the considerable difference in the envelope structure among the 3 organisms. These results substantiate and extend the previous observation of an hydrophobicity bias in the bacterial proteome[9,11,12] by providing an unambiguous link between the subcellular location of the proteins and their amino acid composition.

G+C Content Bias

As noticed in previous works, in addition to the bias driven by the hydrophobicity of the proteins, the overall G+C content of genomes of organisms appears to drive a second bias in the amino acid composition of the proteome. This feature is common to the 3 model prokaryotes studied in the present work. Contrary to expectation, however, this bias is not due to a bias in the second position of the codons, which is driving most of the physicochemical nature of the amino acids for which they code. Remarkably, the effect of the G+C content was apparently correlated to the first codon position (*M. jannaschii*, $r = 0.86$, *E. coli* K-12, $r = -0.59$, *B. subtilis*, $r = 0.71$; $p < 10^{-4}$ for each value). This bias seems to be the major bias for *M. jannaschii* (axis 1 of CA) and appears as the second and the third one for, respectively, *E. coli* K-12 and *B. subtilis* (axes 2 and 3). For *E. coli* K-12, this CA factor is driven by the opposition between by K and N (A+T rich codons) on one side and by L (neither A+T rich nor G+C rich codons) and R (G+C rich codon) on the other side. This probably accounts for the weaker G+C content correlation on that axis. Moreover, we find a correlation between discriminat- ing CA axes and A and C at the first nucleotide position of codons (respectively for A and C: *E. coli*: $r = 0.75$, $p < 10^{-4}$, $r = -0.81$, $p < 10^{-4}$; *B. subtilis*: $r = -0.59$, $p < 10^{-4}$, $r = 0.57$, $p < 10^{-4}$).

Aromaticity of Proteins

As we go from one axis to the next one, the weight of characters retained through evolution is progressively less prominent and influenced by the specific nature of the organism. Indeed, the biases reflected in the third axis differ in the 3 model organisms, splitting into 2 types of behavior: (1) the model Bacteria present a bias based on the aromaticity of proteins, found in axis 2 for *B. subtilis* and in axis 3 for *E. coli* K-12, and (2) the model Archaeon forms a third group of proteins, which is very clearly correlated with the percentage of cysteine in proteins along CA axis 3 ($r = -0.91$, $p < 10^{-4}$) [Fig. 1(c)]. A bias driven by the aromaticity of the proteins is contained in a further axis, with a rather weak contribution to inertia. This third group of *M. jannaschii* proteins was found to be constituted of proteins rich in cysteine residues (< 5% of proteome). To put aside this extreme bias, cysteine was removed from the CA. This resulted in the construction of a cloud of points where groups identical to the A and B groups found in Bacteria, were obtained (Figure 3). Interestingly however, the third CA axis of this new plot, driven on one side by residues F, L and on the other side by residues N, T, did not correlate significantly with any parameters investigated and will deserve further investigation, using other proteomes of Archaea. The fourth axis, having an inertia below 7%, was correlated with the aromaticity of proteins ($r = -0.48$, $p < 10^{-4}$). The low dispersion observed in these and subsequent axes did not warrant further consideration. The bacteria *B. subtilis* and *E. coli* K-12 have a much more pronounced correlation in favor of aromaticity of proteins due to opposition between residues A and G versus Y, W, and F (*B. subtilis*:

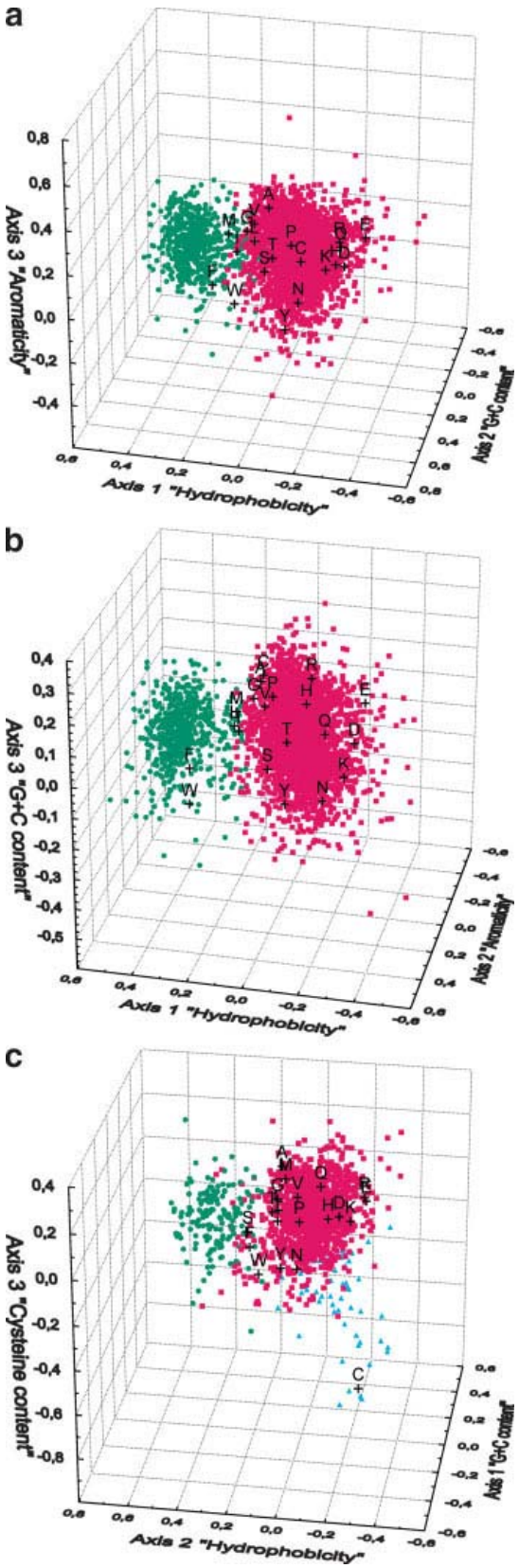


Fig. 1. Distribution of the protein sequences on the CA factorial space determined by the 3 first factors. Green circles represent proteins in group A; pink squares represent proteins in group B; and cyan triangles represent proteins in the third group of *M. jannaschii*. Amino acids are represented by black crosses. (**a**) For *E. coli* K-12, group A represents 18.8% of the analyzed proteome. (**b**) For *B. subtilis*, group A represents 19.8% of the analyzed proteome. (**c**) For *M. jannaschii*, group A represent 10.6% of the analyzed proteome.

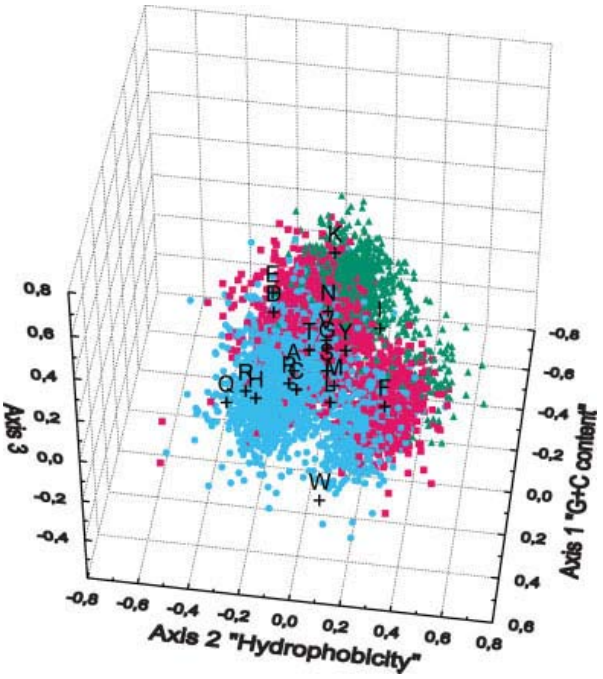


Fig. 2. The 3 first axes of CA of the 3 pooled proteomes: *E. coli* K-12 in blue circles; *B. subtilis* in pink squares; and *M. jannaschii* in green triangles.

r = -0.68, $p < 10^{-4}$; *E. coli* K-12: r = -0.72, $p < 10^{-4}$). In order to analyze the correlation between the bias in aromatic amino acids and protein function, the proteins situated at the extremities of the axis that discriminated aromaticity of proteins (2 x 10% of the total number of proteins) were extracted and compared with each other. We distinguished 2 classes of proteins, termed “high aromatic” and “low aromatic” classes. A large number of proteins of unknown function constituted the group of proteins rich in aromatic amino acids, whereas a number of proteins with housekeeping functions constituted the group of proteins in which aromatic residues are scarce. Especially noteworthy was the presence of ribosomal proteins in the latter group. Two classes of proteins with unknown function exist. They are either shared between a variety of related or less related genomes, or completely original to the genome of interest. In each of the extreme protein groups, the number of orphan proteins (proteins with no resemblance to any other protein in the databases) was studied by sequence alignment comparisons using BLASTP.[20] An orphan protein was defined as a protein with a sequence displaying a similarity score with other

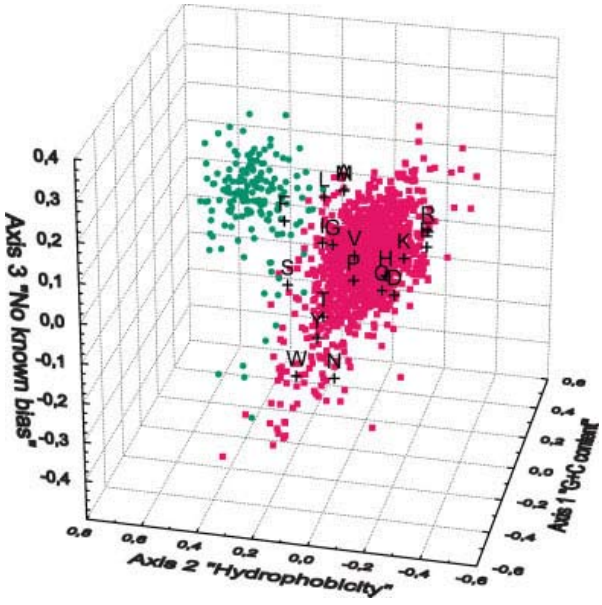


Fig. 3. The 3 first axes of CA of *M. jannaschii* when cysteine is omitted. Green circles represent proteins in group A (12% of analyzed proteome). Pink squares represent proteins in group B.

proteins lower than an *e*-value of less than 1^{e-3} in Swiss-Prot or in TrEMBL. In the case of *B. subtilis* and *E. coli*, this identifies 12.3% and 6.3% of "high aromatic" proteins as orphans, respectively, contrasting with, respectively, 2.02% and 0.55% of orphans in the "low aromatic" proteins. Using the transcriptome data of *B. subtilis* available from our previous work,[21] we checked whether orphan protein genes were expressed. Except for 4 proteins corresponding to genes that had not been included in the transcriptome data set, all of the orphan proteins of *B. subtilis* were found to be expressed (data not shown). These 2 classes were compared using a test of comparison of proportions based on observed proportions. This test revealed that "high aromatic" proteins were enriched in orphan proteins. It was also noted that, for *B. subtilis* and *E. coli* K-12 respectively, 38% and 44% of the orphan proteins were found in the 10% of "high aromatic" proteins retained for the test. A Wilcoxon test allowed us to show that the set of orphan proteins of *B. subtilis* and *E. coli* K-12 was richer in aromatic amino acids than the ensemble of other proteins (non-orphans) (*E. coli* K-12: $p < 5 \cdot 10^{-5}$; *B. subtilis*: $p < 2 \cdot 10^{-9}$).

DISCUSSION

CA With a Very Significant Inertia

In this study, meant to substantiate and extend previous work to correlate amino acid composition of model proteomes to structural and functional properties of their proteins, we restricted most of our analysis to the information given by the 3 first axes of CA. Furthermore, in order to refine exploration, when a property was clearly driven by a single amino acid, we extended the study by performing CA with the set of amino acids lacking the biasing one. Fortunately, the main characteristics of proteomes appeared factor by factor, unfolding in a consistent manner. A first factor linked protein function and hydrophobicity. A second one emphasized the proteomic relationship between the G+C content of the genome and the related genes. And a third factor shifted focus to the aromaticity of proteins in association to the origins of their functions.

Hydrophobicity Versus Electric Charge of Proteins Characteristic of Their Subcellular Location

Previous works either classifying proteomes as bulk entities[13] or analyzing partial or global proteomes[9,22] suggested that the proteome was subject to a bias driven by hydrophobicity. We show here that CA of the amino acid content of the proteome discriminated prokaryotic proteins according to hydropathy in a way that was consistent with the subcellular localization of proteins. Proteins that were known to be integral components of the cytoplasmic membrane all clustered into a uniquely defined group. Remarkably, this group was consistently observed in the 3 model organisms we analyzed despite their enormous phylogenetic divergence. It is known that the envelopes of Gram-positive and Gram-negative bacteria differ mainly in their structure. Gram-positive bacteria usually have a thick peptidoglycan multilayer on their surface, at the exterior of the cytoplasmic membrane, whereas Gram-negative bacteria have a more complex structure. The latter have 2 membranes, of which 1 is at the surface of the peptidoglycan layer, with the following order (from interior to exterior): plasma (inner) membrane, peptidoglycan, and external (outer) membrane. The membrane structure is contradictory; it plays the role of a filter and of a transporter at the same time. It is semipermeable, letting small molecules such as sugars or salts enter, but prevents the entry of large molecules such as proteins. Membranes of Archaea differ from those of Bacteria because they lack peptidoglycan as a constituent of their membranes. Furthermore, their plasma membrane presents a lipid composition which is different from that of all other organisms.[23] Despite these considerable structural differences, we observed that proteins imbedded in the plasma membrane of the 3 model prokaryotes were characterized by proteins that were quite different from the other proteins of the organism in their amino acid composition.[24] This distinction is seen in the composition of groups A and B of the CA, and it is so clear cut that CA followed by dynamic clustering could be suggested as a straightforward tool for the identification of IIMP. As substantiation we analyzed the behavior of 83 proteomes from a variety of genomes and found the exact same splitting of the proteome into at least 2 groups, A and B (data not shown). In group A, the scattering observed in the amino acid composition space shows a very clear separation: hydrophobic residues versus charged residues. This observation substantiated other studies describing the structure and amino acid composition of membrane proteins[25]: Hydrophobic residues (F, L, and I) show a clear preference for the transbilayer region, whereas charged residues (E, K, D, and R) are located on the extracellular side of the membrane (basic residues

preferably inside because of the negative electric potential on the inside of the cell).

Generally, the solubility in water of the lateral chains of the amino acids increases with the electric charge that they possess, and their capacity to establish hydrogen bonds. The hydrophobic transmembrane segments of the polypeptide chains of the IIMPs are essentially composed of amino acid residues with non-polar lateral chains; the peptide bonds between these amino acids are, however, polar themselves, and when they are immersed in the lipid bilayer (away from water), they tend to establish hydrogen bonds between each other, which causes refolding of the segment as an α-helix (20 –30 amino acid residues). α-helix IIMPs make 4 main classes of functions: energy transduction, ion channeling without adenosine triphosphate (ATP) consumption, water diffusion (aquaporin), and active transport of molecules (requiring ATP). This universality in amino acid composition of α-helical IIMPs could be explained by the huge selective constraints imposed by the essential role played by these proteins in the cell, by their lipidic environment, and by the large electrostatic potential (usually 100,000 volts/cm) to which they are submitted.[26] In contrast to the first group, group B is not a homogeneous class, and it is sometimes split into other classes. It contains also some membrane-associated proteins, as the outer membrane proteins in Gram-negative bacteria and some proteins at least partially imbedded in the inner membrane, most of which are sensor and receptor proteins. These latter proteins possess either a large cytoplasmic or external domain, or both, and this accounts for their being excluded from the narrow class of integral inner membrane proteins. The proteins with large transmembrane domains (>30% of the total length of the protein), which are in group B and not in group A, were either integral outer membrane proteins (OmpF, OmpG, OmpX, OmpT, LamB) with β-barrel structures, or unknown proteins (YcdP: probably a permease component for hemin transport, defective in *E. coli* K12; YibH: could belong to an efflux pump). They differ from IIMPs by their code, such as the non-random frequency of the tripeptide motif: aromatic–random–aromatic of porins with β-barrel structures.[27] In summary, the present study brings about a novel feature of a particular consistent class of IIMPs that have key cellular roles, while they can be distinguished from all others membrane protein types. This particular feature will be useful for functional annotation of new genomes, by restricting the domain of hypotheses about putative functions of unknown proteins to a narrower set.

Genome G+C Content Biasing Amino Acid Proteome Composition

The genome's G+C content is known to bias the amino acid proteome composition[28] (i.e., a significant part of the choice of preferred amino acids is not determined by the selection pressure on the proteins, but rather on the genome). Indeed, we observed a corresponding correlation on the discriminant axes for each of the 3 organisms, although the selection pressure, which would have been expected to be mostly driven by the second codon position (the most discriminating one), was found to come from the first position instead. Previous work has documented that, in the majority of thermophiles, the genome's G+C% is the first factor that influences the amino acid composition, as reported, for example, by Lynn et al. in 2002 based on an analysis of usage correspondences of synonymous codons.[29] This may explain why the codon usage bias driven by the genome's G+C% determining a specific amino acid composition of proteins in *M. jannaschii* is greater than the variability due to selection pressure on the subcellular localization of proteins.

The thermophilic biotope of the organism is not the sole cause of this bias, as seen with axis 2 of the CA of *E. coli* K-12 proteome and axis 3 of the CA of *B. subtilis* that split their proteome according to the G+C content of the codons. In previous studies based on incomplete proteomes, Palacios and Wernegreen[7] and Lobry and Gautier[9] identified a somewhat similar axis 2 (conjugated to axis 3 in Palacios and Wernegreen) in the *E. coli* K-12 and interpreted it as the result of the selection pressure acting on the proteins' gene expression. The present results do not substantiate this interpretation. Indeed, the bias due to adenine and cytosine at the first nucleotide position of codons of genes, that would give proteins rich in L, P, H, and Q and poor in I, M, T, N, K, and S amino acids or, inversely, does not appear to have a straightforward interpretation. In any event, this bias does not involve the level of gene expression, as demonstrated by the weak correlation of the Codon Adaptation Index (CAI) with discriminant axes ($r = 0.22$, $p < 10^{-4}$ for axis 2 of *E. coli* K-12 and $r = 0.16$, $p < 10^{-4}$ for axis 3 of *B. subtilis*). The CAI score measures empirically the synonymous codon usage bias. It is positively correlated with the expressivity of a gene.[30]

What could be the impact of a bias of the first nucleotide in the codon for an amino acid in a proteome? Enrichment in A+T at the first position of codons indicates a specific selection pressure that is not simply driven by the hydrophilicity/hydrophobicity controlled by the second codon position. It separates between V, A, D, E, G, P, H, and Q, and F, S, Y, C, W, I, M, T, N, and K, with L and R not discriminating. This partially matches a shift from amino acids with a small volume to those with a large volume, from those with no sulfur to those containing sulfur, from non-aromatic to aromatic and to the acquisition of serine.[31] This trend is at a cost: The richer an organism is in A+T at the first position of codons, the higher the metabolic cost of amino acid production. However, this can also be an advantage. Indeed, it is well established that pathogens, symbionts, as well as phage and plasmid sequences and insertion sequences, are richer in A+T.[32] And, although. the high G+C content of genomes is associated with low carbon content in proteins, pathogens, and bacterial symbionts, even those that are A+T-rich, are less subject to limitations of carbon in the environment, because they are linked to nutritional resources furnished by their host[33] and thus stay competitive in terms of metabolic resources. A+T enrichment may also derive from other selective constraints: Hyperthermophiles such as *M. jannaschii* are

often A + T-rich despite the greater heat stability of the G+C pair. This bias makes these organisms more resistant to cytosine deamination (C ⟶ U), which is activated at high temperature,[34] and they are thus more likely to retain their original sequence. It has also been demonstrated in numerous studies that enrichment in A+T is evidence of recent evolution. Based on results of a study by Sorimachi,[35] serine, considered to be one of the important residues in biological evolution, exhibits increased concentration as evolution proceeds in all organisms studied. It should also be noted that serine contributes to the formation of new protein functions during evolution.[35] This argument is further supported by the work of McDonald et al.,[36] which compares several proteins of mesophiles and thermophiles of *Methanococcus* and *Bacillus* species. This study demonstrates that, in the case of *Methanococcus*, enrichment in A+T is not necessarily due to changes in the temperature of the environment, but rather to the evolution of species.

In summary, the present study uncovers a remarkable impact of the genome G+C content on the corresponding proteome that does not display any significant association either with the optimal growth temperature of the organism or its gene expressivity.

Aromaticity, Source of Novel Proteins

The two first factors explaining the amino acid composition of the prokaryote proteomes were universal. In contrast, the third discriminant factor, at first sight, was not: The Archaeon *M. jannaschii* exhibits a different behavior compared to the other two model Bacteria. Indeed, the third axis of CA of the *M. jannaschii* proteome presented a strong bias driven by the cysteine composition of the proteins. It has been reported that CXXC clusters are more frequent at high temperature, which would allow thermophiles like *M. jannaschii* to somehow protect these residues, as isolated cysteines would otherwise have a tendency to decrease with increasing temperature.[37] It is therefore possible that enrichment in cysteines is due to formation of clusters (perhaps stabilized by metals or formation of sulfur bonds). This would perhaps explain the special bias created by the strong presence of cysteine in a certain number of *M. jannaschii* proteins. Because of this effect, which is particular to this organism, we explored whether CA would once again lead to some kind of universal constraint on the proteomes, omitting cysteine from the analysis. When the cysteine bias was suppressed, we observed that the third axis of the *M. jannaschii* proteome was formed by the opposition of the L and F amino acids to the N and T amino acids. Interestingly, these oppositions suggested a separation between proteins integrated into the membrane (biased in L and F) and excreted proteins. Indeed, Perrière and Thioulouse surmised that in Gram-negative bacteria, the periplasmic proteins were characterized by N, P, Q, and T residues,[16] which are known to slow the folding of proteins, a phenomenon required for proteins that have to be exported. Further exploration is needed to see whether this observation could be used to help predicting excreted proteins in *M. jannaschii*. The next CA axis in *M. jannaschii* correlated with the aromaticity of proteins ($r = 0.48$, $p < 10^{-4}$), suggesting a particular role of aromatic amino acids. This role was much stronger in *E. coli* K-12 and in *B. subtilis*, where an amino acid composition bias opposed aromatic amino acids to A and G residues allowing separation of housekeeping proteins from orphan proteins. It is worth remarking that aromatic amino acids' defective proteins do not constitute a functionally random class of proteins, since ribosomal proteins make the bulk of this class. Ribosomes must belong to the very first organelles that made the first cells, and it is most likely that ribosomal proteins were present in the ancestral proteomes, providing them with a long time for evolution. In this context, it is interesting that ribosomal proteins from Archaea differ significantly in amino acid content from those of Bacteria, because some are common with those of Eukaryotes while they are absent from Bacteria.[38] This might account for the observation that the separation between proteins rich in aromatic amino acids and ribosomal proteins did not stand out prominently in *M. jannaschii*. Ribosomal proteins are essential to the cell's life, and because protein synthesis is at the core of macromolecule metabolism, they must be expressed at a high level. These proteins are enriched in basic amino acids (K, R: they interact with RNA) and in small hydrophobic residues (A, V, G). The latter have a low metabolic cost. This goes in the direction of the selection pressure for high expression of proteins.[39] Furthermore residues A and G are often considered as belonging to the first amino acids present at the origin of the first cells, and this is consistent with the presumably primitive nature of ribosomal proteins. This contrasts with enrichment in aromatic amino acids, which are considered newcomers that progressively invaded the genetic code.[40] Because only a few codons code for aromatic amino acids, and because of their large metabolic cost, they are usually rare in proteins, especially in highly expressed proteins.[5,7] They have, however, extremely interesting physicochemical properties that are witnessed frequently in the 2D and 3D structure of proteins.[9,41,42] Indeed, aromatic residues often interact with one another, in particular in α-helices,[24] usually in an orthogonal interaction (edge-to-face).[43] They have been recruited for a variety of roles in the interaction between protein subunits and with nucleic acids. In particular, it has been observed that transmembrane α-helices are richer in aromatic amino acid residues than cytoplasmic α-helices,[24] strongly suggesting that aromatic amino acids play a particularly important stabilizing role in an hydrophobic context. These residues would play a major role in the structure and function of membrane proteins by allowing interaction between the exposed faces of the α-helices. They would also allow stable anchoring of proteins to the membrane.[24] Membranes must be flexible and must adjust rapidly to variations in the environment, and aromatic amino acids might provide them with the necessary flexibility while conserving stable interactions. Remarkably, we observed not only that small hydrophobic residues were opposed to aromatic amino acids in the overall pattern of protein composition organi-

zation, but also that ancestral housekeeping proteins were opposed to orphan proteins. The latter can either be considered as old, but evolving extremely fast, or recent acquisitions or creations. The opposition between A, G, V, and Y, F and W, as well as the amino acid composition of housekeeping proteins, would strongly argue in favor of the latter hypothesis with orphan proteins as recent acquisitions. This is indeed the general consensus about their origin (some think that they might also be coded by pseudogenes, but pseudogenes, deriving from ancestral genes, should not differ in amino acid composition from the bulk of the genes).[44] Orphan proteins are usually small (150 residues) and their genes are A+T-rich in the third codon position.[45] They are often present in bacteriophages. Furthermore, through the processes of recombination/ excision, they might generate sequences from bits and pieces that would not have counterparts anywhere else. Genomes, like all living processes, are subject to a process of selective stabilization, and those proteins, when recruited for a function increasing somehow the fitness of their host (despite their metabolic cost), would stay there. The very fact that they are "orphans" creates a discrimination between the self and the non-self of the species. One is therefore compelled to uncover a function that would have some mark of "self" for an organism. Proteins are certainly not isolated in the cytoplasm of cells: They must both interact with a precise set of subunits or other factors and avoid interacting with unrelated proteins. We propose that many of those orphan proteins could be factors promoting this stabilization process (i.e., being non-catalytic subunits of complexes). The versatility of aromatic amino acids residues that are present in these "gluons" would provide them with a quick adaptation process that would allow them to be recruited frequently, thus accounting for the apparent ubiquity of orphans in genomes. Once recruited, they would slowly evolve to less costly material by losing their amino acid residues as time elapses, thereby forming classes of well-adapted proteins that would now be transmitted by horizontal transfer from organism to organism, while losing their status of orphan would be helped by intervention of phages, which constitute the major vector of introduction of proteins in genomes.[32,46]

CONCLUSIONS

Despite their essential role in catalysis and protein folding, most amino acid residues in proteins are not subject to such dramatic selection pressure that would link them to the function of the proteins they encode. As a matter of fact, usually less than 10% of the residues are submitted to strict functional constraints. Analyzing the 3 model proteomes of Gram-positive and Gram-negative Bacteria, as well as the model for Archaea, we uncovered two universal biases that affect proteome amino acid composition and a third one that is Bacteria-specific. These results were also observed on a large set of 80 proteomes (data not shown). Proteins integrated in the inner membrane can easily be identified from the bulk; a particular constraint, of still unknown nature, drives the amino acid content of the proteome, as a function of the G+C content of the first position of their codon; and, finally, orphan proteins are unusually rich in aromatic amino acid residues. Preliminary observations suggest that these rules are ubiquitous. Deeper analysis will help us understand the nature of the selection pressure driving these universal biases.

ACKNOWLEDGMENTS

We thank Stephane Cruveiller, Eduardo Rocha, and Cédric Cabau for their critical comments and suggestions, and Susan Cure for her help in writing the manuscript.

This work was initiated as a core genomics program at the HKU-Pasteur Research Centre in Hong Kong.

REFERENCES

1. Beyer A. Sequence analysis of the AAA protein family. Prot Sci 1997;6:2043-2058
2. Ma B, Wolfson HJ, Nussinov R. Protein functional epitopes: hot spots, dynamics and combinatorial libraries. Curr Opin Struct Biol 2001;11:364 –369.
3. Rocha EP, Danchin A. Ongoing evolution of strand composition in bacterial genomes. Mol Biol Evol 2001;18:1789 –1799.
4. Dean AM. Selection and neutrality in lactose operons of *Escherichia coli*. Genetics 1989;123:441– 454.
5. Akashi H, Gojobori T. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc Natl Acad Sci USA 2002;99:3695–3700.
6. Lobry JR. Influence of genomic G C content on average amino- acid composition of proteins from 59 bacterial species. Gene 1997;205:309 –316.
7. Palacios C, Wernegreen JJ. A strong effect of AT mutational bias on amino acid usage in *Buchnera* is mitigated at high-expression genes. Mol Biol Evol 2002;19:1575–1584.
8. Guerdoux-Jamet P, Henaut A, Nitschke P, Risler JL, Danchin A. Using codon usage to predict genes origin: is the *Escherichia coli* outer membrane a patchwork of products from different genomes? DNA Res 1997;4:257–265.
9. Lobry JR, Gautier C. Hydrophobicity, expressivity and aromatic- ity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucleic Acids Res 1994;22:3174 – 3180.
10. Lobry JR, Sueoka N. Asymmetric directional mutation pressures in bacteria. Genome Biol 2002;3(10).
11. Dumontier M, Michalickova K, Hogue CW. Species-specific protein sequence and fold optimizations. BMC Bioinformatics 2002;3.
12. Kreil DP, Ouzounis CA. Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res 2001;29:1608 –1615.
13. Tekaia F, Yeramian E, Dujon B. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. Gene 2002;297:51– 60.
14. Gilks WR, Audit B, De Angelis D, Tsoka S, Ouzounis CA. Modeling the percolation of annotation errors in a database of protein sequences. Bioinformatics 2002;18:1641–1649.
15. Benzecri J-P. L'analyse des donne´ es, L'Analyse des Correspondances. Paris, France: Dunod Edition; 1973.
16. Perrière G, Thioulouse J. Use and misuse of correspondence analysis in codon usage studies. Nucleic Acids Res 2002;30:4548 – 4555.
17. Lebart T, Morineau A, Warwick KA. Multivariate descriptive statistical analysis. New York: Wiley; 1984.
18. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132.
19. Rocha EP, Danchin A, Viari A. Translation in *Bacillus subtilis*: roles and trends of initiation and termination, insights from a genome analysis. Nucleic Acids Res 1999;27:3567–3576.
20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403– 410.
21. Sekowska A, Robin S, Daudin JJ, Henaut A, Danchin A. Extracting biological information from DNA arrays: an unexpected link between arginine and methionine metabolism in *Bacillus subtilis*. Genome Biol 2001;2(6).

22. Rispe C, Delmotte F, van Ham RC, Moya A. Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids. Genome Res 2004;14:44 –53.
23. Cavalier-Smith T. The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. Int J Syst Evol Microbiol 2002;52:7–76.
24. Koshi JM, Bruno WJ. Major structural determinants of transmembrane proteins identified by principal component analysis. Proteins 1999;34:333–340.
25. Ulmschneider MB, Sansom MS. Amino acid distributions in integral membrane protein structures. Biochim Biophys Acta 2001;1512:1–14.
26. Nilsson I, Johnson AE, von Heijne G. How hydrophobic is alanine? J Biol Chem 2003;278:29389 –29393.
27. Wimley WC. Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. Protein Sci 2002;11:301–312.
28. Sandberg R, Branden CI, Ernberg I, Coster J. Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino acid usage and G+C content. Gene 2003;311:35– 42.
29. Lynn DJ, Singer GA, Hickey DA. Synonymous codon usage is subject to selection in thermophilic bacteria. Nucleic Acids Res 2002;30:4272– 4277.
30. Sharp PM, Li WH. The Codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 1987;15:1281–1295.
31. Wilquet V, Van de Casteele M. The role of the codon first letter in the relationship between genomic GC content and protein amino acid composition. Res Microbiol 1999;150:21–32.
32. Rocha EP, Danchin A. Base composition bias might result from competition for metabolic resources. Trends Genet 2002;18:291–294.
33. Baudouin-Cornu P, Schuerer K, Marliere P, Thomas D. Intimate evolution of proteins: proteome atomic content correlates with genome base composition. J Biol Chem 2004;279:5421–5428.
34. Frederico LA, Kunkel TA, Shaw BR. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry 1990;29:2532–2537.
35. Sorimachi K. Evolutionary changes reflected by the cellular amino acid composition. Amino Acids 1999;17:207–226.
36. McDonald JH, Grasso AM, Rejto LK. Patterns of temperature adaptation in proteins from *Methanococcus* and *Bacillus*. Mol Biol Evol 1999;16:1785–1790.
37. Rosato V, Pucello N, Giuliano G. Evidence for cysteine clustering in thermophilic proteomes. Trends Genet 2002;18:278 –281.
38. Lecompte O, Ripp R, Thierry JC, Moras D, Poch O. Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. Nucleic Acids Res 2002;30: 5382–5390.
39. Lin K, Kuang Y, Joseph JS, Kolatkar PR. Conserved codon composition of ribosomal protein coding genes in *Escherichia coli*, *Mycobacterium tuberculosis* and *Saccharomyces cerevisiae*: lessons from supervised machine learning in functional genomics. Nucleic Acids Res 2002;30:2599 –2607.
40. Brooks DJ, Fresco JR. Increased frequency of cysteine, tyrosine, and phenylalanine residues since the last universal ancestor. Mol Cell Proteomics 2002;1:125–131.
41. Thomas A, Meurisse R, Brasseur R. Aromatic side-chain interactions in proteins: II. Near- and far-sequence Phe-X pairs. Proteins 2002;48:635– 644.
42. Thomas A, Meurisse R, Charloteaux B, Brasseur R. Aromatic side-chain interactions in proteins: I. Main structural features. Proteins 2002;48:628 – 634.
43. Hunter CA, Singh J, Thornton JM. Pi–pi interactions: the geometry and energetics of phenylalanine–phenylalanine interactions in proteins. J Mol Biol 1991;218:837– 846.
44. Domazet-Loso T, Tautz D. An evolutionary analysis of orphan genes in Drosophila. Genome Res 2003;13:2213–2219.
45. Daubin V, Ochman H. Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. Genome Res 2004;14:1036 –1042.
46. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, Jacobs-Sera D, Falbo J, Gross J, Pannunzio NR, Brucker W, Kumar V, Kandasamy J, Keenan L, Bardarov S, Kriakov J, Lawrence JG, Jacobs WR Jr, Hendrix RW, Hatfull GF. Origins of highly mosaic mycobacteriophage genomes. Cell 2003; 113:171–182.

C.I.4 Zoom sur le biais aromatique des protéines orphelines.

C.I.4.a Les résultats.

Un des résultats les plus attrayants de cet article est le phénomène d'enrichissement des protéines orphelines par les acides aminés aromatiques. Aussi bien pour *E. coli K-12* que pour *B. subtilis*, un des facteurs de l'AFC corrèle fortement avec l'aromaticité des protéines localisées le long de cet axe (Figure 2). Pour faciliter l'analyse de ce type d'observation, il est usuel de se concentrer plus particulièrement sur les fonctions des protéines situées aux extrémités du nuage de l'AFC, l'extrémité qui contient les protéines très pauvres en acides aminés aromatiques étant opposée à celle qui contient les protéines les plus riches. Ainsi, deux dixièmes de la totalité des protéomes d'*E. coli* K-12 et de *B. subtilis* ont été étudiés scrupuleusement. Contrairement à ces observations, l'aromaticité des protéines de *M. jannaschii* ne corrélaient ni avec le deuxième axe (comme *E. coli* K-12) ni avec le troisième (comme *B. subtilis*), mais avec le cinquième seulement. Ce résultat semblait donc moins probant en ce qui concerne l'archaeabactérie et dans le souci de ne travailler que sur des biais significatifs, l'étude de ce facteur chez *M. jannaschii* n'a pas été poursuivie.

Après analyse, on observe que les protéines sont d'origines très différentes selon qu'elles se situent dans l'une où l'autre des extrémités. Le groupe de protéines pauvres en acides aminés aromatiques mais riches en résidus alanine et glycine, est constitué principalement de protéines de ménage dont les protéines ribosomales (groupe rose). Le groupe opposé, riche en acides aminés aromatiques (groupe vert), se compose presque exclusivement de protéines dont la fonction est inconnue. Nous avons donc cherché à identifier la proportion des protéines orphelines parmi ces protéines de fonction inconnue, une protéine orpheline étant une protéine dont on ne trouve pas de protéines similaires dans les banques de données. Si nous prenons le cas de *B. subtilis* (organisme pour lequel nous avions à disposition des données d'expression), nous observons, après alignement des séquences (Altschul, Gish et al. 1990), que les protéines orphelines représentent 12% du groupe vert, soit 38% des protéines orphelines totales, contre seulement 2% dans le groupe opposé. Afin de confirmer ce premier résultat, un test de moyennes a été réalisé (Figure 3). Il en résulte que l'ensemble des protéines orphelines a une composition en acides aminés aromatiques significativement plus riche que le reste du protéome. A titre indicatif, le Tableau I présente les proportions des différents acides aminés aromatiques dans le protéome de *B. subtilis*, dans les protéines les plus aromatiques et les moins aromatiques. Ces données traduisent que l'incorporation de ces résidus n'est pas du tout homogène.

Tableau I : contenu en acides aminés aromatiques des protéines de *B. subtilis*.

	Protéome	Protéines les plus aromatiques	Protéines les moins aromatiques
% Acides aminés aromatiques	**9,15**	**17,95**	**3,06**
% Phénylalanine	**4,60**	**9,96**	**1,64**
% Tryptophane	**1,03**	**2,33**	**0,16**
% Tyrosine	**3,53**	**5,65**	**1,26**

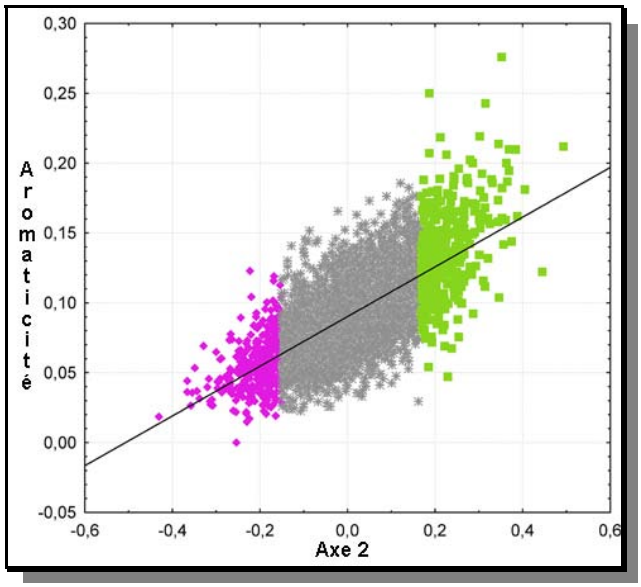


Figure 2 : Corrélation du facteur 2 de l'AFC de *B. subtilis* et de l'aromaticité des protéines situées le long de cet axe. Coefficient de corrélation : r = -0,68 ; p < 10^{-4}

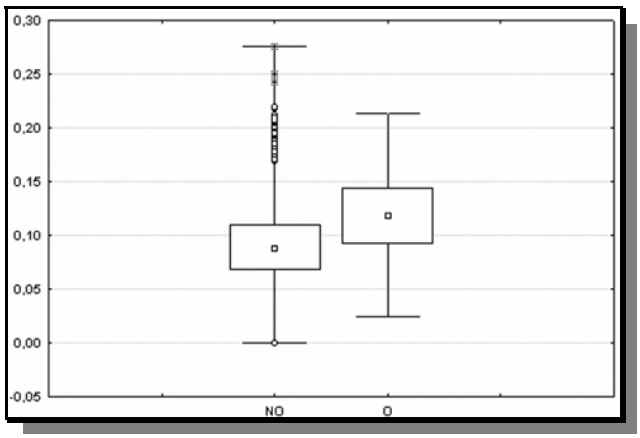


Figure 3 : « Boites à moustaches ». Moyenne de l'aromaticité entre les groupes des protéines non orphelines (NO) et orphelines (O) de *B. subtilis*. Moyenne de NO : 0,0906 ; Moyenne de O : 0,1188 ; t-value : -9,1822 ; p-value : > 10^{-4}

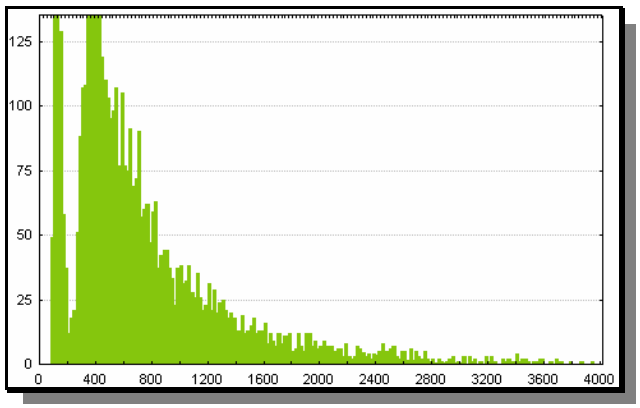


Figure 4 : Distribution de la radioactivité des transcrits de *B. subtilis* (nb d'obs. / unité arbitraire de radioactivité)

C.I.4.b Protéines orphelines ou pseudo-gènes ?

Etant donné le biais si fort en acides aminés aromatiques des protéines détectées orphelines à l'aide du logiciel BLAST (Altschul, Gish et al. 1990), il a été légitime de se demander si ces dernières sont de véritables protéines ou seulement les produits de pseudo-gènes. A l'aide des données de transcriptome de *B. subtilis* (travaux d'Agnieszka Sekowska, Institut Pasteur), nous avons voulu déterminer si les gènes des protéines orphelines détectées étaient exprimés ou non. La Figure 4 représente la distribution de la radioactivité des transcrits de *B. subtilis* (synthèse des ADNc en présence de dCTP marqués au P^{33}). Une zone de bruit est toujours présente, elle s'étend de 0 à 210 unités arbitraires de radioactivité (uar). Les ARN, quant à eux, sont répartis entre 210 et plus de 4000 uar. Les produits des gènes des protéines orphelines ont été trouvés entre 252 et 3064 uar. On peut donc penser que ces derniers sont bien transcrits au sein de l'organisme. Bien entendu, seule la détection, des protéines orphelines dans la cellule pourra certifier que ces transcrits sont bien traduits.

C.I.4.c Pourquoi un tel usage des acides aminés aromatiques dans les protéines orphelines ?

Un des axes de l'AFC corrèle avec l'aromaticité des protéines de *M. jannaschii*, *E. coli* K-12 et *B. subtilis* ce qui suggère un rôle particulier des résidus aromatiques. Ce rôle s'exprime bien plus fortement chez *E. coli* K-12 et *B. subtilis* où l'aromaticité des protéines s'oppose aux acides aminés alanine et glycine. Le biais sépare les protéines de ménage des protéines orphelines, de ce fait l'ensemble des protéines faiblement aromatiques constitue une classe fonctionnelle homogène où se trouvent principalement les protéines ribosomales. Ces dernières sont essentielles à la vie cellulaire, et parce que la synthèse des protéines constitue le cœur du métabolisme des macromolécules, elles doivent être hautement exprimées (Karlin, Mrazek et al. 1998; Karlin et Mrazek 2000). Ces protéines sont enrichies en acides aminés basiques (lysine et arginine qui interagissent avec l'ARN) et en petits résidus hydrophobes (alanine, valine et glycine). Ils ont un faible coût métabolique, critère indispensable à la bonne marche de la production des protéines hautement exprimées (Lin, Kuang et al. 2002). De plus, les résidus alanine et glycine sont souvent considérés comme appartenant à la classe d'acides aminés présents dans les premières cellules, ce qui est compatible avec la nature vraisemblablement ancestrale des protéines ribosomales. Cette observation contraste avec l'enrichissement en acides aminés aromatiques qui sont considérés comme des nouveaux arrivants ayant progressivement envahi le code génétique (Brooks et Fresco 2002). Parce qu'ils sont encodés par peu de codons et qu'ils ont un fort coût métabolique, les résidus aromatiques sont rares dans les protéines et en particulier dans les protéines hautement exprimées (Akashi et Gojobori 2002; Palacios et Wernegreen 2002). Néanmoins, ils ont des propriétés physico-chimiques intéressantes telles que la stabilisation de structures bi- et tridimensionnelles (Lobry et Gautier 1994; Thomas, Meurisse et al. 2002; Thomas, Meurisse et al. 2002). En effet, les résidus aromatiques interagissent souvent ensemble, en particulier dans les hélices α, dans une configuration le plus souvent perpendiculaire (Burley et Petsko 1985; Hunter, Singh et al. 1991), tenant ainsi une place importante dans la stabilisation des protéines dans un contexte hydrophobe. Ces résidus joueraient un rôle majeur dans la structure et la fonction des protéines membranaires leur permettant des interactions sur les faces

exposées des hélices α. Ils faciliteraient également la stabilité des ancrages des protéines dans la membrane (Koshi et Bruno 1999). Les membranes doivent être flexibles et doivent s'adapter rapidement aux variations environnementales, et les acides aminés aromatiques peuvent leur fournir cette flexibilité nécessaire (Fukuchi et Nishikawa 2001).

Nos résultats montrent non seulement une opposition entre les petits résidus hydrophobes alanine et glycine et les résidus aromatiques dans la composition en acides aminés des protéomes, mais ils montrent également que les protéines de ménage sont opposées aux protéines orphelines. Ces dernières peuvent être considérées soit comme anciennes mais évoluant rapidement, soit comme des acquisitions récentes ou des créations. L'opposition entre les résidus alanine, glycine, valine et les résidus tyrosine, phénylalanine et tryptophane, tout comme l'opposition avec des protéines de ménage, argumente fortement en faveur de la seconde hypothèse selon laquelle les protéines orphelines seraient des acquisitions récentes. Ces dernières sont généralement de petites tailles (~150 résidus) et leur gènes sont riches en A+T en troisième base des codons (Daubin et Ochman 2004). Elles sont souvent présentes dans les bactériophages. En outre, à travers les processus de recombinaison/excision, des séquences peuvent être générées à partir de morceaux de génomes dont les protéines n'auront pas d'homologues. Les génomes étant soumis à un processus de stabilisation sélective, ces protéines, si elles sont recrutées pour une fonction apportant d'une manière ou d'une autre un avantage à leur hôte (nonobstant leur coût métabolique), seraient conservées par la cellule hôte. Leur statut d'orpheline crée subséquemment une discrimination entre le « soi » et le « non soi » des espèces, révélant par exemple une fonction qui traduirait une marque du « soi » d'un organisme. Ces protéines ne sont sans doute pas isolées dans le cytoplasme des cellules. Elles doivent à la fois interagir avec un ensemble de sous unités ou d'autres facteurs et éviter d'interagir avec des protéines sans lien de « parenté ». Nous proposons que nombre de ces protéines orphelines pourraient être des facteurs favorisant la stabilisation de processus (i.e., des sous unités non catalytiques de complexes protéiques). La diversité des résidus aromatiques qui sont présents dans ces « gluons » leur fournirait, grâce à un processus d'adaptation rapide, le moyen d'être recrutés fréquemment par l'organisme, ce qui expliquerait l'apparente ubiquité des protéines orphelines dans les cellules. Une fois recrutés, les « gluons » évolueraient lentement vers un coût métabolique moins élevé en perdant leurs acides aminés aromatiques au fur et à mesure du temps. De ce fait, une classe de protéines bien adaptées pourrait être transmise par transfert horizontal d'un organisme à un autre, tout en perdant leur statut de protéines orphelines. Cette transmission pourrait être prise en charge par les bactériophages qui constituent les vecteurs principaux d'introduction de protéines dans les cellules (Rocha et Danchin 2002; Pedulla, Ford et al. 2003).

C.I.5 Conclusion.

Malgré la divergence évolutive entre *E. coli* K-12, *B. subtilis* et *M. jannaschii* qui, d'après la principale théorie actuelle (fondée sur les similarités de l'ARN16S), se situe environ à 1,5 milliards d'années, il est tout a fait remarquable d'observer de telles constances dans les propriétés intrinsèques des différentes classes de protéines.

La plus évidente est sans doute celle impliquant les protéines membranaires. Les différences de compartimentation (double membrane pour les Bactéries à Gram négatif et une unique pour les Bactéries à Gram positif), ainsi que les différences de composition de ces membranes entre ces trois organismes sont établies (liaison ester pour les Eubactéries et liaison éther pour les Archaeabactéries entre les lipides de la membrane). Et pourtant, de manière systématique, il a été observé que les protéines membranaires ont une composition en acides aminés commune entre elles mais bien distincte du reste du protéome, quelque soit d'ailleurs la composition globale en acides aminés du protéome complet.

La propriété la plus surprenante est sans doute l'enrichissement des protéines orphelines en acides aminés aromatiques. A partir de cette observation, il serait très intéressant d'étudier biologiquement certaines protéines orphelines afin de comprendre leur rôle au sein de la cellule. On découvrirait peut être des interactions protéines-protéines, et éventuellement, de nouvelles applications liées à la pharmacologie ou à l'agro-alimentaire. Une des approches pourrait consister à analyser un complexe de protéines dont l'une d'elle serait orpheline. Si la mutation d'une ou de plusieurs des bases du gène cible modifie la structure tridimensionnelle de la protéine et empêche la formation du complexe, il serait intéressant d'observer quelles seraient les modifications fonctionnelles de la cellule. Cependant, la formation de complexe protéique dans les Bactéries est rare rendant cette approche difficilement réalisable. Une autre approche, tout aussi difficile techniquement et très coûteuse, serait, après isolement de l'une de ces protéines orphelines, de cristalliser celle-ci afin de comprendre le rôle fondamental des résidus aromatiques dont elles sont richement composées.

Finalement, cette étude effectuée sur les protéines des organismes modèles *E. coli* K-12, *B. subtilis* et *M. jannaschii*, nous a conduit, tout naturellement, à l'étude d'un plus grand nombre de procaryotes afin de vérifier les observations précédemment établies et d'en découvrir éventuellement de nouvelles.

C.II *L'étude d'un échantillon représentatif du monde procaryote.*

C.II.1 *Préambule.*

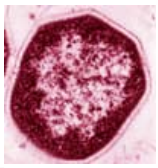
Les méthodes bien acquises, les outils maîtrisés, les premières études effectuées sur les trois modèles procaryotes ont été naturellement étendues à un ensemble d'autres organismes. Le but était de rassembler un maximum d'informations sur les biais compositionnels en acides aminés des protéines et d'en dégager des lois et des spécificités afin de décrire plus précisément les protéines, les protéomes, la cellule et le mode de vie des organismes. Les organismes ont été choisis de manière à obtenir la meilleure représentation possible du monde procaryote, cette sélection restant évidemment dépendante de la diversité de l'ensemble des organismes complètement séquencés et disponibles dans les banques. Ont été choisis six archaeabactéries et 22 bactéries. Elles ont des milieux de croissance très divers au niveau du pH, de la température et de l'environnement (sol,

eau…). Ces organismes ont un contenu en G+C du génome extrêmement variant et présentent parfois des particularités intéressantes telles que la pathogénicité ou le parasitisme.

C.II.2 *Caractéristiques des organismes étudiés.*

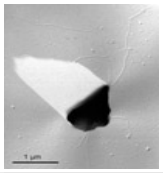
N.B. Les illustrations suivantes (excepté autres notifications) proviennent des sites www.genomenewsnetwork.org et www.ncbi.nlm.nih.gov.

Organismes	Règne	Philae	G+C %	T. O. C.*
Aeropyrum pernix	Archaea	Thermoprotei	56	95



Ce microbe aérobie fut isolé dans une conduite hydrothermale sur le sol océanique près de l'île de Kodakara-Jima au Japon. Il peut croître à une température supérieure à 100°C.

Archaeoglobus fulgidus	Archaea	Archaeoglobi	49	83

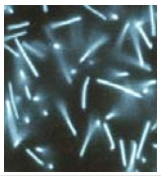


Ce microbe croît à température extrêmement haute et métabolise le soufre. Trouvé dans les puits pétrolifères, il peut diminuer la qualité du pétrole et créer un environnement dangereux pour les travailleurs des plateformes d'extraction car il produit un gaz nocif d'hydrogène sulfurique.

Halobacterium salinarum	Archaea	Halobacteria	68	50

Les halobactéries sont responsables de l'apparence rouge rosé brillant de la mer Morte. Ces organismes vivent dans des environnements à haute concentration en sel dont les étendues de sel elles-mêmes. Elles poussent également dans des denrées alimentaires telles que le poisson salé et la charcuterie. (illustration : www.planete-mars.com)

Methanopyrus kandleri	Archaea	Methanopyri	61	98

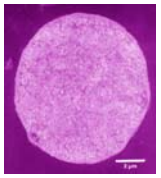


Cet organisme est l'un des organismes les plus tolérants à la chaleur connus jusqu'ici. Il vit à une température avoisinant les 100°C. Il fut isolé, à l'origine, sur le sol océanique du Golfe de Californie, à la base de cheminées sous-marines faites de conduites minérales soufrées crachant activement des fumées noires.

Pyrococcus abyssi	Archaea		45	103

Pyrococcus abyssi est une archaeabactérie isolée à partir de prélèvements effectués aux abords d'une source chaude située à 3500 m de profondeur dans le sud-ouest du Pacifique, et dont les conditions de croissance optimale sont de 103°C sous une pression de 200 atmosphères.

Thermoplasma acidophilum	Archaea	Thermoplasmata	46	59

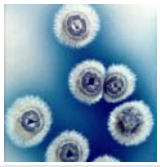
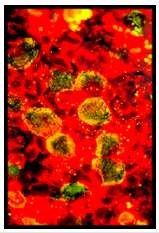
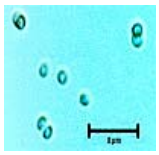
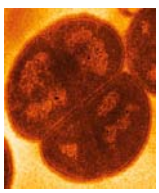
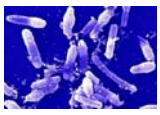


Ce microbe vit dans un environnement acide où la température approche une soixantaine de degrés. Il fut découvert dans un tas de déchets de charbon chauffé et dans un champ de solfatares acides. Atypique des autres organismes vivant dans de telles conditions, cet extrêmophile présente non pas une paroi cellulaire rigide mais une membrane cytoplasmique.

Mycobacterium tuberculosis	Bactérie	Actinobacteria	66	37

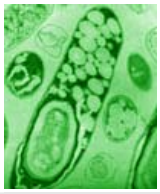
Cet organisme, agent de la tuberculose, a tué et tue toujours des millions d'êtres humains. Il fut même découvert dans les tissus d'anciennes momies. Ces dernières décennies, on observe l'émergence de souches résistantes aux médicaments. Une personne sur trois dans le monde est porteuse de la bactérie sous sa forme latente. (illustration : www.niaid.nih.gov)

* T. O. C. Température Optimale de Croissance

Organismes	Règne	Philae		G+C %	T. O. C.*
Streptomyces coelicolor	Bactérie	Actinobacteria		72	10 - 37
	C'est un des organismes les plus utilisés en pharmacologie car il produit la plupart des antibiotiques naturels prescrits aujourd'hui (tétracycline, érythromycine). La famille des streptomycètes produit également des composants aux propriétés anti-cancéreuses.				
Aquifex aeolicus	Bactérie	Aquificae		43	95
	Cette bactérie vit dans les environnements marins à des températures supérieures à 95°C. Elle a un métabolisme particulier permettant l'utilisation de sources d'énergies inorganiques.				
Chlamydia trachomatis	Bactérie	Chlamydiae		41	37
	La chlamydia est moins connue que les autres MST (maladies sexuellement transmissibles) comme la gonorrhée ou la syphilis, mais est plus fréquente. Les conséquences de la maladie peuvent être, chez la femme, l'infertilité ou les grossesses extra-utérines.				
Synechocystis	Bactérie	Cyanobacteria	Chroococcales	48	mésophile
	C'est un des organismes les plus abondants de la planète. C'est une bactérie photosynthétique vivant dans les océans. Certaines formes peuvent nager à plus de 25 mm/seconde sans avoir pour autant d'appareil de propulsion extérieur. Les bactéries Synechococcus et Prochlorococcus prennent en charge la fixation des deux tiers du carbone des océans.				
Anabaena nostoc	Bactérie	Cyanobacteria	Nostocales	41	mésophile
	*Anabaena nostoc* est une cyanobactérie filamenteuse capable de fixer l'azote. Bien que procaryotes, les cyanobactéries ont un système photosynthétique proche des végétaux chlorophylliens.				
Deinococcus radiodurans	Bactérie	Deinococci		67	25-35
	Cette bactérie, identifiée pour la première fois en 1956 dans des aliments irradiés, a depuis été découverte aux quatre coins du monde dans des zones riches ou pauvres en nutriments. Elle survit à des doses de radiations mille fois supérieures à celles supportées par un être humain. Bien que ces radiations brisent son ADN, elle répare son génome en quelques heures, phénomène très intéressant pour la recherche contre le cancer.				
Bacillus halodurans	Bactérie	Firmicutes	Bacilli	44	45-55
	Cette bactérie vivant en milieu basique produit beaucoup d'enzymes (protéases, cellulases, amylases) utilisées dans l'industrie essentiellement comme additifs aux produits ménagers.				

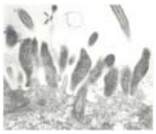
* T. O. C. Température Optimale de Croissance

Organismes	Règne	Philae		G+C %	T. O. C.*
Clostridium acetobutylicum	Bactérie	Firmicutes	Clostridia	31	37



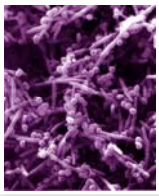
Durant la première guerre mondiale, cette bactérie fut cultivée par les Britanniques pour produire de l'acétone utilisée dans les obus d'artillerie. Elle transforme l'amidon en acétone et en butanol.

Mycoplasma penetrans	Bactérie	Firmicutes	Mollicutes	26	37



Dans les années 90, cette bactérie fût isolée principalement chez les personnes infectées par le VIH. On sait que la bactérie cause des infections des régions urogénitales et respiratoires chez ces mêmes individus (illustration : introduction de *M. penetrans* dans une membrane cellulaire).

Fusobacterium nucleatum	Bactérie	Fusobacteria		27	mésophile

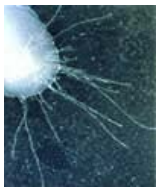


Cette bactérie se trouve dans la plaque dentaire des primates dont l'homme. Elle est fréquemment associée aux périodontites (infections de la tête, du cou, de la poitrine, des poumons et du foie).

Bradyrhizobium japonicum	Bactérie	Proteobacteria	Alpha	64	26

Cette bactérie forme des bosses ou des nodules sur les racines des plants de soja et fournit à son hôte de l'azote. C'est une propriété intéressante car elle pourrait permettre de diminuer la quantité d'engrais utilisé sur les millions d'hectares de culture du soja dans le monde (illustration : nodules d'une plante formés par *B. japonicum*).

Ralstonia solanacearum	Bactérie	Proteobacteria	Bêta	67	30

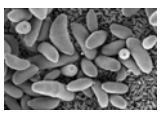


Cette bactérie, appelée auparavant *Pseudomonas solanacearum*, est un pathogène de plantes généralement trouvé dans les sols des pays tropicaux et subtropicaux où elle ravage les cultures. Certaines souches sont adaptées aux conditions des régions tempérées et ont récemment été isolées dans les pays du nord de l'Europe. Cet organisme peut infecter plus de 200 espèces de plantes appartenant à plus de 28 familles botaniques.

Bdellovibrio bacteriovorus	Bactérie	Proteobacteria	Delta	51	mésophile

Cette bactérie se trouve partout et particulièrement dans l'intestin des êtres humains. Elle se reproduit en s'enfouissant à l'intérieur d'autres microbes (*E. coli* K-12), les tuant, et utilisant ensuite leurs nutriments afin de produire leur descendance. C'est un pathogène de l'homme et des plantes. (illustration : *B. bacteriovorus* s'introduisant dans une autre bactérie).

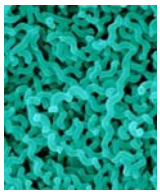
Desulfotalae psychrophila	Bactérie	Proteobacteria	Delta	47	10



Trouvée systématiquement dans les sédiments glacés au pied de l'océan arctique (et ailleurs), cette bactérie peut croître en dessous de -1,8°C. Par sa capacité à réduire le soufre, elle joue probablement un rôle important dans les cycles énergétiques globaux de son écosystème.

* T. O. C. Température Optimale de Croissance

Organismes	Règne	Philae		G+C %	T. O. C.*
Campylobacter jejuni	Bactérie	Proteobacteria	Epsilon	31	42-45

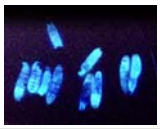


Campylobacter jejuni est une bactérie naturellement présente dans le tube digestif des oiseaux et des animaux, dans le lait cru, la boue et l'eau non traitée. L'infection causée par cette bactérie est connue sous le nom de campylobactériose, maladie gastro-intestinale des plus répandues. La contamination se fait par ingestion de nourriture qui n'a pas été cuite de manière appropriée. (illustration : www.ehagroup.com)

Escherichia coli O157:H7	Bactérie	Proteobacteria	Gamma	50	37

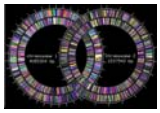
Naturellement présente dans les intestins des animaux (bétail, cochons, brebis), cette bactérie peut causer des empoisonnements alimentaires (viande mal cuite) et d'autres sérieuses maladies. Par le fait qu'elle soit de l'espèce coli et qu'elle soit pathogène, il est très intéressant de la comparer à K-12. (illustration : www.astrographics.com)

Photorhabdus luminescens	Bactérie	Proteobacteria	Gamma	43	30



Cette bactérie vit dans l'estomac d'un ver microscopique qui parasite certains insectes. Une fois le ver enfouit dans l'insecte, la bactérie produit de nombreux antibiotiques et des toxines qui tuent l'insecte.

Photobacterium profundum	Bactérie	Proteobacteria	Gamma	42	10

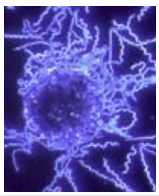


Cet organisme est un excellent modèle pour l'étude de l'adaptation au froid et aux hautes pressions. Bien qu'il soit psychrotolérant et piézophile (il croît de préférence à des pressions supérieures à la pression atmosphérique), il tolère les températures ambiantes et peut croître très rapidement à pression atmosphérique. (illustration : www.tigr.org)

Yersinia pestis	Bactérie	Proteobacteria	Gamma	48	37

Cette bactérie, agent de la peste, infecte les puces parasitant les rongeurs. Elle se transmet d'un rongeur à l'autre, de même qu'à l'être humain, par piqûres de puces infectées. Le contact direct avec des tissus d'animaux malades représente une autre voie de contamination possible pour l'homme. Le taux de létalité de la peste est de 50 à 60 % en l'absence de traitement. Il existe la peste bubonique (infection des ganglions), septicémique (infection sanguine) et pulmonaire (infection des poumons).

Borrelia burgdorferi	Bactérie	Proteobacteria		29	30



Cette bactérie infecte les poux et les tiques et cause chez l'homme la maladie de Lyme ou borréliose, se manifestant par des poussées fébriles successives, 3 à 6 mois après une simple piqûre.

Thermotoga maritima	Bactérie	Proteobacteria		46	80

A l'origine cette bactérie fût isolée dans des sédiments ardents d'un volcan italien. Elle réduit beaucoup de carbohydrates simples et complexes dont la cellulose et le xylan. Transformés en carburant, la cellulose et le xylan représentent potentiellement des sources d'énergies renouvelables.

* T. O. C. Température Optimale de Croissance

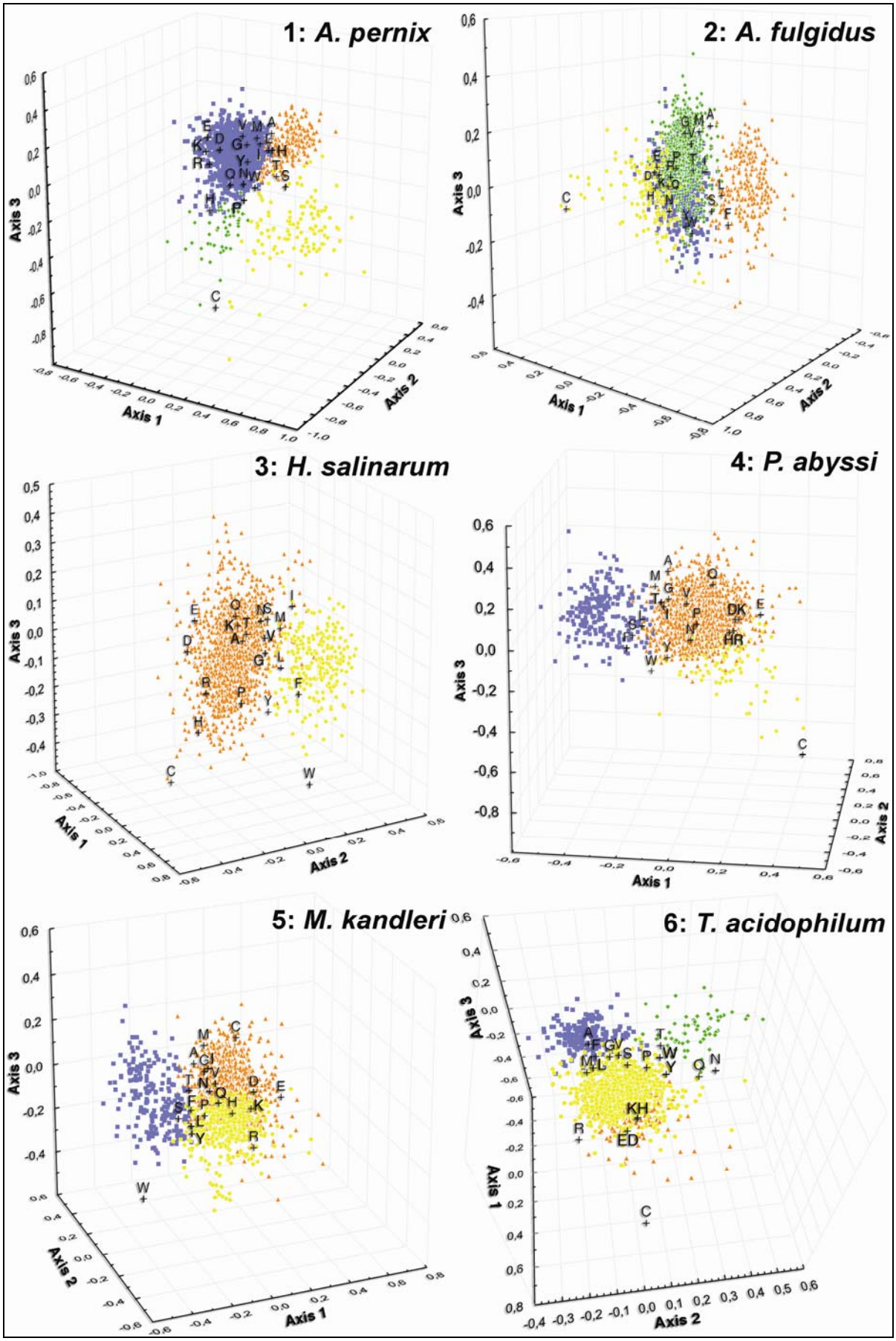
C.II.3 Les biais récurrents dans la composition en acides aminés des protéines procaryotes, article 2 ; résultats et discussions.

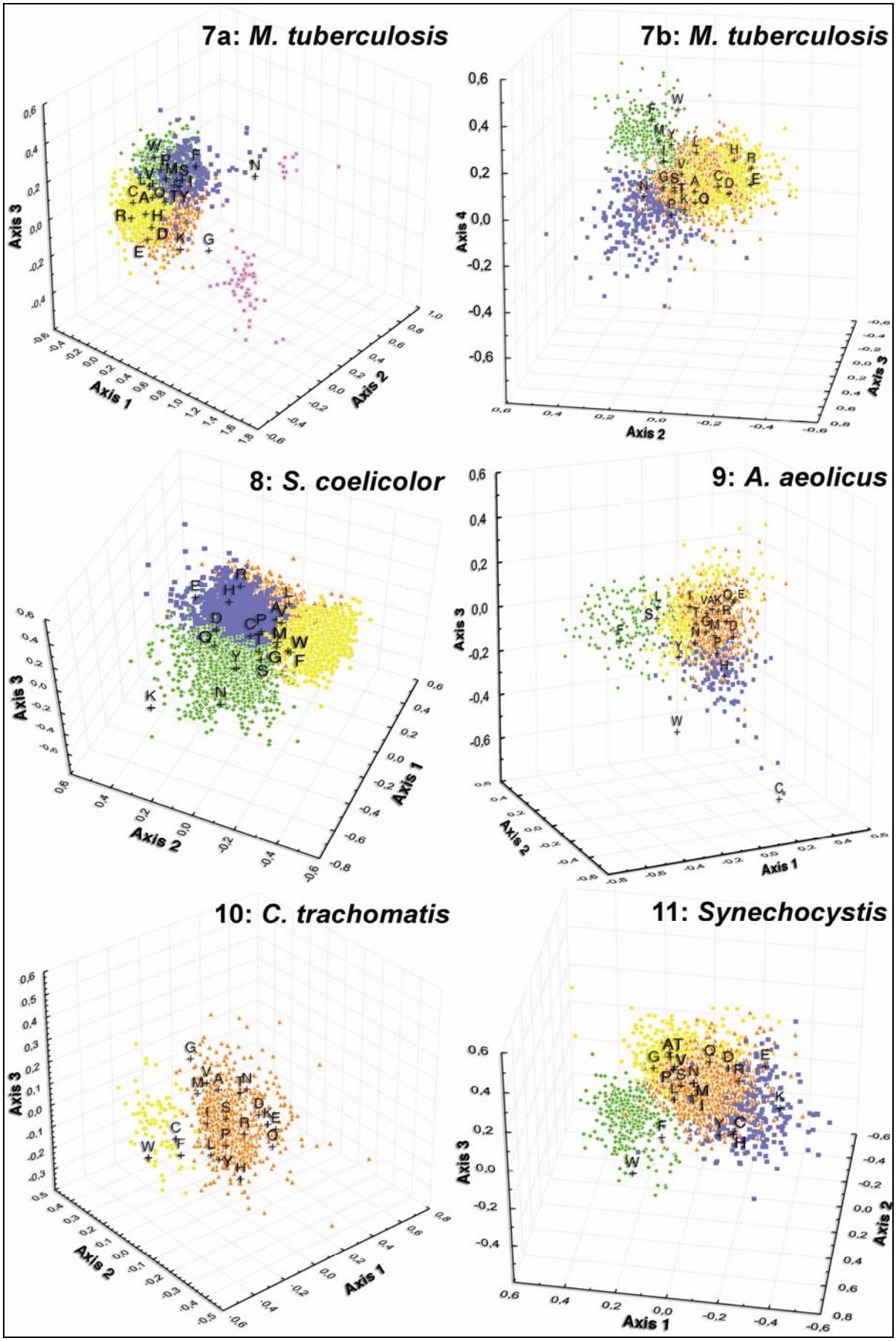
C.II.3.a Les IIMPs : une classe de protéines à part.

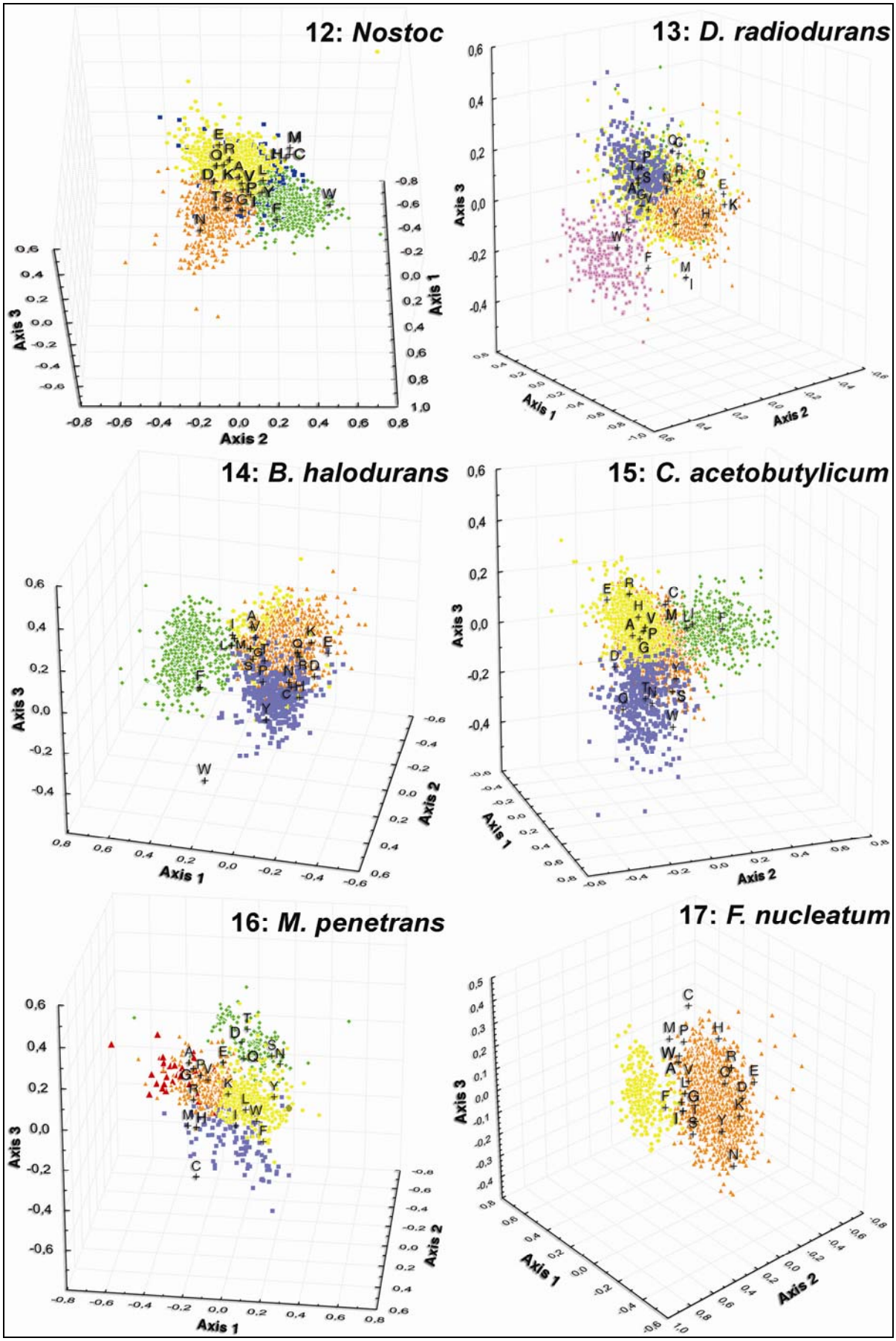
Extrait de l'article « Persistent biases in the amino acid composition of prokaryotic proteins », issu du paragraphe « A universal rule: integral inner membrane proteins cluster together ».

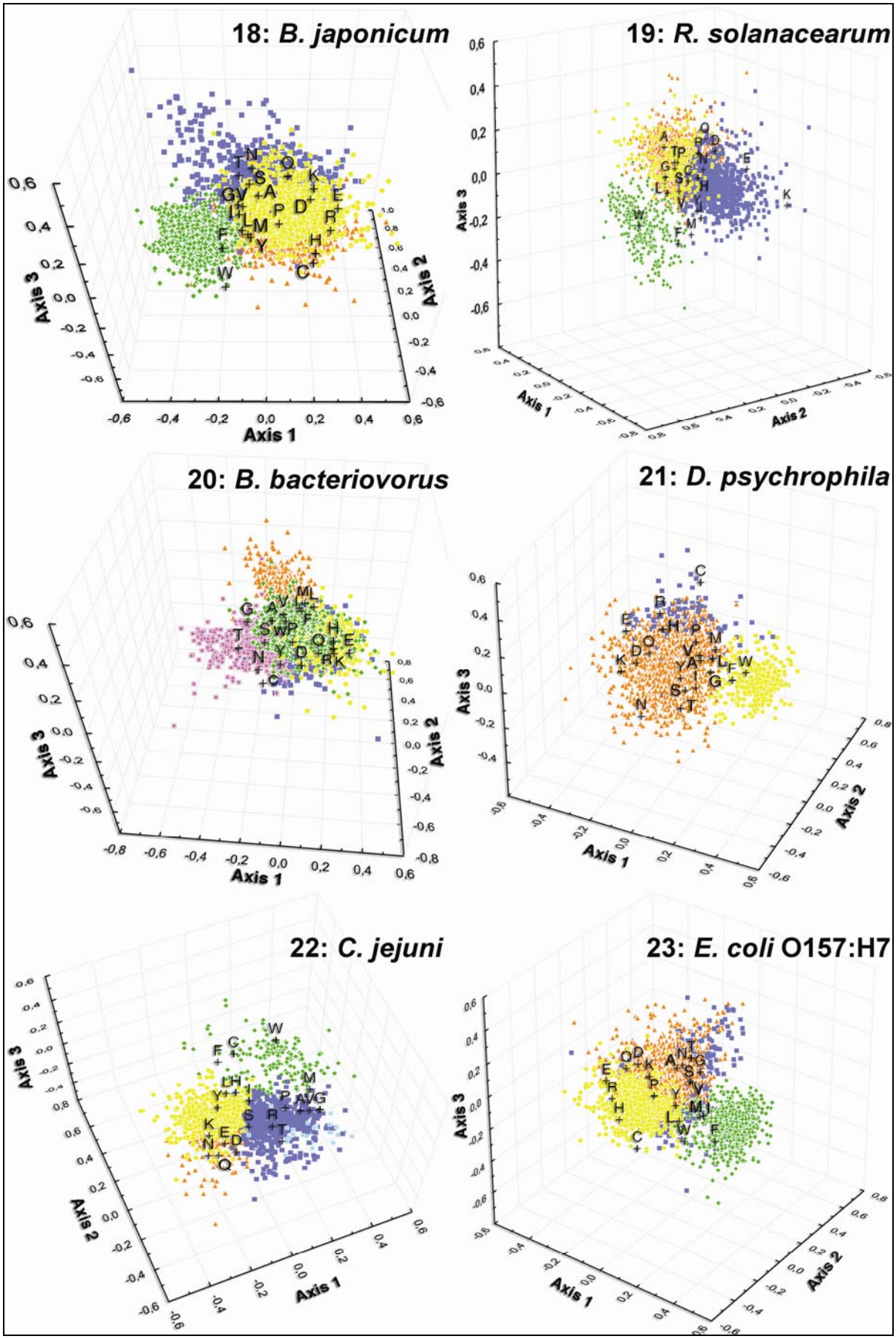
« While this feature seemed a fairly general property, it was of interest to explore whether the corresponding proteins would group together in the specific cloud of proteins of each organism of interest. Remarkably, in all the organisms analysed, a well-separated cloud was observed. This cloud of proteins is distinguished by one single factor, hydrophobicity of proteins, often brought about by leucine and phenylalanine, versus charged residues (Figure 5). In a previous study based on model prokaryotes, this homogeneous class was shown to be constituted exclusively of Integral Inner Membrane Proteins (IIMPs) (Pascal, Medigue et al. 2005). Most of the proteomes studied in the present work have not been experimentally characterized. For this reason, while we can probably be confident that the isolated cluster driven by the hydrophobic vs. charged residues is made of IIMPs, we tentatively named the corresponding proteins Probable Inner Integral Membrane Proteins (PIIMPs).

The presence of a single consistent class of proteins in such a large diversity of organisms is unexpected. This is particularly surprising when considering that the study includes both Bacteria and Archaea, as well as organisms living in extremely different environments. The case of the industrial bacteria *Clostridium acetobutylicum* is of particular interest (Figure 5-15), as these cells produce a mixture of acetone and butanol, which would be supposed to interfere with hydrophobicity of membrane structures. Hydrophobic interactions are dramatically modified in solvents such as acetone that reduce the dielectric constants of environment and reduce, consequently, the strength of hydrophobic interactions in proteins. Furthermore, acetone dissolves fat. The cellular membrane is constituted of lipids and it is remarkable that, despite this important pressure of chemical solvents adverse to bacterial life, the amino acid composition of *C. acetobutylicum* membrane proteins looks similar to that of the membrane proteins of organisms living in more typical conditions. In the same way, it is worth noticing that PIIMPs constitute a consistent class in Archaea, which have a membrane bilayer formed of lipids of a nature completely different from those of Bacteria (ethers instead of esters, in particular) (Pereto, Lopez-Garcia et al. 2004). As a consequence, this feature of the amino acid distribution in the proteins which form the proteome of prokaryotes would be convenient for valid annotation of the corresponding class of genes in genome projects. »


1: A. pernix
2: A. fulgidus
3: H. salinarum
4: P. abyssi
5: M. kandleri
6: T. acidophilum
Axis 1
Axis 2
Axis 3


7a: M. tuberculosis
Axis 1
Axis 2
Axis 3
7b: M. tuberculosis
Axis 2
Axis 3
Axis 4
8: S. coelicolor
Axis 1
Axis 2
Axis 3
9: A. aeolicus
Axis 1
Axis 2
Axis 3
10: C. trachomatis
Axis 1
Axis 2
Axis 3
11: Synechocystis
Axis 1
Axis 2
Axis 3


12: Nostoc
13: D. radiodurans
14: B. halodurans
15: C. acetobutylicum
16: M. penetrans
17: F. nucleatum
Axis 1
Axis 2
Axis 3


18: B. japonicum
19: R. solanacearum
20: B. bacteriovorus
21: D. psychrophila
22: C. jejuni
23: E. coli O157:H7
Axis 1
Axis 2
Axis 3

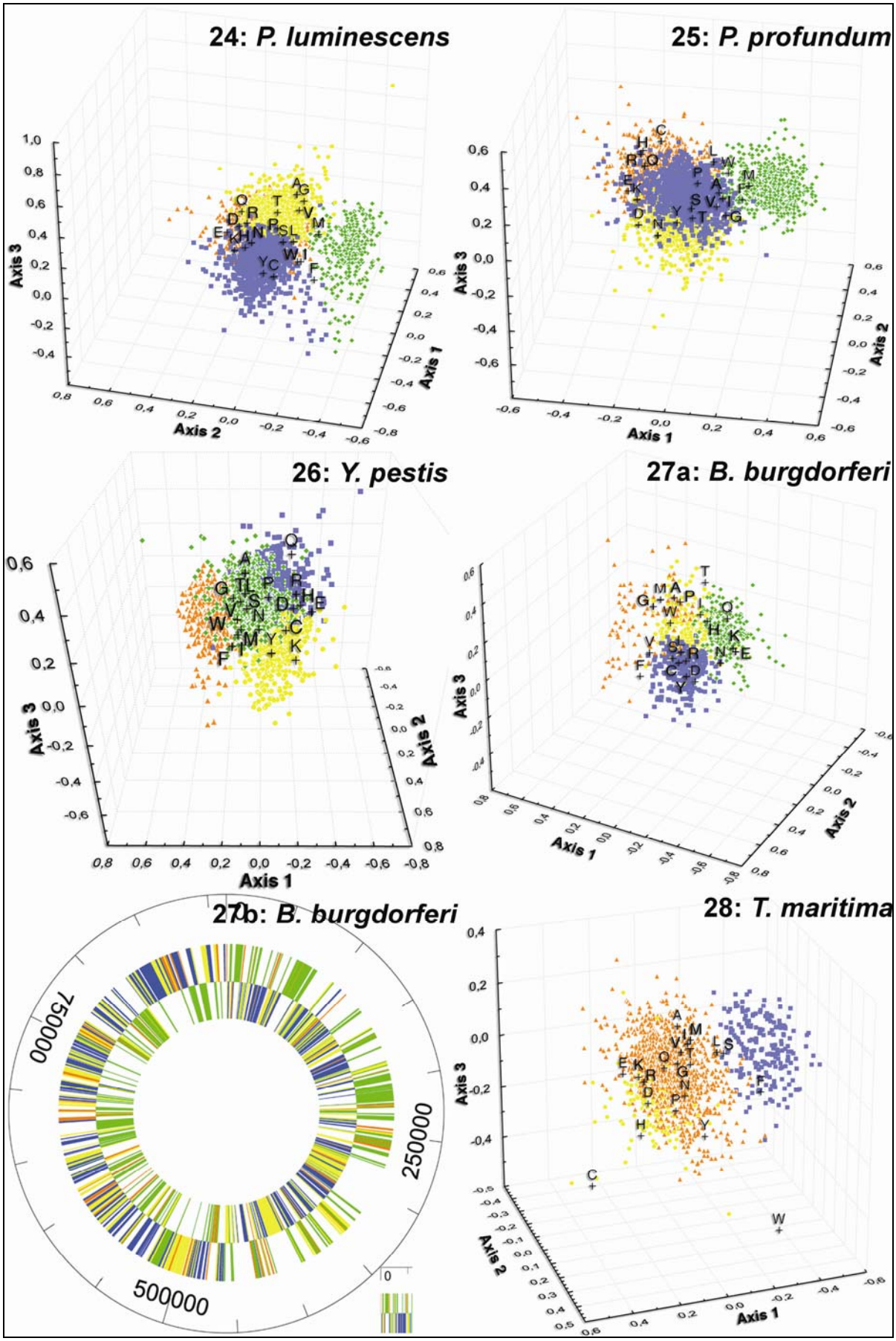

24: P. luminescens
25: P. profundum
26: Y. pestis
27a: B. burgdorferi
27b: B. burgdorferi
28: T. maritima
Axis 1
Axis 2
Axis 3
250000
500000
750000

Figure 5 : First CA space (axis 1, 2 and 3) and results of clustering method, from 2 to 5 clusters are obtained. The PIIMPs group is systematically obtained and its cluster color is indicated in parentheses, except when indicated otherwise 1- *A. pernix* (orange triangles), blue squares and green diamonds represent proteins wich may be false-positive, 2- *A. fulgidus* (orange triangles), 3- *H. salinarum* (yellow circles), 4- *P. abyssi* (blue squares), 5- *M. kandleri* (blue squares), 6- *T. acidophilum* (blue squares), orange triangles represent cysteine rich proteins, yellow circles represent cytoplasmic proteins, green diamonds represent extracellular proteins, 7a- *M. tuberculosis*, PE/PPE/PE-PGRS purple stars, 7b- axis 2, 3 and 4 of *M. tuberculosis* without PE, PPE and PE-PGRS proteins (green diamonds), 8- *S. coelicolor* (yellow circles), 9- *A. aeolicus* (green diamonds), 10- *C. trachomatis* (yellow circles), 11- *Synechocystis* (green diamonds), 12- *Nostoc* (green diamonds), 13- *D. radiodurans* (purple stars), 14- *B. halodurans* (green diamonds), 15- *C. acetobutylicum* (green diamonds), 16- *M. penetrans* (blue squares), green diamonds represent Ser and Thr rich proteins, large red triangles and the large khaki circle represent proteins for which the function includes Gene Ontology as indicated in Annexe I.I.1- Suppl-Tableau I, 17- *F. nucleatum* (yellow circles), 18- *B. japonicum* (green diamonds), 19- *R. solanacearum* (green diamonds), 20- *B. bacteriovorus* (orange triangles), purple stars represent G+C rich gene proteins, 21- *D. psychrophila* (yellow circles), 22- *C. jejuni* (green diamonds), cyan stars represent proteins for which the function includes Gene Ontology as indicated in Annexe I.I.1- Suppl-Tableau I, 23- *E. coli* O157:H7 (green diamonds), 24- *P. luminescens* (green diamonds), 25- *P. profundum* (green diamonds), 26- *Y. pestis* (orange triangles), 27a- *B. burgdorferi* (orange triangles), 27b- distribution along leading and lagging strands of chromosome of genes of CA clustered proteins of *B. burgdorferi*, 28- *T. maritima* (blue squares).

C.II.3.b *Etudes des protéines hautement exprimées.*

Extrait de l'article « Persistent biases in the amino acid composition of prokaryotic proteins », issu du paragraphe « Highly expressed ancestral proteins display common biases ».

« In the majority of the proteomes analysed, the ribosomal proteins were clustered in one or two groups (Annexe I.I.1- Suppl-Tableau I). For example, Figure 5-22 represents CA and clustering of the *Campylobacter jejuni* proteome. In spite of separation of the proteome into four clusters, ribosomal proteins are all clustered into only one of those groups. This systematic gathering in one or two groups in this study shows once again that ribosomal proteins are much conserved in prokaryotes despite phylogenetic divergence or variety of growth environments. »

C.II.3.c *Les contraintes dues aux acides aminés rares.*

Extrait de l'article « Persistent biases in the amino acid composition of prokaryotic proteins », issu du paragraphe « Rare amino acids create specific clusters ».

« As shown in Annexe I.I.2- Suppl-Tableau II, a consequence of this scarcity is that one of the first four axes which best describes the proteome is led by the frequency of the rare residues Cys and Trp. Analysis of the extremities of cysteine-biased axes shows that most Cys-rich proteins systematically belong to the class of metal-binding proteins (iron- or zinc-binding in general), presumably via Cys residues (data not shown). In some organisms (e.g. *Thermoplasma acidophilum*) many of these proteins are annotated as of unknown function; inference from annotation of the other genome sequences suggests that the corresponding proteins might belong to similar classes. As expected perhaps, organisms atypical in their G+C content, such as *Borrelia burgdorferi* or *Fusobacterium nucleatum* (A+T-rich) and *Streptomyces coelicolor* (G+C-rich), show a Trp distribution bias in one of

their factorial first four axes. After analysis of the proteins located at the extremities of Trp bias axes, we observed that this was due to a variety of independent causes and not a single common one (data not shown). In *B. burgdorferi*, the bias mostly affects proteins involved in the translation machinery. This is likely due to the fact that the corresponding genes are located in the leading strand of the organism, which is biased in G+T (Rocha, Danchin et al. 1999). Furthermore, biochemical experiments suggest involvement of Trp residues in RNA binding (Skinner et Jackson 1997; Dresios, Chan et al. 2002). In *F. nucleatum*, proteins are mostly involved, with a significant proportion of proteins predicted to be on the surface of the cell. Interestingly, this seems also to be the case of *S. coelicolor*, where an excess of Trp is to be expected because of the G+C content of the cells, in this case mostly in proteins of the cell surface, and in particular heme-binding proteins (David, Dutt et al. 2000). »

C.II.3.d *Le biais G+C des gènes induit un biais dans la composition en acides aminés des protéines.*

Extrait de l'article « Persistent biases in the amino acid composition of prokaryotic proteins », issu du paragraphe « The G+C-content of the Coding DNA Sequences creates an unexpected bias in amino acid composition ».

« Because the role of nucleotides in codons is not symmetrical (the third position of codons often results in synonymous substitutions in proteins) it is important to assess the impact on CDSs of the G+C content of the genome, depending on the position in the codons. It has long been noticed that the second codon position shows the highest correlation with the specific nature of the amino acid, with T associated with hydrophobic residues and A with hydrophilic residues. When present, the bias due to the second position of the codons' nucleotides was highly correlated with a bias in amino acids (or amino acids similar to each other according to Dayhoff's classification). For example, the *Aquifex aeolicus* G+C bias at the second position of codons (Tableau II) overlaps with the alanine, glycine and proline biases (GCN, GGN and CCN respective codons) and denotes many of cellular metabolism proteins. Remarkably however, we observed that the G+C content of genes influenced the amino acid composition of the corresponding proteins in an unexpected way, frequently following a G+C gradient at the first codon position. To explore whether this resulted from a separation between different classes of functions in proteomes we analysed phylogenetically consistent groups of Bacteria and Archaea for this characteristic (Tableau II). In most cases, different G+C contents reflect different biological functions. The G+C bias at the first position of codons was correlated with the overall G+C content of organism. Two organisms with a genome low in G+C, *B. burgdorferi* and *F. nucleatum*, stood out in our analysis. In the former case (Figure 5-27a), a remarkable separation of proteins into four clusters was observed. Two of those (separated along axis 2 and correlated with the G+C content at the first position of codons) contained almost exclusively proteins encoded by the leading strand (cluster 1) or the lagging strand of the chromosome (cluster 2) (Figure 5-27b). As already noted, this particular split between genes has previously been noticed in several studies (Lafay, Lloyd et al. 1999; Rocha, Danchin et al. 1999). In contrast, the *F. nucleatum* proteome was split into only two clusters (PIIMPs on the one hand and the bulk of other proteins on the other hand, Figure 5-17), precluding

further interpretation at this point. The second case, observed in 19 organisms (typically in Proteobacteria and Archaea, but not in the G+C-rich Actinobacteria), was characterised by a significant discrimination in the CA according to the G+C content at the first position of codons (Tableau II). Analysis of the 10% most biased proteins located at either end of the discriminating axis according to this constraint revealed that their function was more or less randomly distributed among possible functions in the cell.

While the amino acid bias imposed by the G+C content at the first position of the codons seems fairly ubiquitous, its raison-d'être in terms of functionality in proteins is not obvious. This bias creates a discrimination between Asn, Cys, Ile, Lys, Met, Phe, Ser, Thr, Trp, Tyr on the one hand and Ala, Asp, Gln, Glu, Gly, His, Pro, Val on the other hand. Each class has some common properties (aromaticity, sulfur content, hydroxyl group in the first class; negative electric charge, relatively smaller size in the second class) but nothing really compelling (Annexe I.I.3- Suppl-Tableau III). This might account for the lack of discriminating biological functions associated with this G+C bias, which might well be the consequence of a remnant of the origin of translation, associated with some optimisation of the translation machinery, which seems to prefer GNN codons (Brooks et Fresco 2003). »

Tableau II : CA axes built by G+C biases

Organism	Axis	Bias
Archaea		
A. pernix	2 & 3	G1 & G2
A. fulgidus	3	G1, GC1
H. salinarum	2	GC1
M. kandleri	3	G1
P. abyssi	No bias	
T. acidophilum	3	G1
Gram positive bacteria		
B. halodurans	2	G1
C. acetobutylicum	2	G1, GC1
M. penetrans	2	GC1, GC, G1
Actinobacteria		
M. tuberculosis	No bias	
S. coelicolor	No bias	
Proteobacteria		
B. japonicum	2	C1, GC1
R. solanacearum	No bias	
B. bacteriovorus	1	GC2
D. psychrophila	3	GC1, G1, GC
C. jejuni	2	GC1, G1
E. coli O157:H7	2 & 3	C1, GC1 & G1
P. luminescens	1 & 3	GC, GC1 & G1
P. profundum	2	G1, GC1
Y. pestis	2 & 3	C1 & GC1
Cyanobacteria		
Synechocystis	No bias	
Nostoc	3	GC1, G1, GC
Others bacteria		
D. radiodurans	No bias	
T. maritima	2	G1, GC2, GC1
A. aeolicus	2	GC2, GC, C2
C. trachomatis	3	GC2, C2
B. burgdorferi	1 & 2	GC2, GC & GC1, GC, G1
F. nucleatum	1 & 2	GC2, C2, GC & G1, GC1

Tableau III : CA axes built by aromaticity biases and percentage of unknown proteins in 10 % of CA axis aromatic-rich extremity and in the global proteome.

Organism	Axis	% unknown in 10 % aromatic rich	%unknown in proteome
Archaea			
A. pernix	no bias		
A. fulgidus	no bias		
H. salinarum	3	49	33
M. kandleri	no bias		
P. abyssi	no bias		
T. acidophilum	no bias		
Gram positive bacteria			
B. halodurans	2	44	22
C. acetobutylicum	4	25	21
M. penetrans	no bias		
Actinobacteria			
M. tuberculosis	no bias		
S. coelicolor	no bias		
Proteobacteria			
B. japonicum	3	37	30
R. solanacearum	no bias		
B. bacteriovorus	no bias		
D. psychrophila	2	65	26
C. jejuni	4	26	23
E. coli O157:H7	3	46	23
P. luminescens	3	41	30
P. profundum	no bias		
Y. pestis	3	39	21
Cyanobacteria			
Synechocystis	no bias		
Nostoc	3	62	35
Others bacteria			
D. radiodurans	no bias		
T. maritima	2	54	22
A. aeolicus	no bias		
C. trachomatis	3	44	23
B. burgdorferi	2	53	24
F. nucleatum	4	40	22

C.II.3.e *La spécificité des protéines orphelines : leur aromaticité.*

Extrait de l'article « Persistent biases in the amino acid composition of prokaryotic proteins », issu du paragraphe « Aromatic amino acids tag orphan proteins ».

« As new genome sequences continue to be deciphered, a particular class of proteins of unknown function becomes prominent. While the majority of the predicted proteins are similar to counterparts in other genomes a fraction of proteins, approximately 10%, does not look like anything known in other genomes, unless the genomes belong to the same genus. These proteins are usually named orphan proteins. Interestingly, they have been shown to be enriched in aromatic amino acids (Daubin et Ochman 2004; Pascal, Medigue et al. 2005). Because aromatic amino acids are very costly in terms of metabolic requirements, this supports the hypothesis that they are newly created proteins which did

not yet have time to be optimised in terms of cost vs benefit. As a consequence, they are markers of the "self" of a given species. Among their possible functions is that of being stabilising agents for multimeric complexes, and those with that type of hypothetical function have been named "gluons" (Pascal, Medigue et al. 2005). An aromatic amino acid bias was rarely observed in Archaea but frequently in Bacteria. This is shown in Tableau III, where the content of proteins of unknown function (hypothetical or putative proteins) has been computed for the 10% located in the CA cloud at the extremity of the axis driven by aromaticity. This value has subsequently been compared to the percentage of unknown proteins in the whole proteomes. Two interesting features stand out: (i) most aromatic proteins show a significant proportion of unknown proteins (1.1-2.5 times the average, depending on the proteome), (ii) this gradient is more pronounced when the axis driven by aromaticity appears early in the order of the axes organising the CA space. This is most likely due to the contribution of orphan proteins, as we demonstrated in the proteome of model organisms (Pascal, Medigue et al. 2005). »

Tableau IV : CA axes built by lysine/asparagine content biases and their opposite biases.

Organism	Axis	Bias	Opposite
Archaea			
A. pernix	No bias		
A. fulgidus	No bias		
H. salinarum	No bias		
M. kandleri	No bias		
P. abyssi	No bias		
T. acidophilum	2	N A2	T2
Gram positive bacteria			
B. halodurans	No bias		
B. subtilis	3	A1 N	GC1 C1
C. acetobutylicum	3	N S A2	T2
M. penetrans	2	N T1	GC1 GC
Actinobacteria			
M. tuberculosis	1	G N	Arg Glu
	2	N	Ala
M. tuberculosis without PE family proteins	1	K E N	Ala
S. coelicolor	1	K N I	Arg Ala
Proteobacteria			
B. japonicum	2	A1 A2 K	C1 GC1
R. solanacearum	1	K N	Ala
B. bacteriovorus	2	A2 K	T2 GRAVY AROMO
D. psychrophila	3	A1 AROMO N	GC1 G1 GC
C. jejuni	1	A2 K	GC2 Gly GC
	3	N S	Cys
E. coli O157:H7	2	A1 N	C1 GC1
E. coli K-12	2	A1 N	C1 GC1
P. luminescens	1	A1 K N	GC GC1
P. profundum	3	Y N S	GC1
Y. pestis	2	A1 N K	C1 Leu GC
Cyanobacteria			
Synechocystis	4	N Q	Met Val
Nostoc	3	AROMO N	GC1 G1 GC
Others bacteria			
D. radiodurans	1	K I N	Leu Ala
T. maritima	No bias		
A. aeolicus	No bias		
C. trachomatis	No bias		
B. burgdorferi	1	A2 A1 K N	GC2 G GC Val
F. nucleatum	1	A2 K N	GC2 C2 GC
	3	N S	C1

C.II.3.f Un biais récurrent dû aux codons AAN.

Extrait de l'article « Persistent biases in the amino acid composition of prokaryotic proteins », issu du paragraphe « A persistent bias is generated by AAN codons ».

« As we go down the list of importance of the CA axes, we find features that become more and more specific for a given specie. However, we uncovered a remarkable bias that persists in many genomes: numerous organisms (19 of the selected organisms) present an original bias due to a gradient in lysine (Lys) and/or asparagine (Asn) content along one the CA factorial axes (Tableau IV). While this bias exists in *B. subtilis* (41% G+C), it is almost absent in *B. halodurans* (44% G+C), a halophilic Bacillus. It is present in a single Archaeon living in an acidic biotope, *T. acidophilum* (46% G+C).

While this feature is persistent, we could identify several categories of proteins distinct by this bias. In *D. radiodurans*, the AAN bias is the first factor of the CA cloud. The proteins responsible for the bias are clearly linked to the protein biosynthesis machinery. Interestingly, in the case of *C. acetobutylicum* (Asn bias on axis 3) we found that the cluster of Asn-rich proteins (428 proteins) was dominated by enzymes involved in polysaccharide biosynthesis and turnover, and in particular present in the cellulosome (Perret, Belaich et al. 2004). *Clostridium acetobutylicum* is known to have an abundance of polymer degradation systems (Nolling, Breton et al. 2001). In line with a specific relationship between Asn enrichment and the surface of the cell, we also observed that many proteins linked to flagella and the cell wall belonged to this same cluster. Likewise, a Lys-dominated bias was observed on axis 2 of *Bdellovibrio bacteriovorus* and an Asn-dominated bias on axis 3 of *C. jejuni*. In *B. bacteriovorus*, the group formed on axis 2 is mostly composed of hypothetical proteins (two thirds), while most of the rest are linked to proteins involved in the outer surface of the cell: outer membrane proteins, flagella, cell wall, secreted proteases or other extracellular activities. A similar situation holds for *C. jejuni*, with proteins linked to flagella, outer membrane, chemotaxis, proteolysis forming the bulk of the Asn rich proteins. *Ralstonia solanacearum* represents a situation where the bias involves both Lys and Asn. The cluster defined by this bias, the yellow circle (Figure 5-19) as in the preceding examples, comprises mostly outer membrane proteins, porins, siderophore-iron transporter activity, calcium ion binding and iron ion transport. In the same way, the *Y. pestis* proteome also has a bias in Asn and Lys, on axis 2 (Figure 5-26). Proteins situated on the Asn/Lys-enriched extremity are often putatively exported proteins or located in the outer membrane. The most remarkable feature of this particular bias is that it behaves as if Asn and Lys had some common physico-chemical feature that would account for them being coded by the same box (AAN) of the genetic code. Both are hydrophilic, but in general Lys is positively charged. However, when appropriately screened from the environment, the amino-terminal end of Lys can be a doublet electron-donating group. In extant metabolism, the most similar amino acid would be ornithine. This amino acid, however, is not among those present in proteins, the underlying biochemical reason being that upon activation as an amino acid adenylate and loading onto tRNA it would become cyclised. One may therefore conjecture that asparagine might use its end amide group in the same way as lysine does, with a shorter chain, as a substitute for

ornithine, thereby accounting for the selection pressure that has coded both amino acids from a common genetic code box. »

C.II.3.g Zoom sur deux cas atypiques.

Extrait de l'article « Persistent biases in the amino acid composition of prokaryotic proteins », issu du paragraphe « Zoom on interesting cases »

« Most biases described in this study were persistent among prokaryotes, indicative of common trends of selection, probably associated with inevitable physico-chemical constraints. Sometimes however, a bias showing up in one of the CA axes, belonged to only one organism. Three of the corresponding features, which are usually highly revealing of the lifestyle of an organism, are presented in the next paragraphs. »

C.II.3.g.1 Les protéines membranaires de Mycoplasma penetrans.

Extrait de l'article « Persistent biases in the amino acid composition of prokaryotic proteins », issu du paragraphe « Zoom on interesting cases » « Atypical PIIMPs of Mycoplasma penetrans »

« Although *M. penetrans* CA presents a PIIMPs cluster as all other prokaryotes do, this cluster is not driven by an opposition between Phe and Leu and charged amino acids but by an opposition between charged (Lys, Glu) and hydrophilic non-charged (Thr, Ser) residues. At one extremity of the axis, a cluster of proteins rich in Ser and Thr is formed (green diamonds cluster, Figure 5-16), with approximately half annotated as hypothetical proteins, and approximately one third membrane-associated proteins (blue squares cluster contents PIIMPs). Yáñez and co-workers have described the original structure of the membrane of this Mollicute (Antonio Yáñez 1996). *Mycoplasma penetrans* has a typical elongated, flask-shaped morphology, with two internal compartments that permit the cells to adhere, then to penetrate into human cells. Serine-threonine-rich proteins are often involved in adhesion to membranes (Handley, Correia et al. 2005; Plummer, Wu et al. 2005; Siboo, Chambers et al. 2005) and it is not unexpected that the small proteome of this pathogen could be biased by one specific category of proteins. The proteome of *M. penetrans* displayed a second unusual feature: the cluster formed on axis 2 of CA, built by opposition of Ala, Gly and Val to Asn and Tyr, contained almost all the ribosomal proteins of the organism (Suppl-Tableau I, large red triangles, Figure 5-16), except for an interesting exception. Unexpectedly, protein MYPE1290, annotated as a "ribosomal protein" is located completely outside the ribosomal protein cluster, in a cluster comprising many enzymes (large khaki circle, Figure 5-16). This prompted us to check its annotation. In the metabolic pathway database MetaCyc (Krieger, Zhang et al. 2004), MYPE1290 is annotated by homology with counterparts in other organisms as a enzyme (alanine acetyltransferase, putative, EC number 2.3.1.128). Using its sequence to browse the proteome of Firmicutes, we uncovered that it is most probably an enzyme modifying a ribosomal protein (e.g. YdiD in the SubtiList database). Combining this description and the protein localisation in the CA space, we can therefore be confident that this

protein is certainly not a structural constituent of the ribosome. This illustrates that the usage of CA analysis presented in this work is a powerful complement to other methods used for proteome annotation. »

C.II.3.g.2 Les protéines papthogènes de Mycobacterium tuberculosis.

Extrait de l'article « Persistent biases in the amino acid composition of prokaryotic proteins », issu du paragraphe « Zoom on interesting cases » « Pathogen proteins of *Mycobacterium tuberculosis* »

« The CA cloud of *Mycobacterium tuberculosis* is notably atypical, as shown in Figure 5-7a. Axis 1 allows formation of a first cluster (purple stars in Figure 5-7a), opposing Gly and Asn to Arg. It mostly contains the PPE, PE and PE-PGRS family proteins (some of these are also located in the blue cluster located at the positive extremity of axis 1). These proteins are rich in Gly, and certain members of these families could be located in the mycobacterial cell wall (Brennan et Delogu 2002). This cluster is well-separated on axis 2, due to opposition of Asn and Phe versus Ala. Remarkably, this opposition differentiates PPE proteins (positive extremity of axis 2) from PE-PGRS proteins (negative extremity of axis 2). The bias is so intense that it may hide other biases. Indeed, the first CA axis represents approximately 35% of the total information, while the average information of axis 1 of all others organisms is between 15% and 25%. To overcome the contribution of this unwanted bias, we computed a new CA omitting the proteins of the PPE, PE and PE-PGRS families (approximately 140 proteins). The most interesting outcome of the new analysis is the opposition between Pro and Leu, Phe on axis 4, separating two clusters of proteins (Figure 5-7b). All the identified functions of the proteins in one of the clusters correspond to IIMPs; we can therefore reasonably predict that all proteins of this cluster belong to that category. A second one (blue square cluster) is very homogeneous and, when annotated its members are all somehow involved in pathogenicity, allowing us to propose that the whole cluster is composed of proteins involved in pathogenic processes. This comprises: (i) hydrolases, all associated with activity at the cell surface or to proteins anchored in the membrane (proteases, peptidases, mureine hydrolases, complex carbohydrate hydrolases, lipases), (ii) weak complexity proteins, (iii) protein kinase-like regulators and (iv) other proteins present at membrane surface (integral, oxydo-reductases etc.). These results lead to the idea that CA shows several levels of complexity and suppressing some clusters of clouds, we could observe other characteristics due to usage of amino acids. »

C.II.4 Conclusion.

Extrait de l'article « Persistent biases in the amino acid composition of prokaryotic proteins », issu du paragraphe « Conclusion ».

« In addition of the clear characterisation of PIIMPs, CA could be used as a versatile tool for protein functional annotation to refine the annotation of already annotated proteins, and to propose new functional categories for unannotated proteins. Furthermore, as demonstrated in the case of *M.*

penetrans, CA can help to annotate genome sequences or to identify likely annotation errors. This can be illustrated in the CA of *Thermoplasma acidophilum* which is quite unusual. Indeed, CA separates the *T. acidophilum* proteome into four distinct clusters (Figure 5-6). Three clusters are associated with a specific cellular compartment, while the fourth one, driven by cysteine, is composed of iron-binding proteins, mostly proteins containing iron-sulfur clusters. Not surprisingly, the PIIMP cluster is composed of very hydrophobic proteins. In contrast, the yellow circle cluster assembles cytoplasmic proteins, while extracellular proteins that are in the green diamond cluster, are isolated according to their asparagine content.

To investigate further whether the CA analysis of proteomes could be used to improve genome sequence annotation, we chose to work on the proteome of *Aeropyrum pernix* revised by Natale et al. (Natale, Shankavaram et al. 2000). As in the case of *T. acidophilum*, the CA we obtained is atypical. Four well-separated clusters of proteins are formed and correlations of axes are not like those other organisms, excepted for hydrophobicity correlated with axis 1. The orange triangle cluster, shaped on this axis, indeed contains membrane proteins (Figure 5-1). After exploration of the functional annotations of proteins of three other groups, two clusters (green diamond and yellow circle clusters) are composed almost exclusively, of unknown proteins (98% and 90%). And the last one (blue square cluster) has the rest of proteome of which 36% are unknown. We could suggest that, in spite of reannotation efforts, in this genome, the annotated sequence still contains many erroneous proteins, which would be contained in green and yellow clusters. »

C.III Conclusion générale sur l'analyse des procaryotes.

Comme montré à travers l'études des trois modèles procaryotes et de notre échantillon représentatif des génomes procaryotes séquencés, les protéines sont fortement marquées par des biais dus à des pressions de sélection de tout ordre. Ces différentes analyses permettent maintenant d'affirmer qu'un certain nombre de règles régissent le métabolisme des protéines, et par là même, celui de la cellule. Parmi ces règles, nous trouvons l'hydrophobicité des protéines de la membrane interne, l'aromaticité des protéines orphelines ou encore l'abondance des résidus encodés par les triplets AAN (asparagine et lysine) des protéines exportées ou se trouvant dans la membrane externe. Seul le biais d'hydrophobicité des protéines de la membrane interne, qui conduit la caractérisation des IIMPs, semble universel (retrouvé dans chacun des protéomes étudiés) ; un statut « IIMP prediction » a donc été créé au sein du logiciel d'annotation MaGe, créé à l'AGC (Annexe I.II), de manière à apporter une information supplémentaire aux annotateurs. En revanche, parce que chaque procaryote est différent, les autres règles édictées ne sont pas universelles mais quasi universelles. Dorénavant, lors de l'analyse d'un nouveau procaryote, après avoir retrouvé (ou pas) l'ensemble de ces règles, il faudra s'astreindre à étudier ce qui est atypique, spécifique à cet organisme, et ceci afin d'enrichir toujours plus nos connaissances sur la composition en acides aminés des protéines et sur la « vie » de ces dernières dans la cellule.

D Biais compositionnels des bactéries psychrophiles

D.I *Pseudoalteromanas haloplanktis : une bactérie de l'antarctique.*

Les organismes vivants se sont systématiquement protégés des variations de leur environnement en établissant des barrières efficaces. Cependant un paramètre physique, la température, ne peut jamais être contrôlé, sauf chez les organismes multicellulaires complexes homéothermes. Or la température affecte de façon considérable, et pratiquement instantanée, tous les processus physicochimiques de la cellule, depuis la vitesse et la sélectivité des réactions chimiques jusqu'au repliement des macromolécules : ADN, ARN et protéines. Cela a été compris assez tôt, et a motivé une grande partie des travaux concernant les Archaeabactéries, dont beaucoup croissent à très haute température. Curieusement, bien peu se sont rendus compte que si la très haute température pose de solides problèmes (principalement en termes de stabilité chimique des métabolites de base et des polymères), la basse température posait des questions bien plus difficiles. En effet, un grand nombre de processus biologiques dépendent de la formation et de la stabilité des liaisons hydrogènes, et sont orientés par l'accroissement de la baisse de l'énergie disponible (l'entropie) dans le système, qui est en grande partie due à la structure de l'eau. Les propriétés de l'eau, qui restent relativement constantes de 12°C à près de 100°C, changent rapidement et fortement entre 0°C et 10°C. On doit donc attendre que, si la vie se développe dans ces conditions de température, les lois présidant à la formation et au repliement des macromolécules seront altérées, mettant en jeu les propriétés fines de chacun des éléments de base, nucléotides et acides aminés en particulier, à moins qu'un métabolisme particulier ne vienne remédier à la modification de la structure de l'eau (antigels). De même, la structure des membranes devra être modifiée pour permettre l'insertion des protéines essentielles à la vie cellulaire. Enfin, les ARN, dont on reconnaît de plus en plus l'importance, auront une stabilité structurale considérablement accrue, et l'on peut donc espérer, par l'étude d'organismes vivant dans ces conditions, commencer à comprendre la fonction de ces nombreuses "protéines du choc froid", dont on sait qu'elles accompagnent les ARN. N'oublions pas que plus de 90% de la surface terrestre se trouve à moins de 15°C (et souvent beaucoup moins, dans les océans en particulier).

Des expéditions polaires ont rapporté des bactéries de la région antarctique (Figure 6). *Pseudoalteromonas haloplanktis* est un organisme assez voisin des organismes modèles comme *Escherichia coli* ou *Bacillus subtilis*. Ceci a facilité son annotation et a permis d'entreprendre de nombreuses analyses de génomique comparative. La bactérie a un phénotype stable et a pu être cultivée en grande quantité. *P. haloplanktis* semble être omniprésente dans les régions polaires car elle a été trouvée aussi bien en Arctique qu'en Antarctique, dans l'eau de mer mais également dans les sols, les glaciers et sur des restes organiques terrestres. Quelques connaissances structurées étaient déjà acquises à son sujet et les techniques de base de la génétique (conjugaison en particulier) ont pu être réalisées. Le génome n'est ni trop grand, ni trop compliqué ce qui a permis de diminuer le coût du séquençage et de faciliter l'assemblage et le finissage. Le projet d'étude de cette bactérie a bénéficié de l'entourage d'une communauté scientifique concertée qui fût prête à en faire

l'analyse expérimentale, couplée à l'analyse *in silico*. Intéressés par toutes ses considérations, nous avons été conduits à prendre contact avec le Pr. G. Feller de l'Université de Liège, qui avec le Pr. C. Gerday, a rassemblé un grand nombre de bactéries de l'Antarctique. Il a, en collaboration avec le Dr M. Luisa Tutino de l'Université de Naples, fait l'étude comparative d'espèces variées. Et, en accord avec ces chercheurs, nous avons choisi de concentrer notre étude sur la souche *Pseudoalteromonas haloplanktis* TAC 125, qu'ils nous ont fournie pour développer son étude en commun.

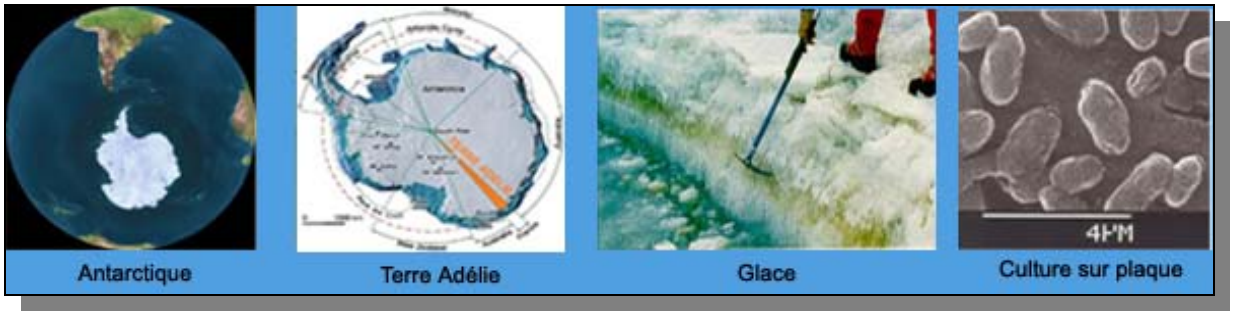


Figure 6 : Lieux de découverte de *P. haloplanktis* et culture en laboratoire.

D.II *L'essentiel de l'article.*

Le séquençage complet du génome de *P. haloplanktis* a permis de révéler que cet organisme exploite de nombreuses solutions pour faire face au froid. Son génome a bénéficié de la validation manuelle des annotations générées automatiquement (annotation syntaxique et fonctionnelle, génomique comparative dont le calcul des groupes de synténies et la reconstruction des voies métaboliques). D'une composition moyenne en G+C de 41%, le génome de cet organisme se partage en deux chromosomes ayant chacun un mode de réplication différent (bi- et unidirectionnel). 3488 CDSs ont pu être identifiées et plus de 65% d'entre elles ont été assignées à une fonction (11% restent uniques à *P. haloplanktis*). L'organisme résiste remarquablement bien aux très hautes concentrations d'H_2O_2 et pousse très rapidement dans un milieu à densité saline élevée. Son protéome a une composition atypique en asparagine (analyse *in silico*), acide aminé enclin à contribuer au vieillissement des protéines. En outre, ce biais a également été constaté, au cours de cette étude, chez d'autres psychrophiles. *P. haloplanktis* n'est pas le seul modèle pour l'étude de l'adaptation aux conditions marines et au froid, mais il promet d'être un outil utilisé par les biotechnologies mises en oeuvre dans la production de protéines.

Les données supplémentaires notées dans l'article se trouvent dans l'Annexe I.IV.

D.III *Article 3.*

Article à paraître dans le journal Genome Research au cours de l'automne 2005.

Article

Coping with cold: The genome of the versatile marine Antarctica bacterium *Pseudoalteromonas haloplanktis* TAC125

Claudine Médigue,[1] Evelyne Krin,[2] Géraldine Pascal,[1,2] Valérie Barbe,[1] Andreas Bernsel,[3] Philippe N. Bertin,[4] Frankie Cheung,[5] Stéphane Cruveiller,[1] Salvino D'Amico,[6] Angela Duilio,[7] Gang Fang,[2] Georges Feller,[6] Christine Ho,[5] Sophie Mangenot,[1] Gennaro Marino,[7] Johan Nilsson,[3] Ermenegilda Parrilli,[7] Eduardo P.C. Rocha,[2] Zoe´ Rouy,[1] Agnieszka Sekowska,[2,8] Maria Luisa Tutino,[7] David Vallenet,[1] Gunnar von Heijne,[3] and Antoine Danchin[2,9]

[1]Genoscope, CNRS-UMR 8030, Atelier de Génomique Comparative, 91006 Evry Cedex, France; [2]Genetics of Bacterial Genomes, Institut Pasteur, 75724 Paris Cedex 15, France; [3]Department of Biochemistry and Biophysics, Stockholm University, S-106 91 Stockholm, Sweden; [4]Dynamique, Evolution et Expression de Génomes de Micro-organismes, Université Louis Pasteur, 67000 Strasbourg, France; [5]Computer Centre, The University of Hong Kong, Hong Kong; [6]Laboratoire de Biochimie, Institut de Chimie B6, Universite´ de Liège, B-4000 Liège-Sart Tilman, Belgium; [7]Dipartimento di Chimica Organica e Biochimica, edificio MB, via Cinthia, Complesso Universitario Monte S. Angelo, 80126 Napoli, Italy; [8]CEA Saclay, Laboratoire Stress Oxydants et Cancer, DSV/DBJC/SBMS, 91191 Gif sur Yvette Cedex, France

A considerable fraction of life develops in the sea at temperatures lower than 15°C. Little is known about the adaptive features selected under those conditions. We present the analysis of the genome sequence of the fast growing Antarctica bacterium *Pseudoalteromonas haloplanktis* TAC125. We find that it copes with the increased solubility of oxygen at low temperature by multiplying dioxygen scavenging while deleting whole pathways producing reactive oxygen species. Dioxygen-consuming lipid desaturases achieve both protection against oxygen and synthesis of lipids making the membrane fluid. A remarkable strategy for avoidance of reactive oxygen species generation is developed by *P. haloplanktis*, with elimination of the ubiquitous molybdopterin-dependent metabolism. The *P. haloplanktis* proteome reveals a concerted amino acid usage bias specific to psychrophiles, consistently appearing apt to accommodate asparagine, a residue prone to make proteins age. Adding to its originality, *P. haloplanktis* further differs from its marine counterparts with recruitment of a plasmid origin of replication for its second chromosome.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to EMBL under accession nos. CR954246 and CR954247. The data are also available at the following Web site: www.genoscope.cns.fr/agc/mage/psychroscope.]

Three quarters of the Earth is covered by sea and >90% of its surface experiences yearly temperatures <15°C, asking for a remarkable adaptation of life to cold conditions. Several marine bacteria have been studied, but so far we possess only limited information about life in the sea at medium and low temperatures (Bartlett 1999; Raven et al. 2002; Thomas and Dieckmann 2002). Furthermore, heterotrophic bacteria in sea ice play a key role in carbon cycling, while little is known about their metabolic features, which are beginning to be deciphered (Moran et al. 2004). Challenges posed by cold to life stem from the slow pace of chemical reactions (for reviews, see Lonhienne et al. 2000; Feller and Gerday 2003; Weber and Marahiel 2003; Georlette et al. 2004), from the constraints induced by the stability of hydrogen bonds (the situation is particularly challenging for folded nucleic acids, i.e., RNA), and from increased solubility of gasses and stability of radicals. Phylogenetic studies based on 16S rRNA sequences indicated close relationships between marine bacteria within two bacterial divisions, Proteobacteria (in the genera Alteromonas, Colwellia, Glaciecola, Octadecabacter, Pseudoalteromonas, Shewanella, and Vibrio) and Cytophaga-Flexibacter-Bacteroides (Cytophaga, Flavobacterium, Gelidibacter, and Polaribacter) (Ivanova et al. 2004). *Pseudoalteromonas haloplanktis* TAC125 has been isolated from an Antarctic coastal sea water sample collected in the vicinity of the French Antarctic station Dumont d'Urville, Terre Adélie (66°40 S; 140° 01 E). By using genome sequencing, corroborated by in silico and in vivo analyses, we have uncovered exceptional genomic and metabolic features of this γ-proteobacterium compared with other bacteria from aqueous environments (Supplemental Table 1). These features, some of which explored by physiological experiments, account for their remarkable versatility and fast growth, showing adaptation to rare but periodic situations of abundance, making it an organism of choice for exploring heterologous protein production at low temperature (Tutino et al. 2001; Duilio et al.

[9]Corresponding author
E-mail adanchin@pasteur.fr; fax 331-45-68-89-48.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.4126905.

Table 1. General features of the *Pseudoalteromonas haloplanktis* genome

	Chromosome 1	Chromosome 2
Size (bp)	3,214,944	635,328
G+C percentage	41.0	39.3
Number of predicted CDSs	2,942	546
Average size of CDSs (bp)	950	1013
Percentage coding	88.6	87.3
Number of rRNA operons (16S-23S-5S)	9	0
5S rRNA (extra copies)	1	0
Number of tRNAs	106	0
CDSs similar to known proteins	1123	157
Putative functions (limited homology/structural features)	759	251
Conserved hypothetical proteins	694	75
Orphan proteins	325	61
Doubful CDS and gene remnant	41	2

2004). Although cold conditions are so prevalent on Earth, we do not possess at the moment a reference set of annotations for the genome of bacteria thriving in such conditions. The present genome sequence was therefore carefully annotated manually, and annotation will be, as much as possible, continuously refined. We endeavored to post to the community reference databases, allowing investigators to rapidly and efficiently retrieve relevant information while comparing it to what is known in other genomes. In the MaGe platform (www.genoscope.cns.fr/agc/mage/psychroscope) investigators can not only see gene annotations in parallel with syntenies with the genomes they chose as relevant but compare the annotations with cognate data from the UniProt knowledgebase, as well as explore possible EC numbers, metabolic pathways reconstruction, COGs, or membrane prediction properties. As a complement, a specialized database, PsychroList (http://bioinfo.hku.hk/PsychroList/), within the GenoChore suite (Fang et al. 2005), allows the user to search for patterns in DNA or protein sequences, taking into account a clustering of genes into formal operons as well as providing extra facilities to query sequences using predefined sequence patterns.

Results and Discussion

Genome organization

As in many marine γ-proteobacteria, the *P. haloplanktis* TAC125 genome is made of two chromosomes (Table 1; Supplemental Table 2). The replication origin of chromosome (chr) I maps near *dnaA* (McLean et al. 1998; Lobry and Louarn 2003) in a region that is highly conserved in γ-proteobacteria (Fig. 1). However, in remarkable contrast with the genomes of the vibrios (Okada et al. 2005), the second chromosome does not display a standard GC skew (Supplemental Fig. 1). The pattern observed is likely to be caused by unidirectional replication. To our knowledge, this is the first time that such a system would be uncovered in an authentic bacterial chromosome. This is supported by the signature of R1 plasmid replication (del Solar et al. 1998): the tus and repA-like genes, the repA and dnaA boxes, and the parA and parB genes (Fig. 1). In addition, kisB and kidB coding for a typical plasmid maintenance system have also been found in chrII. We chose the start at the centre of a TATATA palindrome near the genes coding for the partition system. The G+C content and gene density of chrII match that of chrI (Supplemental Table 2). It contains the essential genes *hisS* and *gcpE*, in addition to a series of genes ubiquitous in γ-proteobacteria (Supplemental Table 3). A third of chrII genes have orthologs in *Escherichia coli*. Remarkably, the whole metabolism of histidine is coded in chrII, in a highly conserved gene cluster (Fig. 1). Nineteen percent of the *P. haloplanktis* chrII genes show high similarities with plasmid-encoded genes, further suggesting that this replicon was a plasmid recruited to become a chromosome encoding essential genes (Fig. 1; Supplemental Table 2).

Genes around the origin of replication in chromosome display a high level of synteny with genes of other known proteobacteria. chrI codes for nine rDNA clusters (23S, 5S, and 16S RNAs, one operon has two copies of the 5S RNA gene), a large number compared with that found in most sequenced γ-proteobacteria (Ussery et al. 2004). The genome sequence of *P. haloplanktis* TAC125 shows some variability in the rRNA genes of ~1% (interestingly, most variations correspond to compensating mutations in regions coding for double stranded RNA) (data not shown). However, this probably does not influence phylogenies (Cilia et al. 1996), and the rRNA sequences are in line with the established phylogeny, placing TAC125 near vibrios and *Shewanella* (Ivanova et al. 2004). In the same way, the number of tRNA genes is quite high (106 genes), a feature in common with that in vibrios and in *Photobacterium profundum* (Table 1). These genes are organized in long runs of repeated sequences. The longest contains 19 tRNA genes in a row, seven of which coding for an identical $tRNA^{glu}$ (TTC anticodon), suggesting a slipped mispair mechanism of expansion in situations of rapid growth. Genomes with an origin of replication display a protein-coding gene distribution bias, with most essential genes located in the leading replication strand (Rocha and Danchin 2003). Interestingly, while the number of CDSs located in the leading strand of the chromosome is 61%, 72% of the tRNA genes are located in that strand. Because the speed of transcription/translation must be limited at low temperature, the large number of rRNA and tRNA genes may participate in the adaptation, allowing fast growth of the organism in the cold. TAC125 is similar to other bacteria in terms of number of tandem repeats. Several genes relevant to adaptation to cold conditions are clustered together in approximate repeats (Supplemental Fig. 2): genes coding for cold-shock proteins, nine paralogs of *cspA*, as in *E. coli* (four in chrII, three of them clustered together), as well as genes coding for a class of putative short secreted proteins that could bind calcium, next to a divalent metal exporter system, most likely used in calcium export. Calcium is known to be involved in cold

Figure 1. Circular representation of the *Pseudoalteromonas haloplanktis* genome. Circles display (from the outside): (1) predicted coding regions transcribed in the clockwise direction; (2) predicted coding regions transcribed in the counterclockwise direction. Genes displayed in 1 and 2 are color-coded according to different functional categories: salmon indicates amino acid biosynthesis; orange indicates purines, pyrimidines, nucleosides, and nucleotides; purple indicates fatty acid and phospholipid metabolism; light blue indicates biosynthesis of cofactors, prosthetic groups, and carriers; light green indicates cell envelope; red indicates cellular processes; brown indicates central intermediary metabolism; yellow indicates DNA metabolism; green indicates energy metabolism; pink indicates protein fate/synthesis; blue indicates regulatory functions; grey indicates transcription; teal indicates transport and binding proteins; and black indicates hypothetical and conserved hypothetical proteins. (3) tRNAs (green) and rRNA (pink) on chrI/genes similar to phage proteins (red) on chrII; (4) and *tonB* and *tonB-like* genes in grey. Chromosome II gene names similar to that of the R1 plasmid replication apparatus (unidirectional) are colored in green.

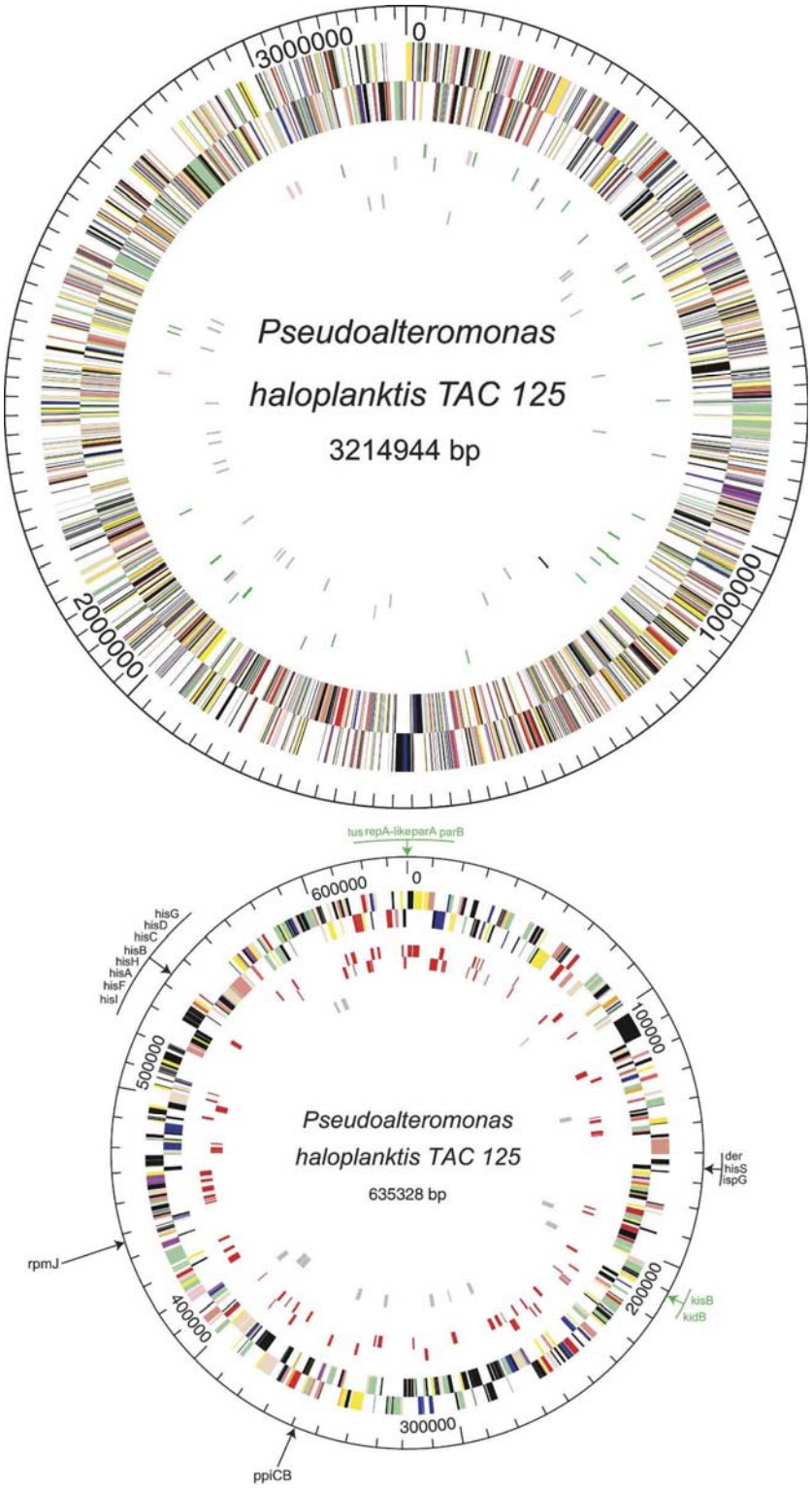


Figure 1. (Legend on facing page)

adaptation and formation of exopolysaccharides (EPS) in bacteria (Kierek and Watnick 2003; Dominguez 2004). The duplication in chrII of the *ast* operon may be involved in adaptation to cold and high osmolarity.

Comparative genomics

Of the 3488 identified protein coding genes (CDSs) (Table 1), a biological function, based on a classification scheme adapted from Riley (Riley 1993; Fang et al. 2005), has been assigned for >65.6% (36.7% with a final assignment and 28.9% with a putative role assignment). More than 63% of the *P. haloplanktis* CDSs are similar to *Shewanella oneidensis* genes (Heidelberg et al. 2002), of which 47% are found in synteny groups, making the comparison with this aquatic organism particularly revealing (Fig. 2A). These mesophilic marine bacteria, together with *Vibrio vulnificus*, share with *P. haloplanktis* several sodium-type flagellar proteins (*mot* genes), sodium-dependent transporters, gene clusters of type IV pilus and flagellin proteins, and some iron ABC transporter and *tonB* system-dependent transport proteins (Fig. 2B; Supplemental Table 4C). A putative cold-shock RNA methyltransferase (PSHAb0516) and a cold-shock regulated carbon starvation protein A (PSHAb0210) have also been found in *S. oneidensis* and *V. vulnificus*. Other *P. haloplanktis* TonB-dependent receptors and calmodulin-like proteins have counterparts in *S. oneidensis* only, while *V. vulnificus* shares with *P. haloplanktis* several integrases/transposases, putative potassium channel proteins and metabolite exporters (Supplemental Table 4A,B). Gene content comparisons have also been performed with the two closest psychrophilic γ-proteobacteria, *Idiomarina loihiensis* and *Photobacterium profundum* (Fig. 2B): apart from the set of genes shared also with the marine bacteria, many TonB-dependent receptors are specific to *P. haloplanktis* and *I. loihiensis*, and two sodium-dependent carbohydrate transporters (a permease and a transporter) have been found in *P. profundum* only (Supplemental Table 4D,E). The set of genes common to *P. haloplanktis* and these two psychrophilic bacteria contain two copper resistance protein genes (*copA* and *copB* genes in chrII) (Supplemental Table 4F). These genes belong to a synteny group made of six genes with *I. loihiensis* and, interest-

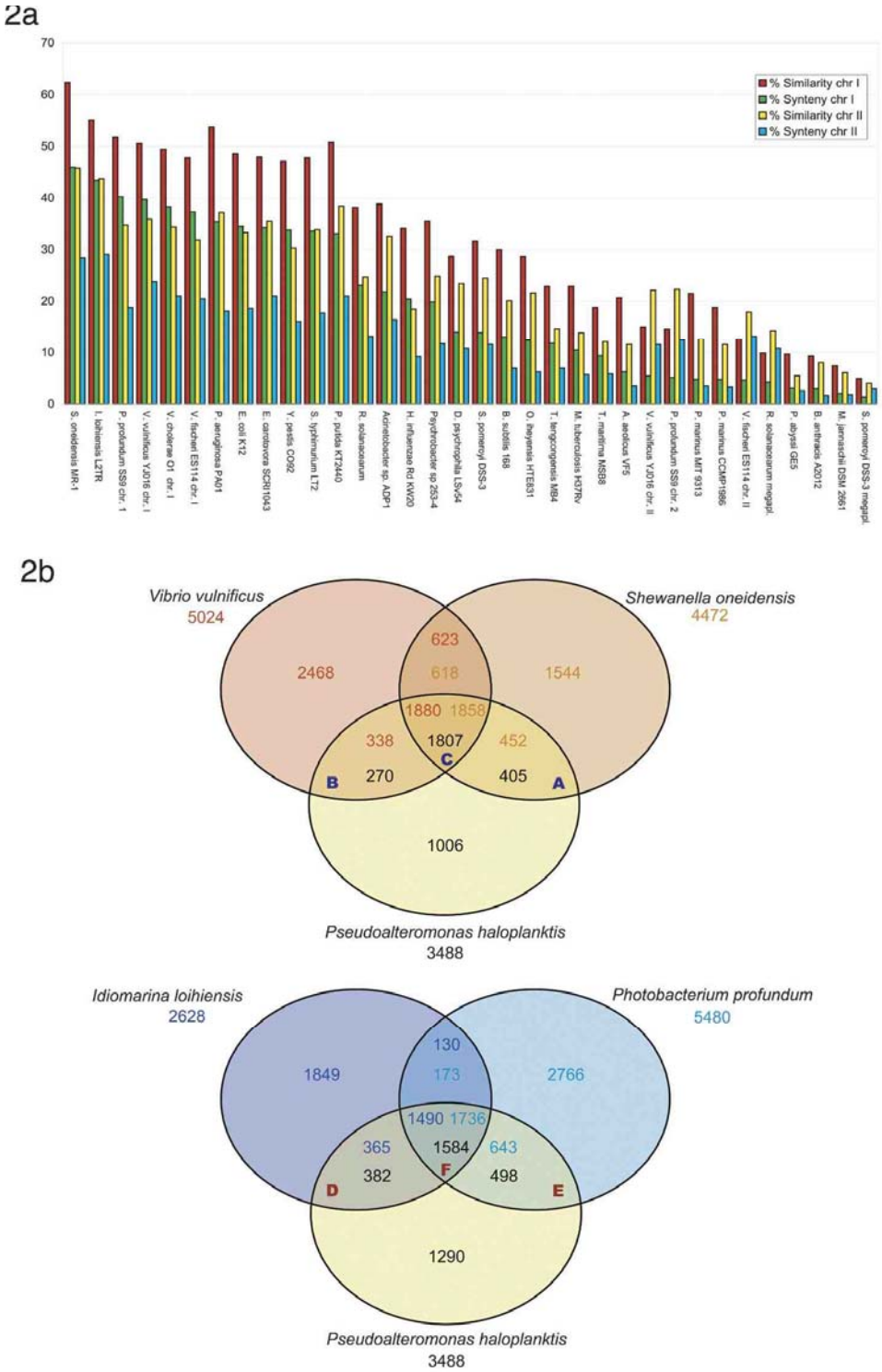


Figure 2. Putative orthologs and syntenies between the genome of *P. haloplanktis* and the genome of related bacteria. The alphabetic letters A to F refer to the Supplemental Table 4 (T4, A to F). (*A*) The percentage of *P. haloplanktis* genes homologs to a selection of 34 complete bacterial proteomes (i.e., 30% identity and a ratio of 0.8 of the length of the smaller protein to that of the larger one) is represented by red bars for chrI and by yellow bars for chrII. The percentage of *P. haloplanktis* genes in synteny groups with a selection of 34 complete bacterial genomes is represented by green bars for chrI and by blue bars for chrII. The closest organism is *S. oneidensis*. (*B*) Comparison of the gene content of *Pseudoalteromonas haloplanktis*, *Shewanella oneidensis*, *Vibrio vulnificus*, *Photobacterium profundum*, and *Idiomarina ioihiensis*. Putative orthologs are defined as genes showing a minimum of 30% identity and a ratio of 0.8 of the length of the smaller protein to that of the larger one, or as two genes included in a synteny group. The intersections between the three circles give the number of genes found in the two or three compared species. Genes outside these areas are specific to the corresponding organism. The total number of annotated genes is also given under each species name.

ingly, with the megaplasmid of *Ralstonia solanacearum*. They are also found in other distant cold marine bacteria such as *Psychrobacter* sp. and *Silicibacter pomeroyi* (in its megaplasmid; Supplemental Table 5).Two synteny groups (a potassium efflux system and an urease operon) are found with *S. pomeroyi* only. A total of 133 *P. haloplanktis* genes (73 hypothetical), including several insertion sequences and genes coding for TonB-dependent receptors, have not been found in the genome of the cold-adapted marine bacteria used for comparison (Supplemental Table 6). In addition, a prophage region (50 kb long, between genes PSHAa1505 and PSHAa1558) is specific to *P. haloplanktis*. We also identified one specific region in chrI coding for several calcium-dependent proteins, as well as a specific gene in chrII that may regulate cell volume and resistance to cold conditions (PSHAb0555).

General features of the proteome

Global properties of the proteins at the level of individual amino acids and motifs integrate all kinds of selection pressure associated to adaptation to cold. The pattern of amino acid distribution in γ-proteobacteria from different biotopes (Supplemental Fig. 3) displays an overall trend similar in the various genomes of interest, with leucine (L) being most abundant, while tryptophan (W), cysteine (C), histidine (H), and to a lesser extent methionine (M) are used infrequently. The proteome of the thermophilic genomes looks significantly different from that of the mesophilic and psychrophilic counterparts (strong avoidance of glutamine [Q] in thermophilic species, preference for alanine [A] in mesophilic and psychrophilic species, except in *Oceanobacillus iheyensis*). The amino acid distribution in mesophilic and psychrophilic species display a few noteworthy differences specifically relevant to growth in the cold (in particular in the relative abundance of N and Q) (Supplemental Fig. 3). Using correspondence analysis (CA) (Benzécri 1984) coupled to dynamic clustering (Delorme and Henaut 1988), to identify subtle differences in this cluster of related objects, five classes with close amino acid composition were found (Fig. 3): (1) integral inner membrane proteins (~12%); (2) proteins involved in the metabolism of small molecules (25%); (3) associated to information transfer pathways (21.5%) (4) associated to the outer membrane or secreted (21.5%); and (5) with unknown functions, or likely to be of phage origin (20%). This biological consistency demonstrates that a relationship exists between amino acid composition and the role of the protein inside the cell. The two first biases have been previously identified: They are driven by the membrane compartmentalization of some proteins and by the G+C-content of codons (Pascal et al. 2005). In contrast, the bias scattering proteins along the third factorial axis was unexpected. It discriminated proteins in *P. haloplanktis* TAC125 according to their asparagine (N) content. This small hydrophilic uncharged amino acid carries an amide group that is often thermolabile (Zhou et al. 2000; Stratton et al. 2001; Weintraub and Manson 2004). Analysis of this remarkable N-driven bias was further submitted to CA using the pool of two psychrophilic (*Desulfotalea psychrophila* and *P. haloplanktis*), two mesophilic (*E. coli* K-12 and *Bacillus subtilis*) and two thermophilic (*Aquifex aeolicus* and *Thermotoga maritima*) bacteria. An N-driven bias was observed in proteins from the psychrophiles, differentiating them from their mesophilic and thermophilic counterparts (Supplemental Fig. 4). The proteins responsible for the bias belong to the following structures or processes: motility of the organism, cell wall, outer membrane, transport, sensor activity, adaptation to atypical conditions, and secretion. Some proteins of DNA metabolism (replication, packaging, segregation, restriction/modification, proof-reading, and repair) and RNA metabolism are also present. Asparagine residues often undergo deamidating cyclisation, a process extremely sensitive to temperature (Daniel et al. 1996). The *P. haloplanktis* TAC125 proteome is enriched in N residues compared with counterparts that prefer growth at higher temperature, making it an organism of choice for foreign protein production when deamidation ought to be put to a minimum (Weintraub and Manson 2004).

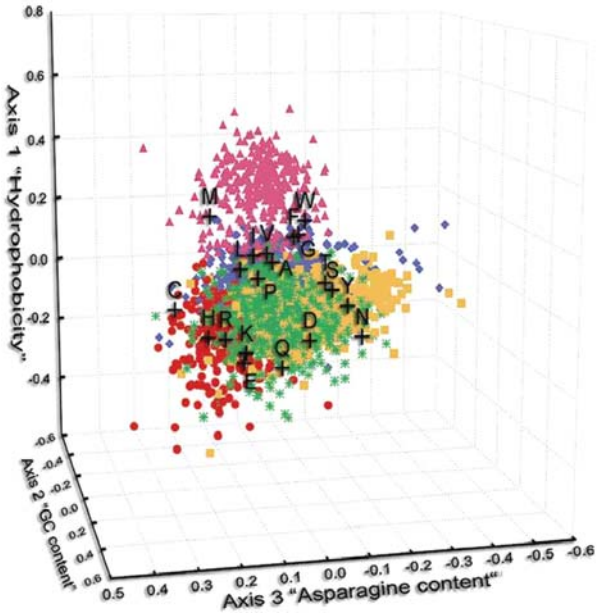


Figure 3. Distribution of the protein sequences on the CA space determined by the three first factors. The first one discriminates proteins by their hydrophobicity; the second one, by the G+C content of genes coding for them; and the third one, by their asparagine content. Five classes of proteins are found by clustering method (see "General Features of Proteome") and are represented: (1) IIMP by pink triangles, (2) proteins involved in the metabolism of small molecules by blue diamonds, (3) proteins associated to information transfer pathways by red circles, (4) proteins associated to the outer membrane or secreted by green stars, and (5) proteins with unknown functions, or likely to be of phage origin, by yellow squares. Amino acids are represented by black pluses.

RNA folding and metabolism

The genome of *P. haloplanktis* TAC125 contains 19 genes presumably coding for known RNA binding proteins or RNA chaperones. The most unexpected feature of the genome is the prominent absence of a RNA/nucleoid-associated cold-shock gene ubiquitous in γ -proteobacteria, *hns*. Analysis in silico failed to uncover a H-NS–like protein gene even when using the highly relaxed comparison criteria that uncovered counterparts in bacteria phylogenetically distant from *E. coli* (Tendeng and Bertin 2003). In contrast, the genes coding for all other nucleoid-associated proteins such as HU, IHF, FIS, and Hfq are present in the genome. The existence of a complete set of genes (including *hns*) was also observed in other marine and cold-adapted bacteria, including *S. oneidensis* (Heidelberg et al. 2002) and *D. psychrophila* (Rabus et al. 2004). In vivo complementation of a *hns* defect of *E. coli* at room temperature led us to repeatedly isolate clones coding for the counterpart of *csrA* as an efficient complementation of the *hns*

defect of *E. coli*. The lack of the *hns*-encoding gene in the psychrophilic *P. haloplanktis* genome shows that H-NS is not sufficient to promote growth at low temperature and that its role is connected to that of the regulatory protein/RNA complex CsrA/CsrB (Fig. 4). The importance of control of RNA folding and degradation at low temperature is visible in the presence of many RNA helicases (three copies of *rhlE*, two in chrI and one in chrII, and possibly a fourth one, PSHAa0641, and two copies of *srmB*, instead of one in *E. coli*). Interestingly, in contrast with the situation reported for *Oleispira antarctica* (Ferrer et al. 2004) the *groES groEL* genes from TAC125 did not permit growth of *E. coli* at low temperature. Other factors are therefore important for cold adaptation in the present organism.

Several RNA motifs indicating the presence of RNA-coding genes have been found in chrI using the Rfam databank (see Methods): a tmRNA at position 2,231,297 bp, the RNA component of RNase P at position 2,674,641 bp, the t44 RNA (of unknown function), and three riboswitch structures (Nudler and Mironov 2004): a lysine riboswitch just in front of *dapA* gene (at position 189,092 bp); the RFN element, located at 1283,710 near *ribB* gene; and the "ubiquitous" TPP riboswitch (THI element) at 496,922 bp (upstream of the *thiC* gene).

Response to oxygen and reactive oxygen species

The solubility of gasses increases rapidly at low temperature. This is the case of dioxygen, which is a very reactive molecule. We expected that the proteome would comprise a vast arsenal of enzymes active against H_2O_2 and superoxide. This was the more so because the organism is indeed remarkably resistant to H_2O_2 (Supplemental Fig. 5). Surprisingly, we found only the gene counterpart coding for the iron superoxide dismutase (*sodB*) and only one clear catalase (*katB*) located in chrII, with a possible paralog in chrI (PSHAa1737). Furthermore, while the oxygen responding OxyR control is present, the SoxR regulation is absent. This unexpected finding was explained when we discovered that *P. haloplanktis* TAC125 lacks a series of activities that result in reactive oxygen species (ROS) production. Despite the availability of molybdate in sea water (Hille 2002), *P. haloplanktis* TAC125 entirely lacks molybdopterin metabolism: not only are the biosynthetic and transport gene absent but genes coding for enzymes using the cofactor are also missing (e.g., TMAO reductase, xanthine oxidase, biotin sulfoxide reductase, or the novel oxido-reductase YedY) (Loschi et al. 2004). All related organisms such as the vibrios or *Shewanella* have molydopterin metabolism, as do almost all Bacteria and Eukarya. The cells, however, must cope with increased oxygen solubility and inevitable interaction with reduced iron, leading to the deleterious Fenton reaction and ROS. A way for the cells to protect their metabolism against those is to use dioxygen directly. This is seen in the large number of putative dioxygenases present in both chromosomes (Table 2). To this list one should add the fatty acids desaturases (see below) that also play a role to increase membrane fluidity at low temperature.

Furthermore, the protective role of methionine against ROS is enhanced by the existence of two MsrA proteins, instead of one as is usual, together with a fused MsrA-MsrB protein, a rare situation, found in *Nitrosomonas europaea*. Interestingly, one MsrA copy, PSHAa2274, is present in all related genomes, while the second copy, PSHAa1583, not present in *P. profundum*, exists in Shewanella and vibrios, and two copies in *D. psychrophila* that develop preferentially at low temperature. A further sign that the bacteria cope with the specific problems posed by ROS is the number of proteins involved in scavenging chemical groups affected: peroxiredoxins such as alkyl hydroperoxide reductase AhpC, AhpCB, thiol-specific cytoplasmic peroxidase BcpA; thiol-specific periplasmic peroxidase Tpx, and their coupled flavoproteins: AhpF, TrxB, PSHAa0892, Gor. All these activities would protect against the inevitable consequences of the Fenton reaction, and the needed sulfur metabolism genes, involving glutathione and *S*-adenosylmethionine, are all present. Furthermore, a control system that would limit the concentration of copper in the cell (Cu^{2+} is particularly reactive toward oxygen) involves at least two transport systems to expel copper out of the cell (CusA, CopAB, in chrI and PSHAb0008, PSHAb0009, CopB, CopA, and PSHAb0325 in chrII), a periplasmic disulfide reductase (DipZ), and a periplasmic putative laccase (PcoA) that could chelate copper ions while also acting as a dioxygenase.

Finally, an interesting gene cluster involved in fatty acid metabolism PSHAa0894–0910 is absent from other γ-proteobacteria, but is largely present in *Mycobacterium tuberculosis* (where it may have a protective role against ROS, contributing to virulence). In short, *P. haloplanktis* TAC125 is remarkably well adapted to protection against ROS under cold conditions, a feature that could be very useful for expression of foreign proteins in the cold.

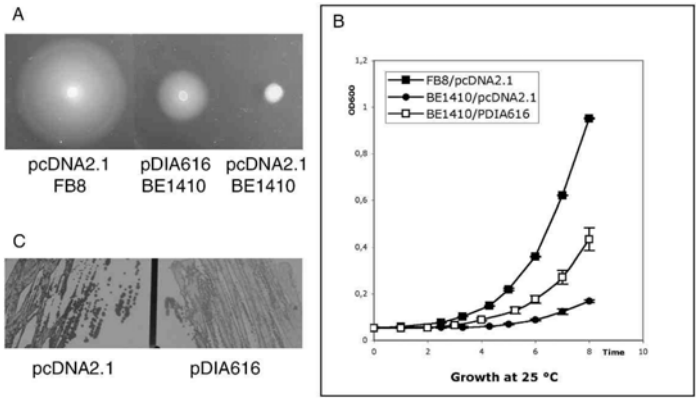


Figure 4. *hns* complementation in *E. coli* with *P. haloplanktis csrA*. A DNA fragment encompassing *P. haloplanktis* gene *csrA* with 140 bp of its upstream region was cloned into plasmid pcDNA2.1 (pDIA616) and the phenotypes of the resulting *E. coli hns* transformants were compared to the *hns* mutant (BE1410) and to the wild-type FB8 parental strain. (*A*) Motility assay: Partial motility is restored with the *csrA* gene of *P. haloplanktis*. Other phenotypes such as serine sensitivity of the *hns* mutant are restored by the *csrA* gene as well. (*B*) Growth at 25°C: overnight cultures were diluted to 0.05 OD_{600} in LB medium with ampicillin 100 µg/mL, and growth was monitored (as in Dersch et al. 1994). Significant improvement of growth is witnessed with the *csrA* gene of *P. haloplanktis*. (*C*) CsrA-dependent storage of glycogen in *E. coli* MG1655. Expression of *P. haloplanktis csrA* inhibits glycogen storage (light iodine color, *right* panel).

Metabolic features

Marine bacteria are facing a medium generally unbalanced in terms of carbon, nitrogen, and phosphorous supply (Moran et al. 2004) but are not depleted

Table 2: proteins involved in cold/salt adaptation

Locus	Proteins size	Gene name	Chr.	Description
Cold-shock protein				
PSHAa0109	463	dbpA	1	Putative ATP-dependent RNA helicase; putative cold- shock protein
PSHAa0114	421	rhlB	1	ATP-dependent RNA helicase, putative cold-shock protein
PSHAa0990	617	deaD	1	ATP-dependent RNA helicase, cold shock protein A
PSHAa1184	68	cspC	1	Cold shock protein, transcription antiterminator
PSHAa1222	209	grpE	1	Putative member of the DnaK/DnaJ/GrpE foldase complex (heat/cold shock protein; nvolved in thermal regulation of folding)
PSHAa1600	204		1	Putative cold-shock DNA-binding domain
PSHAa1726	73	cspD	1	Nucleic acid-binding domain, cold-shock RNA chaperone
PSHAa2978/2979/2980	70	cspE	1	RNA chaperone, transcription antiterminator
PSHAa2981	750		1	Cold-shock RNAse R
PSHAb0078	63	cspX	2	Cold shock protein
PSHAb0210	491	cstA	2	Cold-shock regulated carbon starvation protein A
PSHAb0384/0386/0387	70	cspX	2	Cold shock protein
PSHAb0516	171		2	Putative cold-shock RNA methyltransferase
Dioxygenases				
PSHAa0187	235		1	Conserved protein of unknown function; putative dioxygenase
PSHAa0570	140		1	Conserved protein of unknown function; putative glyoxalase domain
PSHAa0900	297	tesB	1	Putative dioxygenase
PSHAa0904	372		1	Putative dioxygenase
PSHAa2137	365		1	Putative protein with ferredoxin subunits; putative dioxygenase
PSHAa2147	309		1	Putative taurine dioxygenase
PSHAa2168	351	melA	1	4-hydroxyphenylpyruvate dioxygenase (4HPPD) (HPD) (HPPDase)
PSHAa2449	158		1	Conserved protein of unknown function; putative dioxygenase domain
PSHAb0029	167		2	Putative enzyme; dioxygenase superfamily
PSHAb0041	119		2	Putative enzyme; dioxygenase superfamily
PSHAb0115	128		2	Putative enzyme; dioxygenase superfamily
PSHAb0338	434	hmgA	2	Homogentisate 1,2-dioxygenase
Fatty acids (desaturase)				
PSHAa0567	271		1	Conserved protein of unknown function; putative sterol desaturase family protein
PSHAa1269	351		1	Putative fatty acid desaturase
PSHAa2897	378		1	Putative long chain acyl-CoA desaturase
PSHAb0225	264		2	Putative C-5 sterol desaturase
Salt adaptation				
PSHAa0020	459	trkA	1	Potassium transport inner membrane protein subunit; involved in osmoprotection
PSHAa0208	257	cysQ	1	3',5' adenosine diphosphate 3' phosphatase, sodium sensing
PSHAa0325	188	osmY	1	Putative hyperosmotically inducible periplasmic protein
PSHAa0326	53		1	Putative low temperature and salt responsive protein
PSHAa0396	273	mscS	1	Putative mechanosensitive channel protein; protection against hypo-osmotic shock
PSHAa0687	255	surE	1	Putative acid phosphatase SurE; protection against osmotic shock
PSHAa0833	512	putP	1	Major sodium/proline symporter; protection against osmotic shock
PSHAa1072	522		1	Putative glycine betaine transporter
PSHAa1226	439		1	Putative Na+/H+ antiporter
PSHAa1436	531		1	Putative choline/betaine transporter
PSHAa1625	403		1	Putative sodium ABC exporter, permease component
PSHAa1626	263		1	Putative sodium ABC exporter, ATP-binding component
PSHAa1678	675	prc	1	Periplasmic carboxy-terminal protease with specificity for non-polar C-termini; protection against osmotic shock
PSHAa1679	213	proQ	1	Putative post-translational activator of ProP expression; sensing osmotic shock
PSHAa2041	676		1	Putative choline/betaine transporter
PSHAa2202	144	osmC	1	Osmotically inducible protein C; protection against oxidative stress
PSHAa2243	102	bolA	1	Regulator involved in adaptation to osmotic shock
PSHAa2252	561		1	Putative voltage-gated ClC-type chloride channel ClcA
PSHAa2274	660	mdoB	1	Putative phosphoglycerol transferase; shape adaptation to osmotic shock
PSHAa2849	434	envZ	1	Sensor histidine kinase, senses osmolarity
PSHAa2850	240	ompR	1	Response regulator for adaptation to osmolarity
PSHAb0106	576	cvrA	2	Putative Na+/H+ antiporter; adaptation to hypo-osmotic shock
PSHAa0176	289		2	Putative mechanosensitive channel protein; protection against hypo-osmotic shock
PSHAb0127	1536	gltB	2	Glutamate synthase, large subunit, GOGAT
PSHAb0128	497	gltD	2	Glutamate synthase, small subunit
PSHAb0261	534	betC	2	Putative choline dehydrogenase
PSHAb0357	639	dnaK	2	Chaperone protein DnaK; involved in protection
Against osmotic shock				
PSHAb0381	129	betC	2	Putative osmC-like protein
PSHAb0418	556	betA	2	Putative choline dehydrogenase
PSHAb0419	500	betB	2	NAD^+-dependent betaine aldehyde dehydrogenase
PSHAb0420	197	betI	2	Putative transcriptional repressor for the cellular response to osmotic stress

(continued)

Table 2: *Continued*

Locus	Proteins size	Gene name	Chr.	Description
Against osmotic shock				
PSHAb0426	489		2	Putative succinylglutamic semialdehyde dehydrogenase; may be involved in adaptation to high osmolarity
PSHAb0427	345		2	Putative arginine succinyltransferase; may be involved in adaptation to high osmolarity
PSHAb0428	402		2	Putative acetylornithine transaminase; may be involved in adaptation to high osmolarity
Helicases [DEA(DH)]				
PSHAa0109	463	dbpA	1	Putative ATP-dependent RNA helicase; putative cold-shock protein; could be specific for rRNA folding
PSHAa0114	421	rhlB	1	ATP-dependent RNA helicase with nucleoside triP hydrolase domain, putative cold-shock protein
PSHAa0510	409	srmB	1	ATP-dependent RNA helicase
PSHAa0641	433		1	ATP-dependent RNA helicase(rhlE-like); DEAD-box protein family
PSHAa0990	617	deaD	1	ATP-dependent RNA helicase, cold shock protein A
PSHAa1144	1298	hrpA	1	helicase, ATP-dependent
PSHAa1432	441	srmB	1	ATP-dependent RNA helicase, DEAD box family
PSHAa1522	415		1	Conserved protein of unknown function
PSHAa1930	467	rhlE	1	Putative ATP-dependent RNA helicase with P-loop hydrolase domain
PSHAa1991	692	dinG	1	Putative ATP-dependent helicase DinG
PSHAa2216	809	hrpB	1	ATP-dependent helicase
PSHAa2480	831	rnr	1	RNase R, 3'-5' exoribonuclease
PSHAa2572	608	recQ	1	ATP-dependent DNA helicase
PSHAa2624	965	hepA	1	RNA polymerase associated protein (ATP-dependent helicase HepA)
PSHAa2762	676	rep	1	Rep helicase, a single-stranded DNA-dependent ATPase
PSHAa2983	435	rhlE	1	DEAD-box protein family; putative ATP-dependent RNA helicase with P-loop hydrolase domain
PSHAb0003	1049		2	Putative DNA helicase with DEAD/DEAH box helicase domain
PSHAb0039	1352		2	Putative protein with DEAD/DEAH box helicase domain
PSHAb0119	649		2	Putative ATP-dependent DNA helicase (recQ-like)
PSHAb0411	434	rhlE	2	Putative ATP-dependent RNA helicase with P-loop hydrolase domain; DEAD-box protein family
PSHAb0497	639	yoaA	2	Putative ATP-dependent helicase YoaA with nucleotide triphosphate hydrolase domain, SOS repair, DinG family

in sulfur sources. Strain TAC125 is adapted to fast growth, suggesting that it regularly encounters a fairly rich medium (this is probably due to its propensity to make a water/air biofilm [see below], allowing it to live in region full of plankton debris). Excess of several easily metabolized carbon sources present simultaneously is unlikely, making catabolite repression the exception rather than the rule. Indeed, *P. haloplanktis* TAC125 is lacking the cAMP-CAP complex that regulates carbon availability in related organisms such as vibrios and *Shewanella*. Furthermore, in contrast to many γ-proteobacteria (including vibrios), it does not possess a phosphoenolpyruvate-dependent phosphotransferase system for the transport and first metabolic step of carbohydrate degradation. This accounts for its lack of growth on glucose, likely to be phosphorylated by glucokinase (PSHAa1364). When oxygen is present at a high level, the presence of glucose-phosphate isomerase might drive the Embden Meyerhof pathway and subsequently activate NADPH-dependent aldose reductase (PSHAa2392). This would affect the polyol pathway, lowering the amount of NADPH available for the reduction of oxidized glutathione by glutathione reductase. Indeed, growth is inexistent on glucose unless tyrosine is supplemented to the medium (data not shown). The counterpart of PSHAa2392 has been implicated in a process controlling tyrosine bradytrophy in *E. coli* (Timms and Bridges 1998).

An essential step for biomass construction is formation of pyruvate, which, because of the absence of the PTS, must go through an alternative pathway starting with pyruvate kinase. In contrast to *E. coli*, with two such enzymes, and vibrios, with three, TAC125 possesses only one pyruvate kinase. Interestingly, it is homologous to the cold-adapted PykA enzyme of *E. coli*. This is further in line with gluconate as a preferred carbon source, providing pyruvate directly through the Entner-Doudoroff pathway (Edd, Eda), which also provides the level of NADPH needed for protection against oxygen toxicity. As in vibrios, phosphoenolpyruvate synthase (*ppsA* gene) and phosphoenolpyruvate carboxylase (*ppc* gene) genes are located in chrII. However, *ppsA* is located in chrI in *P. profundum*.

TAC125 grows in minimal medium under anaerobic conditions (data not shown), in line with the presence of the *fnr* gene, while the putative aspartate ammonia lyase PSHAa0048 would provide the needed fumarate electron acceptor.

In contrast, the metabolism of nitrogen appears to be highly similar to that in phylogenetically related organisms and be controlled by a phosphorylation cascade involving PtsP (a homolog of the PTS enzyme PtsI) that phosphorylates PtsO and the regulator PtsN, controlling all sigma54-dependent operons. Arginine catabolism could provide a direct source of ammonia under nitrogen-limiting conditions, through the AST pathway, while providing metabolites for adapation to cold (Schneider et al. 1998). The organism can metabolize *N*-acetyl-glucosamine, a carbon and nitrogen source ubiquitously present in marine environment (Riemann and Azam 2002). In the same way, phosphate input in metabolism is controlled by the counterparts of PhoB, PhoR, and PhoU, with several putative transport systems, including one of high affinity.

Most coenzymes can be synthesized in the organism except coenzyme B12 and, as discussed, molybdopterin. There does not appear to exist a selenium metabolism, in line with the high

reactivity of that atom toward oxygen. Other marine bacteria (including the closest one, *S. oneidensis*) do have a selenium metabolism.

Growth, yield, and adaptation to salt

A remarkable feature of TAC125 is that, when provided with sufficient nutrients and aeration, it grows to very high density under laboratory settings, even at 0°C. The very high growth yield indicates that respiration must be particularly efficient in this bacterium. The cells are well adapted to salt, and although they can grow in low osmolarity media, optimal growth is between 1.5% and 3.5% NaCl. We looked for systems that would account for controlling osmolarity in the cell. The common trehalose system does not seem to be present. In contrast, chrII harbors two copies of a choline dehydrogenase (PSHAb0261 and PSHAb0418) for synthesis of glycine betaine, an extremely efficient osmoprotectant (Felitsky et al. 2004). Along the same line, GOGAT glutamate synthase is coded in that same chromosome, allowing synthesis of glutamate that, as potassium glutamate, is the most common response of bacteria to increased osmolarity (Table 2; Lee and Gralla 2004).

All autonomous organisms have at least one pathway to degrade *S*-adenosylhomocysteine (AdoHcy). The MtnN(Pfs)/LuxS pathway leads to synthesis of the quorum sensing (QS) effector autoinducer-2 (AI-2). Xanthomonadales and Pseudomonadales aside, all γ-proteobacteria sequenced use that pathway. LuxS, responsible for the last enzymatic step of AI-2 synthesis is present in bacteria closest to TAC125, and AI-2 produced by heterologous organisms triggers luminescence in reporter strains. The *mtnN* gene is present in TAC125. We failed, however, to identify the *luxS* gene. We looked further for other genes involved in AdoHcy degradation and recycling to homoserine lactones or autoinducer CAI-1. Among other systems, involving quinolones, cyclic dipeptides, or indole, none appears to be present, and in assays, using *Photorhabdus luminescens* as reporter TAC125 supernatant did not trigger luminescence. This does not, however, exclude the presence of less well known QS systems: gene PSHAa0159 codes for a multidomain putative aconitate hydratase that may use aconitate as an iron-sulfur cluster-dependent signal in stationary phase (Kiley and Beinert 2003). Furthermore, TAC125 possesses several enoyl-CoA hydratase-like genes that may be involved in synthesis of a diffusible signaling factor as in plant pathogenic γ-proteobacteria (Barber et al. 1997).

Membranes, motility, biofilms, and secretion

P. haloplanktis TAC125 deals with the membrane fluidity challenge at low temperature by lipid desaturases, which simultaneously protect against dioxygen (Table 2). Two clusters of genes, absent in the closest genomes, may be involved in the degradation of steroids or hopanoids, membrane rigidifying molecules present in the environment of heterotrophic bacteria. Protein export from the cytoplasm is similar to that of proteobacteria, with the long form of SecE. Type II secretion is functional (GSP proteins are present) as is the TAT export system. In contrast, type III secretion is absent. TAC125 possesses two gene clusters in chrI and one in chrII for the biosynthesis of type IV pili and of curli, respectively. Furthermore, chrII encodes elements of the new pathway involved in secretion of a specific amylase composed of a signal peptide, the mature enzyme, and a long C-terminal propeptide without foldase function or action on amylase activity (Claverie et al. 2003). Amylase secretion required two accessory proteins, PSHAb0130 (possibly an outer membrane associated protein) and PSHAb0132, coding for a conserved secreted protein present in several phytopathogenic γ-proteobacteria. Interestingly, the neighboring glutamate synthesis genes were essential to allow amylase secretion when reconstructed in *E. coli*.

P. haloplanktis has several genes and operons that may play an important role in colonization of both biotic and abiotic surfaces. In particular, up to 16 genes involved in the synthesis of mannose-sensitive agglutinin are located on chrI. These genes have been recently demonstrated to facilitate adhesion to the chitin surface in *V. cholerae* (Meibom et al. 2004). When investigating the formation of biofilms on solid surfaces (many genes are compatible with synthesis of a biofilm such as production of EPS), we did not observe synthesis of a strong biofilm. In contrast, the air–water interface was rapidly occupied by a dense layer of compact cells (Supplemental Fig. 6), suggesting that, for this organism, this is the normal way to concentrate cells and occupy a biotope. The formation of such a biofilm compatible with life in water and scavenging of organic particles that concentrate in the foam of waves has been recently demonstrated in *V. parahaemolyticus* (Enos-Berlage et al. 2005). Genes for the synthesis of polar flagellum are present, and these appendages are visible under the microscope. In salty water the cells are highly motile. However, in contrast to the situation with several vibrios, pseudomonads, and related bacteria, the cells have a reduced motility in low salt media, while they still grow well under such conditions (Supplemental Fig. 7). In minimal medium, the strain grew in a large range of NaCl concentration (0% to up 11% NaCl). In rich media, however, the growth of the strain is remarkably sensitive to the presence of salt. At 20°C, no growth occurs in the absence of NaCl. In remarkable contrast, slow but significant growth is observed at 4°C in the absence of salt, suggesting some adaptation to ice or melting ice water.

Conclusions

P. haloplanktis TAC125 has found much unexpected solutions to cope with cold. Not only does it grow fast under such conditions, but it displays remarkable resistance to ROS. Moreover, as seen in silico with its proteome composition, it provides a way to resist to the aging features involving asparagine cyclisation and deamidation. This makes this bacterium not only a model for the study of adaptation to cold marine conditions but also a promising tool for biotechnology production of proteins.

Methods

Bacterial strains, growth media, and assays

The *P. haloplanktis* strain TAC125 is deposited and available at the Institut Pasteur Collection (CIP). *Escherichia coli* strains used in this work are strains MG1655, FB8, and its *hns* defective counterpart BE1410 (Laurent-Winter et al. 1997). TAC125 bacteria were grown in TYP rich media: 16 g/L yeast extract (DIFCO) and 16 g/L bacto-tryptone (DIFCO) supplemented with NaCl as required. For growth on specific carbon sources (0.4%), the following minimum medium was used: 10 g/L Na_2HPO_4, 3 g/L KH_2PO_4, 1 g/L K_2SO_4, 20 g/L NaCl, 0.4 g/L $MgSO_4.7H2O$, 0.1 g/L $CaCl_2$, 0.018 g/L $FeSO_4.7H_2O$, and 3 g/L NH_4Cl (pH 7). Motility assay of *E. coli* cells with plasmids carrying *P. haloplanktis* genes was performed on semisolid plates with 0.5% NaCl 16 h. For CsrA-dependent storage of glycogen analysis, wild-type MG1655 *E. coli* colonies growing on Kornberg medium with 1% glucose and 100

µg/mL ampicillin with (pDIA616) and without (pcDNA2.1) *csrA* were assayed twice for glycogen accumulation. The plates were stained with iodine vapor after 48 h growth at 25°C as in Liu and Romeo (1997). Oxidative stress adaptation was assayed as follows: after overnight growth in TYP medium with 1% NaCl, bacteria are washed in fresh TYP medium with 1% NaCl and diluted to OD_{600} = 0.15. When the OD_{600} reached 0.5, the culture was separated into four equal parts and then exposed to 0, 15, 20, and 25 mM of H_2O_2. The OD_{600} was measured at various times during growth. Experiments were performed twice at 37°C for MG1655 *E. coli* strain and at 15°C for *P. haloplanktis* TAC125. For electron microscopy cells were stained with 0.1% (v/v) osmium tetroxide prepared in water: 20 µL cells were deposited onto a 300-mesh copper grids coated with Formvar (Electron Microscopy Sciences). Excess sample was removed by using Whatman 3MM paper. Bacteria were examined at 75 keV under a Hitachi H600 transmission electron microscope. Images acquisition was performed with a CCD Advantage HR Hamamatsu camera and the AMT 542 software (Advanced Microscopy Techniques).

Genome sequencing, assembly, and annotation

DNA was isolated from *P. haloplanktis* TAC125 grown in rich medium supplemented with 20 g/L NaCl. The complete genome sequence was determined by using the whole shotgun method (10 x coverage, using two plasmid libraries and one BAC library to order contigs). Finishing was performed by PCR amplification from contigs extremities. All rDNA clusters were sequenced individually. After a first round of annotation, regions of lower quality as well as regions with putative frameshifts were resequenced from PCR amplification of the dubious regions.

A first set of potential coding sequences (CDSs) was identified by using the AMIGene software (Annotation of MIcrobial Genes) (Bocs et al. 2003) trained with a set of CDSs >500 bp from the genomic sequence. Three gene models, computed from the three gene classes identified by codon usage analysis (see below), were then subsequently used together in the core of AMIGene with minimum CDSs length set to 60 bp. This second set of putative genes (made of 3488 CDSs) was submitted to functional annotation: exhaustive BLAST searches against the UniProt databank were performed to determine significant homology. Protein motifs and domains were documented by using the InterPro databank. In parallel, genes coding for enzymes were classified by using the PRIAM software (Claudel-Renard et al. 2003). PRO-TMHMM and PRODIV-TMHMM were used to identify transmembrane domains (Viklund and Elofsson 2004), and SignalP 3.0 was used to predict signal peptide regions (Bendtsen et al. 2004). Finally, tRNAs were identified by using tRNAscan-SE (Lowe and Eddy 1997).

Sequence data for comparative analyses were obtained from the National Center for Biotechnology Information (NCBI) databank. Putative orthologs between *P. haloplanktis* and the 228 other genomes were defined as genes showing a minimum of 30% identity and a ratio of 0.8 of the length of the smallest protein. Orthology relations were strengthened by synteny detection (i.e., conservation of the chromosomal colocalization between pairs of orthologous genes from different genomes) using the Syntonizer software, in which all possible kinds of chromosomal rearrangements are allowed (inversion, insertion/deletion). A "*gap*" parameter, representing the maximum number of consecutive genes not involved in a synteny group was set to five genes. Species-specific genes were identified as having no ortholog in the compared species. Specific regions are defined by at least two consecutive specific genes. Insertion of genes with similarities in the compared species was allowed. A *gap* parameter, representing the maximum number of consecutive genes with similarities, was set to two genes.

All the data (i.e., syntactic and functional annotations, and results of comparative analysis) were stored in a relational database (using the MySQL DBMS software). Each predicted gene was assigned a unique identifier prefixed with "PSHAa" for chrI and "PSHAb" for chrII. Manual validation of the automatic annotation was performed by using the Web interface MaGe (Magnifying Genomes), which allows graphic visualization of the *P. haloplanktis* annotations enhanced by a synchronized representation of synteny groups in other genomes chosen for comparisons. Translation start codons were corrected based on protein homology, proximity of ribosome-binding site, and relative position to predicted signal peptides when present. To this purpose, we used the Artemis sequence Viewer connected to the MaGe system. The *P. haloplanktis* nucleotide sequence and annotation data have been deposited at EMBL databank under accession number CR954246 and CR954247. The PseudoList database is constructed by using the MySQL DBMS, as previously described (Fang et al. 2005).

Correspondence analysis (Benzécri 1984) was used to analyze the data table with the relative synonymous codon usage values of each annotated gene as well as the table of distribution of amino acids in the proteome of *P. haloplanktis*. Clustering into consistent classes used a second method (dynamic clouds) (Delorme and Henaut 1988) that automatically clusters the objects located close to one another.

Acknowledgments

This work was supported by the European Union Network of Excellence BioSapiens, the French Ministry of Research ACI IMPBio Blastsets and MicroScope, and the Hong Kong Innovation and Technology Commission BIOSUPPORT program. G.M. and M.L.T. thank the Programma Nazionale di Ricerche in Antartide 2004, and grants L.R. 05/03 and CRdC-ATIBB, Regione Campania, Italy. S.D.A. and G.F. acknowledge the support of the Fonds National de la Recherche Scientifique, Belgium.

References

Barber, C.E., Tang, J.L., Feng, J.X., Pan, M.Q., Wilson, T.J., Slater, H., Dow, J.M., Williams, P., and Daniels, M.J. 1997. A novel regulatory system required for pathogenicity of *Xanthomonas campestris* is mediated by a small diffusible signal molecule. *Mol. Microbiol.* **24:** 555–566.

Bartlett, D.H. 1999. Microbial adaptations to the psychrosphere/piezosphere. *J. Mol. Microbiol. Biotechnol.* **1:** 93–100.

Bendtsen, J.D., Nielsen, H., von Heijne, G., and Brunak, S. 2004. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340:** 783–795.

Benzécri, J.P. 1984. *L'analyse des données. L'analyse des correspondances.* Dunod, Paris.

Bocs, S., Cruveiller, S., Vallenet, D., Nuel, G., and Medigue, C. 2003. AMIGene: Annotation of MIcrobial Genes. *Nucleic Acids Res.* **31:** 3723–3726.

Cilia, V., Lafay, B., and Christen, R. 1996. Sequence heterogeneities among 16S ribosomal RNA sequences, and their effect on phylogenetic analyses at the species level. *Mol. Biol. Evol.* **13:** 451–461.

Claudel-Renard, C., Chevalet, C., Faraut, T., and Kahn, D. 2003. Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.* **31:** 6633–6639.

Claverie, P., Vigano, C., Ruysschaert, J.M., Gerday, C., and Feller, G. 2003. The precursor of a psychrophilic α-amylase: Structural characterization and insights into cold adaptation. *Biochim. Biophys. Acta* **1649:** 119–122.

Daniel, R.M., Dines, M., and Petach, H.H. 1996. The denaturation and degradation of stable enzymes at high temperatures. *Biochem. J.*

317: 1–11.
Delorme, M.O. and Henaut, A. 1988. Merging of distance matrices and classification by dynamic clustering. *Comput. Appl. Biosci.* **4:** 453–458.
del Solar, G., Giraldo, R., Ruiz-Echevarria, M.J., Espinosa, M., and Diaz-Orejas, R. 1998. Replication and control of circular bacterial plasmids. *Microbiol. Mol. Biol. Rev.* **62:** 434–464.
Dersch, P., Kneip, S., and Bremer, E. 1994. The nucleoid-associated DNA-binding protein H-NS is required for the efficient adaptation of *Escherichia coli* K-12 to a cold environment. *Mol. Gen. Genet.* **245:** 255–259.
Dominguez, D.C. 2004. Calcium signalling in bacteria. *Mol. Microbiol.* **54:** 291–297.
Duilio, A., Tutino, M.L., and Marino, G. 2004. Recombinant protein production in Antarctic Gram-negative bacteria. *Methods Mol. Biol.* **267:** 225–237.
Enos-Berlage, J.L., Guvener, Z.T., Keenan, C.E., and McCarter, L.L. 2005. Genetic determinants of biofilm development of opaque and translucent *Vibrio parahaemolyticus*. *Mol. Microbiol.* **55:** 1160–1182.
Fang, G., Ho, C., Qiu, Y., Cubas, V., Yu, Z., Cabau, C., Cheung, F., Moszer, I., and Danchin, A. 2005. Specialized microbial databases for inductive exploration of microbial genome sequences. *BMC Genomics* **6:** 14.
Felitsky, D.J., Cannon, J.G., Capp, M.W., Hong, J., Van Wynsberghe, A.W., Anderson, C.F., and Record Jr., M.T. 2004. The exclusion of glycine betaine from anionic biopolymer surface: Why glycine betaine is an effective osmoprotectant but also a compatible solute. *Biochemistry* **43:** 14732–14743.
Feller, G. and Gerday, C. 2003. Psychrophilic enzymes: Hot topics in cold adaptation. *Nat. Rev. Microbiol.* **1:** 200–208.
Ferrer, M., Lunsdorf, H., Chernikova, T.N., Yakimov, M., Timmis, K.N., and Golyshin, P.N. 2004. Functional consequences of single:double ring transitions in chaperonins: Life in the cold. *Mol. Microbiol.* **53:** 167–182.
Georlette, D., Blaise, V., Collins, T., D'Amico, S., Gratia, E., Hoyoux, A., Marx, J.C., Sonan, G., Feller, G., and Gerday, C. 2004. Some like it cold: Biocatalysis at low temperatures. *FEMS Microbiol. Rev.* **28:** 25–42.
Heidelberg, J.F., Paulsen, I.T., Nelson, K.E., Gaidos, E.J., Nelson, W.C., Read, T.D., Eisen, J.A., Seshadri, R., Ward, N., Methe, B., et al. 2002. Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis*. *Nat. Biotechnol.* **20:** 1118–1123.
Hille, R. 2002. Molybdenum and tungsten in biology. *Trends Biochem. Sci.* **27:** 360–367.
Ivanova, E.P., Flavier, S., and Christen, R. 2004. Phylogenetic relationships among marine Alteromonas-like proteobacteria: Emended description of the family Alteromonadaceae and proposal of Pseudoalteromonadaceae fam. nov., Colwelliaceae fam. nov., Shewanellaceae fam. nov., Moritellaceae fam. nov., Ferrimonadaceae fam. nov., Idiomarinaceae fam. nov. and Psychromonadaceae fam. nov. *Int. J. Syst. Evol. Microbiol.* **54:** 1773–1788.
Kierek, K. and Watnick, P.I. 2003. The *Vibrio cholerae* O139 O-antigen polysaccharide is essential for Ca^{2+}-dependent biofilm development in sea water. *Proc. Natl. Acad. Sci.* **100:** 14357–14362.
Kiley, P.J. and Beinert, H. 2003. The role of Fe-S proteins in sensing and regulation in bacteria. *Curr. Opin. Microbiol.* **6:** 181–185.
Laurent-Winter, C., Ngo, S., Danchin, A., and Bertin, P. 1997. Role of *Escherichia coli* histone-like nucleoid-structuring protein in bacterial metabolism and stress response: Identification of targets by two-dimensional electrophoresis. *Eur. J. Biochem.* **244:** 767–773.
Lee, S.J. and Gralla, J.D. 2004. Osmo-regulation of bacterial transcription via poised RNA polymerase. *Mol. Cell* **14:** 153–162.
Liu, M.Y. and Romeo, T. 1997. The global regulator CsrA of *Escherichia coli* is a specific mRNA-binding protein. *J. Bacteriol.* **179:** 4639–4642.
Lobry, J.R. and Louarn, J.M. 2003. Polarisation of prokaryotic chromosomes. *Curr. Opin. Microbiol.* **6:** 101–108.
Lonhienne, T., Gerday, C., and Feller, G. 2000. Psychrophilic enzymes: Revisiting the thermodynamic parameters of activation may explain local flexibility. *Biochim. Biophys. Acta* **1543:** 1–10.
Loschi, L., Brokx, S.J., Hills, T.L., Zhang, G., Bertero, M.G., Lovering, A.L., Weiner, J.H., and Strynadka, N.C. 2004. Structural and biochemical identification of a novel bacterial oxidoreductase. *J. Biol. Chem.* **279:** 50391–50400.
Lowe, T.M. and Eddy, S.R. 1997. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25:** 955–964.
McLean, M.J., Wolfe, K.H., and Devine, K.M. 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J. Mol. Evol.* **47:** 691–696.
Meibom, K.L., Li, X.B., Nielsen, A.T., Wu, C.Y., Roseman, S., and Schoolnik, G.K. 2004. The *Vibrio cholerae* chitin utilization program. *Proc. Natl. Acad. Sci.* **101:** 2524–2529.
Moran, M.A., Buchan, A., Gonzalez, J.M., Heidelberg, J.F., Whitman, W.B., Kiene, R.P., Henriksen, J.R., King, G.M., Belas, R., Fuqua, C., et al. 2004. Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. *Nature* **432:** 910–913.
Nudler, E. and Mironov, A.S. 2004. The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.* **29:** 11–17.
Okada, K., Iida, T., Kita-Tsukamoto, K., and Honda, T. 2005. Vibrios commonly possess two chromosomes. *J. Bacteriol.* **187:** 752–757.
Pascal, G., Médigue, C., and Danchin, A. 2005. Universal biases in protein composition of model prokaryotes. *Proteins* **60:** 27–35.
Rabus, R., Ruepp, A., Frickey, T., Rattei, T., Fartmann, B., Stark, M., Bauer, M., Zibat, A., Lombardot, T., Becker, I., et al. 2004. The genome of *Desulfotalea psychrophila*, a sulfate-reducing bacterium from permanently cold Arctic sediments. *Environ. Microbiol.* **6:** 887–902.
Raven, J.A., Johnston, A.M., Kubler, J.E., Korb, R., McInroy, S.G., Handley, L.L., Scrimgeour, C.M., Walker, D.I., Beardall, J., Clayton, M.N., et al. 2002. Seaweeds in cold seas: Evolution and carbon acquisition. *Ann. Bot. (Lond.)* **90:** 525–536.
Riemann, L. and Azam, F. 2002. Widespread *N*-acetyl-D-glucosamine uptake among pelagic marine bacteria and its ecological implications. *Appl. Environ. Microbiol.* **68:** 5554–5562.
Riley, M. 1993. Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.* **57:** 862–952.
Rocha, E.P. and Danchin, A. 2003. Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res.* **31:** 6570–6577.
Schneider, B.L., Kiupakis, A.K., and Reitzer, L.J. 1998. Arginine catabolism and the arginine succinyltransferase pathway in *Escherichia coli*. *J. Bacteriol.* **180:** 4278–4286.
Stratton, L.P., Kelly, R.M., Rowe, J., Shively, J.E., Smith, D.D., Carpenter, J.F., and Manning, M.C. 2001. Controlling deamidation rates in a model peptide: Effects of temperature, peptide concentration, and additives. *J. Pharm. Sci.* **90:** 2141–2148.
Tendeng, C. and Bertin, P.N. 2003. H-NS in Gram-negative bacteria: A family of multifaceted proteins. *Trends Microbiol.* **11:** 511–518.
Thomas, D.N. and Dieckmann, G.S. 2002. Antarctic Sea ice: A habitat for extremophiles. *Science* **295:** 641–644.
Timms, A.R. and Bridges, B.A. 1998. Reversion of the tyrosine ochre strain *Escherichia coli* WU3610 under starvation conditions depends on a new gene tas. *Genetics* **148:** 1627–1635.
Tutino M.L., Duilio, A., Parrilli, E., Remaut, E., Sannia G., and Marino, G. 2001. A novel replication element from an Antarctic plasmid as a tool for the expression of proteins at low temperature. *Extremophiles* **5:** 257–264.
Ussery, D.W., Binnewies, T.T., Gouveia-Oliveira, R., Jarmer, H., and Hallin, P.F. 2004. Genome update: DNA repeats in bacterial genomes. *Microbiology* **150:** 3519–3521.
Viklund, H. and Elofsson, A. 2004. Best -helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci.* **13:** 1 1908–1917.
Weber, M.H. and Marahiel, M.A. 2003. Bacterial cold shock responses. *Sci. Prog.* **86:** 9–75.
Weintraub, S.J. and Manson, S.R. 2004. Asparagine deamidation: A regulatory hourglass. *Mech. Ageing Dev.* **125:** 255–257.
Zhou, F.X., Cocco, M.J., Russ, W.P., Brunger, A.T., and Engelman, D.M. 2000. Interhelical hydrogen bonding drives strong interactions in membrane proteins. *Nat. Struct. Biol.* **7:** 154–160.

Web site references

www.genoscope.cns.fr/agc/mage/psychroscope; PsychroScope

Received May 13, 2005; accepted in revised form August 4, 2005.

D.IV Psychrophile[1] ou psychrotrophe ?

Concernant le caractère psychrophile de la souche, nous considérons que *P. haloplanktis* est une véritable psychrophile parce qu'elle croît très rapidement à basse température (principal point ayant été pris en considération). Que la souche puisse ou ne puisse pas croître à 20 ou 30°C n'a finalement pas grande importance mais l'état physiologique altéré à ces températures (pas en termes de taux de croissance) est indicatif de son caractère psychrophile (Feller et Gerday 2003). Le terme psychrotrophe (qui peut vivre en dessous de 20°C) n'est plus utilisé parce qu'il est fallacieux.

D.V Principaux résultats des analyses factorielles des correspondances.

D.V.1 Description.

Le protéome de *P. haloplanktis* (3499 protéines), après formatage des données (suppression des 10 premiers et 5 derniers acides aminés de chaque séquence et élimination des protéines de longueur inférieure à 100 résidus, *cf.* chapitre B.II), est réduit à 2983 protéines. Les 2502 et 481 gènes, codant pour ces protéines, se trouvent respectivement sur les chromosomes 1 et 2. Devant l'homogénéité des résultats des AFCs obtenus sur les chromosomes 1 et 2, nous avons choisi de traiter systématiquement les deux chromosomes ensemble plutôt qu'individuellement (Figure 7). L'inertie cumulée de cette AFC sur les quatre premiers axes est de plus de 46% ce qui est tout à fait semblable aux inerties des AFCs calculées sur la plupart des protéomes des organismes étudiés et représente une quantité d'information suffisamment significative pour être exploitée.

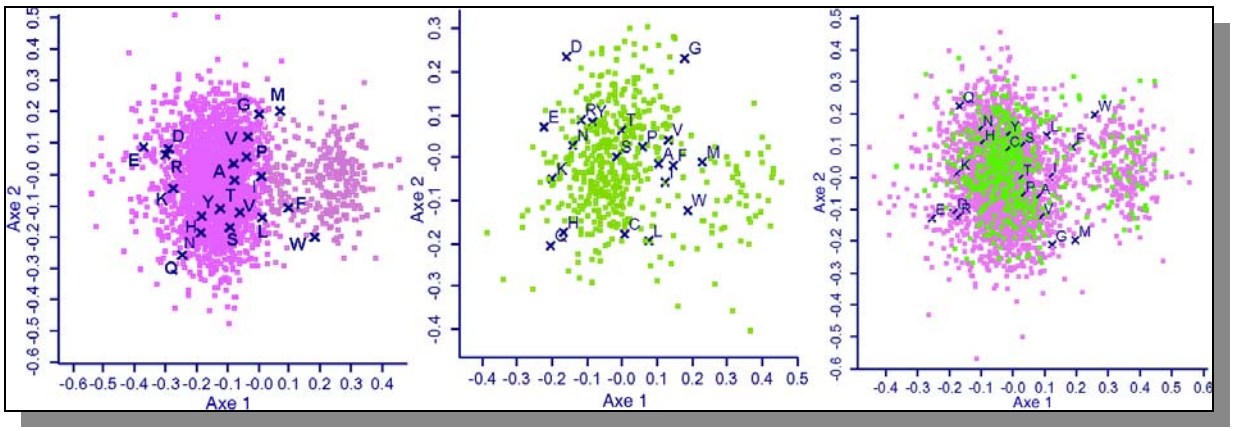


Figure 7 : Premiers plans des AFCs de *P. haloplanktis* points violets : protéines encodées sur chromosome 1, points vert : protéines encodées sur chromosome 2, croix marines : acides aminés.

[1] Définition tirée de « le Petit Robert » – édition Dictionnaires Le Robert – Paris, 2003.
psychrophile : adj. et n. m. – 1963 ; du grec *psukhros* « froid » et *–phile* « ami » ; Qui vit et se reproduit dans des conditions optimales, à des températures inférieures à 20°C (en parlant d'un micro-organisme).

D.V.2 *Analyse des clusters.*

L'AFC de *P. haloplanktis*, couplée à la méthode de partitionnement, les nuées dynamiques (Diday 1971; Delorme et Hénaut 1988), a isolé cinq groupes bien définis (Figure 8). Grâce aux annotations suffisamment précises (les annotations manuelles de cet organisme ont été organisées selon un classement particulier et spécifique à *P. haloplanktis*, Annexe I.V), nous avons pu analyser assez rapidement le contenu des différents groupes de protéines. Après la programmation d'un utilitaire en Perl, les données d'annotations formatées ont été transformées en un tableau de résultats ordonnés facilement exploitable (Annexe I.VI). Cette analyse a révélé un cloisonnement assez précis des différentes fonctions ou localisations subcellulaires de la bactérie selon le cluster d'appartenance des protéines. Un cluster (carrés jaunes, Figure 8) apparaît comme atypique aux vues des analyses précédemment faites sur d'autres organismes. Ce cluster est discriminé par l'axe 3 de l'AFC, axe représentant en fait un gradient en asparagine des protéines. Son contenu et ses spécificités seront détaillés au cours des chapitres suivants.

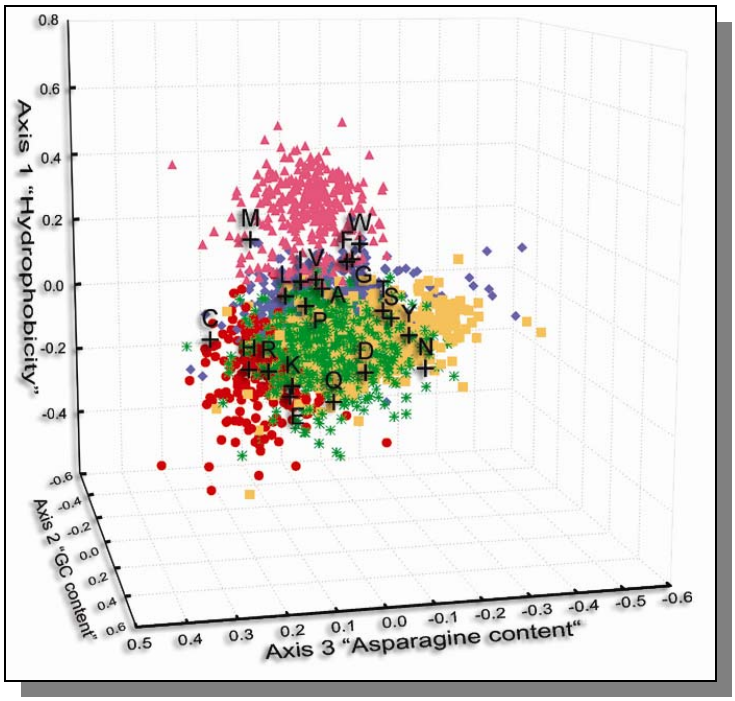


Figure 8 : Distribution des protéines sur le premier espace factoriel résultant du calcul de l'AFC. Le premier axe discrimine les protéines par leur hydrophobicité, le deuxième par le contenu en G+C des gènes qui les encodent et le troisième par leur contenu en asparagine. 5 classes sont trouvées par la méthode de partitionnement ; elles sont composées globalement de : (i) IIMPs, triangles roses, (ii) protéines impliquées dans le métabolisme de petites molécules, losanges bleus, (iii) protéines associées à l'information des voies de transfert, cercles rouges, (iv) protéines associées à la membrane externe ou sécrétées, étoiles vertes et (v) protéines dont la fonction est inconnue ou ayant vraisemblablement des origines phagiques, carrés jaunes. Les acides aminés sont représentés par des croix noires.

D.VI *Ses camarades de jeu.*

Afin d'étudier la composition en acides aminés du protéome de la bactérie, nous avons réalisé non seulement une AFC individuelle de *P. haloplanktis*, mais également une AFC de génomique comparative regroupant à la fois des bactéries psychrophiles, mésophiles et thermophiles (Figure 9). Ont été choisies les bactéries *Desulfotalae psychrophila* (psychrophile), *E. coli* K-12, *B. subtilis* (mésophiles), *Thermotoga maritima* et *Aquifex aeolicus* (thermophiles), car leur génome est complètement séquencé, un certain nombre de données d'annotations les concernant sont

disponibles et à l'exception d'*E. coli* K-12 et de *B. subtilis*, les autres organismes sont issus du milieu marin comme *P. haloplanktis*. L'analyse du produit de la méthode de partitionnement (les nuées dynamiques (Delorme et Hénaut 1988)), a révélé que l'un des groupes de protéines constitué majoritairement de protéines des organismes psychrophiles suivait un biais en asparagine (Figure 10).

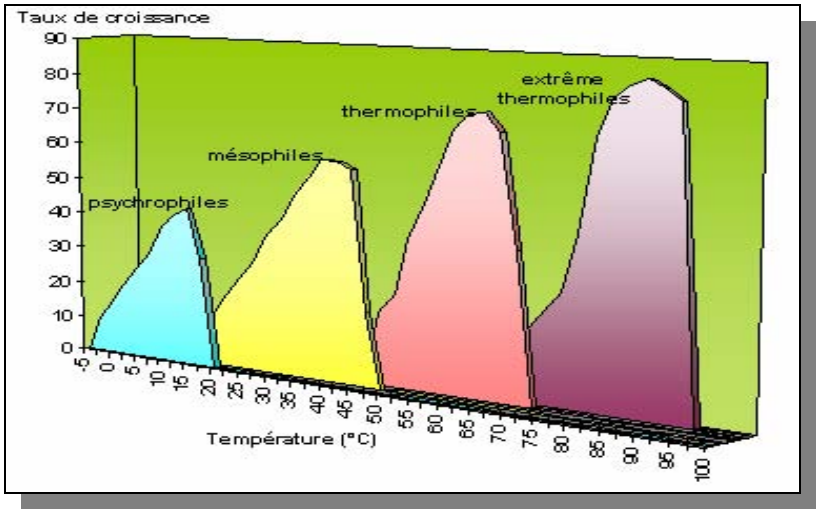


Figure 9 : Taux de croissance des bactéries selon leur température optimale de croissance. Source Internet non nominative.

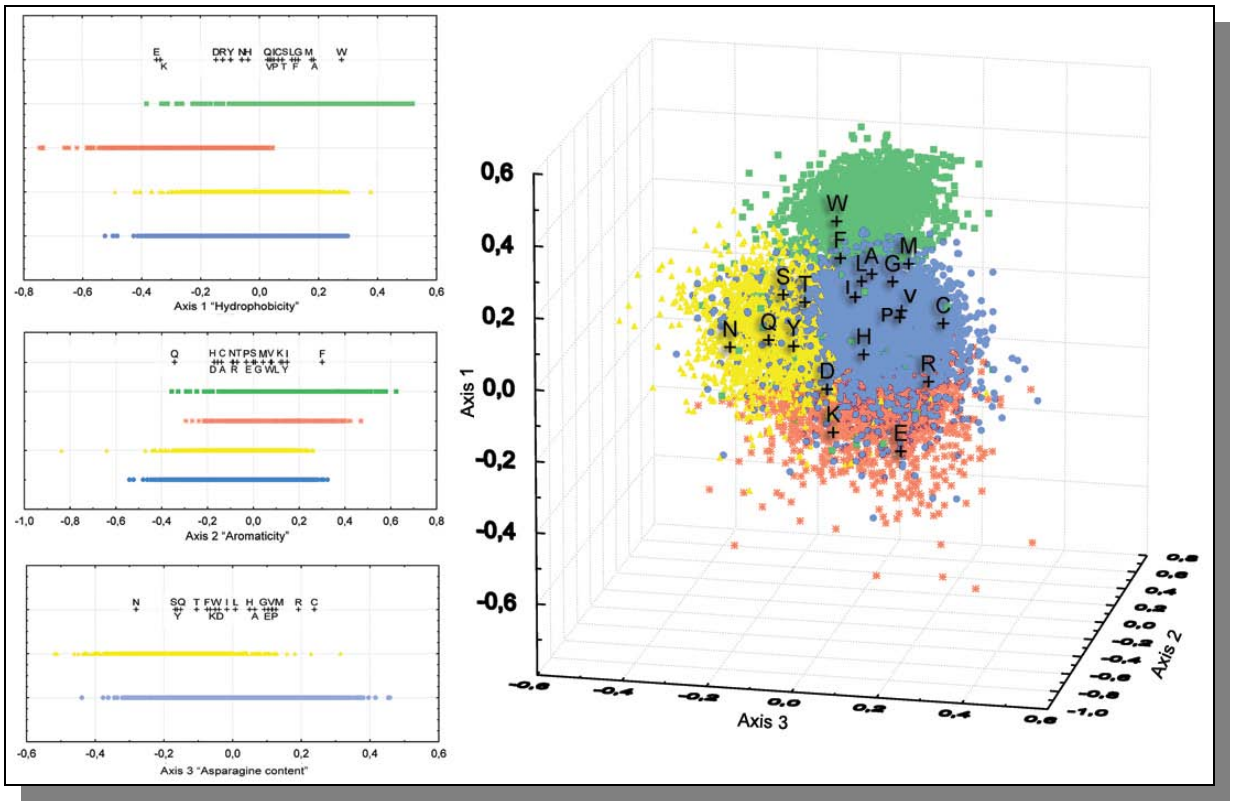


Figure 10 : Représentation graphique à 1 et à 3 dimensions du résultat de l'AFC sur l'ensemble des organismes liés à différentes températures. L'axe 1 discrimine les protéines par leur hydrophobicité, l'axe 2 par leur aromaticité et l'axe 3 par leur contenu en Asn. 62% du groupe orange est constitué de protéines de thermophiles, 53 % du groupe bleu est constitué de protéines de mésophiles, 55% du groupe jaune est constitué de protéines de psychrophiles. Le groupe vert contient les IIMPs. Les acides aminés sont représentés par des croix noires.

D.VII Zoom sur le biais en asparagine des bactéries du froid.

D.VII.1 Les protéines psychrophiles biaisées en asparagine.

Aussi bien dans l'AFC individuelle de *P. haloplanktis* (Figure 8) que dans l'AFC globale de génomique comparative, certaines protéines de *P. haloplanktis* présentent un biais particulier dû à leur contenu en asparagine (Asn) (Figure 11). Seul biais semblant caractériser plus particulièrement les protéines des organismes psychrophiles, ce biais en asparagine a retenu toute notre attention. Après calculs, un cluster de l'AFC (carrés jaunes, Figure 8) se révèle composé de protéines environ 1,4 fois plus riches en résidu Asn que le reste du protéome. Après l'analyse des classes fonctionnelles données par l'annotation, le cluster jaune de l'AFC individuelle est clairement formé de protéines dont la fonction est inconnue ou ayant vraisemblablement des origines phagiques (Annexe I.VI). Si nous partons de l'hypothèse que ces protéines riches en Asn sont caractéristiques de la vie au froid, et étant donnée le nombre d'études et de programmes de séquençage complet des bactéries psychrophiles est encore faible, il est vraisemblable qu'un nombre important de protéines typiques de ces organismes soit de fonction hypothétique ou inconnue. Le grand nombre de protéines d'origines phagiques nous laisse penser que *P. haloplanktis* pourrait contenir de nombreuses protéines orphelines. En effet, ces dernières pourraient être caractérisées comme telles (Rocha et Danchin 2002; Pedulla, Ford et al. 2003; Pascal, Medigue et al. 2005). En analysant de plus près les protéines les plus biaisées par le contenu en Asn de *P. haloplanktis*, nous trouvons de manière très probante un grand nombre de protéines en relation avec la protéine TonB (Tableau V).

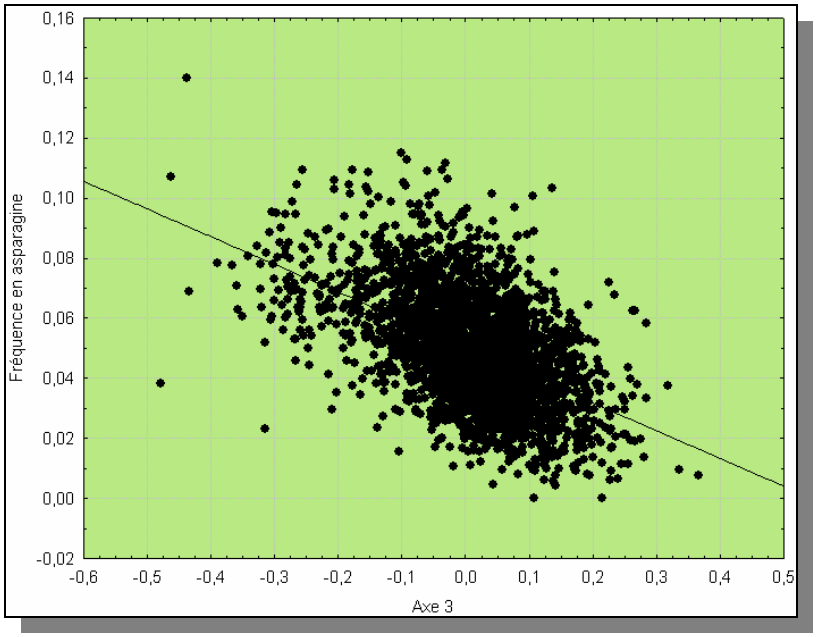


Figure 11 : Corrélation des coordonnées des protéines sur l'axe 3 de l'AFC de *P. haloplanktis* avec leur fréquence relative en asparagine ($r = -0{,}5091$; $p < 10^{-4}$)

Tableau V : Fonctions et produits (Annexe I.V) (selon leur cluster d'appartenance) des 100 protéines se trouvant à l'extrémité de l'axe 3 de l'AFC de *P. haloplanktis* (les couleurs utilisées sont celles de la Figure 8).

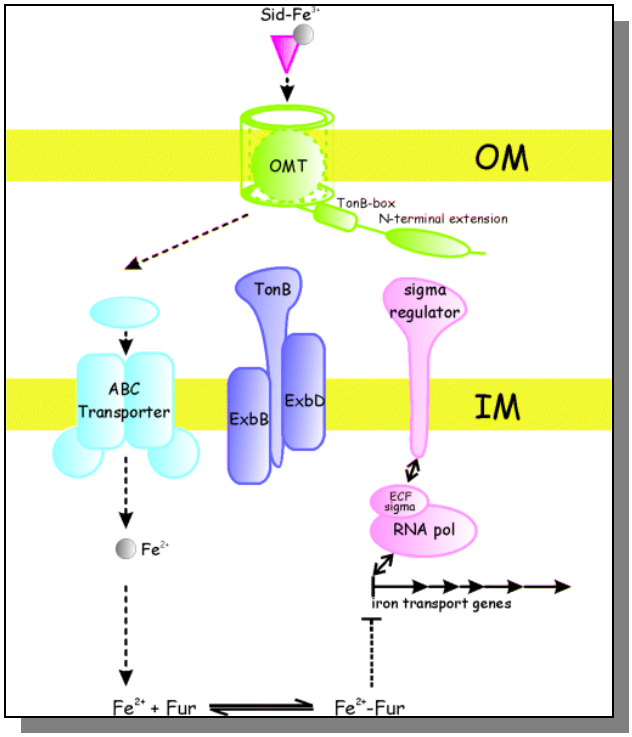
identifiant	axe 3	cluster	fonction	produit
PSHAa2824	-0,477413	bleu	5.2 : From other organisms	conserved protein of unknown function
PSHAa1921	-0,457686	jaune	5.2 : From other organisms	conserved protein of unknown function ; putative signal peptide
csgA	-0,440725	bleu	1.1.3 : Surface structures (S-layers?)	putative secreted major subunit of curlin, may bind calcium
PSHAb0416	-0,429427	jaune	1 : Cell envelope and membrane-associated cellular processes	putative periplasmic or outer membrane protein
PSHAb0329	-0,394159	bleu	5.2 : From other organisms; 5.1 : From P. haloplanktis	putative unknown lipoprotein
flgE	-0,36825	bleu	1.5.3 : Motility	flagellar biosynthesis; hook protein
PSHAa2125	-0,364013	bleu	6.2 : Hypothetical	putative orphan protein
PSHAb0526	-0,35987	bleu	5.2 : From other organisms	conserved protein of unknown function
PSHAa1862	-0,347217	jaune	1.3 : Sensors (signal transduction)	putative TonB-dependent receptor
PSHAa2649	-0,34603	bleu	2.2.3.1 : Degradation of proteins, peptides, and glycopeptides	Serine protease precursor
PSHAb0340	-0,324863	jaune	1.2 : Transport/binding proteins and lipoproteins	putative Outer membrane protein with a TonB box
PSHAb0113	-0,320613	jaune	1.3.1 : Sensors for chemicals	putative TonB-dependent receptor for Fe
PSHAa1840	-0,319644	jaune	1.2 : Transport/binding proteins and lipoproteins	putative TonB-dependent receptor
PSHAb0341	-0,319073	jaune	1.2 : Transport/binding proteins and lipoproteins	putative outer membrane protein with a TonB box
PSHAa1987	-0,313206	jaune		putative TonB-dependent receptor protein
PSHAb0232	-0,312309	jaune	1.3 : Sensors (signal transduction)	putative TonB dependent receptor
PSHAa2198	-0,311924	bleu	6.2 : Hypothetical	putative orphan protein
PSHAa1621	-0,311758	jaune	1 : Cell envelope and membrane-associated cellular processes	putative TonB-dependent receptor protein with OmpA-like transmembrane domain
PSHAa1502	-0,308575	jaune		conserved protein of unknown function
PSHAa1271	-0,306402	jaune		putative TonB-dependent receptor
PSHAa1340	-0,304291	jaune	1.2.4 : Cations	putative calcium binding protein
PSHAa2567	-0,303915	jaune	1.2.4 : Cations	putative Outer membrane porin
PSHAb0375	-0,30292	jaune	6.2 : Hypothetical	putative orphan protein
PSHAa1575	-0,300545	jaune		putative outer membrane porin
PSHAa2119	-0,299259	jaune	1.2.7 : Unknown substrates	putative TonB-dependent receptor
PSHAa2968	-0,298099	jaune	5.2 : From other organisms; 4.4 : Phage-related functions	conserved protein of unknown function ; putative secreted protein
PSHAa0289	-0,297417	jaune	5.2 : From other organisms	conserved protein of unknown function ; putative signal peptide
PSHAb0165	-0,293273	jaune	1.2 : Transport/binding proteins and lipoproteins	putative TonB-dependent receptor; outer membrane
PSHAa2275	-0,291024	jaune	1.2 : Transport/binding proteins and lipoproteins	putative TonB-dependent Outer membrane receptor
PSHAa1305	-0,290993	bleu	5.2 : From other organisms	conserved protein of unknown function
PSHAa1347	-0,288481	bleu	5.2 : From other organisms	conserved protein of unknown function
PSHAa0428	-0,287538	jaune		putative TonB-dependent receptor
flgG	-0,285844	bleu	1.5.3 : Motility	flagellar biosynthesis; cell-distal portion of basal-body rod
PSHAa2268	-0,282243	jaune	1 : Cell envelope and membrane-associated cellular processes	putative Fimh-like protein
PSHAa2269	-0,281725	jaune	6.2 : Hypothetical	putative orphan protein
PSHAa1989	-0,279717	bleu	6.2 : Hypothetical	conserved protein of unknown function
PSHAa1664	-0,279158	jaune	6.2 : Hypothetical	putative orphan protein
PSHAa2947	-0,279118	jaune	1.3 : Sensors (signal transduction)	putative TonB-dependent receptor
PSHAa2973	-0,277175	jaune	1.2.4 : Cations	putative secreted TonB-dependent receptor
PSHAa1479	-0,277076	jaune	6.2 : Hypothetical	conserved protein of unknown function
PSHAb0299	-0,276917	jaune	1 : Cell envelope and membrane-associated cellular processes	putative TonB-dependent outer membrane receptor

identifiant	axe 3	cluster	fonction	produit
PSHAb0512	-0,275505	jaune	1.2.4 : Cations	putative Ton-B dependent protein (could be involved in iron transport)
PSHAa2457	-0,273974	jaune	1.2.7 : Unknown substrates	putative TonB-dependent receptor
osmY	-0,272136	bleu	4.1 : Adaptation to atypical conditions	putative hyperosmotically inducible periplasmic protein
PSHAb0279	-0,271329	jaune		putative outer membrane receptor
PSHAa1782	-0,270266	jaune		conserved protein of unknown function
PSHAa0695	-0,269097	jaune	1.2.4 : Cations	putative outer membrane receptor for ferric iron uptake ; putative TonB-dependent receptor
PSHAa0696	-0,267948	jaune	5.2 : From other organisms	conserved protein of unknown function
flgK	-0,267802	bleu	1.5.3 : Motility	putative flagellar hook-associated
PSHAa0876	-0,267162	jaune	5.2 : From other organisms	conserved protein of unknown function; could be a lipoprotein or a secreted protein; could be a binding protein for a transporter
PSHAa0476	-0,265504	jaune	1 : Cell envelope and membrane-associated cellular processes	putative TonB-dependent receptor
PSHAa1824	-0,264931	jaune	1.5 : Motility and chemotaxis	putative ferric enterobactin receptor
ssb	-0,264388	vert	3.1.2 : DNA recombination; 3.1.4 : DNA proof-reading and repair	Single-strand binding protein (SSB) (Helix-destabilizing protein)
PSHAa0505	-0,264125	jaune	5.2 : From other organisms	conserved protein of unknown function
PSHAa2436	-0,263519	jaune	5.2 : From other organisms	conserved protein of unknown function
PSHAb0527	-0,261754	jaune	1.3 : Sensors (signal transduction)	putative TonB-dependent receptor
PSHAa2138	-0,261422	jaune	1 : Cell envelope and membrane-associated cellular processes	putative TonB-dependent receptor
PSHAb0200	-0,260388	bleu	2.2.3.1 : Degradation of proteins, peptides, and glycopeptides	putative secreted serine protease, subtilisin family, possibly excreted
flgC	-0,259926	vert	1.5.3 : Motility	flagellar biosynthesis; cell-proximal portion of basal-body rod
PSHAb0254	-0,258656	jaune	1 : Cell envelope and membrane-associated cellular processes	putative Outer membrane TonB-dependent receptor
mipA	-0,258307	jaune	1.1 : Cell wall	putative scaffolding protein for murein-synthesizing holoenzyme, outer membrane protein
PSHAa1740	-0,25828	jaune	1.2 : Transport/binding proteins and lipoproteins	putative TonB-dependent receptor
PSHAb0158	-0,257255	bleu	5.2 : From other organisms; 1.2 : Transport/binding proteins and lipoproteins	putative TonB-dependent receptor with TonB-box
PSHAa2206	-0,25495	jaune		putative TonB-dependent receptor
PSHAb0330	-0,252714	bleu	5.2 : From other organisms	conserved protein of unknown function ; putative secreted protein highly similar to Gloeobacter violaceus Glr2979 protein
PSHAa1273	-0,252409	jaune	2.2.3.1 : Degradation of proteins, peptides, and glycopeptides	Prolyl endopeptidase
PSHAa0397	-0,251236	jaune	6.2 : Hypothetical	putative orphan protein
PSHAa1557	-0,250503	jaune		conserved protein of unknown function
PSHAa2703	-0,24831	jaune	5.2 : From other organisms	conserved protein of unknown function
PSHAa1352	-0,246442	jaune		putative TonB-dependent receptor protein
PSHAb0072	-0,24522	jaune	1.2.8 : Other	putative Hemin receptor protein HmuR
PSHAa2936	-0,245195	bleu	5.2 : From other organisms	conserved protein of unknown function ; putative periplasmic calcium binding protein
fliD	-0,244174	bleu	1.5.3 : Motility	putative flagellar hook-associated protein
PSHAb0286	-0,243993	jaune	1 : Cell envelope and membrane-associated cellular processes	putative TonB dependent outer membrane receptor
PSHAa1298	-0,240588	jaune	5.2 : From other organisms	conserved protein of unknown function ; putative periplasmic protein
PSHAa0840	-0,240249	jaune	5.2 : From other organisms; 5.1 : From P. haloplanktis	conserved protein of unknown function ; putative TonB-dependent receptor

identifiant	axe 3	cluster	fonction	produit
mshQ	-0,236089	jaune	1.2 : Transport/binding proteins and lipoproteins; 1.6 : Protein secretion	putative Mannose-sensitive agglutinin (MSHA) biogenesis protein MshQ (pilus type IV)
flgL	-0,232972	vert	1.5.3 : Motility	putative flagellar synthesis; flagellar regulon; hook-associated protein
PSHAa1674	-0,231751	jaune	1 : Cell envelope and membrane-associated cellular processes; 5.2 : From other organisms	putative membrane protein
PSHAa0108	-0,230037	jaune	1.2.4 : Cations	putative TonB-dependent receptor; putative outer membrane bound protein involved in iron chelated transport
PSHAa0512	-0,229327	jaune	5.2 : From other organisms	conserved protein of unknown function
malB	-0,229293	jaune	1.2.3 : Carbohydrates, organic alcohols, and acids	putative maltoporin, high-affinity receptor for maltose and maltoseoligosaccharides
PSHAb0414	-0,226974	jaune	5.2 : From other organisms	putative membrane associated hydrolase
irgA	-0,226331	jaune	1.2.4 : Cations	Iron-regulated outer membrane virulence protein homolog
flgH	-0,224705	jaune	1.5.3 : Motility	flagellar biosynthesis; basal-body outer-membrane L (lipopolysaccharide layer) ring protein
PSHAa1584	-0,223209	jaune	1.3.1 : Sensors for chemicals	putative TonB-dependent receptor for ferrichrome transport
PSHAa2180	-0,222789	jaune	1.3.4 : Other	putative TonB-dependent receptor
PSHAb0025	-0,220268	jaune	6.2 : Hypothetical	putative orphan protein
fadL	-0,219297	jaune	1 : Cell envelope and membrane-associated cellular processes	Long-chain fatty acid transport protein precursor (Outer membrane fadL protein)
PSHAa0930	-0,219012	jaune	1 : Cell envelope and membrane-associated cellular processes	putative Type IV pilus biogenesis protein
PSHAa0927	-0,218347	jaune	5.2 : From other organisms	conserved protein of unknown function ; putative pilin biogenesis protein
PSHAb0115	-0,218325	bleu	5.2 : From other organisms	putative enzyme protein
PSHAa1242	-0,218258	jaune	6.2 : Hypothetical	putative orphan protein
PSHAa2478	-0,218107	jaune	1.2.4 : Cations	putative TonB-dependent receptor protein
PSHAa0863	-0,217664	jaune	5.2 : From other organisms	conserved protein of unknown function ; putative outer membrane protein
PSHAa1953	-0,217291	jaune		putative peptidase
lolA	-0,21719	vert	1.6 : Protein secretion	putative periplasmic chaperone effects translocation of lipoproteins from inner membrane to outer membrane
fliC	-0,21491	vert	1.5.3 : Motility	flagellin
PSHAb0347	-0,21101	vert	3.3.7 : Protein folding	putative Peptidyl-prolyl cis-trans isomerase

D.VII.2 *Les protéines associées à TonB.*

Le fer est essentiel à la croissance aussi bien bactérienne qu'eucaryote. Contrôler la quantité de fer libre en solution est souvent une stratégie utilisée par les organismes hôtes afin de limiter une invasion de microbes pathogènes ; ceci peut se faire en liant hermétiquement le fer à l'intérieur de protéines (transferrine, lactoferrine). Pour contrer ce phénomène, certaines bactéries expriment en surface des récepteurs permettant de capturer ces protéines hôtes liant le fer, tandis que d'autres ont développé des sidérophores afin de récupérer le fer libre en solution et/ou de détacher le fer des chélateurs de l'hôte (Guerinot 1994). Une faible concentration intracellulaire de fer déclenche la transcription de clusters de gènes qui encodent à la fois les sidérophores et les récepteurs chargés d'internaliser les ferri-sidérophores (Nahlik, Brickman et al. 1989). Un exemple chez *E. coli* K-12 est fepA qui se situe dans l'enveloppe externe et capture l'entérobactine liant le fer (Buchanan, Smith et al. 1999). Afin de compléter le transport lié au fer à travers la membrane interne, un second complexe est nécessaire. Le principal composant de ce complexe est TonB, une protéine de 27kDa qui facilite le transfert de l'énergie à partir d'une force motrice à protons vers les récepteurs externes (Postle et Good 1983; Moeck et Coulton 1998; Braun et Braun 2002), (Figure 12).



D'après (Schalk, Yue et al. 2004)

Figure 12 : Siderophore transport system for OMT_Ns. Ferric-siderophore (Sid-Fe^{3+}) from the extracellular medium is recognized by an OMT_N, which serves two functions. First, the OMT_N transports Sid-Fe^{3+} into the periplasm, which is further transported into the cytoplasm by an ABC transporter (light blue). Second, the OMT_N regulates the transcription induction of iron uptake genes. The latter process is initiated by the binding of Sid-Fe^{3+} to the OMT_N, and involves several components: the N-terminal extension of the OMT_N, the inner membrane sigma-regulator protein and the cytoplasmic ECF sigma factor (pink). Both transport and induction functions require energy transduction from the TonB–ExbB–ExbD complex in the inner membrane (purple). Sid-Fe^{3+}-bound OMT_N is believed to interact with TonB via its TonB-box motif. When the intracellular iron level is high, the transcriptional repressor Fur is bound with ferrous ion (Fe^{2+}). The Fe^{2+}–Fur complex represses the transcription of iron uptake genes. RNA pol, RNA polymerase; OM, outer membrane; OMT_N, OM Transporters have N-terminal extension about 70 residues; IM, inner membrane; ECF, extracytoplasmic function.

Concernant les protéines les plus biaisées par le contenu en Asn de *P. haloplanktis*, un premier constat à faire est que l'identité des protéines associées à TonB dans le groupe isolé par l'AFC n'est pas explicite. En effet, malgré les efforts d'annotation consentis, la définition de ces protéines reste assez vague puisque peu d'informations sont disponibles à leur propos. Ainsi, nous ne connaissons pas précisément les protéines qui se cachent derrière les annotations « TonB-dependent ». Cependant, nous savons que le fer ne se trouve qu'à l'état de traces dans l'océan et que 99% du fer dissout est lié à des ligands organiques se trouvant alors sous forme d'ions ferriques (Fe^{3+}) (Wu, Boyle et al. 2001).

La stratégie de capture du fer de *P. haloplanktis* semble donc fondée sur l'utilisation de sidérophores peut être amphiphiles à l'instar d'autres bactéries marines (Martinez, Carter-Franklin et al. 2003). La grande affinité de ces sidérophores pour le fer permet de détacher ce dernier des ligands organiques présents dans l'océan. Toutefois, ceci n'explique pas la richesse atypique en acides aminés Asn de ces protéines en relation avec TonB.

D.VII.3 *Le biofilm.*

Outre les protéines en relation avec TonB, le cluster jaune, qui est discriminé par le contenu en Asn des protéines qu'il contient, regroupe les seules protéines impliquées dans la formation du biofilm ainsi qu'un grand nombre de protéines de la biosynthèse et de la dégradation des polysaccharides (Annexe I.IV.6, I.VI et Figure 8).

Un biofilm est une couche de micro-organismes contenus dans une matrice solide, se formant sur des surfaces en contact avec l'eau. Cette matrice est généralement composée de polysaccharides extracellulaires ou exopolysaccharides (EPS). Les études sur la croissance de bactéries dans des systèmes aquatiques montrent que presque toutes les cellules sont entourées par des EPS (arabinose, ribose, galactose, rhamnose, glucose…) (Decho 1990; Costerton, Stewart et al. 1999) et que la plupart de ces cellules sont enfermées dans un biofilm (White 1986). Il a été suggéré qu'une forte concentration d'EPS en eau salée pourrait jouer le rôle de tampon lors de conditions d'hivers rigoureux et de haute salinité. Ces EPS offriraient une "cryoprotection" aux microbes vivant dans ces milieux contre la formation de cristaux de glaces (Krembs, Eicken et al. 2002). De plus, il a été montré que la production d'EPS décroît lorsque la température augmente (Mancuso Nichols, Garon et al. 2004). En outre, le biofilm permet aussi aux bactéries d'absorber des molécules organiques et minérales comme le fer. Etant donné que la disponibilité du fer est extrêmement faible dans les eaux du pôle sud, la production d'EPS bactériens peut avoir un rôle important dans la survie des bactéries marines et avoir d'importantes implications au sein de la communauté microbienne des eaux glacées.

Bien que la fonction du biofilm dans la psychrotolérance soit évidente, la raison de l'atypique richesse en acides aminés Asn des protéines le composant n'est cependant pas encore élucidée.

D.VII.4 *La déamidation de l'asparagine : une réaction thermosensible.*

L'exploration fonctionnelle des protéines du groupe jaune de l'AFC (Figure 8) n'ayant pas permis d'expliquer cet enrichissement en Asn, nous avons recherché d'éventuelles causes biochimiques et/ou métaboliques. Une des principales modifications post-traductionnelles que subissent les protéines d'organismes procaryotes et eucaryotes est la déamidation des acides aminés Asn. Elle entraîne le vieillissement prématuré des protéines. Cette réaction est le résultat de la conversion d'un résidu Asn en isoaspartate et aspartate (Figure 13). Le rôle de cette modification dans la cellule est encore mal connu mais il semblerait qu'elle puisse affecter la structure des protéines et par conséquent leur fonction. Elle pourrait jouer un rôle dans la régulation du repliement et être le signal de dégradation de la protéine régulée au niveau intracellulaire (Wright 1991; Lindner et Helliger 2001). Plusieurs conditions favorisent la déamidation de l'Asn (Robinson et Robinson 2001) ; les principales étant :

1. une haute fréquence d'une glycine juxtaposée en aval d'une asparagine (en N+1). En fait, sur la protéine où la déamidation est identifiée, le site caractérisé en N+1 est généralement une glycine. Moins fréquemment, les acides aminés polaires à relativement petite chaîne latérale (sérine, thréonine, aspartate) se trouvent également en N+1 de l'Asn. Par contre, l'Asn est rarement suivie de résidus hydrophobes ou à grande chaîne latérale. Ceci peut s'expliquer par le fait que l'encombrement stérique de la grande chaîne latérale peut limiter l'accessibilité au groupe amino de la chaîne latérale amide de l'Asn.
2. un milieu alcalin. En effet, la déprotonation du groupe amide étant facilitée à pH élevé, le taux de formation de succinimide s'accroît.
3. une température élevée (Patel et Borchardt 1990; Daniel, Dines et al. 1996).

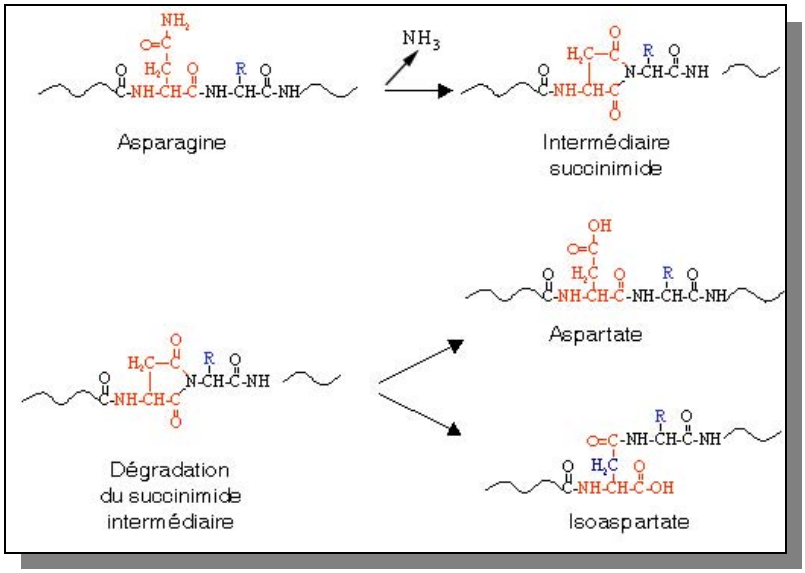


Figure 13 : Déamidation de l'asparagine au sein d'une protéine.

Si les températures élevées favorisent la déamidation de l'Asn, on peut imaginer que ce résidu a tendance à être contre sélectionné chez les mésophiles et les thermophiles, ceci afin d'éviter les

problèmes post-traductionnels que cette réaction engendre. Il serait alors observé chez les psychrophiles non pas une sélection positive envers l'introduction d'Asn dans les protéines mais plutôt une « levée » de la contre sélection observée chez les mésophiles et les thermophiles. Cette hypothèse est confortée par les statistiques présentées en Figure 14. Non seulement l'Asn est en surabondance chez les psychrophiles mais la glycine (Gly), la sérine (Ser) et la thréonine (Thr) le sont également. Or, si le bouclier contre la déamidation de l'asparagine est levé, il n'existe plus de raison d'éviter la présence de ces trois acides aminés en aval de l'Asn (en N+1), et ceci modifierait le métabolisme global de ces acides aminés. On remarque tout de même que l'aspartate (Asp) n'est pas plus fréquent chez les psychrophiles que chez les autres. De plus, la glutamine (Gln) semble elle aussi en sur-expression chez les psychrophiles (Figure 14), ceci s'explique peut être par le fait que la déamidation touche également ce résidu.

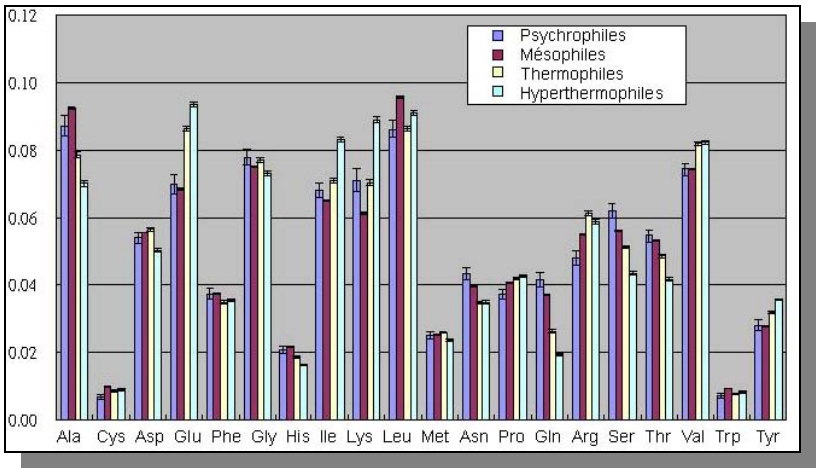


Figure 14 : fréquences de l'ensemble des acides aminés dans les organismes psychrophiles, mésophiles, thermophiles et hyperthermophiles de la base de données PGTdb qui organise les données des organismes selon leur température optimale de croissance (Huang, Wu et al. 2004).

D.VII.5 *Et l'analyse des dipeptides ?*

Désireux de confirmer cette observation, nous nous sommes ensuite focalisés non plus sur la quantité d'acides aminés mais sur un éventuel différentiel de fréquences de dipeptides entre psychrophiles, mésophiles et thermophiles. Pour cela, nous avons choisi de comparer les occurrences des dipeptides au sein d'un groupe d'orthologues (547 protéines) entre les bactéries marines *Pseudoalteromonas haloplanktis*, *Desulfotalae psychrophila* (psychrophile), *Photobacterium profundum* (psychrophile), *Vibrio vulnificus* (mésophile), *Shewanella oneidensis* (mésophile) et *Aquifex aeolicus* (thermophile). Nous avons tenté d'inclure dans ce groupe les séquences d'une seconde bactérie thermophile, *Thermotoga maritima*, mais le nombre de protéines orthologues obtenu était insuffisant. De cette comparaison d'occurrences résulte que les fréquences de la plupart des dipeptides sont significativement différentes entre psychrophiles et thermophiles, entre mésophiles et thermophiles mais peu de différences ont été constatées entre psychrophiles et mésophiles. En particulier, les différences d'occurrences des dipeptides NG, NS, NT et ND, quand elles existent, ne sont pas statistiquement significatives. Plusieurs tentatives ont été effectuées en associant différents organismes afin de former des groupes d'orthologues différents mais aucun résultat probant n'est apparu. Il s'avère que l'analyse, par l'AFC, d'un grand ensemble de protéines ne nous a pas permis

de déceler suffisamement finement les raisons du biais en Asn, en conséquence, nous avons décidé de ne traiter qu'un petit groupe de protéines ciblées.

D.VII.6 *Études préliminaires sur la mise en évidence de l'inhibition de la déamidation*

Le but de cette analyse est d'observer d'éventuelles surabondances du dipeptide NG dans des protéines de psychrophiles par rapport à des protéines de mésophiles et de thermophiles. Pour cette étude, les protéines ribosomiques ont été choisies. Ces protéines sont généralement décrites comme protéines ancestrales, elles sont essentielles et sur-exprimées dans les organismes (Woese et Fox 1977; Woese, Kandler et al. 1990). De génération en génération, l'ubiquité de ces protéines leur contraint une forte conservation du schéma de leur séquence et celle-ci reste globalement stabilisée au fil du temps, n'évoluant que lentement.

Parmi l'ensemble des organismes disponibles dans les banques de données, cinq bactéries psychrophiles (*P. haloplanktis, D. psychrophila, P. profondum, Psychrobacter sp., Colwellia psychrerythraea*), cinq bactéries mésophiles (*B. subtilis, E. coli K-12, Pseudomonas aeruginosa, B. bacteriovorus, S. coelicolor*) et cinq bactéries thermophiles (*A. aeolicus, T. maritima, Symbiobacterium thermophilum, Thermoanaerobacter tencongensis, Streptococcus thermophilus*) ont été choisies. D'après les travaux de G. Fang sur les protéines conservées au sein d'un grand ensemble d'organismes (Fang, Rocha et al. 2005), 15 protéines ribosomiques conservées par bactéries ont pu être analysées.

En guise de travaux préliminaires de simples statistiques ont été calculées. Ont été comparés entre les trois différentes classes de bactéries (Tableau VI) (i) la moyenne du nombre d'occurrences de peptides NG dans l'ensemble des 15 protéines ribosomiques des cinq bactéries de chacun des trois groupes (ii) le pourcentage de la fréquence moyenne d'occurrences NG par rapport au nombre total de dipeptides NX (tout dipeptide dont le premier acide aminé est une asparagine) et (iii) le pourcentage de la fréquence moyenne d'occurrence de dipeptides NG par rapport au nombre total de dipeptides (dipeptides XX). Les résultats sont présentés à travers les trois graphiques de la Figure 15 : avec la baisse des températures du milieu de croissance des bactéries choisies est observé l'accroissement systématique du nombre résultant des statistiques sur les dipeptides NG. Ainsi, bien que la présence en N+1 de l'Asn d'un résidu glycine soit très favorable à la réaction de déamidation de l'Asn provoquant le vieillissement des protéines, la contre sélection de ce résidu en aval de l'Asn s'estompe lorsque les protéines proviennent d'organismes psychrophiles. Ces calculs préliminaires ne présente en aucun cas une preuve suffisante de l'inhibition de la déamidation des acides aminés asparagines au sein des organismes psychrophiles, néanmoins ces premiers résultats, qui demandent à être complétés, suggèrent éventuellement, que le biais en Asn, observé grâce à l'AFC, discriminant typiquement les protéines psychrophiles (groupe jaune, Figure 8 et Figure 10), est dû à l'inhibition de la réaction thermosensible de déamidation des acides aminés Asn des protéines au sein de ces organismes.

Tableau VI : Statistiques sur le nombre d'occurrence du dipeptide NG dans chaques groupes de bactéries choisies selon leur température optimale de croissance.

Bactéries	Thermophiles	Mésophiles	Psychrophiles
Moyenne ΣNG	6,4	7,8	8,2
(%) Moyenne ΣNG/ΣNX	6,59	7,54	7,78
(%) Moyenne ΣNG/ΣXX	0,23	0,28	0,30

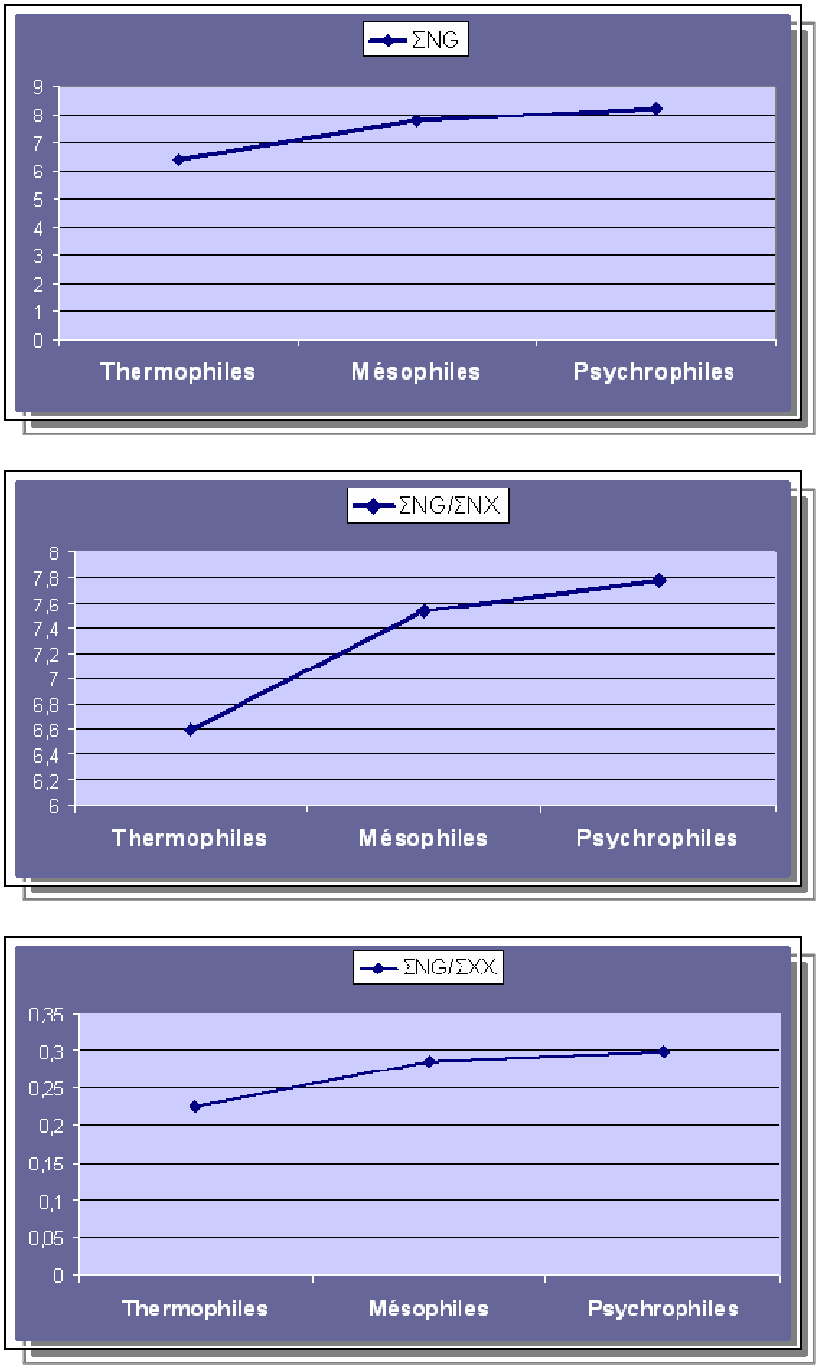


Figure 15 : Représentation graphique des données du Tableau VI.

D.VIII *Conclusion.*

L'adaptation au froid des organismes psychrophiles et en particulier de *P. haloplanktis* résulte probablement de la combinaison d'effets de mécanismes complexes existants dans la cellule. Néanmoins, le biais asparagine détecté à la fois dans les protéines de la membrane externe confrontées tout particulièrement aux basses températures extérieures et dans des protéines intracellulaires est une caractéristique spécifique de ces organismes. Les bactéries psychrophiles se révèlent alors comme un extraordinaire outil biotechnologique. De fait, la déamidation spontanée étant inhibibée par le froid, ces organismes deviendraient, en culture à basses températures, d'efficaces « producteurs » de protéines permettant de s'affranchir des réactions post-traductionnelles de déamidation qui engendrent le vieillissement prématuré de ces protéines. Ces bactéries pourraient en ce sens rapidement devenir les cibles de laboratoires pharmaceutiques et industriels.

E Une étude spécifique sur *Photorhabdus luminescens*

E.I Intérêt.

Depuis toujours, certains insectes sont considérés comme nuisibles pour l'agriculture, soit parce qu'ils se nourrissent des récoltes, soit parce qu'ils détruisent les denrées stockées. Afin de prévenir ces dommages, plusieurs stratégies sont employées telles que l'utilisation d'insecticides chimiques (DTT), d'insecticides biologiques (toxines de *Bacillus thuringiensis*) ou de prédateurs naturels (petits animaux ou autres insectes). Ces techniques sont plus ou moins efficaces et ne sont pas sans conséquences écologiques notamment dans le cas de l'utilisation d'insecticides chimiques. La découverte d'un couple biologique « *Photorhabdus luminescens* et *Heterorhabditis bacteriophora* **»** ne s'attaquant qu'à certains insectes cibles a ouvert une nouvelle voie dans la lutte contre les ravageurs de cultures.

P. luminescens est une bactérie à Gram négatif mobile (porteuse de flagelles) dont les colonies ont la particularité d'être pigmentées. Elle se caractérise par sa bioluminescence, sa production de protéases, de lipases, d'antibiotiques, de toxines entomopathogènes ainsi que de catalases (ffrench-Constant, Waterfield et al. 2003). *P. luminescens* vit en symbiose exclusive avec le nématode *Heterorhabditis bacteriophora* et ce tandem est utilisé comme agent insecticide naturel (Georgis et Poinar 1990) dont l'efficacité est comparable à celui des insecticides chimiques. Les caractéristiques biologiques de *P. luminescens* sont très intéressantes pour la recherche fondamentale et médicale. Les agents anti-microbiens qu'elle sécrète pourraient fournir de nouveaux antibiotiques et antifongiques à large spectre.

Les cycles de vie du nématode et de *P. luminescens* sont fortement liés et se déroulent en trois principales étapes (Figure 16). *P. luminescens* vit à l'intérieur de l'intestin du nématode lorsque ce dernier se trouve au stade juvénile, stade pendant lequel le nématode infecte les insectes cibles. Pour ce faire, le nématode pénètre dans l'hémocoele[1] de l'insecte via les ouvertures naturelles comme la bouche, l'anus, les spiracles respiratoires ou à travers la cuticule (Forst et Clarke 2002). *P. luminescens* est alors régurgité dans l'hémolymphe[2] de l'insecte (ffrench-Constant, Waterfield et al. 2003), et doit se protéger des attaques du nouvel hôte. En effet, l'insecte déploie ses hémocytes qui phagocytent ou encapsulent les bactéries et les molécules anti-microbiennes qu'elles produisent. Mais 24 à 48 heures après l'entrée du nématode dans l'insecte, celui-ci meurt. Les agents anti-microbiens produits par les bactéries protègent le cadavre de toute autre contamination externe (bactéries commensales de l'insecte, bactéries opportunistes, autres nématodes …), ce qui établit une pression sélective favorisant la survie de *P. luminescens* et *H. bacteriophora*. De plus, des exoenzymes de la bactérie vont permettent la dégradation des tissus de l'insecte qui serviront de garde-manger aux vers. Finalement, après plusieurs générations bactériennes et la reproduction des nématodes dans le corps de l'hôte, les microbes sont à nouveau ingérés par ces derniers qui quittent l'insecte 7 à 21 jours après la première pénétration pour redémarrer un nouveau cycle.

[1] Vaste cavité interne remplie de sang, occupant tout ou partie du corps. Les organes baignant donc dans le sang. Circulation ouverte par battements cardiaques et mouvements de déformation du corps et des appendices.
[2] Nom que l'on donne au sang des insectes.

Ce cycle de vie implique une symbiose parfaite entre la bactérie et le nématode aboutissant à une pathogénicité ciblée. Ces différents stades de vie doivent être finement contrôlés afin de garantir le bon déroulement du cycle. La mort de l'insecte parasité et le développement de *H. bacteriophora* sont sous la gouverne de *P. luminescens*, ce qui suggère que cette dernière produit une multitude de molécules permettant à *H. bacteriophora* de survivre dans ces deux environnements (ffrench-Constant, Waterfield et al. 2003).

Disposant de données expérimentales et du génome complet, il était extrêmement intéressant d'analyser la composition en acides aminés de certaines protéines supposées être impliquées dans la symbiose avec le nématode ou dans la pathogénicité vis-à-vis de l'insecte.

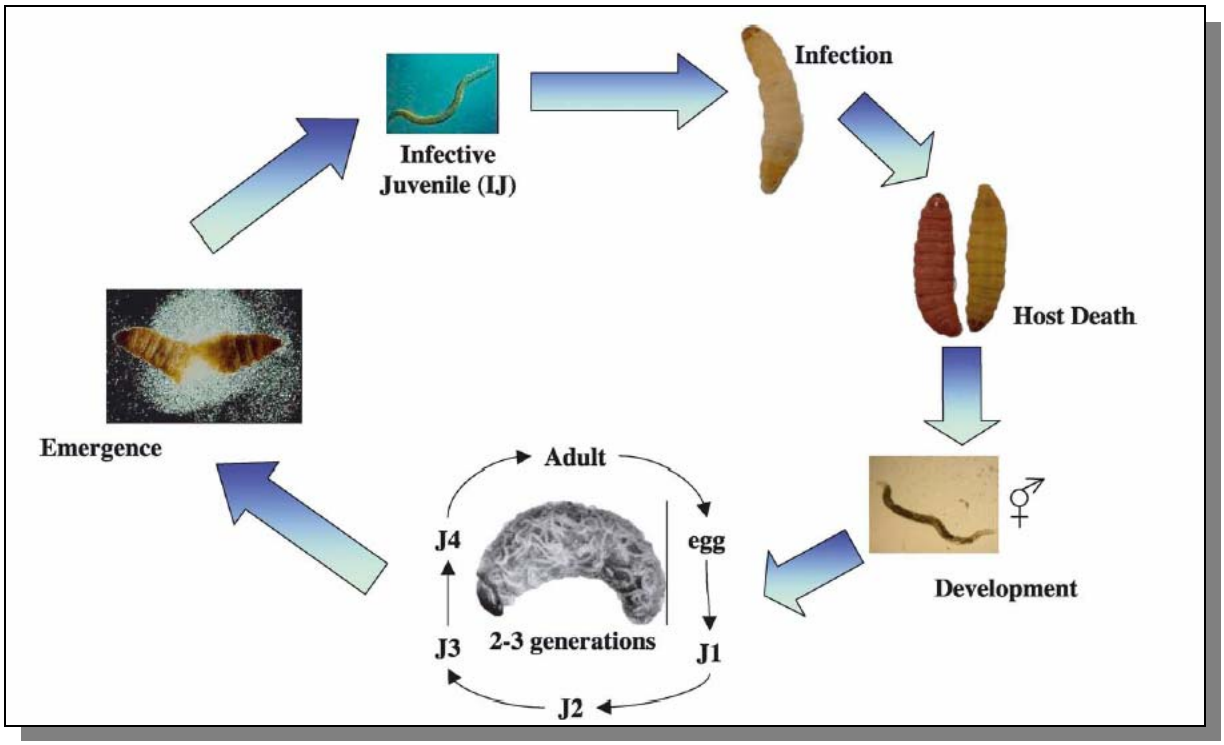


Figure 16 : Cycle de vie du tandem symbiotique *H. bacteriophora – P. luminescens.*[1]

E.II *Du in silico au in vitro.*

Les méthodes, les outils et l'expertise mis en œuvre durant cette thèse ont pu être transposés à une recherche appliquée. En effet, en collaborant aux travaux de Sylvianne Denzelle (INRA) et d'Evelyne Turlin (Institut Pasteur) sur *P. luminescens*, la bioinformatique s'est combinée à la biologie expérimentale. Le but était de montrer si les protéines supposées liées à la membrane interne étaient

[1] Tirée de ffrench-Constant, R., N. Waterfield, P. Daborn, S. Joyce, H. Bennett, C. Au, A. Dowling, S. Boundy, S. Reynolds et D. Clarke (2003). "Photorhabdus: towards a functional genomic analysis of a symbiont and pathogen." FEMS Microbiol Rev **26**(5): 433-56.

détectées comme IIMPs[1] (protéines intégrées à la membrane interne) et également d'observer la position dans les nuages de l'AFC des protéines extraites par les méthodes d'extractions de protéines mises au point par E. Turlin (méthodes décrites dans l'article 4). Cette application a permis à la fois :

- de détecter les principaux biais compositionnels de la bactérie ;
- de conforter les résultats expérimentaux en montrant que les protéines sécrétées ne se trouvaient pas dans le groupe des IIMPs ;
- d'affiner l'approche expérimentale en observant que les protéines membranaires extraites ne se trouvaient pas dans le groupe des IIMPs (cela signifie sans doute que ces protéines sont faiblement ancrées dans la membrane ce qui facilite leur extraction) ;
- et d'enrichir les jeux de données expérimentales afin de valider les méthodes statistiques utilisées. En effet, les annotations présentent dans les banques de données sont si pauvres, si hétérogènes et souvent si erronées que de nouvelles données expérimentales deviennent des mines d'informations précieuses.

E.III *L'essentiel de l'article.*

En laboratoire, deux populations de *P. luminescens* de souche sauvage sont observées. L'une a été isolée dans les insectes parasités - c'est le variant I pathogène - et l'autre n'a été observée que dans les conditions de laboratoire, après une période d'incubation prolongée du variant I (~10 jours) - c'est le variant II. Le passage du variant I au II a toujours été observé comme irréversible. Le variant II semble avoir perdu toutes les caractéristiques de la pathogénicité. Le protéome de la souche sauvage a été analysé par AFC afin de caractériser les biais possibles de sa composition en acides aminés. Le contenu en G+C du génome et l'hydrophobicité des protéines ont été montrés comme les deux principaux facteurs de déviation des protéines. Après culture de *P. luminescens* variant I et II en milieu Schneider (milieu riche, contrôlé et spécifique à la culture de cellules d'insecte), les protéines associées à la membrane, les protéines extracellulaires et cellulaires furent solubilisées et déposées sur gels d'électrophorèse à deux dimensions. L'électrophorèse bidimensionnelle permet de séparer dans un premier temps les protéines en fonction de leur point isoélectrique (gradient de pH) puis en fonction de leur masse (SDS-PAGE). L'identification de 450 taches isolées après migration, fut réalisée par spectrométrie de masse. Elles correspondaient en fait à 231 protéines différentes de *P. luminescens*. Une analyse comparative des cartes protéiques des deux états (variant I et II) fut entreprise et a permis de mettre en évidence d'importantes différences et ce, uniquement en phase stationnaire de croissance. Bien que, durant cette phase, le variant II ne sécrète qu'extrêmement peu de protéines, certaines classes sont sur-exprimées par rapport au variant I. Il s'agit de protéines impliquées dans le stress oxydatif, le métabolisme énergétique, la traduction, la transcription ainsi que le métabolisme des nucléotides. Les protéines de transport du fer, des sucres et des acides aminés sont également affectées dans le même sens. A l'inverse, les protéines chaperonnes sont fortement

[1] IIMPs : Integral Inner Membrane Proteins, *cf.* Article 1

réprimées chez le variant II. Au cours de ces études, il a, par ailleurs, été montré que la protéine H-NS (régulateur de stress) pourrait être impliquée dans la différence phénotypique entre les variants I et II.

E.IV *Article 4.*

Article soumis le 6 septembre au journal Proteomics

Proteome analysis of the phenotypic variation process in Photorhabdus luminescens

Evelyne TURLIN[1*], Géraldine PASCAL[1,2], Jean-Claude ROUSSELLE[3], Pascal LENORMAND[3], Saravuth NGO[1], Antoine DANCHIN[1] & Sylviane DERZELLE[1‡]

[1]*Unité de Génétique des Génomes Bactériens, Département de Structure et Dynamique des Génomes, Institut Pasteur, 75724 Paris Cedex 15, France*

[2]*Genoscope/CNRS UMR 8030, Atelier de Génomique Comparative, 2 rue Gaston Crémieux, 91006 Evry Cedex, France*

[3]*Plateau technique Protéomique, Institut Pasteur, 75724 Paris Cedex 15, France*

‡*Present address: AFSSA-LERQAP, 23 avenue du général de Gaulle, 94706 Maisons-Alfort cedex, France*

Running title: Proteome analysis of *P. luminescens*

*Corresponding author. Mailing address: Unité de Génétique des Génomes Bactériens, Institut Pasteur, 28 rue du Dr. Roux, 75724 Paris cedex 15. Phone: +33 (0) 1 45 68 84 44. Fax: +33 (0) 1 45 68 89 48. E-mail : eturlin@pasteur.fr

Abbreviations: MALDI-TOF, matrix-assisted laser desorption/ionization time of flight; IJ, infective juvenile; CA, Correspondence analysis; CDS, coding DNA sequence.

Abstract

Photorhabdus luminescens is an insect pathogen associated with specific soil nematodes. The bacterium has a complex life cycle with a symbiotic stage in which bacteria colonize the intestinal tract of the nematodes, and a pathogenic stage against susceptible larval-stage insect. Symbiosis-'deficient' phenotypic variants (known as secondary forms) arise during prolonged incubation. Correspondence analysis of the in silico proteome translated from the genome sequence of *P. luminescens* strain TT01 identified two major biases in the amino acid composition of the proteins. We analyzed the proteome in vivo, separating three classes of extracts: cellular, extracellular and membrane-associated proteins, resolved by two-dimensional gel electrophoresis using one pH gradient (pH 4-7). Approximately 450 spots matching the translation products of 231 different coding DNA sequences were identified by peptide mass fingerprinting. A comparative analysis was performed to characterize the protein content of both variants. Differences were evident during stationary growth phase. Very few proteins were found in variant II supernatants and numerous proteins were lacking in the membrane-associated fraction. Proteins up-regulated by the phenotypic variation phenomenon were involved in oxidative stress, energy metabolism, translation and transcription as well as nucleotide metabolism. The transport and binding of iron, sugars and amino acids were also affected and molecular chaperones were strongly down-regulated. A potential role for H-NS in phenotypic variation control is discussed.

1. Introduction

The Gram-negative gamma-proteobacterium *Photorhabdus luminescens* forms an entomopathogenic cyclic symbiosis with soil nematodes of the *Heterorhabditis* genus. Bacteria are carried in the gut of the infective stage of the nematode, known as the infective juvenile (IJ). The latter persists free in the soil, actively seeking out larval insect prey to invade. After locating a susceptible insect host, the nematodes access the insect haemocoel where they regurgitate their bacterial symbionts. Inside the haemolymph, the bacteria proliferate rapidly and produce exo- and endotoxins that kill the insect within 24-48 h. At this stage, the bacteria have reached a high cell density and produce different combinations of exoenzymes to exploit the dying insect and establish suitable conditions for nematode reproduction. When resources in the insect cadaver have been consumed, nematodes terminate the propagative phase of their life cycle and develop to a new generation of infective juveniles. These IJs no longer digest their symbiotic bacteria, but retain their partner bacteria in the intestine before leaving the insect carcass in search of new hosts (1).

Particularly under in vitro conditions, Photorhabdus spp. produce two phenotypically distinct phase variants, designated wild-type primary (variant I) and secondary forms (variant II) (1, 2). Both forms seem equally pathogenic for the insects but differ strikingly in the success of their relation with nematodes. Variant I is the bacterium found in association with IJ nematodes. The secondary form neither supports normal nematode development in the insect cadaver, nor is retained in the IJs intestine (3,4). This form appears spontaneously upon prolonged culturing and occasionally in vivo after resources in the insect cadaver have been consumed (5). Beside involvement in symbiosis, the two variants also differ in a wide range of characteristics, including their biochemical properties and their colonial and cellular morphologies. They are distinguished based on dye absorption, pigmentation, production of antibiotic substances and degradative enzymes, occurrence of crystalline inclusion proteins and bioluminescence. Secondary variant either lacks or has reduced levels of the previous properties (6,7). Little is known about the biological role of this secondary form in nature but it might provide the bacterium with a strategy for adapting to more than one particular environment, especially low osmotic strength or anoxic environments. The mechanism of phase variation in *P. luminescens* is unknown (8)

Photorhabdus pathogenicity, symbiosis and phenotypic variation seem tightly correlated to the growth stage of the bacteria. *P. luminescens* multiplication does not occur in juveniles. The bacteria multiply in natural conditions in the body of the parasitized insect. After reaching a high cell density in this medium, the bacteria enter the post-exponential phase of growth, equivalent to stationary phase in laboratory culture. During this stage, the bacteria release several exoenzymes to exploit the dying insect. The production of many different primary variant-specific proteins during the post-exponential phase of bacterial growth is associated with successful symbiosis (9). Degradative enzymes break down the macromolecules of the insect cadaver to

provide the developing nematode with a nutrient supply, while the antibiotics suppress contamination with other microorganisms and therefore secure their transmission to the nematode offsprings. Finally, phase variation spontaneous occurrence has only been observed with in vitro or in vivo stationary phase *P. luminescens* populations.

Using the genome sequence of *P. luminescens* TT01 (10) in an attempt to better characterize the functions involved in symbiosis and virulence, the present work reports bi-dimensional electrophoresis reference maps of the most abundant cytosolic, membrane-associated and extracellular proteins of strain TT01 grown in rich medium, with emphasis on phase variation. Using comparative approaches, this work identifies important proteins associated to the phenotypic variation process. The function of proteins that might be important for pathogenicity and symbiosis is discussed.

2. Materials and methods

2.1 Bacterial strains, media and growth conditions

Permanent stocks of *P. luminescens* were kept at -80 °C in Luria-Bertani medium (LB) broth supplemented with glycerol. Strain TT01 variant I (11) and variant II (12) were used in this study. Bacteria were routinely grown at 30°C in Schneider medium (Bio-Whittecker). Late exponential phase cells from overnight cultures were used to inoculate 50 or 200 ml of broth in 1 l Erlenmeyer flasks. Cells were cultivated in duplicate at 30°C with shaking. Samples from exponential (OD_{650} = 0.8), early- and late-stationary phases were collected after approximately 6 h, 24 h and 4 days, respectively. All experiments were performed in compliance with the European regulation requirements concerning the contained use of Genetically Modified Organisms of Group-I (agreement no. 2736 CAII).

2.2 Preparation of cellular, extracellular and membrane proteins

Cells were harvested by centrifugation (7000 g, 10 min, 4°C). The cell pellets were washed with cold ED minimal medium (120 mM potassium phosphate buffer, 3 mM tri-sodium citrate) and resuspended in 1 ml distilled water containing 2.5% w/v Protease Inhibitor Cocktail (Complete, Roche). After DNase and RNase treatment, cells were disrupted with a "FP120 FastPrep Cell disruptor" (Bio101) (two times 30-sec at maximum speed with 1-min intervals on ice). Cell debris was removed by ultracentrifugation for 60 min at 90,000 g. The supernatant (crude cytoplasmic extract) was divided into aliquots and stored at –20°C. The resulting pellet was suspended in 1 ml Tris/HCl buffer (10mM) containing 1% v/v Triton X-100. This suspension was stirred for 2 h at 4°C, centrifuged at 80,000 g for 90 min and the supernatant containing the solubilized membrane fraction was divided into aliquots and stored at –20°C until use. Extracellular proteins were precipitated overnight at –4°C with 10% w/v TCA before being centrifuged and washed twice in ice-cold ethanol (10,000 g, 10 min, 4 °C). After air-drying, the proteins in the pellet were solubilized.

2.3 Analytical two-dimensional gel electrophoresis

Isoelectric focusing (IEF) was conducted using the horizontal Multiphor II system (Pharmacia) at a temperature of 20°C (13, 14). For analytical gels, 60 µg of the protein extract was solubilized in 400 µl of rehydratation solution (0.5 % v/v Pharmalyte 3-10, 8 M urea, 65 mM DTT, 2 % v/v Nonidet P40), and loaded onto a 18 cm pH 4-7 L immobilized pH gradient strip (IPG) using the in-rehydratation technique (15). For preparative gels, 150 µg of the protein extract was solubilized as mentioned above. For both analytical and preparative gels, focusing was performed for 3 h at 300 V, 1 h at 750 V, 30 min at 1500 V, 16 h at 2500 V and 2 h at 3500 V (total = 50 kVh). The IPGs were equilibrated as previously described (14). The second dimension was performed with 11.5 % (w/v) SDS-polyacrylamide gels using the Protean II xi 2D Multicell system (Biorad). Proteins were stained with silver nitrate and gels were digitized using a JX-330 scanner (Sharp). Digitized 2-D gel patterns were edited and matched using the PDQUEST software package (PDI, Humington Station). To account for unspecific variations, a minimum of four gels was run for both variants and growth conditions using two independent protein preparations extracted from two independent cultures. Quantification of proteins was expressed as percentage volume, which corresponds to the percentage ratio between the volume

of a single spot and the total volume of all spots present in a gel. The mean intensity value of each spot was calculated using at least three gels. Spots showing important variations were not considered.

2.4 MALDI-TOF mass spectrometry and database searches

Mass spectrometry analyses were performed using a matrix-assisted laser desorption ionization-time-of-flight (MALDI-TOF) Voyager-DE-STR mass spectrometer (Applied Biosystems, Framingham, MA), operated in positive ion reflector mode. Spots of interest were cut out and digested with trypsin (Roche) as described previously (16). Peptide mixtures were desalted with ZipTip C18 (Millipore) and analyzed using a saturated solution of -cyano-4-hydroxycinnamic acid (Sigma) in acetonitrile containing 1% trifluoroacetic acid (Sigma) (50/50 v/v). The trypsin autolysis peptides were used as internal calibrants. Peptides were selected in the mass range of 800–3000 Da. A local copy of the MS-FIT program, developed by the University of California at San Francisco, was used to search the *P. luminescens* database (10). Search parameters were as follows: monoisotopic masses, maximum allowed peptide mass error of 50 ppm, consideration of one incomplete cleavage per peptide, and oxidation of methionine. No restrictions on M_r or pI were made. A minimum of four matching peptides was required for protein identification. To identify low M_r proteins, post source decay (PSD) experiments were performed with the MALDI instrument. Amino acid sequence similarity search was carried out using the BLASTP software (17).

2.5. Correspondence Analysis, statistics and data clustering

Correspondence analysis (CA) (18) was used to identify the major factors that shape variation in amino acid usage in the proteins of *P. luminescens*. The analysis was based on absolute frequencies in order to avoid introducing systematic biases (19). CA was applied to a data table including the amino-acid usage of *P. luminescens* proteins, to determine an orthogonal space, or factorial space, with dimension 19. The axes (or factors) are represented in a decreasing order of importance as quantified by their corresponding “inertia” (20). An automatic clustering method, the dynamic clouds method (21, 22), was subsequently used to interpret the graphical representation in terms of clusters with common properties. Proteins and amino acids are represented jointly in the obtained factorial space. Sequences that have a similar amino acid composition appear as neighbors. To avoid biases in amino acid composition linked to the molecular processes of initiation and termination of translation, protein sequences were truncated by 10 amino acids from their N-terminal end, and 5 amino acids from their C-terminal end, as there is an over-representation of hydrophilic residues near both termini of proteins (23). In addition, to reduce influence of stochastic variations that may occur in small proteins, only proteins longer than 100 residues (after truncation) were kept for analysis. After formatting, 3,910 proteins were included in the study.

3. Results

3.1. 2-DE reference maps of *P. luminescens*: identification of highly expressed proteins

Putative virulence or symbiose factors being presumably expressed in exponential and stationary growth phase, respectively, samples were taken at different time points (mid-exponential, early- and late-stationary phases) in bacteria growing in Schneider medium. Cellular, membrane-associated or extracellular proteins were extracted and 2-DE gels performed over the pH range 4 to 7.

Image analysis showed little pattern deviation between replica gels. A total of 927 spots that could be visualized following silver staining was obtained from the cellular (658 spots; Fig1A and B), membrane-associated (144 spots; Fig 2) and extracellular (135 spots; Fig 3) fractions. Most were digested with trypsin and analyzed by MALDI-TOF spectrometry. Of these proteins, 451 (200 cytoplasmic, 116 membrane-associated and 125 extracellular) provided PMM spectra of sufficient quality to perform a database search for the corresponding translated CDSs in *P. luminescens* genome (10). They matched 321 translated CDSs (159 cytoplasmic, 73 membrane-associated and 89 extracellular) corresponding to 321 different proteins. It was observed that many proteins displayed isoelectric heterogeneity (charged isoforms) and even mass variations. Isoelectric heterogeneity results from post-translational modifications, such as phosphorylation, glycosylation and acylation. It is also often the result of

systematic process-induced modifications including urea-mediated carbamylation or deamidation. The nature of the modification of proteins observed in the current analysis remains to be determined. After matching, a putative function was assigned to each protein following sequence similarity and conserved domain searches (17, 24). Each protein was categorized according to this function (Table 1).

3.1.1. Cytoplasmic proteins

Three proteins contributed to more than 30 % of the total protein content during exponential growth (Fig 1A) : the molecular chaperone GroEL (spots n°139), a protein similar to putative lipase which contains the Lipase (class 3) domain pfam01764 (Plu1518, spots n°54), and an unknown protein (Plu4286, n°141) similar to protein Pa2318 from *Pseudomonas aeruginosa* which belongs to the cupin family and is possibly involved in virulence (InterPro IPR011051). Other highly expressed proteins at this growth stage belong to four functional groups involved in biomass construction: (i) molecular chaperones, in particular DnaK (spots n°22), GrpE (n°108) and Tig (n°127), in addition to GroEL; (ii) enzymes involved in carbon assimilation and energy production among which EnoA (spot n°33), Fba (spot n°36) and GpmA (spot n°53) Icd (spot n°99), SucD (spot n°52), Mdh (spot n°147) and FumC (spot n°85); (iii) enzymes involved in nucleotides biosynthesis and metabolism, i.e. Dut (spot n°158), PurD (spot n°17), Ndk (spot n°46) and Udp (spot n°145); and (iv) proteins implicated in translation such as RpsA (spot n°64), RplI (spot n°150) and the elongation factors FusA (spot n°12), Efp (spot n°138), TufA and TufB (spots n°153 and 13).

Several proteins whose putative function might be linked to symbiosis or pathogenic processes were also identified. (i) two proteins involved in iron-uptake: Plu1174 (spot n°42) and Plu2853 (spot n°102). (ii) three antioxidant enzymes which might contribute to protection against the free radicals produced by the insect immune system: AhpC (spot n°130), SodA (spot n°4), Tpx (spot n°88). An OsmC-like protein of unknown function (Plu2338, spot n°83) might also play a role in this respect. Indeed, the OsmC pfam02566 family contains an organic hydroperoxide detoxification protein with a novel pattern of oxidative stress regulation. (iii) one protein showing similarity to cell adhesion protein (Plu1561, spot n°60). (iv) an Ail/Lom-like protein of the pfam06316 family (Plu1967, spot n°68) similar to the known virulence factors *Yersinia enterocolitica* Ail protein and *Escherichia coli* O157:H7 PagC membrane protein. The cytoplasmic localisation of this putative outer membrane protein was however unexpected. And finally, (v) the global regulator H-NS (spot n°229).

In stationary phase, cells have somewhat redirected their metabolism (Fig. 1B). High amount of PTS glucose-specific IIA component Crr (spot n°48), GabD (spot n°37), LeuA (spot n°116), RibH (spot n°129) and AldB (spot n°118), a predicted aldehyde dehydrogenase highly similar to several Lactaldehyde dehydrogenases, was detected. The abundance of KdsA (spot n°78) and Tig (spot n°127) has similarly increased, as those of several unknown proteins and one putative tail fiber assembly protein (Plu4369, spot n°144). But the most expressed proteins were still molecular chaperones GroEL (spots n°139) and GroES (spot n°140), while new isoforms of both proteins appeared. Some changes were also observed for the antioxidant proteins Tpx (spot n°88) and SodA (spot n°4).

As expected, several proteins suspected to be involved in secondary metabolites biosynthesis were also up-regulated during the stationary growth-phase. Three putative antibiotic biosynthesis proteins were identified. It included a predicted epimerase similar to phenazine biosynthesis-like proteins PhzC/PhzF (Plu2271, spot n°82), a putative monooxygenase related to polyketide biosynthesis enzymes (Plu0947, spot n°34) and a probable short chain dehydrogenase/reductase similar to the 3-oxoacyl- acyl-carrier protein reductase YusR of Bacillus sp. (Plu1541, spot n°58).

3.1.2. Membrane-associated proteins

The protein content of membrane-associated fractions was found to be quite similar in exponential or stationary growth phase (Fig 2 and data not shown). Many proteins identified in this fraction, as well as in the extracellular culture fluid, would classically be considered to be associated with the cytoplasm or the inner surface of the cytoplasmic membrane. These included abundant cellular proteins, such as molecular chaperones or housekeeping metabolic proteins such as AldB, EnoA, Tpx, Crr, TufA and TufB. Because of their high abundance in the cell, they could be contaminating proteins from residual cell lysis. However, it has

recently become apparent that a number of proteins thought to be restricted to the cytosol were also associated with cell-surface or secreted into the external medium.

Besides major cytoplasmic proteins, few proteins predicted to be inner membrane compounds using the PSORT program (http://psort.nibb.ac.jp) were resolved in the membrane-associated fraction. Among them, the more expressed proteins were Plu2059 and Plu2060 (spots n°74 and n°169), two unknown proteins similar to each other but exhibiting no other significant similarity to proteins in databases. Both Plu2059 and Plu2060 were also detected in lesser abundance in the extracellular and cytoplasmic fractions. Others inner membrane proteins included AtpD (spot n°2), HemX (spot n°173) and two probable aminotransferase present in low amount: an aminotransferase of class 3 showing high similarity to beta-alanine-pyruvate transaminase of *Pseudomonas aeruginosa* and *Ralstonia eutropha* (Plu2260, spot n°172) and a transaminase (Plu3517, spot n°176) similar to the homocysteine synthase MetY of *Pseudomonas putida*.

Numerous proteins predicted to be periplasmic or outer membrane proteins were also identified. The major one was the receptor TolB (spot n°170). Four proteins were involved in transport and binding of substrate: MetQ (spot n°178), OppA2 (spot n°168) and two iron chelator ABC transporter substrate-binding proteins Plu1174 (spot n°42) and Plu2853 (spot n°102). Others included a protein similar to acriflavin resistance protein A precursor, AcrA (spot n°171), an unknown protein (Plu3611, spot n°180), a component of the pyruvate dehydrogenase complex, LpdA (spot n°114), and one enzyme responsible for the hydrolysis of deacylated phospholipids, GlpQ (spot n°175). Finally, seven other proteins were detected in both membrane-associated and culture supernatant fractions , i.e. spot n°187, n°56, n°67, n°186, n°182, n°119 and n°34. They are discussed below (see point 3.1.3.).

3.1.3. Extracellular proteins

We noted that the production of many extracellular proteins occurred during the post-exponential growth phase. Indeed, very few proteins were extracted from exponentially growing culture supernatants (data not shown). In contrast, 89 different proteins were isolated from the supernatant of stationary phase cultures (Fig 3, Table 2).

Among those, one protein resolved as several isoforms of various pI and M_r and represented more than 30% of the total protein content. This prevalent polypeptide, identified as Plu1537 (spot n°56), is a small hypothetical toxin of 136 amino acid residues also detected in lesser amount in both other fractions. Sequence similarity searches revealed a low similarity (E value below the 0.005 significance cutoff) to Cry34A toxins from *Bacillus thuringiensis*. Previous analysis using iterated PSI-BLAST and *B. thuringiensis* Cry34Ab1 as query had already highlighted such similarity. Plu1537 was reported to be 27% identical to Cry34Ab1 (25). In *B. thuringiensis*, crystal proteins of Cry34 and Cry35 classes function together as binary insecticidal toxins showing activity on western corn rootworm (25, 26).

P. luminescens is known to develop large intracellular protein crystal inclusions in stationary phase which may represent a stored source of nutrients for the nematode host (8). The crystalline inclusion proteins CipA (spot n°222) and CipB (spot n°187) were found among the highly abundant extracellular and membrane-associated proteins, but not in the cytoplasmic extracts.

Other proteins which might be important for the pathogenic or symbiotic relationships of the bacterium with its hosts were also identified. They were among other functions related to adhesion, proteolysis and antioxidant defense. Putative adhesion molecules included Plu1561 (spot n°60), Plu2963 (spot n°211) and Plu2096 (spot n°218), respectively similar to Dictyostelium discoideum calcium-dependent cell adhesion molecule-1, regions of a putative adhesin of *Escherichia coli* and galactophilic lectin PA-I of *Pseudomonas aeruginosa*. The two proteases detected were Plu1382 (spot n°47) and Plu2455 (spot n°197). The former is similar to extracellular metalloproteinase precursors. The latter shows weak similarities with calpain cysteine proteases. Four enzymes were putatively involved in oxidative stress: AhpC (spot n°130), Tpx (spot n°88), SodA (spot n°4) and Bcp (spot n°97). All four were also detected in the cytoplasmic fraction. Besides those proteins, we also resolved one component of the Pir toxin of *P. luminescens*, Plu4093 (spot n°188), in both extracellular and membrane-associated fractions (10,27). Another protein (Plu2534, spot n°186) showing some similarity to this component was also found in the same fractions. Two putative sialidases (Plu0734, spot n°201; Plu0735, spot n°202) and a protein similar to hemolysin co-regulated proteins

Hcp (Plu373, spot n°8) might also be part of *P. luminescens* virulence factors. Identification of the UMP kinase PyrH (spot n°227), whose role in virulence has been shown in *Vibrio vulnificus* (Kim et al., 2003), was also of interest. Indeed UMP kinase is an important enzyme for the de novo synthesis of pyrimidine nucleotides that may be essential for in vivo cell growth and division. This enzyme has been shown to be associated in vivo to the cell's envelope (28).

Several proteins with no significant similarity to proteins in databases were also present in high amount: Plu1840 (spot n°67), Plu2972 (spot n°182) and Plu3795 (spot n°119) were found in both extracellular and membrane-associated fractions, while protein Plu2256 (spot n°220) was only detected in the culture supernatant. Gene plu2256 codes for a very short protein with a prominent putative signal sequence. A last class of proteins well represented in the extracellular proteome of TT01 were those related to phage proteins, the most abundant being proteins Plu0014 (spot n°213), Plu0015 (spot n°1) and Plu2959 (spot n°212). Three extracellular proteins present in lower amount worth also to be mentioned: a probable amidinotransferase highly similar to L-arginine:lysine amidinotransferase from *Pseudomonas syringae* (Plu0158, spot n°6); a porin termed OmpN (Plu1752, spot n°199) and a probable monooxygenase putatively involved in polyketide biosynthesis (Plu0947, spot n°34) which has been detected in all three fractions.

3.2 Correspondence Analysis of the *P. luminescens* proteome

Correspondence Analysis (CA) was used to analyze the overall amino acid composition of the proteome (3,910 individual proteins) of *P. luminescens*. The predictive power of CA is characterised by the distribution of the inertia (percentage of variance) along each axis. The higher the total inertia of one axis, the higher its information content. With the *P. luminescens* proteome, more than half of the inertia was distributed into the four first factors (while a random distribution of residues would have predicted an average inertia slightly larger than 5%). Analysis of the information carried in the CA was limited to the first three axes, carrying the most significant part of the whole information.

3.2.1. The amino acids distribution in the proteome is biased by the G+C-content of the genes

In the projection of the cloud of points on the factorial plane made of the two first axes, two well-separated clouds, more or less elliptical in shape, were observed (Fig 4 A and B). Along the first axis, alanine (A, codons GCN) was opposed to two amino acids with A-rich codons: asparagine (N, codons AAY) and lysine (K, codons AAR), contributing to most of the inertia (data not shown). This suggested that the cloud's shape might result from some contribution of the G+C-content of the genome to the amino acid composition of its proteome. This prompted us to compute the correlation between the G+C content of each protein's coding sequence and its coordinate on axis 1. These parameters are highly correlated ($r=0.82$, $p<10^{-4}$). This demonstrated that a significant pressure for amino acid preference is not determined by the selection pressure on the proteins, but rather on the genome.

3.2.2. Hydrophobic amino acids bias of membrane proteins

The strong inertia of the second CA axis revealed another driving force in the selection pressure on the amino acid composition of the proteome. As shown in Figure 4B, this axis splits the distribution of proteins into two well-separated groups (A and B). Overlying the dual space on the space of proteins, the factorial coordinates of amino acids provided us with an interpretation of the graph. Large non-aromatic hydrophobic amino acids (I and L) oppose to some polar residues (E, Q, R and D), accounting for most of the contribution to the inertia of this axis. A very strong correlation ($r =-0.91$, $p<10^{-4}$) was observed between axis 2 and hydrophobicity of *P. luminescens* proteins, substantiating that hydrophobicity vs polarity biases the *P. luminescens* proteome. The smallest group (group A) was shown to contain only proteins located to the inner membrane with transmembrane segments contributing to more than 30% of their length (Pascal et al., 2005). This cluster represents approximately 10% of the *P. luminescens* proteome and contains, indeed, only Integral Inner Membrane Proteins (IIMP) among known proteins.

Dynamic clustering was used as a tool for straightforward identification of IIMPs in complete bacterial proteomes (29). Applied to the three class of proteins experimentally identified with 2D-electrophoresis gels no extracellular or cytoplasmic protein belong

to group A (Fig 4A). Furthermore, no protein identified as membrane-associated belonged to group A. We can safely conclude that those which were experimentally identified contain less than 30% of their amino acid residues imbedded within the membrane. This demonstrated that the method used to isolate the membrane fraction could not extract IIMPs from the cells or that they failed to enter gels.

3.2.3. Extracellular proteins and aromaticity

Remarkably, the proteins identified in the 2D electrophoresis study were not distributed randomly along the third CA axis of the proteome cloud of points: the extracellular proteins were clustered together (Fig 4C). This axis is negatively correlated with the aromatic amino acids content of proteins ($r=-0.56$, $p<10^{-4}$), as shown by the opposition between residues A and G versus Y, W and F. Positive coordinates along the axis being associated to a low content in aromatic amino acids, this indicated that extracellular proteins are comparatively poor in aromatic amino acid residues and rich in the small amino acids A and G.

3.3 Comparison of the 2-DE profiles of phenotypic variants

To get insight into the mechanisms involved in phenotypic variation and to correlate the phenomenon with expression of particular proteins, proteomic profiles of the secondary phenotypic variant of TT01 were performed. To compare 2D-electrophoresis gels, variant II was similarly grown in Schneider medium to exponential or early and late stationary phase. For each condition, two independent proteins preparations were made and at least two gels were run, silver-stained and analysed for each preparation. Representative patterns of silver-stained proteins in 2-DE are shown in Fig 5 and 6.

3.3.1. Extracellular and Membrane-associated proteins

The supernatant of secondary cultures was found to contain considerably less proteins than that of its primary counterpart. We had to proceed to 10 to 100-fold concentration of the supernatant to get faint visible spots in 2D-electrophoresis gels (data not shown). Therefore all proteins found in primary supernatants (see point 3.1.3 and Table 2) were lacking or in strongly reduced amount. Variant II is therefore defective in protein secretion.

Huge modifications were also observed in the membrane-associated proteins content (Fig 5, Table 3). At least thirty three polypeptides were affected, most of them being down-regulated or lacking in the secondary variant. Proteins whose abundance were altered between both phenotypic forms were of the following families of function: (i) Adaptation to stress. The amount of ProQ (spot n°231) and, an effector involved in osmosensing and activation of solute transport, and Plu3820 (spot n°228), a thioredoxin-like protein, was higher in the secondary variant. (ii) Molecular chaperones. The HSP-cofactor GrpE (spot n°108) was found to be down-regulated while HtpG was up-regulated. (iii) Transport and binding of nutrients. MetQ (spot n°178) and a putative iron-binding protein of an ABC transport system (Plu1174, spot n°42) were down-regulated, in contrast to ManX (spot n°230), the IIAB component of mannose-specific PTS system. (iv) Secondary metabolites. A dramatic decrease in putative Cry3A-like toxin Plu1537 (spot n°56) and crystal protein CipB (spot n°187) was noted. (v) Cell envelope-related proteins. Lower levels of HemX (spot n°173) were detected while more TolB receptors (spot n°170) were observed. (vi) Energy metabolism. Higher levels of LpdA (spot n°114), Ppa (spot n°149) or GpmA (spot n°53) were detected while several isoforms of the predicted aldehyde dehydrogenase AldB (spot n°118) were lacking. (vii) Translation. Some proteins such as the elongation factors Ts (spot n°25) and TuA (spot n°153) or the ribosomal protein RpsA (spot n°64) were up-regulated. (viii) Unknown. Most of the proteins of unknown function disappeared from the secondary variant 2D-pattern, especially the predicted inner membrane associated proteins Plu2059 and Plu2060 (spots n°74-169), highlighting their potential importance for symbiosis or virulence. When detected in other fractions (either cytoplasmic and/or extracellular), similar trends were observed for those proteins.

3.3.2. Cytoplasmic proteins

Consistent with a previous report (12), the overall profile of total soluble proteins in exponential growth phase was found to be identical in both variants (data not shown). In contrast, major differences were detected between phenotypic variants in stationary growth phase (Fig. 6, Table 4).

Proteins whose abundance was modified belong to the same functional families as in the case of membrane-associated proteins. (i) Several proteins involved in adaptation to stressful conditions, in particular oxidative stress, were up-regulated in the secondary variant. This included SodA (spot n°4), TrxB (spot n°61), AhpC (spot n°130) and UspA (spot n°217). (ii) Numerous isoforms of major molecular chaperones disappeared in the secondary 2D-pattern, including DnaK (spot n°22), GroEL (spot n°139), GrpE (spot n°108) and Tig (spot n°127). A putative thioredoxin-like protein (Plu0198, spot n°7) was also down-regulated. (iii) The level of several proteins implicated in transport and binding of nutrients was modified. The leucine-binding protein LivK (spot n°137) was up-regulated in the variant II, while both the IIA component of the glucose-specific PTS, Crr (spot n°48), and the iron compound ABC transporter substrate-binding protein (Plu1174, spot n°42), were down-regulated. (iv) Soluble proteins involved in secondary metabolites biosynthesis were also down-regulated. This included a putative polyketide ketoreductase (Plu1541, spot n°58), a protein homologue to the C-terminal endonuclease domain of S-type pyocin killer protein (Plu887, spot n°31), the Cry34-like toxin Plu1537 (spot n°56) and a probable monooxygenase involved in polyketide biosynthesis (Plu0947, spot n°34). (v) Cell envelope-related proteins such as KdsA (spot n°78), an aldolase involved in LPS biosynthesis, or the Ail-like protein Plu1967 (spot n°68) were found in lower amount in variant II. (vi) Several enzymes involved in energy metabolism, mainly those of the TCA cycle or glycolysis pathway, were up-regulated: Mdh (spot n°147), FumC (spot n°85), SucD (spot n°52), GpmA (spot n°53), Eno (spot n°33), Ppa (spot n°149) and DeoB (spot 19). In contrast, AldB (spot n°118), the main expressed soluble proteins during stationary phase, was down-regulated, as GabD (spot n°37) or CarA (spot n°23). (vii) Some proteins associated with translation and transcription processes were more abundant in the secondary variant cells among which the elongation factors Ts (spot n°25) and Efp (spot n°138), ribosomal protein RplJ (spot n°15) or the antiterminaison factor NusG (spot n°14). An increased amount of PpiB (spot n°121), which is involved in posttranslational modification, was also noted. (viii) The level of some proteins of unknown function was modified. Plu4286 (spot n°141) and Plu3826 (spot n°123) were up-regulated, while Plu1840 (spot n°67) was down-regulated.

Additional functions up-regulated by the phenotypic variation process were also pointed out by this analysis. It involved proteins related to nucleotides biosynthesis (PyrB, spot n°146), nucleotide and carbohydrate transport and metabolism (HIT family of hydrolase, Plu2825, spot n°100) and metabolism of coenzymes (RibH, spot n°129-140). Finally, the abundance of one global regulator, H-NS (spot n°229), was found to be higher in the secondary variant cells.

4. Discussion

This study established the first comprehensive proteomic reference map of cytoplasmic, membrane-associated and extracellular proteins of *P. luminescens* in a medium rich in amino acids and peptides attempting to mimick growth inside insects. It also highlighted the possible role played by several uncharacterized proteins in virulence (among others the putative Cry34-like toxin Plu1537 and PagC-like protein Plu1967) or nematode-bacterium symbiosis (particularly AldB and a putative adhesion molecule Plu1561) that are worth further investigation.

A further focus of the 2 DE-analysis on the phenomenon of phenotypic variation, demonstrated an increased level of proteins implicated in energy metabolism (TCA cycle, glycolysis and pyruvate dehydrogenase complex), translation and, to a lesser extent, nucleotide metabolism in the secondary variant. Up-regulation of these proteins suggested that variant II may be more active with respect to cellular metabolism and/or may accumulate “stock” proteins to get ready to grow into new environments. This is consistent with the fact that variant II grows faster than the primary and restarts growth more rapidly after periods of starvation (1). Secondary cells were also reported to differ in their assimilation of nutrients and to incorporate proline faster than primary cells (1, 30). Several proteins responsible for the transport and binding of compounds such as iron, sugar or amino acids were indeed found to be up- or down-regulated. In this respect, we may wonder whether variant II does not represent a robust form of the bacteria that may, upon appropriate interaction with a specific medium, reverse to the variant I symbiotic form. *P. luminescens* is not found as a free-living organism in the wild, and, while it undergoes an efficient symbiotic cycle, it is likely that it should exist in a persistent form that would be able to start again a symbiotic cycle when relevant conditions are found in

the environment. Previous experiments, however, failed to reverse the variant I – variant II transition. This may be due to lack of identification of an essential factor needed for this reversion. The increase in the PTS mannose transport system might provide a clue for a required component, as mannose derivatives are often involved in symbiotic interactions.

Another interesting finding of this analysis was the up-regulation of the global regulator H-NS in secondary cells. A key role for H-NS in the control of phenotypic variation shift is an attractive possibility. Several aspects are consistent with this hypothesis. Firstly, this phenomenon occurs in nature after resources in the insect cadaver have been consumed. As cases in point, most modifications characterized in the proteomic profile of variant II are detected during the stationary growth phase. Then, H-NS level was specifically increased during this period in the secondary variant (Fig 6, Table 4). hns expression is growth phase-dependent in both *P. luminescens* variant I (this study and data not shown) and *E. coli*, being highly expressed during the exponential growth phase. Secondly, H-NS has been shown to control many genes involved in responses to osmolarity and oxygen starvation in *E. coli* (31). The presence in higher amount in the variant II of several proteins involved in oxidative stress, as well as of the regulator ProQ, suggested that osmosensing and the oxidative status of cells are important factors in the decision to switch from variant I to variant II. Moreover, both stresses induce the formation of secondary cells in *P. luminescens* (1, 32). Thirdly, H-NS is known to act, in particular, in derepressing genes. In *E. coli*, it represses 80% of its regulon (33). In this respect, we have shown that an important feature of the secondary variant of *P. luminescens* was the lack of synthesis of numerous proteins, specially extracellular and membrane-associated polypeptides. At last, similarities exist between both *E. coli* H-NS targets (33) and proteins associated to phenotypic variation in *P. luminescens*. Both included cell envelope components, proteins involved in adaptation to environmental challenges, translation and central metabolism.

Except the previously mentioned up-regulated proteins, the shift from variant I to variant II was above all associated with down-regulation of proteins. All extracellular proteins and most secondary metabolites biosynthesis proteins, including putative toxins or antibiotic biosynthesis proteins, were missing or present in lower amount in the secondary variant. Modifications of cell envelope components were also largely associated with lack or down-regulation of membrane-associated proteins, in particular the predicted envelope-associated proteins Plu2059 and Plu2060. This study demonstrated that variant II did not secrete numerous polypeptides. Several processes might account for that observation. First, as reported by Joyce and Clarke (9) in the case of P. temperata, expression or synthesis of a wide range of primary-specific compounds might be shut off during the post-exponential phase of bacterial growth in the secondary variant. Secondly, inaccurate post-translational modifications and degradation can also be suggested. A higher proteolytic activity resulting in the appearance of fragments of proteins was indeed detected in the proteome patterns of secondary cells. This could be consistent with the strong decrease in amount observed for several isoforms of major molecular chaperones, as GroEL, responsible for folding, repair and degradation of proteins particularly when export and translocation processes are altered. Thirdly, a defect in export process itself due to membrane modifications or altered protein targeting can be suspected. Composition of membrane-associated proteins was found to deeply differ between both forms, suggesting that both the organisation as well as the functioning (transport) of the bacteria may be changed in the secondary form.

Appearance of high amount of RpsA in the stationary growth phase membrane-associated fraction of variant II was in this respect of interest. Indeed a mutation suppressor of a preprotein translocase complex mutation (secY24), causing protein export defect and accumulation of precursors of periplasmic and outer membrane proteins within the cell, has been mapped within rpsA in *E. coli* (34, 35). Moreover, in Yersinia, a S1 RNA binding domain has been shown to be required for the optimal functioning of the type three secretion system (36). Up-regulation of S1 in variant II might therefore pointed out a possible secretion modulator in this form. We can however not exclud that S1 might play pleiotropic role in variant II. Beside its activity in translation initiation, the homologue of rpsA has been shown to modulate the expression of virulence genes in Bordetella pertussis (37). S1 is also thought to act as an RNA-binding protein presenting mRNA to the degradosome complex (38).

In *P. luminescens*, phenotypic variation is associated with loss of symbiosis relationship with the nematode host (9). The lack or reduced synthesis of numerous proteins in the secondary form is likely of significance in its symbiose-deficient behaviour. It may affect the availability of essential nutrients and developmental signaling factors required for *Heterorhabditis* growth and reproduction. Inactivation of either cip gene encoding putative stored source of nutrients for the nematode hosts has been shown to create a variant cell type resembling the secondary variant in many respects. Like variant II, the cip strain is unable to support growth of the nematode (39). Bioconversion of the insect cadaver by degradative enzymes produced by *P. luminescens*

are also important for the developing nematode but form II did not export or produce a variety of secondary metabolites such as protease Plu1382 (spot n°47). A lesser abundance of several uncharacterised iron-binding proteins might also be detrimental to the mutualistic relationship. Aspects of iron metabolism in Photorhabdus have been reported to be important during the symbiosis with the nematode (40). Beside this nutritional aspect, modifications of the (outer) surface of the bacterium in variant II might also be detrimental for the symbiotic interaction. *P. luminescens* is expected to require functions for attachment to the nematode intestine. A role in this process might be speculated for the putative adhesion molecule Plu1561 (spot n°60) which almost disappeared in the variant protein content. Glycocalix produced by the primary form are also thought to be involved in the specific association with the epithelial cells of the nematode gut (6). Differences observed in cell surface proteins might therefore explain in some extent why IJs do not retain variant II cells in their intestine and why variants II cells show distinct phenotypic traits linked to envelope as dye absorption, pigmentation or colony morphology.

Acknowledgements

Financial support came from the Institut Pasteur, the Centre National de la Recherche Scientifique (URA 2171) and the French ASG program involving Bayer CropScience, the Institut Pasteur and INRA, supported by the Ministry of Industry.

References

1. Boemare, N., Givaudan, A., Brehelin, M., Laumond, C., *Symbiosis* 1997, *22*, 21-45. 2.
2. Boemare, N.E., Akhurst, R.J., *J. Gen. Microbiol.* 1988, *134*, 751-761.
3. Han, R., Ehlers, R.-U., *J. Invert. Pathol.* 2000, *75*, 55-58.
4. Han, R., Ehlers, R.-U., *FEMS Microbiol. Ecology* 2001, *35*, 239-247.
5. Hurlbert, R.E., ASM News 1994, 60, 473-478.
6. Forst, S., Dowds, B., Boemare, N., Stackebrandt, E., *Annu. Rev. Microbiol*. 1997, *51*, 47-72.
7. FfrenchConstant, R., Waterfield, N., Daborn, P., Joyce, S. *et al. FEMS Microbiol. Rev.* 2003, *26*, 433-456.
8. Forst, S., Clarke, D., in: Gaugler, R. (Eds), *Entomopathogenic Nematology*, CABI publishing, Wallingford 2002, pp. 57-77.
9. Joyce, S.A., Clarke, D.J., *Mol. Microbiol*. 2003, *47*, 1445-1457.
10. Duchaud, E., Rusniok, C., Frangeul, L., Buchrieser, C. *et al.*, *Nat. Biotechnol*. 2003, *21*, 1307-1313.
11. Fischer-Le Saux, M., Viallard, V., Brunel, B., Normand, P., Boemare, N. E., *Int. J. Syst. Bacteriol.* 1999, *49*, 1645-1656.
12. Derzelle, S., Ngo, S., Turlin, E., Duchaud, E. *et al*., *Microbiology* 2004, *150*, 897-910
13. Gorg, A., Postel, W., Gunther, S., *Electrophoresis* 1988, *9*, 531-546.
14. Gorg, A., Postel, W., Weser, J., Günther, S. *et al.* Electrophoresis 1987, *8*, 122-124.
15. Rabilloud, T., Valette, C., Lawrence, J. J., *Electrophoresis* 1994, *15*, 1552-1558.
16. Shevchenko, A., Wilm, M., Vorm, O., Mann, M., *Anal. Chem.* 1996, *68*, 850-858.
17. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J. *et al*., *Nucleic Acids Res.* 1997, 25, 3389-3402.
18. Benzecri, J.-P., in: Dunod Edition (Eds), L'analyse des données, L'Analyse des Correspondances, Vol. 2., Paris 1973
19. Perrière, G., Thioulouse, J., *Nucleic Acids Res.* 2002, *30*, 4548-4555.
20. Lebart, T., Morineau, A., Warwick, K. A., *Multivariate Descriptive Statistical Analysis*, John Wiley and Sons, 1984.
21. Delorme, M. O., Hénaut, A. *Comput. Appl. Biosci.* 1988, *4*, 453-458.
22. Diday, E. *Rev. Stat. Appliquée* 1971, *19*, 19-33.
23. Rocha, E. P., Danchin, A., Viari,, A., *Nucleic Acids Res*. 1999, *27*, 3567-3576.
24. Marchler-Bauer, A., Bryant, S. H., *Nucleic Acids Res*. 2004**,** *32*, 327-331.
25. Schnept, H. E., Lee, S., Dojillo, J. A., Burmeister, P. *et al.*, *Appl. Environ. Microbiol*. 2005, *71*, 1765-1774.
26. Moellenbeck, D. J., Peters, M. L., Bing, J. W., Rouse, J. R. *et al.*, *Nature Biotech.* 2001, *19*, 668-672.
27. Waterfield, N., Kamita, S.G., Hammock, B.D., Ffrench-Constant, R., *FEMS Microbiol. Lett.* 2005, *245*, 47–52.
28. Landais, S., Gounon, P., Laurent-Winter, C., Mazie, J. C. *et al., J. Bacteriol.* 1999, *181*, 833-840.
29. Pascal, G., Medigue, C., Danchin, A., *Proteins*, 2005, *60*, 27-35.
30. Smigielski, A. J., Akhurst, R. J., Boemare, N. E., *Appl. Environ. Microbiol.* 1994, *60*, 120-125.
31. Atlung, T., Ingmer, H., *Mol. Microbiol.* 1997, *24*, 7-17.
32. Krasomil-Osterfeld, K.C., Appl. Environ. Microbiol. 1995, *61*, 3748-3749.
33. Hommais, F., Krin, E., Laurent-Winter, C., Soutourina, O. *et al.*, *Mol. Microbiol*. 2001, *40*, 20-36.
34. Shiba, K., Ito, K., Yura, T., *J. Bacteriol.* 1984, *160*, 696-701.
35. Artamonova, V.S., Boni, I.V., *Bioorg Khim.* 1996, *22*, 941-3.
36. Rosenzweig, J.A., Weltman, G., Plano, G.V., Schesser, K., *J. Biol. Chem.* 2005, *280*, 156-63.
37. Fuchs, T. M., Deppisch, H., Scarlato, V., Gross, R. *J. Bacteriol.* 1996, *178*, 4445-52.
38. Danchin, A., *DNA Res.* 1997, *4*, 9-18.
39. Bintrim, S. B., Ensign, J. C., *J. Bacteriol.* 1998, *180*, 1261-69.
40. Watson, R. J., Joyce, S. A., Spencer, G. V., Clarke, D.J., *Mol. Microbiol*. 2005, *56*, 763–773.

Legends to Figures

Fig I : Global 2-DE maps of cytoplasmic proteins of the wild-type *P. luminescens* strain, TT01, grown in Schneider medium. Proteins were focused in the first dimension using mid-range pH 4.0-7.0 IPG strips, prior to separation on 11.5% w/v polyacrylamide gels in the second dimension and silver staining. Numbers indicate spots that were identified by MALDI-TOF MS (Table1). (A) Proteins harvested in exponential phase (6 h-old). (B) Proteins harvested in early stationary phase (24 h-old).

Fig 2: Two-dimensional silver-stained map of membrane-associated proteins of TT01 grown in Schneider medium. Proteins were focused in the first dimension using mid-range pH 4.0-7.0 IPG strips, prior to separation on 11.5% w/v polyacrylamide gels in the second dimension. Numbers indicate spots that were identified by MALDI-TOF MS (Table1).

Fig 3: Two-dimensional silver-stained map of TCA precipitated proteins from the culture fluid of TT01 grown in Schneider medium. Proteins were focused in the first dimension using mid-range pH 4.0-7.0 IPG strips, prior to separation on 11.5% w/v polyacrylamide gels in the second dimension. Numbers indicate spots that were identified by MALDI-TOF MS (Table1). .

Fig 4: Distribution of the protein sequences on the CA space determined by the three first factors. Red squares represent proteins in well-separated group (IIMPs), yellow diamonds represent experimental membrane proteins, pink triangles represent extracellular proteins, dark blue circles represent cytoplasmic proteins and cyan squares represent all other proteins. Amino acids are represented by black crosses.

Fig 5: Comparison of the membrane protein synthesis patterns of *P. luminescens* TT01 variant I (A) and TT01 variant II (B). Cells were grown in Schneider medium at 30°C for 24 h (early stationary phase). Proteins were separed in IPG pH 4-7 gels in the first dimension and in11.5 % (w/v) polyacrylamide gels in the second dimension. After silver staining, proteins induced (□) or repressed (O) in the TTO1 variant II were identified by MALDI-TOF. Numbers refer to proteins that have been identified (see Table 3). 2-DE gels were repeated at least three times for each strain condition.

Fig 6: Comparaison of the cytoplasmic protein synthesis patterns of *P. luminescens* TT01 variant I (A, C) and TT01 variant II (B, D). Cells were grown in Schneider medium at 30°C for 24 h (A, B) and 96 h (C ,D). Proteins were separed in IPG pH 4-7 gels in the first dimension and in 11.5 % (w/v) polyacrylamide gels in the second dimension. After silver staining, proteins induced (□) or repressed (O) in the TTO1 variant II were identified by MALDI-TOF. Numbers refer to proteins that have been identified (see Table 4). 2-DE gels were repeated at least three times for each strain and condition.

Figures

Figure 1:

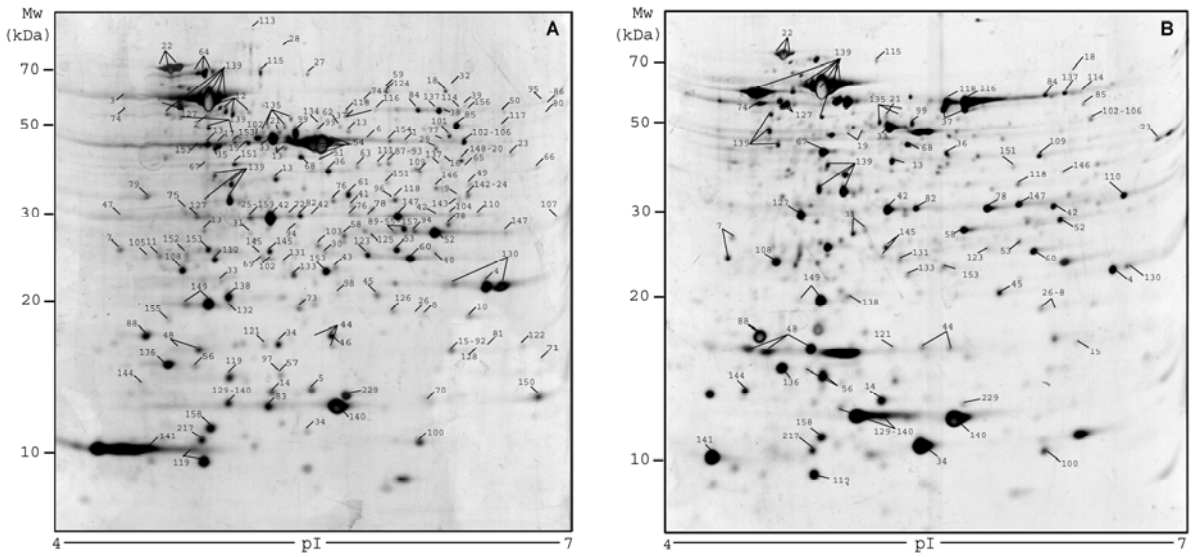


Figure 2:

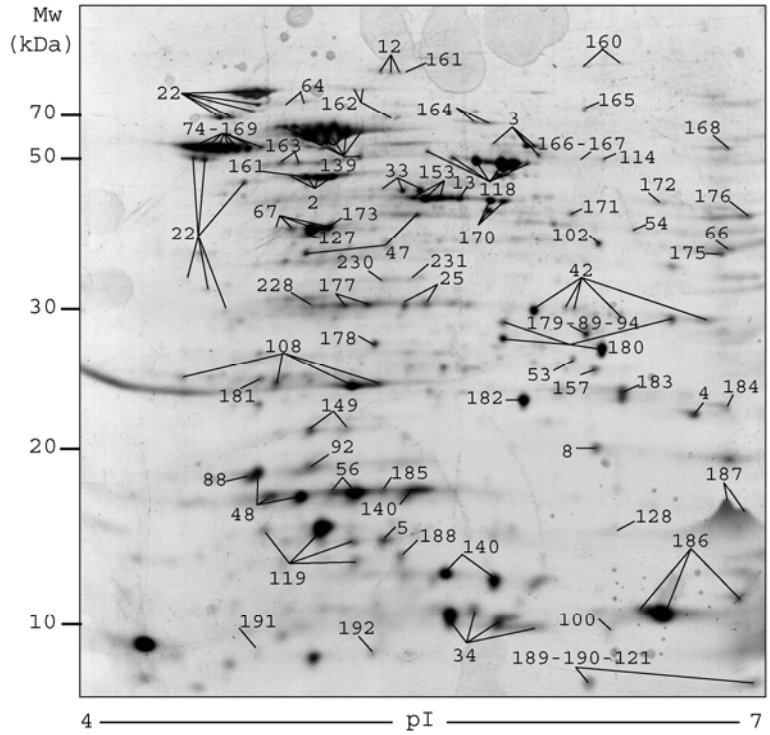


Figure 3:

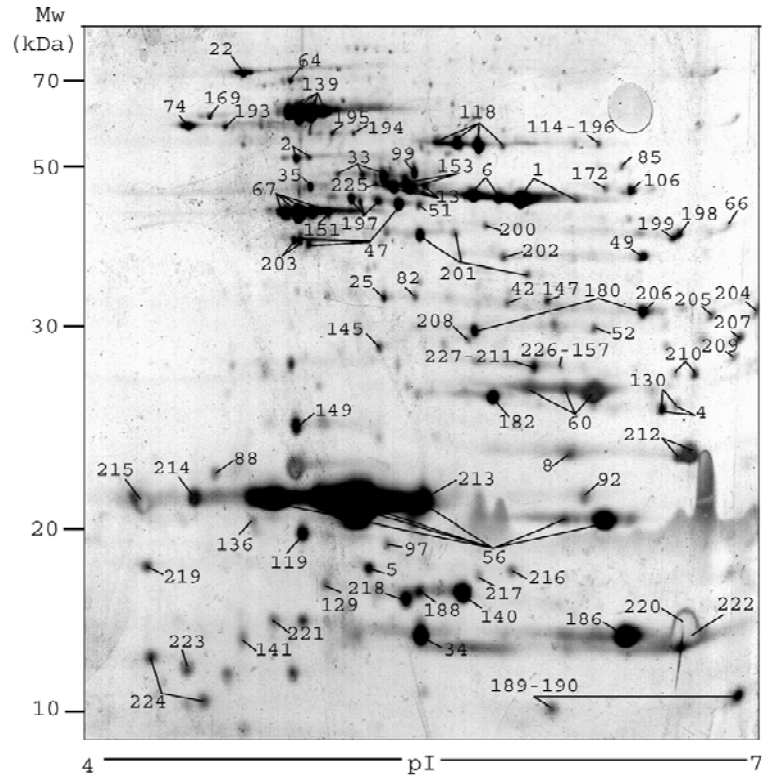


Figure 4:

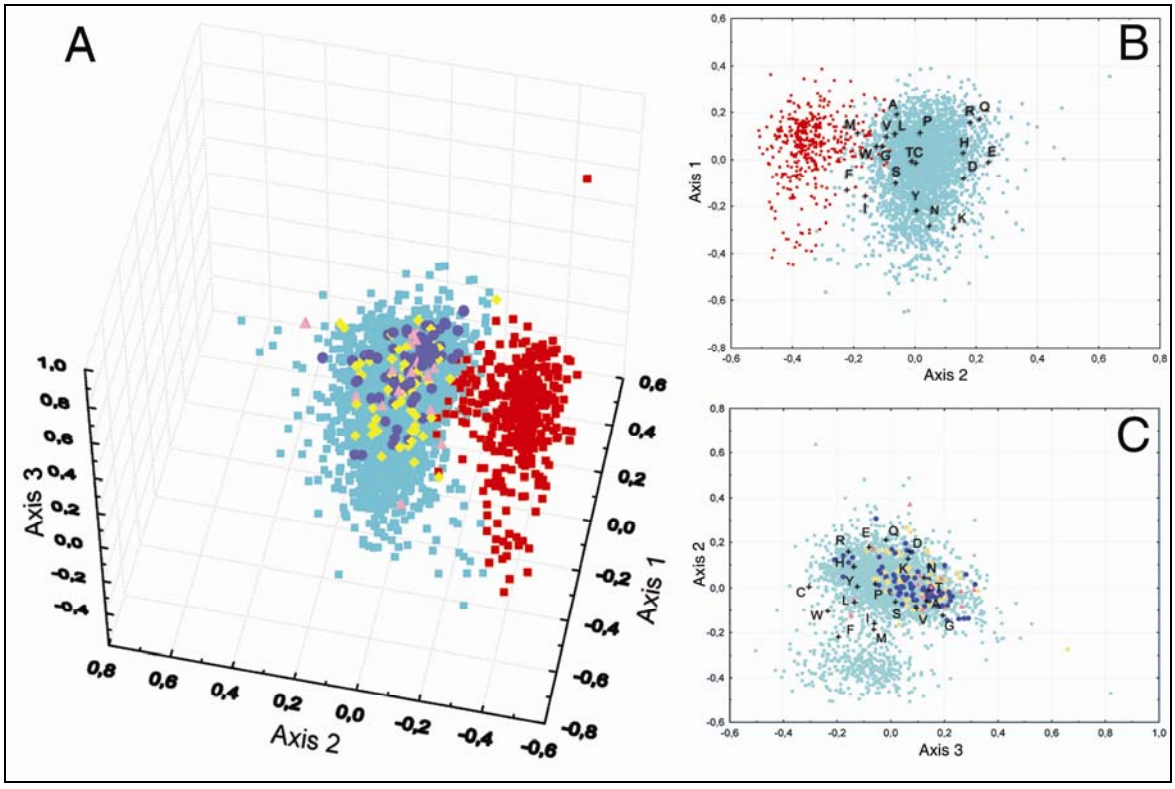


Figure 5:

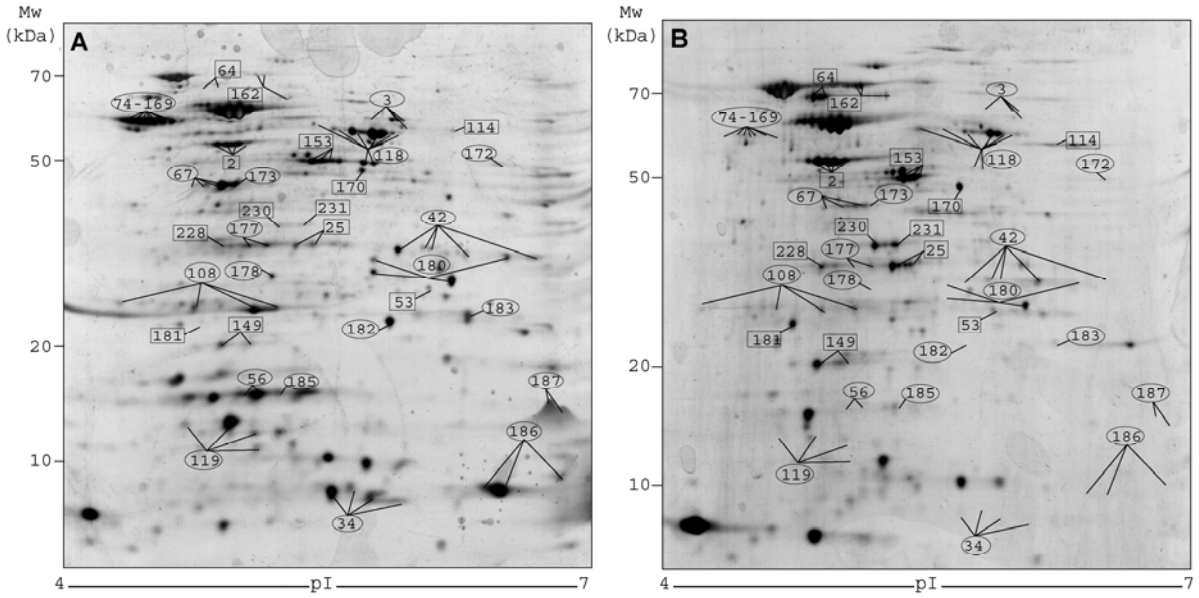
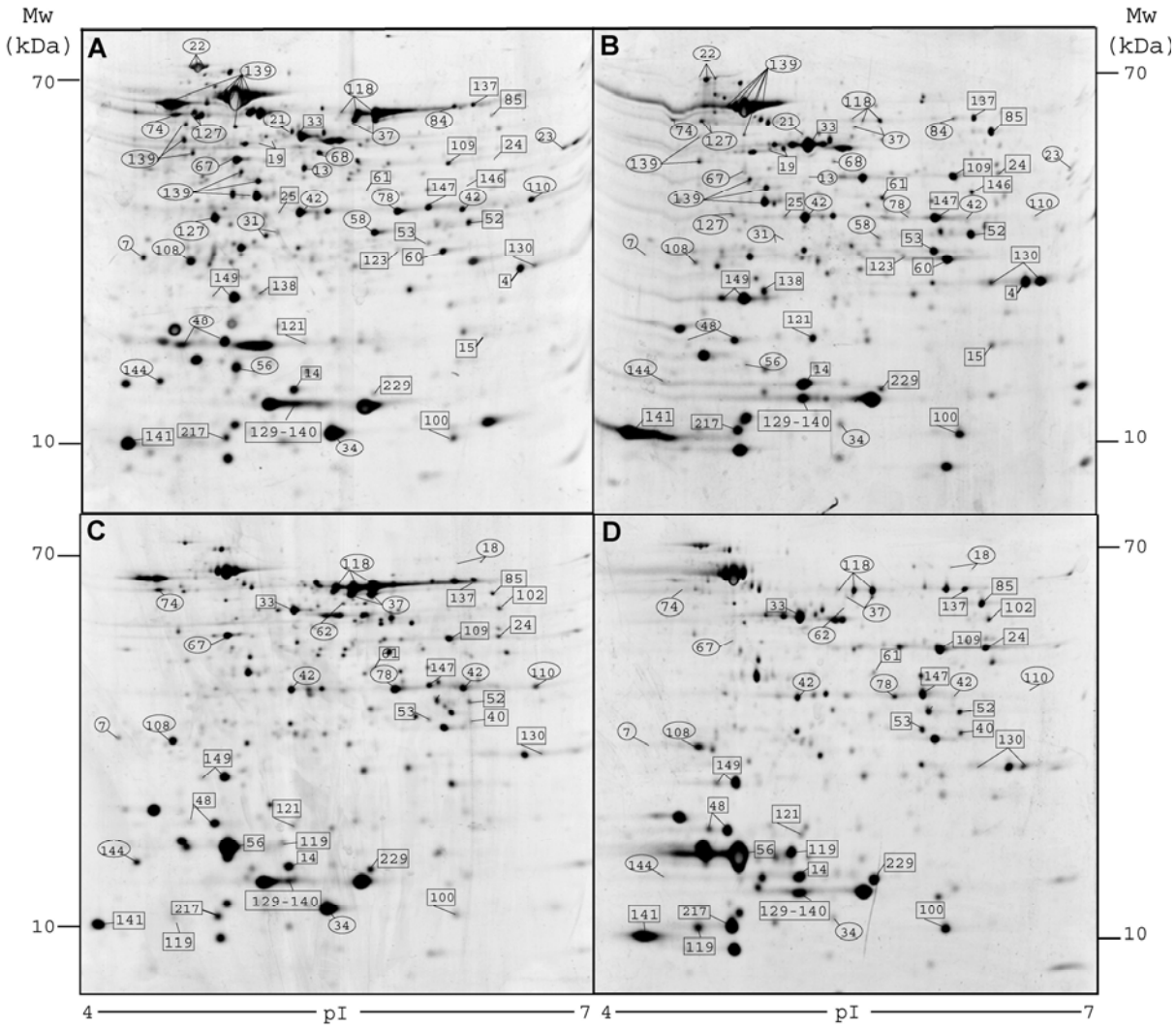


Figure 6:



Table

Table 1: List of the identified proteins of *P. luminescens* TT01 with their gene name, putative function, sequence coverage, theorical pI and Mr values. Localisation of proteins in cellular (1), membrane-associated (2) or extracellular (3) fractions are indicated.

N°	plu	function (CD search)	gene	pI	MW	recovery rate	Gel
		Energy production and conversion					
2	40	ATP synthase beta chain	atpD	4,64	50,037	40-51	1; 2; 3
3	42	ATP synthase alpha chain	atpA	5,41	55,435	40-46	1; 2
114	3621	Dihydrolipoamide dehydrogenase	lpdA	6,07	50,502	23-39	1; 2; 3
115	3622	Dihydrolipoamide acetyltransferase component of pyruvate dehydrogenase complex (E2)	aceF	4,85	56,814	21	1
51	1432	Succinyl-CoA synthetase beta chain	sucC	5,07	41,306	30-33	1; 3
52	1433	Succinyl-CoA synthetase alpha chain	sucD	6,25	29,905	17-30	1; 3
85	2359	fumarate hydratase class II	fumC	6,26	50,118	22-29	1; 3
99	2801	isocitrate dehydrogenase	icd	5,03	45,765	53-55	1; 3
113	3619	aconitate hydratase 2	acnB	4,83	94,283	21	1
147	4547	Malate dehydrogenase	mdh	5,50	32,635	45-46	1; 3
143	4360	quinone oxidoreductase	qor	6,25	35,214	31	1
84	2349	putative aldehyde dehydrogenase (pfam00171: Aldedh, COG1012: PutA)		5,95	51,401	50	1
118	3739	Aldehyde dehydrogenase B	aldB	5,28	54,210	54-62	1; 2; 3
142	4332	Alcohol dehydrogenase class III (COG1062)	adhC	6,23	39,259	28	1
37	984	succinate-semialdehyde dehydrogenase	gabD	5,2	52,370	41	1
59	1546	malate dehydrogenase	maeA	5,68	62,999	18	1
156	4768	glycerol kinase	glpK	6,13	56,030	28	1
149	4551	inorganic pyrophosphatase	ppa	4,68	19,813	22-50	1; 2; 3
9	375	Glutathione oxidoreductase	gor	6,24	49,544	35	1
		Metabolism of carbohydrate and related molecules					
33	913	Enolase	eno	4,88	45,973	41-56	1; 2; 3
35	956	phosphoglycerate kinase	pgk	4,65	41,455	46-61	1; 3
36	957	fructose 1,6-bisphosphate aldolase	fba	5,23	39,069	26	1
50	1407	phosphoglucomutase	pgm	6,6	59,406	29	1
53	1471	phosphoglycerate mutase 1	gpmA	5,69	28,414	44-55	1; 2
80	2118	Pyruvate kinase II	pykA	6,87	51,688	33	1
157	4772	triosephosphate isomerase	tpiA	5,93	26,808	24-32	1; 2; 3
148	4550	Fructose-1,6-bisphosphatase	fbp	6,31	36,713	22	1
112	3606	ribose 5-phosphate isomerase A, pentose phosphate pathway	rpiA	4,67	23,122	11	1
19	521	phosphopentomutase	deoB	4,87	44,465	43	1
		nucleotide and carbohydrate transport and metabolism					
100	2825	similar to HIT family of hydrolase (cd01276: Protein kinase C interacting protein related, COG0537:Hit)		6,23	13,332	52	1; 2
		nucleosides and nucleotides biosynthesis and metabolism					
124	3828	Protein UshA precursor,UDP-sugar-hydrolase 5'-nucleotidase	ushA	6,06	61,242	20	1
17	494	phosphoribosylglycinamide synthetase	purD	4,69	45,510	38	1
18	495	Phosphoribosylaminoimidazolecarboxamide formyltransferase and IMP cyclohydrolase	purH	6,16	57,682	48	1
32	912	CTP synthetase	pyrG	6,14	60,193	27	1
46	1372	nucleoside-diphosphate kinase	ndk	5,24	15,601	70	1
76	2066	Ribose-phosphate pyrophosphokinase	prsA	5,23	34,213	47	1
95	2713	inosine-5'-monophosphate dehydrogenase (pfam00478: IMPDH) biosynth guanosine nucleotide	guaB	5,24	15,601	70	1
98	2759	Uracil phosphoribosyltransferase	upp	5,35	22,503	41	1

N°	plu	function (CD search)	gene	pI	MW	recovery rate	Gel
122	3807	phosphoribosylaminoimidazole carboxylase catalytic subunit	purE	6,63	18,539	24	1
125	3836	adenylate kinase	adk	5,41	24,009	58	1
145	4417	uridine phosphorylase	udp	4,94	26,894	41-46	1; 3
146	4492	aspartate carbamoyltransferase catalytic chain	pyrB	6,18	34,141	19	1
158	4867	deoxyuridine 5'-triphosphate nucleotidohydrolase	dut	4,73	16,301	44	1
227	674	pyrH, UMP Kinase	pyrH	5,67	26,08	23	3
		Metabolism of amino acids and related molecules					
20	523	Unknown, probable cystathionine gamma-lyase		6,32	41,641	34	1
21	565	Threonine synthase	thrC	4,91	47,803	32	1
49	1395	Cysteine synthase A	cysK	6,18	34,130	44-59	1; 3
63	1619	3-phosphoserine aminotransferase	serC	5,47	40,201	38	1
65	1750	Aspartate aminotransferase	aspc	6	43,575	41	1
96	2746	dihydrodipicolinate synthase	dapA	5,68	31,789	25	1
106	3291	Serine hydroxymethyltransferase	glyA	6,32	45,258	19-33	1; 3
110	3556	Putative aminomethyltransferase related to GcvT (COG0354)		6,17	36,359	16	1
111	3598	Aminomethyltransferase	gcvT	5,60	40,083	27	1
116	3673	2-Isopropylmalate synthase	leuA	5,52	57,483	24	1
117	3675	3-isopropylmalate dehydratase large subunit	leuC	6,39	50,477	36	1
154	4742	Arginosuccinate synthase	argG	5.36	44,799	36	1
155	4743	acetylglutamate kinase	argB	4,63	27,358	23	1
23	603	carbamoyl-phosphate synthase	carA	6,56	43,397	15	1
176	3517	probable homocysteine synthase MetY (COG0626: MetC, pfam01053: Cys/Met metabolism)		6,73	45,932	20	2
		Metabolism of lipids					
89	2592	Enoyl-[acyl-carrier-protein] reductase [NADH]	fabI	5,67	28,183	30-62	1; 2
101	2831	probable Beta-ketoacyl-acyl carrier protein (ACP) synthase (KAS), type I and II (cd00834: KAS_I_II)	fabB/fabF	6,18	43,347	22	1
136	4074	biotin carboxyl carrier protein of acetyl-CoA carboxylase	accB	4,48	16,923	21	1; 3
175	4120	glycerophosphoryl diester phosphodiesterase	glpQ	6,69	41,181	43	2
		metabolism of cofactors, prosthetic groups and carriers					
30	872	3-Methyl-2-oxobutanoate hydroxymethyltransferase	panB	5,35	28,840	21	1
57	1539	molybdopterin biosynthesis protein	moeB	6,25	27,342	15	1
77	2069	Glutamyl-tRNA reductase	hemA	6,16	46,992	21	1
107	3337	Pyridoxal phosphate biosynthetic protein PdxJ	pdxJ	6,68	26,656	61	1
129	3898	6,7-dimethyl-8-ribityllumazine synthase	ribH	4,67	16,100	34	1; 3
173	4646	uroporphyrin-III C-methyltransferase	hemX	4,73	42,121	35	2
172	2260	probable Adenosylmethionine-8-amino-7-oxononanoate aminotransferase (pfam00202, COG0161: BioA)		6,23	48,286	34	2; 3
		Transcription					
14	434	transcription antitermination factor	nusG	6,29	20,580	18	1
151	4702	RNA polymerase alpha subunit	rpoA	4,7	36,480	27-41	1; 3
		Translation, ribosomal structure and biogenesis					
161	4525	polyribonucleotide nucleotidyltransferase, member of mRNA degradosome	pnp	4,83	76,896	19-26	2
24	671	methionine aminopeptidase	map	6,17	29,331	22	1
221	4498	unknown, (pfam01042:endoribonuclease L-PSP, COG0251:TdcF)		4,70	13,79	16	3
		translation factor					
12	431	Elongation factor G	fusA	4,78	77,698	25-46	1; 2
13	432	translation elongation factor EF-Tu.B	tufF	4,97	43,14	27-44	1; 2; 3
25	673	elongation factor EF-Ts	tsf	4,92	30,366	24-64	1; 2; 3
26	675	ribosome releasing factor	frr	5,53	20,842	26	1
153	4730	Elongation factor TufA	tufA	4,90	43,163	32-51	1; 2; 3
103	2861	Elongation factor P-like protein		4,97	21,285	42	1

N°	plu	function (CD search)	gene	pI	MW	recovery rate	Gel
138	4130	Elongationfactor P	efp	4,58	20,700	27	1
		ribosomal proteins					
15	437	50S ribosomal subunit protein L10	rlpJ	9,07	18,191	66	1
64	1622	30S ribosomal subunit protein S1	rpsA	4,62	61,216	30-43	1; 2; 3
73	2055	peptidyl-tRNA hydrolase	pth	9,39	21,573	18	1
152	4703	30S ribosomal subunit protein S4	rpsD	10,62	23,574	23	1
150	4570	50S ribosomal subunit protein L9	rplI	6,54	15,882	78	1
223	430	rpsG, 30S ribosomal subunit	rpsG	10,98	17,64	30	3
		aminoacyl-tRNA synthetase					
27	692	Prolyl-tRNA synthetase	proS	5,15	63,698	42	1
62	1604	seryl-tRNA synthetase	serS	5,17	48,412	31	1
93	2665	Phenylalanyl-tRNA synthetase alpha chain	pheS	5,73	37,256	53	1
		Cell wall/membrane biogenesis					
78	2073	2-dehydro-3-deoxyphosphooctonate aldolase (LPS, KDO biosynthesis)	kdsA	6,15	30,674	25	1
199	1752	putative porine OmpN (pfam00267: Porin 1, COG3203: OmpC)		6,81	42,27	18	3
171	3851	acrA, acriflavin resistance protein A precursor	acrA	7,49	42,646	51	2
87	2501	UTP-glucose-1-phosphate uridylyltransferase	galU	5,64	36,503	23	1
60	1561	Some similarity to Dictyostelium discoideum calcium-dependent cell adhesion molecule-1		5,94	23,717	44-45	1; 3
68	1967	Unknown (pfam06316: Ail-Lom-like protein; COG3637) similar to PagC of E. coli		6,25	20,119	39	1
211	2963	similar to putative adhesins of Burkholdera cepacia and E. coli		5,87	21,87	58	3
198	4238	unknown (C-terminal part similar to pfam07472: Fucose-binding lectin II PA-IIL)		6,61	34,80	54	3
218	2096	Similar to galactophilic lectin PA-I of *Pseudomonas aeruginosa*		5,06	12,96	25	3
		Intracellular trafficking and secretion					
170	1455	TolB protein precursor	tolB	6,38	46,384	39	2
		Transport and binding proteins					
16	458	periplasmic maltose-binding protein precursor	malE	7,09	43,472	38	1
42	1174	Iron compound ABC transporter substrate-binding protein (cd01140: Siderophore binding protein FatB)		8,63	34,202	31-39	1; 2; 3
48	1392	PTS system, glucose-specific IIA component	crr	4,61	18,342	23-28	1; 2
86	2493	Periplasmic-binding protein precursor OppA2 (COG4166, OppA)		7,01	62,067	48	1
94	2672	Iron (chelated) ABC transporter, periplasmic-binding protein YfeA	yfeA	7,09	34,336	37-43	1; 2
102	2853	Solute-binding periplasmic protein of iron-siderophore ABC transporter (cd01139:TroA-f)		7	40,676	35-52	1; 2
137	4098	Leucine-specific binding protein precursor	livK	6,79	39,564	35	1
168	2493	Periplasmic-binding protein precursor OppA2	oppA	7,01	62,067	47	2
178	695	metQ, D-methionine-binding periplasmic protein precursor MetQ	metQ	5,47	29,636	49	2
230	2697	PTS system, mannose-specific IIAB component	manX	4,83	35,682	23	2
		inorganic ion transport and metabolism					
4	75	Superoxide dismutase	sodA	6,52	23,523	27-39	1; 2; 3
55	1522	3-mercaptopyruvate sulfurtransferase (COG2897: Rhodanese-related sulfurtransferase, cd01448)	sseA	5,97	30,838	38	1
5	112	cyanate lyase (COG1513: CynS)	cynS	4,9	17,008	31-40	1; 2; 3
		Posttranslational modification, protein turnover, chaperones					
11	422	peptidyl-prolyl cis-trans isomerase	slyD	4,48	20,203	44	1
121	3805	Peptidyl-prolyl cis-trans isomerase B	ppi	4,95	18,272	34	1; 2
10	381	Disulfide interchange protein DsbA precursor	dsbA	8,17	22,969	27	1
22	579	chaperone protein (heat shock protein 70)	dnaK	4,49	68,860	16-54	1; 2; 3
127	3870	Trigger factor (chaperone)	tig	4,46	48,629	26-51	1; 2
139	4134	60kDa Chaperonin	groEL	4,60	57,462	29-57	1; 2; 3

N°	plu	function (CD search)	gene	pI	MW	recovery rate	Gel
140	4135	10 kDa chaperonin	groES	5,16	10,293	37-59	1; 2; 3
108	3372	GrpE protein	grpE	4,48	21,789	24-29	1; 2
162	3837	heat shock protein htpG	htpG	4,75	72,307	35-41	2
160	1270	clpB, heat shock protein F84.1	clpB	6,01	95,662	37-47	2
126	3869	ATP-dependent proteolytic subunit of clpA-clpP serine protease,heat shock protein F21.5	clpP	5,94	23,291	38	1
192	4664	TrxA, thioredoxin 1	trxA	4,71	11,603	34	2
228	3820	putative thioredoxin-like protein (pfam00085, COG3118, pfam06057:VirJ)		4,61	32,040	35	2
7	198	putative thioredoxin-like protein GntY, N-term(pfam01521: HesB) C-term (COG0694:Trx-like)		4,21	20,910	28	1
164	4565	putative carbamoyl transferase of the NodU family (pfam02543: carbamoyltransferase, COG2192)		5,37	63,706	29	2
130	3907	Alkyl hydroperoxide reductase, small subunit (antioxidant)	ahpC	6,40	22,273	48-64	1; 3
88	2579	Thio peroxidase	tpx	4,36	17,719	43-57	1; 2; 3
97	2748	bacterioferritin comigratory protein, hydroperoxide peroxidase (peroxiredoxin)	bcp	4,99	17,627	60-74	1; 3
61	1599	thioredoxin reductase	trxB	5,57	34,244	54	1
		secondary metabolites biosynthesis, transport and catabolism					
6	158	highly similar to L-arginine:lysine amidinotransferase (pfam02274: Amidinotransferase, COG1834)		5,43	42,185	38-63	1; 3
34	947	similar to protein involved in polyketide biosynthesis related to monooxygenase (pfam03992: Antibiotic biosynthesis monooxygenase, COG2329)		5,16	13,806	37-65	1; 2; 3
58	1541	Similar to granaticin polyketide ketoreductase (pfam00106: short chain dehydrogenase, COG1028: FabG)		5,34	26,259	23	1
81	2164	Similar to □-ketoacyl synthase III -like protein (cd00827: "initiating" condensing enzymes, COG0332: FabH)		5,67	42,533	21	1
82	2271	putative epimerase (pfam02567: Phenazine biosynthesis-like proteins PhzC/PhzF, COG0384),		5,14	31,424	39-46	1; 3
131	4007	enhancing lycopene biosynthesis protein 2	elbB	4,96	23,604	42	1
134	4060	PmbA protein, involved in the maturation of antibiotic	pmbA	5,29	47,995	41	1
195	3567	putative oxidase highly similar to N-formimidoyl fortimicin A synthase (pfam01266: DAO, COG0665: DadA)		4,75	52,35	23	3
167	246	probable 4-hydroxyphenylacetic acid hydroxylase (pfam03241: HpaB, COG2368)		6,04	58,452	21	2
196	4258	similar to 4-hydroxyphenylacetate 3-hydroxylase family (pfam03241: HpaB, COG2368) put antibio???		6,11	54,47	27	3
38	986	5-carboxymethyl-2-hydroxy-muconic acid isomerase (pfam02962: CHMI, COG3232: HpaF)	hpcD	6,79	14,732	54	1
39	988	5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase (COG1012: PutA)	HpcC	6,26	53,308	54	1
40	990	Unknown, probable 4-hydroxyphenylacetate catabolism (COG0179: MhpD)	hpaG1	6,01	22,704	22	1
8	373	Highly similar to Hcp protein (Hemolysin co-regulated protein) of Yersinia Pestis		5,91	19,017	22-28	1; 2; 3
201	734	similar to putative sialidase (neuraminidase) of Clostridium tetani and Streptomyces coelicolor		5,27	40,12	25	3
202	735	similar to putative sialidase (neuraminidase) of Clostridium tetani and Streptomyces coelicolor		5,72	40,48	22	3
188	4093	Component of toxin loci Pir		5,12	14,849	37-69	2; 3
186	2534	unknown, some similarity with Pir toxin component Plu4093		6,78	13,053	30	2; 3
56	1537	Unknown, weak similarity with 13.6 kDa insecticidal crystal proteins Cry34 of B. thuringiensis		4,76	14,894	34-41	1; 2; 3
165	840	putative toxin, highly similar to heat-stable cytotonic enterotoxin Ast of Aeromonas hydrophila		6,25	72,609	20	2
31	887	similar to endonuclease domain of colicins, klebicins or pyocins		10,78	16,979	31	1
47	1382	Similar to extracellular metalloproteinase precursor		5,11	41,447	33-45	1; 2; 3
135	4064	predicted Zn-dependent proteases and their inactivated homologs (COG0312: TldD, pfam01523)	tldD	5,10	51,406	42	1
29	831	Beta-lactamase class C	ampC	6,43	42,520	21	1

N°	plu	function (CD search)	gene	pI	MW	recovery rate	Gel
179	2238	Highly similar to AHL-lactonase AttM/AiiB of Agrobacterium tumefaciens (pfam00753: Lactamase B)		6,07	28,919	28	2
197	2455	similar to calpain cysteine protease (cd00044: CysPc): apoptosis, signal transduction, etc…		5,01	43,19	45	3
222	1576	cipA, cristalline inclusion protein CipA	cipA	6,49	11,72	37	3
187	1575	unknown, similar to crystalline inclusion protein type II	cipB	7,33	12,379	26	2
54	1518	similar to lipase (pfam01764: Lipase 3)		6,18	41,580	16-17	1; 2
		Information and regulatory pathways					
45	1253	Autoinducer-2 (AI-2) production protein LuxS	luxS	5,77	19,180	35	1
69	2016	Weakly similar to transcriptional regulator, LuxR family		4,9	26,855	55	1
104	3147	AI-2 Processing aldolase	lsrF	6,14	31,823	40	1
105	3219	DNA-binding HTH domain-containing protein, putative transcriptional regulator of the LuxR family		4,37	26,986	24	1
185	2885	unknown, Weakly similar to putative regulatory protein of salmonella enterica		5,90	10,390	46	2
216	171	Uxu operon transcriptional regulator	uxuR	6,11	28,56	19	3
229	2498	DNA binding protein H-NS	hns	5,13	15,32	53	1
231	2683	ProP effector	proQ	10,33	26,739	36	2
		adaptations and atypical conditions					
70	2030	similar to universal stress protein (pfam00582: Usp)		6,68	15,917	62	1
71	2032	similar to universal stress protein (pfam00582: Usp)		6,68	15,917	62	1
132	4012	Stringent starvation protein B	sspB	4,73	19,127	32	1
133	4013	Stringent starvation protein A	sspA	4,98	24,435	56	1
189	1289	cspE, cold shock-like protein	cspE	7,54	7,560	71	2; 3
190	2783	cspC, cold shock-like protein	cspC	5,76	7,35	62-72	2; 3
217	121	universal stress protein A	uspA	5,41	15,77	37	1; 3
159	4871	similar to stress-induced protein of Yersinia pestis and E. coli (pfam03755: YicC-like family)		4,78	33,571	32	1
83	2338	OsmC-like protein (pfam02566:OsmC; COG1765: predicted redox protein)		5,59	16,128	PSD	1
		phage-related proteins					
1	15	putative Phage tail sheath protein (pfam04984: phage sheath 1, COG3497)		5,51	42,703	22-28	1; 3
144	4369	similar to tail fiber assembly protein (pfam02413: caudoTAP)		4,31	16,946	31	1
194	2023	similar to tail fiber assembly protein (pfam02413: caudoTAP)		4,82	55,37	41	3
200	1666	putative Phage tail sheath protein (pfam04984: phage sheath 1, COG3497)		5,44	39,52	33	3
203	23	similar to phage-related baseplate assembly protein (pfam04865:Baseplate J, COG3948)		4,67	37,24	46	3
204	2036	unknown, putative tail fiber protein		4,90	47,45	47	3
205	2022	similar to tail fiber assembly protein (pfam02413: caudoTAP)		6,69	33,30	55	3
206	2303	unknown, putative tail fiber protein		6,30	31,92	47	3
207	2960	similar to tail fiber assembly protein (pfam02413: caudoTAP)		6,81	27,31	57	3
208	1464	unknown, putative tail fiber protein		5,32	29,50	31	3
210	2024	similar to tail fiber assembly protein (pfam02413: caudoTAP)		6,52	27,15	46	3
212	2959	unknown, similar to bacteriophage protein		7,04	21,16	45	3
213	14	putative Phage tail tube protein FII (pfam04985: Phage tube)		5,06	19,11	41	3
219	2958	similar to tail fiber assembly protein (pfam02413: caudoTAP)		4,16	16,12	61	3
209	2034	unknown, putative tail fiber protein		6,80	26,21	41	3
		unknown					
28	804	Unknown		5,01	85,323	26	1
41	1025	Unknown		4,65	35,628	15	1
43	1185	unknown (pfam04452, DUF558)		7,16	27,052	18	1
44	1232	Unknown		5,28	17,815	55	1
66	1795	Unknown		6,54	37,361	31-48	1; 2; 3
67	1840	Unknown		4,67	39,976	25-69	1; 2; 3

N°	plu	function (CD search)	gene	pI	MW	recovery rate	Gel
72	2046	Unknown		5,43	31,461	28	1
74	2059	Unknown		4,30	54,790	20-40	1; 2; 3
75	2064	Unknown		4,51	35,073	24	1
79	2109	unknown (pfam01709, DUF28)		4,36	26,315	24	1
92	2639	Unknown		6,24	14,816	17-47	1; 2; 3
109	3393	Unknown		5,3	15,363	27	1
119	3795	Unknown		4,72	16,515	48-62	1; 2; 3
123	3826	Unknown (pfam07446: GumN, COG3735)		6,33	30,997	45	1
128	3881	Unknown (pfam04461, DUF520)		5,84	18,322	35-42	1; 2
141	4286	Unknown (pfam05899, DUF861)		4,66	13,251	39	1; 3
163	2063	unknown		4,51	47,932	18	2
166	3249	unknown		9,52	8,934	32	2
169	2060	unknown		4,46	56,315	17-21	2; 3
177	1574	Unknown		4,73	31,787	28	2
180	3611	unknown (pfam04402; DUF541)		6,29	25,877	41-47	2; 3
181	2143	unknown		4,89	23,411	19-24	2
182	2972	unknown		5,23	24,303	28-29	2; 3
183	1461	Unknown		7,21	13,322	42	2
184	2444	unknown		6,63	23,222	19-42	2
191	4334	unknown		4,67	11,745	18	2
193	1665	unknown		4,34	52,02	31	3
214	1009	unknown		6,66	16,78	PSD	3
215	480	unknown (pfam02018: Carbohydrate binding domain)		4,20	17,38	58	3
220	2256	unknown		8,63	10,49	37	3
224	4200	unknown (COG3521)		10,11	19,35	22	3
225	3063	unknown		4,17	20,59	57	3
226	1424	unknown (pfam01784:NIF3, COG0327)		6,12	27,28	19	3

Table 2: List of the identified extracellular proteins of *P. luminescens* TT01 in stationary phase.

N°	PLU	function	ene
2	40	ATP synthase beta chain	*atpD*
114	3621	Dihydrolipoamide dehydrogenase	*lpdA*
51	1432	Succinyl-CoA synthetase beta chain	*sucC*
52	1433	Succinyl-CoA synthetase alpha chain	*sucD*
85	2359	fumarate hydratase class II	*fumC*
99	2801	isocitrate dehydrogenase	*icd*
147	4547	Malate dehydrogenase	*mdh*
118	3739	Aldehyde dehydrogenase B	*aldB*
149	4551	inorganic pyrophosphatase	*ppa*
33	913	Enolase	*eno*
35	956	phosphoglycerate kinase	*pgk*
157	4772	triosephosphate isomerase	*tpiA*
145	4417	uridine phosphorylase	*udp*
227	674	pyrH, UMP Kinase	*pyrH*
49	1395	Cysteine synthase A	*cysK*
106	3291	Serine hydroxymethyltransferase	*glyA*
129	3898	6,7-dimethyl-8-ribityllumazine synthase	*ribH*
172	2260	probable Adenosylmethionine-8-amino-7-oxononanoate aminotransferase (pfam00202, COG0161: BioA)	
151	4702	RNA polymerase alpha subunit	*rpoA*
221	4498	unknown, (pfam01042:endoribonuclease L-PSP, COG0251:TdcF)	
13	432	translation elongation factor EF-Tu.B	*tufF*
25	673	elongation factor EF-Ts	*tsf*
153	4730	Elongation factor TufA	*tufA*
64	1622	30S ribosomal subunit protein S1	*rpsA*

N°	PLU	function	ene
223	430	rpsG, 30S ribosomal subunit	*rpsG*
199	1752	putative porine OmpN (pfam00267: Porin 1, COG3203: OmpC)	
60	1561	Some similarity to *Dictyostelium discoideum* calcium-dependent cell adhesion molecule-1	
211	2963	similar to putative adhesins of *Burkholdera cepacia* and *E. coli*	
198	4238	unknown (C-terminal part similar to pfam07472: Fucose-binding lectin II PA-IIL)	
218	2096	Similar to galactophilic lectin PA-I of *Pseudomonas aeruginosa*	
42	1174	Iron compound ABC transporter substrate-binding protein (cd01140: Siderophore binding protein FatB)	
4	75	Superoxide dismutase	*sodA*
5	112	cyanate lyase (COG1513: CynS)	*cynS*
22	579	chaperone protein (heat shock protein 70)	*dnaK*
139	4134	60kDa Chaperonin	*groEL*
140	4135	10 kDa chaperonin	*groES*
130	3907	Alkyl hydroperoxide reductase, small subunit (antioxidant)	*ahpC*
88	2579	Thio peroxidase	*tpx*
97	2748	bacterioferritin comigratory protein, hydroperoxide peroxidase (peroxiredoxin)	*bcp*
6	158	highly similar to L-arginine:lysine amidinotransferase (pfam02274: Amidinotransferase, COG1834)	
34	947	similar to protein involved in polyketide biosynthesis related to monooxygenase	
82	2271	putative epimerase (pfam02567: Phenazine biosynthesis-like proteins PhzC/PhzF, COG0384),	
195	3567	putative oxidase highly similar to N-formimidoyl fortimicin A synthase (pfam01266: DAO, COG0665: DadA)	
196	4258	similar to 4-hydroxyphenylacetate 3-hydroxylase family (pfam03241: HpaB, COG2368) put antibio???	
8	373	Highly similar to Hcp protein (Hemolysin co-regulated protein) of *Yersinia Pestis*	
201	734	similar to putative sialidase (neuraminidase) of *Clostridium tetani* and *Streptomyces coelicolor*	
202	735	similar to putative sialidase (neuraminidase) of *Clostridium tetani* and *Streptomyces coelicolor*	
188	4093	Component of toxin loci Pir	
186	2534	unknown, some similarity with Pir toxin component Plu4093	
56	1537	Unknown, weak similarity with 13.6 kDa insecticidal crystal proteins Cry34 of *B. thuringiensis*	
47	1382	Similar to extracellular metalloproteinase precursor	
197	2455	similar to calpain cysteine protease (cd00044: CysPc): apoptosis, signal transduction, etc…	
222	1576	cipA, cristalline inclusion protein CipA	*cipA*
216	171	Uxu operon transcriptional regulator	*uxuR*
189	1289	cspE, cold shock-like protein	*cspE*
190	2783	cspC, cold shock-like protein	*cspC*
217	121	universal stress protein A	*uspA*
1	15	putative Phage tail sheath protein (pfam04984: phage sheath 1, COG3497)	
194	2023	similar to tail fiber assembly protein (pfam02413: caudoTAP)	
200	1666	putative Phage tail sheath protein (pfam04984: phage sheath 1, COG3497)	
203	23	similar to phage-related baseplate assembly protein (pfam04865:Baseplate J, COG3948)	
204	2036	unknown, putative tail fiber protein	
205	2022	similar to tail fiber assembly protein (pfam02413: caudoTAP)	
206	2303	unknown, putative tail fiber protein	
207	2960	similar to tail fiber assembly protein (pfam02413: caudoTAP)	
208	1464	unknown, putative tail fiber protein	
210	2024	similar to tail fiber assembly protein (pfam02413: caudoTAP)	
212	2959	unknown, similar to bacteriophage protein	
213	14	putative Phage tail tube protein FII (pfam04985: Phage tube)	
219	2958	similar to tail fiber assembly protein (pfam02413: caudoTAP)	
209	2034	unknown, putative tail fiber protein	
66	1795	Unknown	
67	1840	Unknown	
74	2059	Unknown	
119	3795	Unknown	
141	4286	Unknown (pfam05899, DUF861)	
180	3611	unknown (pfam04402; DUF541)	
182	2972	unknown	
193	1665	unknown	
214	1009	unknown	

N°	PLU	function	ene
215	480	unknown (pfam02018: Carbohydrate binding domain)	
220	2256	unknown	
224	4200	unknown (COG3521)	
225	3063	unknown	
226	1424	unknown (pfam01784:NIF3, COG0327)	

Table 3: Membrane Proteins with altered level of synthesis in the TT01 variant II in stationary phase.

N°	PLU	Function	Gene	Induction ratio pI/pII
230	2697	PTS system, mannose-specific IIAB component	manX	<0,04
231	2683	ProP effector	proQ	<0,04
64	1622	30S ribosomal subunit protein S1	rpsA	<0,04
114	3621	dihydrolipoamide dehydrogenase	lpdA	0,13
162	3837	heat shock protein htpG	htpG	0,19
53	1471	phosphoglycerate mutase 1	gpmA	0,19
153	4730	Elongation factor TufA	tufA	0,22
228	3820	putative thioredoxin-like protein (pfam00085, COG3118, pfam06057:VirJ)		0,23
2	40	atp synthase beta chain	atpD	0,25
149	4551	inorganic pyrophosphatase	ppa	0,25
181	2143	unknown		0,34
170	1455	TolB protein precursor	tolB	0,36
25	673	elongation factor EF-Ts	tsf	0,36
182	2972	unknown		>25
119	3795	unknown		>25
178	695	D-methionine-binding periplasmic protein precursor MetQ	metQ	>25
185	2885	unknown, Weakly similar to putative regulatory protein of *salmonella enterica*		>25
56	1537	Unknown, weak similarity with 13.6 kDa insecticidal crystal proteins Cry34 of *B. thuringiensis*		>25
186	2534	unknown, some similarity with Pir toxin component Plu4093		>25
187	1575	unknown, similar to crystalline inclusion protein type II		>25
74	2059	unknown		23,7
169	2060	unknown		23,7
183	1461	unknown		23,09
34	947	similar to protein involved in polyketide biosynthesis related to monooxygenase		15,88
67	1840	unknown		13,55
42	1174	Iron compound ABC transporter substrate-binding protein (cd01140: Siderophore binding protein FatB)		7,3
173	4646	uroporphyrin-III C-methyltransferase	hemX	5,72
118	3739	aldehyde dehydrogenase B	aldB	3,5
108	3372	GrpE protein (HSP-70 cofactor)	grpE	3,33
177	1574	Unknown		2,77
180	3611	unknown (pfam04402; DUF541)		2,63
172	2260	probable Adenosylmethionine-8-amino-7-oxononanoate aminotransferase (pfam00202, COG0161: BioA)		2,08

Table 4: Cytoplasmic Proteins with altered level of synthesis in the TT01 variant II in stationary phase.

N°	PLU	Function	Gene	Induction ratio pI/pII (24h)	(96h)
129	3898	6,7-dimethyl-8-ribityllumazine synthase	*ribH*	<0,04	<0,04
85	2359	Fumarate hydratase class II	*fumC*	0.057	0.177
53	1471	phosphoglycerate mutase 1	*gpmA*	0,1	0,26
130	3907	Alkyl hydroperoxide reductase, small subunit (antioxidant)	*ahpC*	0,1	0,434
19	521	phosphopentomutase	*deoB*	0,11	
146	4492	aspartate carbamoyltransferase catalytic chain	*pyrB*	0,13	
25	673	elongation factor EF-Ts	*tsf*	0,17	
109	3393	unknown		0.17	0.30
121	3805	Peptidyl-prolyl cis-trans isomerase B	*ppiB*	0,2	0,49
14	434	Transcription antitermination factor	*nusG*	0,20	0,49
149	4551	inorganic pyrophosphatase	*ppa*	0,205	0,368
52	1433	Succinyl-CoA synthetase alpha chain	*sucD*	0,246	0,45
60	1561	Some similarity to *Dictyostelium discoideum* calcium-dependent cell adhesion molecule-1		0,25	
138	4130	Elongationfactor P	*efp*	0,3	
15	437	50S ribosomal subunit protein L10	*rlpJ*	0,34	
61	1599	thioredoxin reductase	*trxB*	0,34	0,29
4	75	Superoxide dismutase	*sodA*	0,35	
229	2498	DNA binding protein H-NS	*hns*	0.35	0,20
141	4286	Unknown (pfam05899, DUF861)		0.36	0,119
24	671	methionine aminopeptidase	*map*	0,4	0,297
147	4547	Malate dehydrogenase	*mdh*	0.42	0.45
33	913	Enolase	*eno*	0,48	<0,04
100	2825	similar to HIT family of hydrolase		0,48	0,182
123	3826	Unknown (pfam07446: GumN, COG3735)		0,49	
217	121	universal stress protein A	*uspA*	0,49	0,135
137	4098	Leucine-specific binding protein precursor	*livK*	0,34	0,86
119	3795	Unknown			0,11
102	2853	Solute-binding periplasmic protein of iron-siderophore ABC transporter (cd01139:TroA-f)			0,22
40	990	Unknown, probable 4-hydroxyphenylacetate catabolism (COG0179: MhpD)	*hpaG1*		0,318
37	984	succinate-semialdehyde dehydrogenase	*gabD*	>25	5
42	1174	Iron compound ABC transporter substrate-binding protein		>25	6,33
21	565	Threonine synthase	*thrC*	>25	
31	887	similar to endonuclease domain of colicins, klebicins or pyocins		>25	
78	2073	2-dehydro-3-deoxyphosphooctonate aldolase (KDO biosynthesis)	*kdsA*	>25	3,25
34	947	similar to protein involved in polyketide biosynthesis related to monooxygenase		17,8	24
144	4369	Similar to tail fiber assembly protein (pfam02413)		>25	3,8
48	1392	PTS system, glucose-specific IIA component	*crr*	15,82	0,3
118	3739	Aldehyde dehydrogenase B	*aldB*	13,22	5,13
67	1840	Unknown		10,31	>25
74	2059	Unknown		10,16	11,7
7	198	putative thioredoxin-like protein GntY, N-term(pfam01521: HesB) C-term (COG0694:Trxlike)		7,56	>25
23	603	carbamoyl-phosphate synthase	*carA*	5,33	
58	1541	Similar to granaticin polyketide ketoreductase (pfam00106)		4,16	
110	3556	Putative aminomethyltransferase related to GcvT		5,33	>25
108	3372	GrpE protein	*grpE*	4,86	2,85
56	1537	Unknown, weak similarity with 13.6 kDa insecticidal crystal proteins Cry34 of *B. thuringiensis*		4,16	0,483
139	4134	60kDa Chaperonin	*groEL*	4,34	
22	579	chaperone protein (heat shock protein 70)	*dnaK*	4,24	
68	1967	Unknown (pfam06316: Ail-Lom-like protein; COG3637) similar to PagC of E.coli		3,88	
127	3870	Trigger factor (chaperone)	*tig*	2,9	
13	432	translation elongation factor EF-Tu.B	*tufB*	2,34	
84	2349	putative aldehyde dehydrogenase (pfam00171: Aldedh, COG1012: PutA)		2,07	
62	1604	seryl-tRNA synthetase	*serS*		>25
18	495	Phosphoribosylaminoimidazolecarboxamide formyltransferase and IMP cyclohydrolase	*purH*		2,2

F Etudes préliminaires d'un eucaryote

F.I Descriptif de *Penicillium marneffei* un champignon dimorphique.

Penicillium marneffei est un champignon dimorphique, principal vecteur des mycoses systémiques respiratoires et de la peau en Asie du sud-est. Il pousse à 25°C sous la forme d'un champignon mycélien. Il forme des colonies duveteuses, de couleur blanche à jaune vert, devenant rouge en vieillissant. On observe également un pigment rouge vif caractéristique diffusant dans le milieu de culture (Agar Sabouraud) (Figure 17). Ces cultures sont dangereuses à manipuler. *P. marneffei* est la seule espèce de *Penicillium* qui possède un dimorphisme thermique. En effet, à température ambiante (env. 25°C), il se trouve sous la forme d'un champignon alors qu'il prend une forme de levure à 37°C, température corporelle. Les caractères microscopiques sont la tête sporulante typique du genre *Penicillium* avec des conidies attachées en chaînes sur un conidiophore[1] ramifié (*Penicillium* du latin *penicillus* signifiant pinceau).

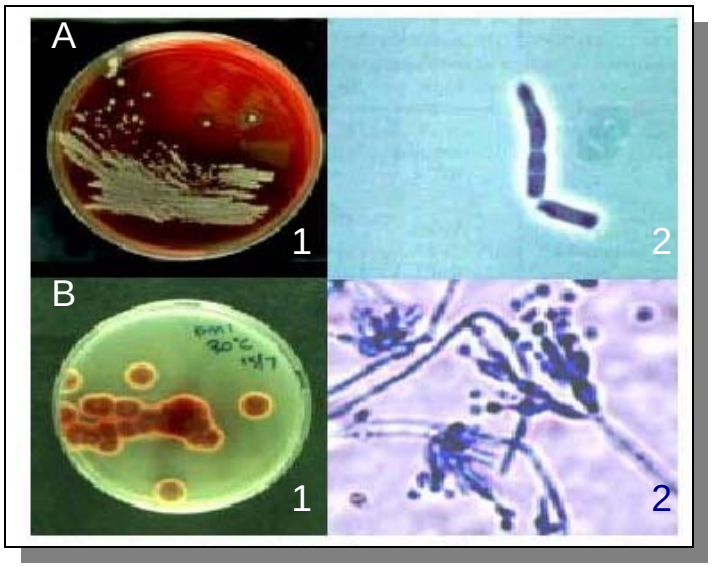


Figure 17 : Culture de P. marneffei à A - 37°C, B – 25°C (1 – sur boite, 2 – grossissement d'isolat)

Malgré sa découverte en 1956 dans l'abcès hépatique de *Rhizomus sinenesis*, le rat chinois du bambou, seulement 18 cas de maladies humaines ont été déclarés jusqu'en 1985 (chez des patients VIH négatif) (Deng et Connor 1985). Puis une explosion de maladies apparaît chez les patients VIH positifs ayant résidé ou voyagé en Asie du sud-est. Environ 10 % des patients atteints du SIDA à Hong Kong sont maintenant infectés par ce champignon dimorphique (Wong et Lee 1998), qui apparaît comme la troisième maladie indicatrice la plus commune du SIDA à Hong Kong, après la tuberculose extra pulmonaire et la cryptococcose méningée (Marty, Brun et al. 2000). L'infection apparaît, dans la plupart des cas, après l'inhalation de spores de *P. marneffei*. Par ailleurs, quelques cas ont été décrits chez des occidentaux séropositifs pour le VIH ayant séjourné en zone d'endémie,

[1] Les conidiophores supportent et produisent les conidies (terme général désignant une spore issue de la multiplication végétative d'un champignon).

dont une douzaine en France. Il existe également de rares cas rapportés chez des sujets immunocompétents. Aujourd'hui en Chine, où plus d'un million de personnes sont porteuses du VIH, on s'attend à une augmentation dramatique des cas d'infection par *P. marneffei*.

Malgré l'importance médicale et la propriété dimorphe peu commune de ce champignon, il n'a suscité que de rares études au niveau moléculaire. Seulement un gène de mannoprotéine de la paroi cellulaire a été caractérisé et employé avec succès dans le diagnostic sérologique de cette infection (Cao, Chan et al. 1998; Cao, Chen et al. 1998). Les études fondées sur les ARNr nucléaires et les ARNr mitochondriaux ont permis aux investigateurs de suggérer un rapport phylogénétique fort avec *Talaromyces sp*. (LoBuglio et Taylor 1995). L'infectiosité des conidies de la phase champignon ainsi que certaines étapes sexuelles de son développement sont encore énigmatiques. Aussi, peu de choses sont connues concernant la biologie cellulaire et moléculaire de base des aspects métaboliques ou structurels, des facteurs de virulence, de la structure du génome et de la biologie liée au développement de cet organisme.

Nous décrivons dans l'article présenté ci-après, les résultats d'une première exploration du génome de *Penicillium marneffei*. Ces derniers ont facilité la recherche moléculaire et ont posé les fondations pour un projet de séquençage du génome complet de ce champignon unique. Une connaissance complète du génome devrait permettre aux chercheurs de mieux comprendre les mécanismes de base du dimorphisme, de la pathogénicité du champignon, de sa virulence et de la défense immune humaine, ouvrant ainsi la voie pour traiter ou éradiquer cette maladie.

La localisation géographique proéminente de ce champignon a donc amené le Pr. K. Y. Yuen et son équipe à s'intéresser plus particulièrement à l'étude de cet organisme en collaboration avec l'Institut Pasteur de Hong Kong dirigé à l'époque par Antoine Danchin et avec l'AGC dirigé par Claudine Médigue à Evry. C'est ainsi que mes premiers travaux de recherche ont été consacrés à l'étude du dimorphisme génétique de *P. marneffei* aboutissant par la suite, à mon départ pour Hong Kong pour une période de 18 mois. Ce fut d'ailleurs l'amorce des travaux plus poussés sur les biais compositionnels en acides aminés des protéines procaryotes présentés précédemment.

F.II *Des procaryotes aux eucaryotes.*

Le choix de ne pas traiter immédiatement des eucaryotes, même unicellulaires, s'est imposé en grande partie e raison de la grande taille de leur génome et de leur compartimentation cellulaire. En effet, la compartimentation cellulaire chez les eucaryotes est indispensable au fonctionnement correct de leur métabolisme complexe. Dans une cellule eucaryote, plus de 10000 espèces de protéines sont présentes, la plupart sont synthétisées dans le cytosol, mais leur site d'action peut être la membrane plasmique, le noyau ou encore l'un des organites cytoplasmiques. Le "tri" ou l'adressage des protéines vers leur compartiment d'activité se fait selon des processus très précis, qui diffèrent selon les compartiments cellulaires et la fonction de la protéine. Ainsi, par exemple, de nombreux processus biochimiques s'effectuent sur des surfaces membranaires (métabolisme des lipides, phosphorylation

oxydative dans les mitochondries ou photosynthèse dans les chloroplastes). Dans le cas des organelles clos, un système d'adressage spécifique des protéines qui devront traverser les membranes est nécessaire. Chez les eucaryotes, on distingue trois types de compartiments fonctionnels : (i) le noyau et le cytosol, qui communiquent par les pores nucléaires ou même qui ne forment qu'un seul compartiment physique lors de la division chez les eucaryotes supérieurs ; (ii) les organelles impliqués dans les mécanismes endocellulaires et sécrétoires : le réticulum endoplasmique, l'appareil de Golgi, les endosomes et les lysosomes ; (iii) les mitochondries et chloroplastes, d'origine endosymbiotique. Pour une première étape, la tâche semblait trop lourde (beaucoup de protéines devant passer d'un compartiment à un autre sont donc fortement marquées au niveau de leur composition en acides aminés ; trop grande quantité de protéines) et nous nous sommes naturellement tournés vers l'analyse des Eu- et Archaeabactéries. Cependant, maintenant que les fondations de ce type d'analyse ont été établies, il serait très intéressant dans un futur proche de s'intéresser aux eucaryotes unicellulaires.

F.III *L'essentiel de l'article.*

Le génome de *P. marneffei* se condense en au moins trois chromosomes de 2,2, 4 et 5 Mb. La taille globale du génome étant comprise entre 17,8 et 26,2 Mb, il doit exister plusieurs chromosomes de tailles identiques ou très proches non séparables sur simples gels d'électrophorèse. Le contenu en G+C du génome est de 48,8 %. L'exploration aléatoire du génome de *P. marneffei* donna 2303 étiquettes de séquences (dites Random Sequence Tags - RSTs), correspondant à 9% du génome, avec 11,7, 6,3 et 17,4% de RSTs ayant des similarités de séquences (par BlastX) avec respectivement des séquences spécifiques (i) de la levure, (ii) du champignon excluant le phylum levure et (iii) des deux c'est-à-dire des champignons dont la levure. Bien que le dimorphisme n'ait pas été décelé à travers l'analyse de l'usage des codons du génome de *P. marneffei*, un grand nombre de fonctions essentielles ont pu être trouvées. L'analyse des RSTs a révélé des gènes codant le transfert d'information (gènes de protéines ribosomales, sous unités d'ARNt synthétase, initiateurs de la traduction et facteurs d'élongation), le métabolisme et la compartimentation, incluant plusieurs gènes de protéines de résistance multi-drogue et des homologues aux gènes de résistance au fluconazole, un antifongique. En outre, plusieurs indices tels que la présence de gènes codant des homologues de phéromones d'algues et d'autres champignons suggèrent fortement la présence d'une phase sexuelle existant vraisemblablement dans la nature mais non observée en laboratoire.

F.IV *Article 5.*

Article paru dans le journal « Archives of microbiology » en mai 2003.

Arch Microbiol (2003) 179 : 339–353
DOI 10.1007/s00203-003-0533-8

ORIGINAL PAPER

**Kwok-yung Yuen Géraldine Pascal Samson S. Y. Wong
Philippe Glaser Patrick C. Y. Woo Frank Kunst
James J. Cai Elim Y. L. Cheung Claudine Médigue
Antoine Danchin**

Exploring the *Penicillium marneffei* genome

Received: 2 September 2002 / Revised: 17 February 2003 / Accepted: 17 February 2003 / Published online: 15 March 2003

Abstract *Penicillium marneffei* is a dimorphic fungus that intracellularly infects the reticuloendothelial system of humans and bamboo rats. Endemic in Southeast Asia, it infects 10% of AIDS patients in this region. The absence of a sexual stage and the highly infectious nature of the mould-phase conidia have impaired studies on thermal dimorphic switching and host-microbe interactions. Genomic analysis, therefore, could provide crucial information. Pulsed-field gel electrophoresis of genomic DNA of *P. marneffei* revealed three or more chromosomes (5.0, 4.0, and 2.2 Mb). Telomeric fingerprinting revealed 6–12 bands, suggesting that there were chromosomes of similar sizes. The genome size of *P. marneffei* was hence about 17.8–26.2 Mb. G+C content of the genome is 48.8 mol%. Random exploration of the genome of *P. marneffei* yielded 2303 random sequence tags (RSTs), corresponding to 9% of the genome, with 11.7, 6.3, and 17.4% of the RSTs having sequence similarity to yeast-specific sequences, non-yeast fungus sequences, and both (common sequences), respectively. Analysis of the RSTs revealed genes for information transfer (ribosomal protein genes, tRNA synthetase subunits, translation initiation, and elongation factors), metabolism, and compartmentalization, including several multi-drug-resistance protein genes and homologues of fluconazole-resistance gene. Furthermore, the presence of genes encoding pheromone homologues and ankyrin repeat-containing proteins of other fungi and algae strongly suggests the presence of a sexual stage that presumably exists in the environment.

Keywords *Penicillium marneffei* Genome Genomic analysis

K. Yuen (✉) A. Danchin
HKU-Pasteur Research Centre, 8 Sassoon Road, Hong Kong
Tel.: +852-2816-8403, Fax: +852-2872-2782,
e-mail: kyyuen@hkucc.hku.hk

K. Yuen S. S. Y. Wong P. C. Y. Woo J. J. Cai
E. Y. L. Cheung
Department of Microbiology, The University of Hong Kong,
University Pathology Building, Queen Mary Hospital,
Hong Kong

G. Pascal A. Danchin
Unité Génétique des Génomes Bactériens, Institut Pasteur,
28 rue du Docteur Roux, 75724 Paris Cedex 15, France

P. Glaser F. Kunst
Laboratory of Pathogenic Microbial Genomes, Institut Pasteur,
25 rue du Docteur Roux, 75724 Paris Cedex 15, France

C. Médigue
Genoscope and CNRS UMR-8030,
2 Rue Gaston Cremieux, CP 5706, 91057 Evry Cedex, France

Introduction

Penicillium marneffei is the most important thermal dimorphic fungus causing respiratory, skin and systemic mycosis in Southeast Asia (Yuen et al. 1994; Lo et al 1995; Kwan et al. 1997; Chim et al. 1998; Wong et al. 1999; Wong et al. 2001). Discovered in 1956 in hepatic abscesses of the Chinese bamboo rat *Rhizomys sinensis*, only 18 cases of human diseases were reported (in HIV-negative patients) until 1985 (Deng and Connor 1985). The appearance of the HIV pandemic, especially in Southeast Asian countries, saw the emergence of the infection as an important opportunistic mycosis in immunocompromised patients. About 10% of AIDS patients in Hong Kong are infected with *P. marneffei* (Wong and Lee 1998). In northern Thailand, penicilliosis is the third most common indicator disease of AIDS following tuberculosis and cryptococcosis (Supparatpinyo et al. 1994). Clinically, penicilliosis manifests as a systemic febrile illness, which results from intracellular infection of the reticuloendothelial cells by the yeast phase of the fungus and the associated inflammatory response of the host.

Despite its medical importance and its unusual thermal dimorphic capability, a large part of the ecology and epidemiology of *P. marneffei* remains unknown. The natural habitat of the fungus and its exact route of transmission have not been described. Molecular studies of this fungus at the molecular level have been limited. Only one cell-wall mannoprotein gene has been characterized and suc-

cessfully used in serodiagnosis and prevention of this infection (Cao et al. 1998a, b, 1999; Wong et al. 2001; Wong et al. 2002). Based on the mitochondrial and spacer rRNA, which allowed investigators to suggest a strong phylogenetic connection with *Talaromyces* species (LoBuglio and Taylor 1995), a PCR/hybridization assay was designed for molecular identification of this fungus in positive cultures (Vanittanakom et al. 1998).

P. marneffei is a model organism for understanding the molecular basis of thermal dimorphism. Given its propensity to cause disease in AIDS patients, the genome of *P. marneffei* may also provide insights into its pathogenic mechanisms and its possible interactions with the immune system. We describe in this report a random analysis of the genome of *P. marneffei*, which will facilitate further molecular research and lay the foundation for the complete genomic sequencing project of this fungus. A comprehensive knowledge of the genome will enable researchers to understand the basic mechanisms of thermal dimorphism, disease pathogenesis, virulence, and immune defense.

Materials and methods

Strains and DNA preparation

P. marneffei strain PM1 was isolated from an HIV-negative patient suffering from culture-documented penicilliosis in Hong Kong. In addition, ten additional strains of *P. marneffei*, isolated from the tissue and blood cultures of ten (7 HIV-positive and 3 HIV-negative) patients in Hong Kong, were used for electrokaryotyping. The arthroconidia ("yeast form") of PM1 was used throughout the DNA sequencing experiments. Genomic DNA was prepared from the arthroconidia grown at 37 °C. A single colony of the fungus grown on Sabouraud dextrose agar at 37 °C was inoculated into yeast peptone broth and incubated in a shaker at 30 °C for 3 days. Cells were cooled in ice for 10 min, harvested by centrifugation at 2,000 x g for 10 min, washed twice and resuspended in ice-cold 50 mmol EDTA/l buffer (pH 7.5). Subsequently, 20 mg novazym/ml was added and incubated at 37 °C for 1 h followed by digestion in a mixture of 1 mg proteinase K/ml, 1% *N*-lauroylsarcosine, and 0.5 mol EDTA/l pH 9.5 at 50 °C for 2 h. Genomic DNA was then extracted by phenol, phenol-chloroform, and finally precipitated and washed in ethanol. After digestion with RNase A, a second ethanol precipitation was washed with 70% ethanol, air-dried and dissolved in 500 µl of TE (pH 8.0).

Electrokaryotyping

The numbers and sizes of *P. marneffei* chromosomes in the 11 *P. marneffei* isolates were determined using pulsed-field gel electrophoresis. *P. marneffei* arthroconidia were grown for 3 days at 37 °C in yeast peptone broth as above. The harvested arthroconidia were washed and then inoculated into lyticase buffer containing 20 U of lyticase (Sigma). The protoplasts were then embedded in low melting point agarose plugs (2% prepared in isotonic solution and warmed to 50 °C) which were incubated in lyticase buffer at 37 °C for 1 h before treatment with lysis solution containing 1% *N*-laurylsarcosine and 1 mg proteinase K/ml. The mixture was incubated at 50 °C for 48 h.

Gels were cast using chromosomal-grade agarose (0.8%) in 0.5 x TBE buffer and pulsed-field gel electrophoresis was carried out for 120 h at 12 °C and 2 V per cm, using a contour-clamped homogenous electric field and a CHEF Mapper XA System (Bio-Rad Laboratories, Hercules, Calif., USA). In the first 96 h, an included angle of 120° and pulse times of 3–60 min were used, and in the next 48 h, an included angle of 106° and pulse times of 3–20 min were used. *Saccharomyces cerevisiae* and *Hansenula wingei* molecular mass standards were used. Gels were stained with ethidium bromide and photographed with the Eagle System (Strategene, La Jolla, Calif., USA).

Telomeric fingerprinting

Genomic DNA of strain PM1 was digested to completion with *Eco*RI, *Eco*RV, *Hae*III, or *Sal*I and separated on a 0.8% (w/v) agarose gel. For southern hybridization, Hybond N^+ membranes (Amersham) were used according to the manufacturer's instructions. After hybridization at 50 °C for 5 h with the DIG-labeled (DIG Oligonucleotide 3' End Labeling Kit, Roche) probe $(TTAGGG)_6$, the nylon membrane was washed twice with 2 x SSC/0.1% SDS (1 x SSC is 0.15 M NaCl with 0.015 M sodium citrate) at room temperature for 10 min, followed by washing twice with 0.1 x SSC/ 0.1% SDS at 50 °C for 15 min. Signal was detected according to the manufacturer's instructions.

Library construction

In order to make a library of *P. marneffei* DNA, 10 µg of genomic DNA was partially digested by *Sau*3A. Fragments of 1–2 kb were gel purified and ligated onto the pBK-CMV vector at the *Bam*HI site. The ligation mix was used to transform *Escherichia coli* DH5α cells by electroporation. Bacteria were plated on LB medium containing 100 mg ampicillin/l; 5,0000 clones were obtained. One hundred clones were randomly picked and checking by miniprep, which confirmed that 98% of the clones has inserts of 800–2,300 bp.

DNA sequencing

DNA was sequenced using the chain-termination reaction according to published protocols (Frangeul et al. 1999) with slight modifications. To ensure high and even sequence quality for the entire set of RSTs, each sequencing profile was inspected on a Sparc II workstation using the Alfsplit and Ted programmes of the Staden package; sequences containing any base-calling ambiguity, or <100 nucleotides were eliminated. Suspected frameshifts-issued form errors in our single-read sequences detected by BLASTX comparisons or by using DNA-Strider dot plot matrices were corrected according to the sequence alignments. The average error rate of the RSTs was determined to be 0.5% for nucleotide substitution and 0.3% for base insertion or deletion, by repeated sequencing of the pCMV vector from empty clones.

Sequence analysis

All sequences were submitted as a batch file (FASTA format) to the LASSAP software (Glemet and Codani 1997) for a general BLAST (Altschul et al. 1990) study against the aggregated SwissProt+GenePept+PIR Banks. For the identification of the best hits, the non-redundant library SwissProt+TREMBL+TREMBL_New was used. The output was analyzed by a software constructed ad hoc, filtering data of interest (all sequences similar to known sequences were retained for further studies).

For the reconstitution of the rDNA genes, a specialized library of rDNA sequences was constructed using the World-Wide DNA Data Library (GenBank/EBI-EMBL/DDBJ). The rDNA sequences were collected and assembled for further study. The phylogenetic relationships of *P. marneffei* to other related species were determined using PileUp method with GrowTree (Genetics Computer Group). A total of 1,726 nucleotide positions of the 18S rRNA genes were included in the analysis.

Nucleotide sequence accession numbers

Nucleotide sequences reported in this article are deposited in the GenBank database under accession numbers AL683898 to AL686199.

Results

Physical characteristics of the genome and telomeric fingerprinting

The genomes of all 11 *P. marneffei* strains consist of three or more chromosomes (Fig. 1). The sizes of the three bands are about 5.0, 4.0, and 2.2 Mb. The results of telomeric fingerprinting of *P. marneffei* are shown in Fig. 2. Six to 12 bands were detected after digestion with *Eco*RI, *Eco*RV, *Hae*III, or *Sal*I, suggesting that chromosomes of similar sizes co-migrated (Wu et al. 1996). Assuming 12 telomeric fragments representing six chromosomes, the genome size of *P. marneffei* was hence about 17.8 (5.0+4.0+2.2+2.2+ 2.2+2.2) to 26.2 (5.0+5.0+5.0+5.0+4.0+2.2) Mb.

Random sequences

The overall G+C content of the genome sequences identified was 48.8 mol%. The 2,303 sequence tags that were generated had an average length of 773±103 bp. These

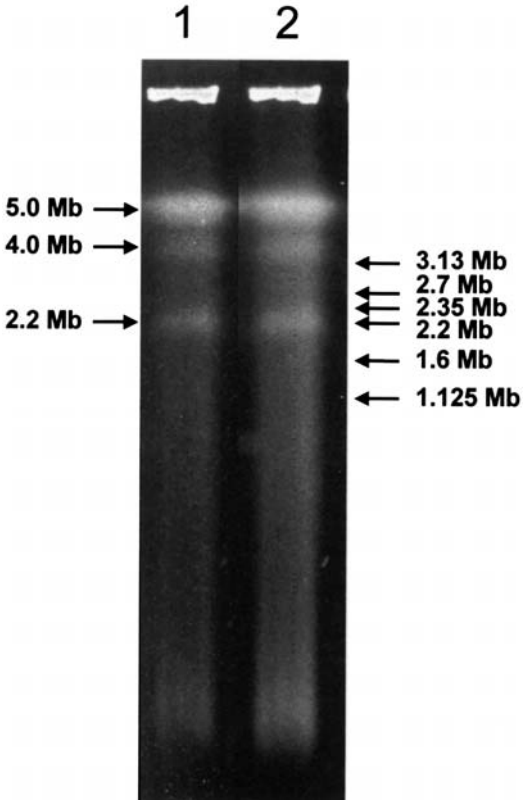


Fig. 1 Pulsed field gel electrophoresis of genomic DNA of *Penicillium marneffei*. *Lanes 1, 2 P. marneffei* strains PM1 and PM2, respectively. Sizes of markers (*Saccharomyces cerevisiae* and *Hansenula wingei* standards) are indicated on the *right*

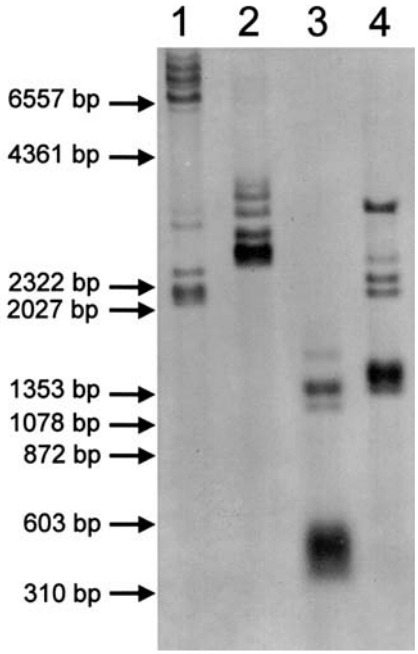


Fig. 2 Telomeric fingerprinting of *P. marneffei*. Six to 12 bands were detected after digestion with *Eco*RI (*lane 1*), *Eco*RV (*lane 2*), *Hae*III (*lane 3*), or *Sal*I (*lane 4*). Sizes of markers (λ *Hin*dIII digest and ΦX174 *Hae*III digest) are indicated on the *left*

2,303 sequence tags represented about 9% of the whole genome, if the genome size is 20 Mb. Table 1 lists the sequence tags of *P. marneffei* that show significant similarities to gene sequences in public databases. All major processes characterizing life are represented (Danchin 1989): metabolism (33%), information transfer (59%), and compartmentalization (8%) (Fig. 3).

One third of the genes are classified as primary metabolic genes, including genes coding for membrane-bound enzymes. They consist of genes for metabolism. Examples of genes involved in energy metabolism include cytochrome P450 enzymes, glucokinase, and enzymes of the glycolytic pathway. This is in line with the phenotypic observations showing that *P. marneffei* uses β-glucosides as carbon sources (Wong et al. 2001). Amino acid metabolism is represented by various aminotransferases and amino acid synthases; a number of enzymes involved in fatty-acid and phospholipids synthesis are also present, including a homologue of the lovastatin nonaketide synthase from *Aspergillus terreus*. Genes coding for an endoglucanase (which may be a cellulose) and a chitinase precursor are noted; these are likely to be involved in cell wall synthesis of the fungus. An interesting feature of the sequence tags identified here is that 3.4% code for secondary metabolism genes for non-ribosomal peptide synthesis and polyketide synthesis (this is likely to be an underestimate since polyketide synthases are usually highly repeated, and may have been removed from our sample as putative duplicates).

In the category of information transfer, six ribosomal protein genes as well as three tRNA synthetase subunits and five translation initiation and elongation factors (two

Table 1: Random sequence tags of penicillium marneffei matched to know sequences in public databases. *sp* SWISS-PROT, *gb* Genbank

Functions	Organism	Description	GenBank accession no.	Accession no. of closest hit	*Penicillium marneffei* RST	E value
Cell cycle	*Aspergillus nidulans*	G1/S regulator	AL684013	sp/O93843	PM10E8.B	$7.57\ e^{-25}$
	Aspergillus nidulans	Chromosome segregation SepB protein	AL684194	sp/Q00210	PM11E2.G	$1.13\ e^{-57}$
	Saccharomyces cerevisiae	Ubiquitin-like protein DSK2	AL683940	sp/P48510	PM10B7.G	$8.32\ e^{-10}$
	Saccharomyces cerevisiae	SIT4-associating protein SAP185	AL684015	sp/P40856	PM10E9.B	$3.59\ e^{-29}$
	Saccharomyces cerevisiae	SIT4-associating protein SAP190	AL684016	sp/P36123	PM10E9.G	$3.63\ e^{-16}$
	Schizosaccharomyces pombe	Cyclosome subunits	AL683936	sp/O42839	PM10B5.G	$2.94\ e^{-5}$
	Schizosaccharomyces pombe	Pelota protein	AL684046	sp/Q9USL5	PM10G11.G	$2.92\ e^{-22}$
Cell envelope	*Arabidopsis thaliana*	Mycolic acid methyl-transferase-like protein	AL684059	sp/Q9LUH5	PM10G7.B	$4.55\ e^{-10}$
	Arabidopsis thaliana	Mycolic acid methyl-transferase-like protein	AL684060	sp/Q9LUH5	PM10G7.G	$1.11\ e^{-6}$
	Arabidopsis thaliana	Putative pectin esterase	AL684185	sp/Q9SIJ9	PM11E1.B	$9.44\ e^{-11}$
	Arabidopsis thaliana	Multi-spanning membrane protein	AL684292	sp/Q9LIC2	PM12A3.G	$2.13\ e^{-42}$
	Beta vulgaris	Chitinase precursor	AL684111	sp/Q42421	PM11A9.B	$3.89\ e^{-14}$
	Beta vulgaris	Chitinase precursor	AL684288	sp/Q42421	PM12A12.G	$3.30\ e^{-41}$
	Beta vulgaris	Chitinase precursor	AL684297	sp/Q42421	PM12A6.B	$2.05\ e^{-38}$
	Beta vulgaris	Chitinase precursor	AL684305	sp/Q42421	PM12B1.B	$2.43\ e^{-28}$
	Beta vulgaris	Chitinase precursor	AL684310	sp/Q42421	PM12B11.G	$2.03\ e^{-28}$
	Beta vulgaris	Chitinase precursor	AL684311	sp/Q42421	PM12B12.B	$5.39\ e^{-31}$
	Beta vulgaris	Chitinase precursor	AL684319	sp/Q42421	PM12B5.B	$1.63\ e^{-34}$
	Beta vulgaris	Chitinase precursor	AL684326	sp/Q42421	PM12B8.G	$5.20\ e^{-25}$
	Beta vulgaris	Chitinase precursor	AL684351	sp/Q42421	PM12C9.B	$2.58\ e^{-26}$
	Beta vulgaris	Chitinase precursor	AL684364	sp/Q42421	PM12D3.G	$2.75\ e^{-10}$
	Candida glabrata	Cell wall synthesis protein KNH1 (cell wall β-1,6-glucan synthesis)	AL683994	sp/O74684	PM10E1.G	$1.27\ e^{-7}$
	Entamoeba histolytica	Surface antigen ariel1	AL684259	sp/O96609	PM11H10.B	$9.43\ e^{-4}$
	Homo sapiens	Mucin 2 precursor	AL684069	sp/Q02817	PM10H11.B	$1.55\ e^{-14}$
	Volvox carteri	Sulfated surface glycoprotein SSG185	AL684056	sp/P21997	PM10G5.G	$8.34\ e^{-4}$
Cellular metabolism						
Biosynthesis of cofactors	*Clostridium thermocellum*	Acetate kinase	AL683898	sp/O52594	PM10A1.B	$2.96\ e^{-24}$
Carbohydrate metabolism	*Aspergillus nidulans*	Trehalose-6-phosphate synthase subunit 1	AL684282	sp/O59921	PM12A1.G	$1.08\ e^{-36}$
Amino acid metabolism	*Aeropyrum pernix*	Dihydroxy-acid dehydratase	AL684193	sp/Q9YG88	PM11E2.B	$5.98\ e^{-12}$
	Arabidopsis thaliana	NADH-dependent glutamate synthase	AL684347	sp/Q9LV03	PM12C7.B	$1.51\ e^{-55}$
	Aspergillus nidulans	Homogentisate dioxygenase	AL683907	sp/Q00667	PM10A2.G	$2.97\ e^{-37}$
	Aspergillus nidulans	Pentafunctional arom polypeptide (polyaromatic amino acid biosynthesis)	AL684053	sp/P07547	PM10G4.B	$4.32\ e^{-99}$
	Aspergillus nidulans	Pentafunctional arom polypeptide (polyaromatic amino acid biosynthesis)	AL684054	sp/P07547	PM10G4.G	$3.15\ e^{-68}$

Table 1: (continued)

Functions	Organism	Description	GenBank accession no.	Accession no. of closest hit	*Penicillium marneffei* RST	E value
	Cochliobolus carbonum	Branched chain amino acid aminotransferase	AL684166	sp/Q9Y885	PM11D11.G	$8.64\ e^{-16}$
	Cochliobolus carbonum	Putative branched chain amino acid aminotransferase	AL684201	sp/Q9Y885	PM11E6.B	$6.11\ e^{-16}$
	Cochliobolus heterostrophus	Polyketide synthase	AL684273	sp/Q92217	PM11H6.B	$1.09\ e^{-55}$
	Escherichia coli	Arginase family	AL684071	sp/P16936	PM10H12.B	$6.17\ e^{-25}$
	Lactococcus lactis	Aminotransferase	AL684114	sp/Q9CE18	PM11B1.G	$1.72\ e^{-15}$
	Legionella pneumophila	Homogentisate dioxygenase	AL683911	sp/Q9S4T0	PM10A4.G	$3.74\ e^{-18}$
	Saccharomyces cerevisiae	Acetolactate synthase small subunit	AL683948	sp/P25605	PM10C10.G	$9.11\ e^{-11}$
	Saccharomyces cerevisiae	Glutamate synthase [NADPH] precursor	AL684348	sp/Q12680	PM12C7.G	$2.64\ e^{-47}$
	Vibrio cholerae	2-Hydroxyacid dehydrogenase family	AL684117	sp/Q9KP72	PM11B11.B	$5.73\ e^{-27}$
Energy metabolism	*Aspergillus nidulans*	Cytochrome P450	AL684154	sp/Q9Y7G5	PM11C6.G	$1.39\ e^{-13}$
	Bacillus halodurans	N-acetylglucosamine-6-phosphate deacetylase	AL683980	sp/Q9KFQ7	PM10D3.G	$5.26\ e^{-17}$
	Bacillus halodurans	Acetamidase	AL684270	sp/Q9KGN3	PM11H4.G	$5.25\ e^{-28}$
	Boophilus microplus	Cytochrome P450	AL683972	sp/Q9Y1T8	PM10D10.G	$4.64\ e^{-23}$
	Burkholderia cepacia	2,4-D dioxygenase	AL684226	sp/P96312	PM11F6.G	$6.73\ e^{-7}$
	Coprinus cinereus	Cytochrome P450	AL684373	sp/O74643	PM12D8.B	$6.93\ e^{-7}$
	Globodera pallida	NADH-ubiquinone oxidoreductase subunit 4	AL684327	sp/Q9T6M3	PM12B9.B	$7.35\ e^{-8}$
	Glycine max	Cytochrome P450	AL684153	sp/O48928	PM11C6.B	$2.29\ e^{-7}$
	Neurospora crassa	Mitochondrial NADH dehydrogenase	AL684228	sp/Q9Y7G7	PM11F7.G	$2.15\ e^{-56}$
	Neurospora crassa	64 kDa mitochondrial NADH dehydrogenase	AL684296	sp/Q9Y7G7	PM12A5.G	$9.74\ e^{-59}$
	Saccharomyces cerevisiae	6-Phosphogluconate dehydrogenase	AL684361	sp/P38720	PM12D2.B	$4.63\ e^{-25}$
	Vibrio cholerae	Glucokinase regulatory protein	AL684022	sp/Q9KVE0	PM10F11.G	$5.99\ e^{-6}$
Fatty acid and phospholipid metabolism	*Aspergillus fumigatus*	Inositol phosphorylceramide synthase	AL683939	sp/Q9Y745	PM10B7.B	$1.13\ e^{-68}$
	Aspergillus terreus	Lovastatin nonaketide synthase	AL684225	sp/Q9Y8A5	PM11F6.B	$1.36\ e^{-17}$
	Aspergillus terreus	Lovastatin nonaketide synthase	AL684235	sp/Q9Y8A5	PM11G10.B	$1.05\ e^{-28}$
	Aspergillus terreus	Lovastatin nonaketide synthase	AL684236	sp/Q9Y8A5	PM11G10.G	$4.62\ e^{-63}$
	Avena sativa	UDP-glucose:sterol glucosyltransferase	AL684217	sp/O22678	PM11F2.B	$2.22\ e^{-35}$
	Caenorhabditis elegans	Acyl-CoA dehydrogenase	AL684224	sp/Q19057	PM11F5.G	$2.44\ e^{-21}$
	Neurospora crassa	Long-chain fatty acid CoA ligase	AL684087	sp/Q9P3D2	PM10H9.B	$9.87\ e^{-34}$
	Neurospora crassa	Long-chain fatty acid CoA ligase	AL684088	sp/Q9P3D2	PM10H9.G	$1.11\ e^{-41}$
Purines or pyrimidines	*Schizosaccharomyces pombe*	Ribonucleoside-diphosphate reductase small chain	AL684332	sp/P36603	PM12C10.G	$1.39\ e^{-56}$

Table 1: (continued)

Functions	Organism	Description	GenBank accession no.	Accession no. of closest hit	*Penicillium marneffei* RST	E value
Others	*Acetobacter pasteurianus*	Carboxylesterase	AL684119	sp/O66374	PM11B12.B	$1.80\ e^{-5}$
	Arthrobacter globiformis	Choline oxidase	AL683979	sp/Q59117	PM10D3.B	$1.62\ e^{-5}$
	Aspergillus fumigatus	Catalase	AL683912	sp/P78574	PM10A5.B	$1.95\ e^{-78}$
	Aspergillus nidulans	Acetyl-CoA carboxylase	AL684151	sp/O60033	PM11C5.B	$3.41\ e^{-59}$
	Aspergillus nidulans	Alcohol dehydrogenase I	AL684354	sp/P08843	PM12D1.G	$6.02\ e^{-13}$
	Bacillus stearothermophilus	Alcohol dehydrogenase	AL683998	sp/P42328	PM10E11.G	$6.89\ e^{-7}$
	Bacillus subtilis	YVRD protein (short chain dehydrogenase/reductase)	AL683976	sp/O34782	PM10D12.G	$4.83\ e^{-5}$
	Cladosporium herbarum	Aldehyde dehydrogenase	AL684360	sp/P40108	PM12D12.G	$3.07\ e^{-24}$
	Homo sapiens	Short-chain dehydrogenases/reductase	AL684281	sp/Q9NRW0	PM12A1.B	$4.16\ e^{-6}$
	Nectria haematococca mpVI	Aromatic-ring hydroxylase	AL684135	sp/Q01446	PM11B9.B	$2.02\ e^{-23}$
	Neurospora crassa	Folylpolyglutamate synthetase	AL683975	sp/O13492	PM10D12.B	$4.07\ e^{-7}$
	Ophiostoma novo-ulmi	Polygalacturonase	AL684159	sp/O59934	PM11C9.B	$1.85\ e^{-11}$
	Pyrococcus abyssi	Chlorohydrolase	AL683982	sp/Q9V0Y5	PM10D4.G	$3.12\ e^{-4}$
	Schizosaccharomyces pombe	Probable acid phosphatase	AL684049	sp/Q9USS6	PM10G2.B	$1.63\ e^{-16}$
	Shewanella frigidimarina	Fumarate reductase flavoprotein subunit precursor	AL684051	sp/Q9Z4P0	PM10G3.B	$9.31\ e^{-17}$
	Thermotoga maritima	Folylpolyglutamate synthetase	AL684143	sp/Q9WY13	PM11C12.B	$1.21\ e^{-10}$
Cell signalling						
Receptor and their associated proteins	*Saccharomyces cerevisiae*	Pheromone receptor	AL683922	sp/P06842	PM10B1.G	4.75^{-14}
Protein kinase and phosphatase	*Arabidopsis thaliana*	Dual-specificity protein phosphatase	AL683901	sp/Q9M8K7	PM10A10.G	$2.38\ e^{-5}$
	Arabidopsis thaliana	Histidine kinase 1 osmosensor	AL684195	sp/Q9SXL4	PM11E3.B	$3.69\ e^{-5}$
	Candida albicans	Serine/threonine-protein kinase	AL683946	sp/Q92212	PM10C1.G	$5.46\ e^{-6}$
	Saccharomyces cerevisiae	Probable serine/threonine-protein kinase	AL684172	sp/Q12399	PM11D3.G	$8.90\ e^{-5}$
	Schizosaccharomyces pombe	Protein-tyrosine phosphatase (dual specificity protein phosphatase)	AL684242	sp/O13819	PM11G2.G	$2.75\ e^{-35}$
Others	*Aspergillus nidulans*	Nuclear migration protein NUDF	AL684050	sp/Q00664	PM10G2.G	$5.70\ e^{-6}$
	Neurospora crassa	GTP-binding protein YPT1	AL684157	sp/P33723	PM11C8.B	$2.22\ e^{-56}$
	Neurospora crassa	G-protein β WD-40 repeats	AL684170	sp/Q9P6V7	PM11D2.G	$2.44\ e^{-4}$
	Saccharomyces cerevisiae	Ankyrin repeat-containing protein Akr1p	AL683959	sp/P39010	PM10C5.B	$1.06\ e^{-7}$
	Saccharomyces cerevisiae	Ankyrin repeat-containing protein Akr1p	AL683960	sp/P39010	PM10C5.G	$5.18\ e^{-7}$
	Volvox carteri	Pherophorin-S precursor	AL684293	sp/P93797	PM12A4.B	$4.75\ e^{-8}$
	Volvox carteri	Pherophorin-S precursor	AL684333	sp/P93797	PM12C11.B	$1.98\ e^{-32}$
	Volvox carteri	Pherophorin-S precursor	AL684337	sp/P93797	PM12C2.B	$2.85\ e^{-25}$
	Volvox carteri	Pherophorin-S precursor	AL684349	sp/P93797	PM12C8.B	$6.45\ e^{-18}$
	Volvox carteri	Pherophorin-S precursor	AL684353	sp/P93797	PM12D1.B	$2.95\ e^{-18}$
	Volvox carteri	Pherophorin-S precursor	AL684362	sp/P93797	PM12D2.G	$2.75\ e^{-6}$
	Volvox carteri	Pherophorin-S precursor	AL684363	sp/P93797	PM12D3.B	$1.90\ e^{-13}$

Table 1: (continued)

Functions	Organism	Description	GenBank accession no.	Accession no. of closest hit	*Penicillium marneffei* RST	E value
DNA replication and metabolism						
DNA replication, modification	*Aspergillus flavus*	O-methyltransferase	AL684183	sp/Q9P900	PM11D9.B	1.44 e^{-15}
	Saccharomyces cerevisiae	Mitochondrial membrane GTPase	AL683977	sp/P38297	PM10D2.B	5.45 e^{-40}
	Saccharomyces cerevisiae	Mitochondrial membrane GTPase	AL683978	sp/P38297	PM10D2.G	1.94 e^{-4}
	Schizosaccharomyces pombe	Replication factor-A protein 2	AL684123	sp/Q92373	PM11B3.B	4.09 e^{-17}
DNA repair	*Arabidopsis thaliana*	Ubiquitin-protein ligase 2	AL684025	sp/P42745	PM10F2.B	7.48 e^{10}
	Pyrococcus kodakaraensis	Methylated DNA protein cysteine methyltransferase	AL683969	sp/O74023	PM10D1.B	2.25 e^{-9}
	Saccharomyces cerevisiae	DNA repair protein RAD5	AL684331	sp/P32849	PM12C10.B	5.53 e^{-9}
DNA binding	*Gallus gallus*	RING zinc-finger protein	AL684278	sp/Q90972	PM11H8.G	3.62 e^{-10}
	Trypanosoma cruzi	Kinetoplast-associated protein	AL684304	sp/Q26938	PM12A9.G	4.45 e^{-18}
Chromosomal structure	*Ensis minor*	Nuclear protein (linker histone H1)	AL684302	sp/Q24898	PM12A8.G	1.29 e^{-16}
	Ensis minor	Nuclear protein (linker histone H1)	AL684308	sp/Q24898	PM12B10.G	8.21 e^{-15}
	Ensis minor	Nuclear protein (linker histone H1)	AL684368	sp/Q24898	PM12D5.G	4.24 e^{-11}
	Schizosaccharomyces pombe	Chromosome region maintenance protein 1	AL683937	sp/P14068	PM10B6.B	2.79 e^{-54}
	Schizosaccharomyces pombe	Chromosome region maintenance protein 1	AL683938	sp/P14068	PM10B6.G	2.15 e^{-43}
	Schizosaccharomyces pombe	Histone H4	AL684175	sp/P09322	PM11D5.B	3.24 e^{-4}
RNA-directed DNA polymerase	*Aedes aegypti*	Pol-like protein (similar to RNA-directed DNA polymerase)	AL684213	sp/Q9U4W1	PM11F11.B	6.60 e^{-8}
	Neurospora crassa	Similar to RNA-directed DNA polymerase	AL684140	sp/Q01375	PM11C10.G	9.92 e^{-6}
	Neurospora crassa	Similar to RNA-directed DNA polymerase	AL684334	Q01375	PM12C11.G	1.42 e^{-5}
Intracellular trafficking	*Homo sapiens*	Transmembrane protein Tmp21 precursor	AL684256	sp/P49755	PM11G9.G	4.91 e^{-12}
	Neurospora crassa	Probable mitochondrial membrane dicarboxylate carrier protein	AL684082	sp/Q9P5U2	PM10H6.G	9.79 e^{-6}
	Saccharomyces cerevisiae	Protein transport protein Sec23	AL683984	sp/P15303	PM10D5.G	2.7 e^{-69}
	Schizosaccharomyces pombe	Importin subunit	AL683935	sp/O13864	PM10B5.B	4.02 e^{-64}
	Schizosaccharomyces pombe	Putative vacuolar biogenesis protein	AL684131	sp/Q9P6N4	PM11B7.B	2.02 e^{-6}
	Schizosaccharomyces pombe	Putative vacuolar biogenesis protein	AL684132	sp/Q9P6N4	PM11B7.G	4.91 e^{-9}
	Schizosaccharomyces pombe	Putative involvement in vesicular transport	AL684221	sp/Q9P6K0	PM11F4.B	3.36 e^{-8}
	Xenopus laevis	Nucleolar phosphoprotein	AL684222	sp/Q91803	PM11F4.G	1.51 e^{-6}
Membrane transport	*Amanita muscaria*	Sugar transporter	AL684177	sp/O13411	PM11D6.B	1.34 e^{-24}
	Arabidopsis thaliana	Sugar transporter	AL684261	sp/O23213	PM11H11.B	1.28 e^{-8}
	Botrytis cinerea	ABC transporter	AL684028	sp/O60034	PM10F3.G	3.29 e^{-91}
	Cochliobolus heterostrophus	Fatty-acid transporter protein	AL684113	sp/O42633	PM11B1.B	1.28 e^{-20}

Table 1: (continued)

Functions	Organism	Description	GenBank accession no.	Accession no. of closest hit	*Penicillium marneffei* RST	E value
	Gibberella pulicaris	MFS-multidrug resistance transporter	AL684204	sp/Q9P8F5	PM11E7.G	4.13 e^{-26}
	Homo sapiens	Calcium channel β2a subunit	AL684181	sp/Q9Y341	PM11D8.B	2.42 e^{-4}
	Mus musculus	Peroxisomal protein ALDR (ABC transporter)	AL684298	sp/Q61285	PM12A6.G	6.85 e^{-23}
	Mycobacterium avium	Molybdate uptake secreted protein	AL684161	sp/Q48919	PM11D1.B	3.47 e^{-4}
	Neurospora crassa	Amino acid permease 2	AL684102	sp/O59942	PM11A4.G	2.70 e^{-12}
	Saccharomyces cerevisiae	Purine-cytosine permease	AL683910	sp/P17064	PM10A4.B	1.18 e^{-23}
	Saccharomyces cerevisiae	Possible small-molecule transporter	AL683990	sp/P53134	PM10D8.G	2.93 e^{-37}
	Saccharomyces cerevisiae	Purine-cytosine permease	AL684099	sp/P17064	PM11A3.B	1.43 e^{-30}
	Saccharomyces cerevisiae	Fluconazole resistance protein 1	AL684103	sp/P38124	PM11A5.B	6.74 e^{-13}
	Schizosaccharomyces pombe	ABC transporter	AL683915	sp/P36619	PM10A6.G	6.43 e^{-22}
	Schizosaccharomyces pombe	Membrane transporter	AL684104	sp/O59700	PM11A5.G	1.74 e^{-39}
	Schizosaccharomyces pombe	Amino acid permease	AL684147	sp/Q9US40	PM11C3.B	1.39 e^{-60}
	Schizosaccharomyces pombe	Putative transporter of the allantoate permease family	AL684187	sp/Q10097	PM11E10.B	4.28 e^{-13}
	Schizosaccharomyces pombe	Probable membrane transporter	AL684188	sp/Q9US44	PM11E10.G	6.07 e^{-21}
	Schizosaccharomyces pombe	HMSF membrane transporter	AL684203	sp/O43081	PM11E7.B	4.95 e^{-33}
	Streptomyces fradiae	Transporter	AL684042	sp/Q9RP97	PM10G1.G	6.91 e^{-11}
Chaperone system	*Saccharomyces cerevisiae*	T-complex protein 1, α-subunit	AL683995	sp/P12612	PM10E10.B	2.62 e^{-23}
	Thermotoga maritima	Chaperone protein Dnaj	AL683961	sp/Q9WZV3	PM10C6.B	4.44 e^{-6}
Protein synthesis and degradation						
Ribosomal proteins	*Homo sapiens*	60S acidic ribosomal protein PO	AL684033	sp/Q9UKD2	PM10F6.B	4.12 e^{-11}
	Neurospora crassa	60S ribosomal protein L28	AL684045	sp/P08978	PM10G11.B	1.97 e^{-15}
	Schizosaccharomyces pombe	50S ribosomal protein	AL683930	sp/O94345	PM10B2.G	1.95 e^{-12}
RNA polymerase	*Caenorhabditis elegans*	DNA-directed RNA polymerase II largest subunit	AL684065	sp/P16356	PM10H1.B	2.13 e^{-5}
	Drosophila melanogaster	DNA-directed RNA polymerase II largest subunit	AL684211	sp/P04052	PM11F10.B	1.73 e^{-7}
RNA-binding proteins	*Mycobacterium tuberculosis*	Translation initiation factor 2	AL684306	sp/P71613	PM12B1.G	4.23 e^{-8}
	Nicotiana glutinosa	RNA-binding protein	AL683904	sp/O24106	PM10A12.B	2.66 e^{-6}
Aminoacyl-tRNA synthetase, tRNAs	*Homo sapiens*	Bifunctional aminoacyl-tRNA synthetase	AL683996	sp/P07814	PM10E10.G	2.48 e^{-83}
	Neurospora crassa	Valyl-tRNA synthetase	AL684355	sp/P28350	PM12D10.B	5.13 e^{-33}
	Neurospora crassa	Valyl-tRNA synthetase	AL684356	sp/P28350	PM12D10.G	5.34 e^{-33}
	Saccharomyces cerevisiae	tRNA synthetase	AL683919	sp/P38707	PM10A8.G	1.45 e^{-47}
Protein synthesis	*Tolypocladium inflatum*	Cyclosporin synthetase	AL683903	sp/Q09164	PM10A11.G	5.1 e^{-20}

Table 1: (continued)

Functions	Organism	Description	GenBank accession no.	Accession no. of closest hit	*Penicillium marneffei* RST	E value
Protein modification and translation factors	*Acanthamoeba castellanii*	Myosin I heavy-chain kinase	AL684328	sp/Q93107	PM12B9.G	1.88 e^{-18}
	Homo sapiens	Geranylgeranyl transferase type I β-subunit	AL684199	sp/P53609	PM11E5.B	1.35 e^{-5}
	Metarhizium anisopliae	Peptide synthetase	AL683902	sp/Q01135	PM10A11.B	1.75 e^{-60}
	Schizosaccharomyces pombe	Eukaryotic translation initiation factor	AL683957	sp/Q10425	PM10C4.B	2.74 e^{-37}
Degradation of proteins	*Aspergillus oryzae*	Alanyl dipeptidyl peptidase	AL684260	sp/Q9Y8E3	PM11H10.G	3.24 e^{-30}
	Caenorhabditis elegans	Ubiquitin fusion degradation protein 1 homologue	AL683963	sp/Q19584	PM10C7.B	9.33 e^{-4}
	Mus musculus	Peptide:N-glycanase	AL684043	sp/Q9JI78	PM10G10.B	3.13 e^{-35}
	Saccharomyces cerevisiae	Ubiquitin fusion degradation protein 1	AL684064	sp/P53044	PM10G9.G	7.48 e^{-6}
	Schizosaccharomyces pombe	26S Proteasome regulatory subunit	AL684128	sp/O74762	PM11B5.G	1.03 e^{-11}
	Schizosaccharomyces pombe	26S Proteasome regulatory subunit 12	AL684238	sp/O74440	PM11G11.G	8.14 e^{-12}
	Schizosaccharomyces pombe	26S Proteasome regulatory subunit MTS3	AL684344	sp/P50524	PM12C5.G	1.78 e^{-9}
	Sclerotinia sclerotiorum	Acid protease	AL684037	sp/Q9P8R1	PM10F8.B	6.0 e^{-4}
Transcription and mRNA regulation	*Aspergillus niger*	Transcription factor pacC	AL684253	sp/Q00203	PM11G8.B	1.21 e^{-31}
	Aspergillus niger	Transcription factor pacC	AL684254	sp/Q00203	PM11G8.G	2.62 e^{-37}
	Mus musculus	GA binding protein β-1 chain	AL683986	sp/Q00420	PM10D6.G	2.50 e^{-4}
	Saccharomyces cerevisiae	Transcriptional adaptor	AL683966	gb/NP010736	PM10C8.G	1.47 e^{-6}
	Schizosaccharomyces pombe	Transcriptional adaptor	AL683965	sp/Q9P7J7	PM10C8.B	6.83 e^{-24}
	Schizosaccharomyces pombe	Transcription factor ATF1 required for sexual development	AL684078	sp/P52890	PM10H4.G	2.84 e^{-5}
	Schizosaccharomyces pombe	BTB domain and ankyrin repeat-containing protein	AL684127	sp/O74881	PM11B5.B	2.52 e^{-23}
DEAD box proteins	*Saccharomyces cerevisiae*	ATP-dependent RNA helicase DOB1	AL684084	sp/P47047	PM10H7.G	2.27 e^{-26}
	Leishmania amazonensis	ATP-dependent RNA helicase	AL684124	sp/P90549	PM11B3.G	1.25 e^{-62}
Cytoskeleton	*Acanthamoeba castellanii*	Myosin IC heavy chain	AL684318	sp/P10569	PM12B4.G	2.30 e^{-9}
	Acanthamoeba castellanii	Myosin IA heavy chain	AL684320	sp/O77202	PM12B5.G	2.20 e^{-15}
	Acanthamoeba castellanii	Myosin IA	AL684372	sp/O77202	PM12D7.G	1.62 e^{-20}
	Aspergillus nidulans	Myosin I heavy chain	AL684174	sp/Q00647	PM11D4.G	3.74 e^{-35}
	Bos taurus (Bovine)	N-WASP	AL684375	sp/Q95107	PM12D9.B	5.56 e^{-19}
	Homo sapiens	WASP interacting protein	AL684365	sp/O43516	PM12D4.B	6.99 e^{-15}
	Homo sapiens	Diaphanous protein homologue 1	AL684367	sp/O60610	PM12D5.B	1.09 e^{-23}
	Mus musculus	Lymphocyte-specific formin-related protein	AL683968	sp/Q9Z2V7	PM10C9.G	9.83 e^{-6}
	Nectria haematococca mpVI	Kinesin	AL684125	sp/P78718	PM11B4.B	9.42 e^{-84}

348

Table 1: (continued)

Functions	Organism	Description	GenBank accession no.	Accession no. of closest hit	*Penicillium marneffei* RST	E value
Hypothetical proteins	*Arabidopsis thaliana*	46.1-kDa protein	AL683905	sp/Q9LF56	PM10A12.G	$1.09\ e^{-5}$
	Arabidopsis thaliana	61.8-kDa TRP-ASP repeats-containing protein	AL684058	sp/O22212	PM10G6.G	$5.55\ e^{-23}$
	Arabidopsis thaliana	gb|AAF34307.1	AL684289	sp/Q9LIR9	PM12A2.B	$4.22\ e^{-15}$
	Caenorhabditis elegans	Coded for by *C. elegans* cDNA YK165E3.3	AL683934	sp/O02173	PM10B4.G	$4.94\ e^{-5}$
	Caenorhabditis elegans	52.8-kDa protein	AL684257	sp/P30640	PM11H1.B	$7.00\ e^{-4}$
	Escherichia coli	47.3-kDa protein	AL684137	sp/P75791	PM11C1.B	$4.69\ e^{-10}$
	Mus musculus	Octapeptide-repeat protein T2	AL684321	sp/Q06666	PM12B6.B	$1.10\ e^{-4}$
	Neurospora crassa	93.3-kDa protein	AL683947	sp/Q9P6B2	PM10C10.B	$1.77\ e^{-39}$
	Saccharomyces cerevisiae	77.7-kDa protein	AL683992	sp/P47077	PM10D9.G	$6.21\ e^{-14}$
	Saccharomyces cerevisiae	Possible role in ribosome biogenesis	AL684068	sp/Q04660	PM10H10.G	$1.16\ e^{-7}$
	Saccharomyces cerevisiae	19.7-kDa protein	AL684077	sp/P36088	PM10H4.B	$1.68\ e^{-22}$
	Saccharomyces cerevisiae	110.9-kDa protein	AL684081	sp/P53920	PM10H6.B	$6.01\ e^{-14}$
	Saccharomyces cerevisiae	Integral membrane protein	AL684098	sp/P40468	PM11A2.G	$6.48\ e^{-28}$
	Saccharomyces cerevisiae	Possible isomerase	AL684109	sp/Q12177	PM11A8.B	$2.36\ e^{-30}$
	Saccharomyces cerevisiae	81.5-kDa integral membrane protein	AL684291	sp/P40071	PM12A3.B	$4.09\ e^{-28}$
	Schizosaccharomyces pombe	52.3-kDa integral membrane protein	AL683916	sp/Q10254	PM10A7.B	$1.85\ e^{-9}$
	Schizosaccharomyces pombe	21.3-kDa protein	AL683917	sp/O94329	PM10A7.G	$7.39\ e^{-10}$
	Schizosaccharomyces pombe	170.7-kDa integral membrane protein	AL683943	sp/Q10250	PM10B9.B	$4.43\ e^{-76}$
	Schizosaccharomyces pombe	68.1-kDa protein	AL683985	sp/O94419	PM10D6.B	$1.42\ e^{-6}$
	Schizosaccharomyces pombe	170.7-kDa transmembrane protein	AL683987	sp/Q10250	PM10D7.B	$2.53\ e^{-76}$
	Schizosaccharomyces pombe	WD-repeat protein	AL684011	sp/Q9USZ0	PM10E7.B	$1.15\ e^{-10}$
	Schizosaccharomyces pombe	181.5-kDa protein	AL684023	sp/Q09853	PM10F12.B	$2.92\ e^{-24}$
	Schizosaccharomyces pombe	16.6-kDa protein	AL684034	sp/Q9UUC8	PM10F6.G	$1.43\ e^{-13}$
	Schizosaccharomyces pombe	Integral membrane protein	AL684063	sp/O94348	PM10G9.B	$3.38\ e^{-15}$
	Schizosaccharomyces pombe	27.4-kDa protein	AL684133	sp/O14359	PM11B8.B	$1.22\ e^{-18}$
	Schizosaccharomyces pombe	63.7-kDa protein	AL684146	sp/O94667	PM11C2.G	$2.27\ e^{-5}$
	Schizosaccharomyces pombe	55.8-kDa protein	AL684271	sp/Q9URX1	PM11H5.B	$2.77\ e^{-29}$

for the mitochondrial translation apparatus) were found. A significant number of regulatory protein genes involved in transcription, control of the cell cycle or in differentiation, in particular serine/threonine and tyrosine kinases and phosphatases as well as, interestingly, a class III adenylyl cyclase (Barzu and Danchin 1994), were also observed. The former part of this category, to which the significant number of sequences extracted from rDNA regions can be added, allowed a rough evaluation of the genome length (this cannot be done with the category of

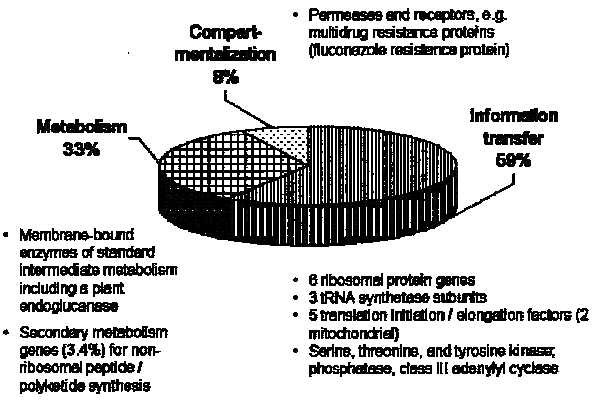


Fig. 3 Distribution of *P. marneffei* genes with regard to cellular functions

regulatory genes, which can be extremely variable from one organism to another).

Similar to other organisms, 8% of the sequences with identified function annotations presumably code for proteins involved in cell compartmentalization, in particular permeases and receptors. A large number of membrane transporters and permeases were found that are involved in the transport of carbohydrates, proteins, and fatty acids. Among them were several multi-drug-resistance protein genes, including a gene similar to that of a fluconazole resistance gene of *S. cerevisiae*.

One most intriguing observation from this sequence tags collection is the indications for the presence of a mating type and of mating pheromones in the genome. Similarities to the *S. cerevisiae* pheromone receptor gene *STE2* were found in some sequence tags, as well as similarities to pherophorins that bear sequence similarities to algal (*Volvox carteri*) sexual pheromone. Further evidence comes from the similarities to the *S. cerevisiae* ankyrin repeat-containing protein Akr1p which is involved in the yeast's pheromone response pathway and contributes to the control of cell shape and signal transduction (Kao et al. 1996; Pryciak and Hartwell 1996), although the expected values of $1.06\ e^{-7}$ and $5.18\ e^{-7}$ were not as high.

The genome appears to contain a large proportion of unknown and repeated sequences. However, there were not many transposase-related genes except for a counterpart of a transposase found in *Talaromyces stipitatus* (WWDDL accession number: CAA09449). Repeated sequences were often similar to segments of genes coding for proteins of the cell wall, or proteins involved in the cytoskeleton, as in the case of other yeasts and fungi. This is also a characteristic feature of pathogenic microorganisms. Among the repeated sequences were those that may code for proteins similar to the PE-PGRS glycine-rich proteins of *Mycobacterium tuberculosis* (Espitia et al. 1999). At this stage, however, it is not certain that these GC-rich regions indeed code for proteins (translation termination codons are statistically rare in such regions).

After removing duplicates, the random sequences were submitted to the LASSAP software against two specific databanks built for the purpose of this work—using the SRS tool (Stoesser et al. 2001)—written in the appropriate format: one containing all the known yeast sequences (YeastDB), and one containing the available fungus-non-yeast sequences (FungiDB). Taking the best hits, 269 matched only with yeast-specific sequences, 144 only with fungus-non-yeast sequences, and 400 matched with both (common sequences). The remaining sequence tags (1,183) matched with other sequences in the aggregated databanks.

Ribosomal DNA

The various fragments collected from rDNA loci can be assembled into contigs, allowing identification of most the sequence of the 18S and 28S RNAs of *P. marneffei*. Based on this knowledge, *P. marneffei* is likely to be an anamorph of a *Talaromyces* species. This substantiates the observation that the spacer regions of the rDNA loci are highly similar to that found in *Talaromyces* species (Kappe et al. 1996; Verweij et al. 1995). Indeed the sequence is almost identical with that of *T. flavus* (Fig. 4). It is also very similar to that of *Chromocleista cinnabarina*, a soil fungus that produces a red pigment, as does *P. marneffei* (Udagawa et al. 1973).

Discussion

Little is known about the ecology, transmission, or pathogenesis of *P. marneffei* infection. Although several species of bamboo rats are known to be carriers of the fungus, and, on rare occasions, the fungus has been isolated from bamboo rat burrows, it appears that they do not serve as significant reservoirs for human infection but are coincidentally infected from a common environmental source (Chariyalertsak et al. 1996, 1997). It is generally believed that *P. marneffei* exists in the soil of endemic areas and susceptible hosts—humans and bamboo rats—acquired the infection by inhaling the infectious conidia of the fungus, as in the case of other thermal dimorphic fungi.

The route of transmission of *P. marneffei*, however, has not been convincingly established. The main obstacle lies in the fact that the fungus has never been consistently isolated from any environmental samples, even in highly endemic areas. Conventional mycological culture relies on the identification of the fungus using the criteria of: (1) microscopic morphology of the reproductive structures, (2) elaboration of a diffusible red pigment, and (3) thermal dimorphism. However, this rests on the premise that the fungus exists in the familiar mould form in nature. Morphological identification of *P. marneffei* in environmental samples would fail if the saprophytic form is indeed a teleomorph, which may have a very different morphology. A teleomorph for *P. marneffei* has never been described, although our analysis of the genome does suggest this possibility. The presence of homologues for fun-

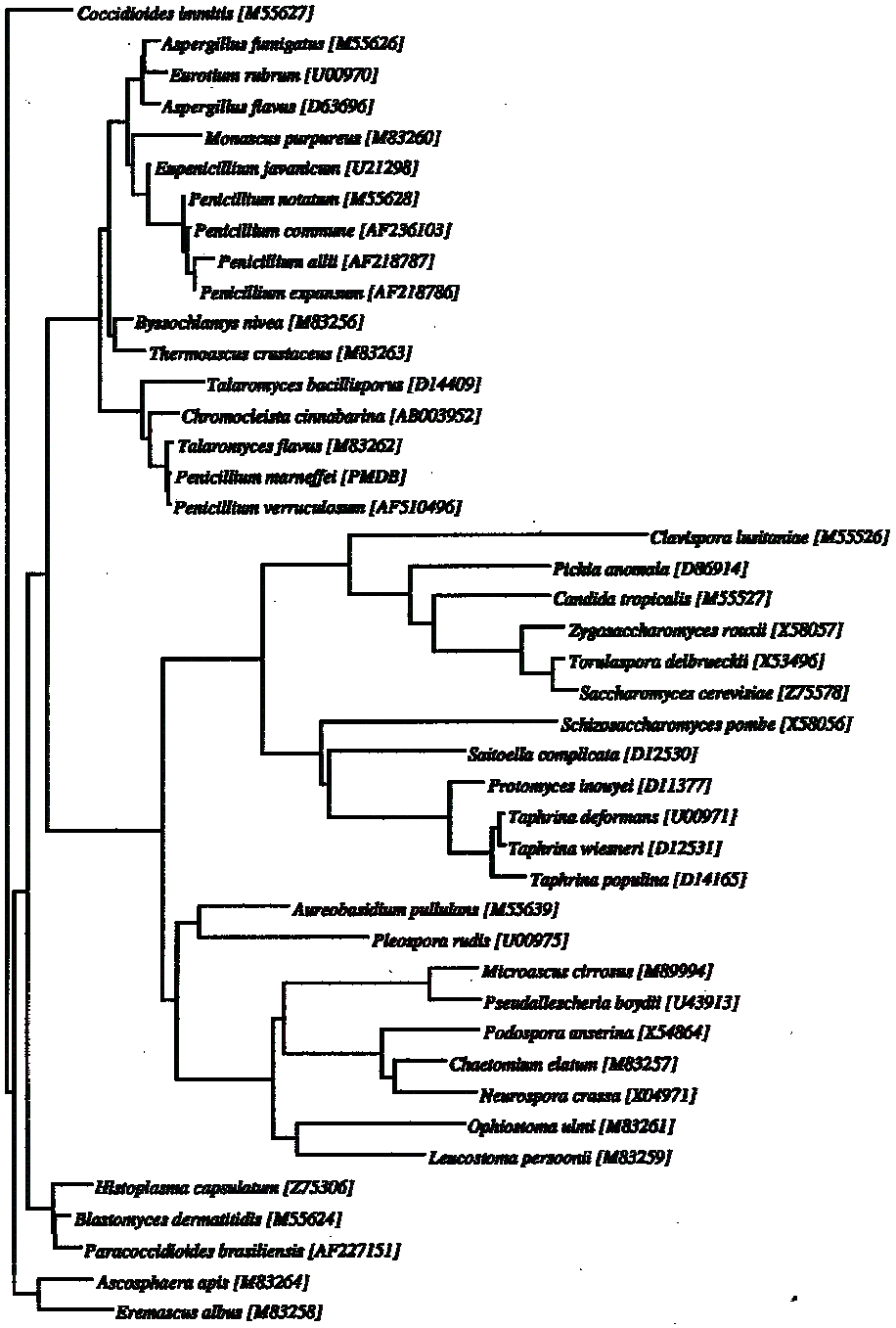


Fig. 4 Phylogenetic tree showing the relationships of *P. marneffei*. to other *Penicillium* and *Talaromyces* species. The tree was inferred from 18S rRNA data by the neighbor-joining method. *Scale bar* Estimated number of substitutions per 100 bases using the Jukes-Cantor correction. Names and accession numbers are given as cited in the GenBank database

gal and algal pheromone as well as molecules involved in the *S. cerevisiae* pheromone response pathway is highly suggestive of the presence of a sexual stage in the life cycle of *P. marneffei*. A further line of evidence comes from the finding that an *STE12* homologue of *P. marneffei* restores the defect in sexual development of *Aspergillus nidulans steA* mutant (Borneman et al. 2001). It is generally accepted that organisms without some sort of sexuality are rare, because sex-associated gene recombination or reassortment processes are the only way to escape the fate of Muller's ratchet, leading to degeneracy and ultimately to disappearance (Kondrashov 1994). It is therefore to be expected that some form of *P. marneffei* should involve mating. Indeed, teleomorphs of several other *Penicillium* species have been described and classified under *Eupenicillium* and *Talaromyces*, an example being *Penicillium*

Table 2 Karyotypes and genome sizes of *P. chrysogenum*, *P. notatum*, *P. nalgiovense*, *P. janthinellum*, *P. paxilli*, *P. purpurogenum*, and *P. marneffei*

Penicillium species	Number of chromosomes	Genomes sizes	References
P. chrysogenum	4	34.1	Fierro et al. 1993
P. notatum	4	32.1	Fierro et al. 1993
P. nalgiovense	4	26.5	Farber and Geisen 2000
P. janthinellum	8	39.0-49.0	Kayser and Schulz 1991
P. paxilli	8	-	Young et al. 1998
P. purpurogenum	5	21.2	Chavez et al. 2001
P. marneffei	3-6	17.8-26.2	Present study

emersonii which is the anamorph of *Talaromyces emersonii*, a thermophilic fungus usually isolated from soil (Cimon et al. 1999). The phylogenetic position of *P. marneffei* was recently studied using nuclear and mitochondrial ribosomal DNA sequences. Results of the study placed *P. marneffei* closely related to *Talaromyces*. It is also known that in Ascomycetes (e.g. *Talaromyces*) the haploid forms may differ widely, with the diploid form being transient. As a consequence it is possible that the pathogenic stage of the fungus is just the haploid form of a thermotolerant saprophyte with a short-lived sexual reproduction.

The genome size of *P. marneffei* is relatively small compared to other *Penicillium* species (Chavez et al. 2001). The karyotypes and genome sizes of six *Penicillium* species (*P. chrysogenum*, *P. notatum*, *P. nalgiovense*, *P. janthinellum*, *P. paxilli*, and *P. purpurogenum*) have been published (Table 2). The chromosome number varies from four to eight while the estimated genome size varies from 21.2 to 49.0 Mb. The genome size of *P. marneffei* is small compared to these six *Penicillium* species. The significance of this is unknown. One possible explanation could be that *P. marneffei*, at some stage of its life cycle, is an obligate parasite in susceptible hosts; in fact, it is the only species of *Penicillium* that consistently causes disease in humans. Compared to purely saprophytic species of *Penicillium*, some of the genes may have been lost during evolution. Other obligate parasites such as *Mycoplasma* (Fraser et al. 1995, 1998; Himmelreich et al. 1996), *Treponema pallidum* (Fraser et al. 1998), and *Mycobacterium leprae* (Cole et al. 2001) also characteristically possess relatively small genome sizes compared to their free-living counterparts or species with more elaborate life cycles. Whether this is the case for *P. marneffei* needs to be confirmed by comparative genomic studies with other related *Penicillium* and *Talaromyces* species.

The unique virulence and pathogenicity of *P. marneffei* amongst the otherwise saprophytic *Pencillium* genus are unexplained. The crux of this is the propensity to cause disease in patients with impaired cellular immunity, with AIDS patients being the largest at-risk population. The thermal tolerance of *P. marneffei* is definitely essential for its virulence, but detailed pathogenic mechanisms have not been described, apart from reports on its interactions with leukocytes and adhesion of the conidia to laminin (Hamilton et al. 1998, 1999). The presence of a large number of thioester-mediated non-ribosomal protein synthesis or reduced carbon-chain carboxylate intermediates (polyketides or related molecules) suggests a very rich secondary metabolism (Stachelhaus et al. 1995), as found in *Streptomyces* and other saprophytic organisms where these metabolites are presumably used in complex regulatory pathways. While these molecules may have their role in signaling pathways with concomitant adaptation in their hosts, they may play a more novel role in modulating the immune responses of the host and hence have a crucial role in pathogenesis. Important examples include homologues to lovastatin nonaketide synthase of *Aspergillus terreus* and, more interestingly, cyclosporin synthetase of *Tolypocladium niveum*. Polyketide and the non-ribosomal peptides are two large families of compounds that include many clinically important antimicrobials (e.g. erythromycin, oleandomycin, vancomycin) and immunosuppressants (e.g. cyclosporin, tacrolimus, sirolimus), and cytotoxic agents (e.g. doxorubicin, bleomycin, epothilones). Most of these compounds, including the macrolides, possess immunomodulating (predominantly immunosuppressive) or cytotoxic activities. Cyclosporin A, for example, is noted for its T-cell immunosuppressive activities and hence is clinically important for various anti-rejection treatments in transplant recipients. It is well known that many intracellular pathogens (e.g. *Leishmania* species) actively modulate host cytokine production and/or Th1/Th2 cellular immune responses to enhance their survival (Alexander et al. 1999). It would, therefore, not be surprising if *P. marneffei* also utilizes similar strategies to facilitate its persistence inside susceptible hosts.

Finally, there are a large number of membrane transport proteins in *P. marneffei*. Among these is fluconazole resistance protein 1, encoded by the *FLU1* gene. Fluconazole resistance protein is a member of the major facilitator superfamily of multidrug efflux transporter. This is in accord with the observed fluconazole resistance of *P. marneffei* (Imwidthaya et al. 2001).

Acknowledgements This work was supported by the collaboration between The University of Hong Kong and the Institute Pasteur and was partly funded by the University Development Fund, Research Grants Council, and AIDS Trust Fund, Hong Kong.

References

Alexander J, Satoskar AR, Russell DG (1999) Leishmania species: models of intracellular parasitism. J Cell Sci 112:2993–3002

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Barzu, O, Danchin A (1994) Adenylyl cyclases: a heterogeneous class of ATP-utilizing enzymes. Prog Nucleic Acid Res Mol Biol 49:241–283

Borneman AR, Hynes MJ, Andrianopoulos A (2001) An STE12 homolog from the asexual, dimorphic fungus *Penicillium marneffei* complements the defect in sexual development of an *Aspergillus nidulans* steA mutant. Genetics 157:1003–1014

Cao L, Chan CM, Lee C, Wong SS, Yuen KY (1998) *MP1* encodes an abundant and highly antigenic cell wall mannoprotein in the pathogenic fungus *Penicillium marneffei*. Infect. Immun 66:966–973

Cao L, Chen DL, Lee C, Chan CM, Chan KM, Vanittanakom N, Tsang DN, Yuen KY (1998) Detection of specific antibodies to an antigenic mannoprotein for diagnosis of *Penicillium marneffei* penicilliosis. J Clin Microbiol 36:3028–3031

Cao L, Chan KM, Chen D, Vanittanakom N, Lee C, Chan CM, Sirisanthana T, Tsang DN, Yuen KY (1999) Detection of cell wall mannoprotein Mp1p in culture supernatants of *Penicillium marneffei* and in sera of penicilliosis patients. J Clin Microbiol 37:981–986

Chariyalertsak S, Vanittanakom P, Nelson KE, Sirisanthana T, Vanittanakom N (1996) *Rhizomys sumatrensis* and *Cannomys badius*, new natural animal hosts of *Penicillium marneffei*. J Med Vet Mycol 34:105–110

Chariyalertsak S, Sirisanthana T, Supparatpinyo K, Praparattanapan J, Nelson KE (1997) Case-control study of risk factors for *Penicillium marneffei* infection in human immunodeficiency virus-infected patients in northern Thailand. Clin Infect Dis 24:1080–1086

Chavez R, Fierro F, Gordillo F, Francisco Martin J, Eyzaguirre J (2001) Electrophoretic karyotype of the filamentous fungus *Penicillium purpurogenum* and chromosomal location of several xylanolytic genes. FEMS Microbiol Lett 205:379–383

Chim CS, Fong CY, Ma SK, Wong SS, Yuen KY (1998) Reactive hemophagocytic syndrome associated with *Penicillium marneffei* infection. Am J Med 104:196–197

Cimon, B, Carrere J, Chazalette JP, Vinatier JF, Chabasse D, Bouchara JP (1999) Chronic airway colonization by *Penicillium emersonii* in a patient with cystic fibrosis. Med Mycol 37: 291–293

Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D, Mungall D, Basham D, Brown D, Chillingworth T, Connor R, Davies RM, Devlin K, Duthoy S, Feltwell T, Fraser A, Hamlin N, Holroyd S, Hornsby T, Jagels K, Lacroix C, Maclean J, Moule S, Murphy L, Oliver K, Quail MA, Rajandream MA, Rutherford KM, Rutter S, Seeger K, Simon S, Simmonds M, Skelton J, Squares R, Squares S, Stevens K, Taylor K, Whitehead S, Woodward JR, and Barrell BG (2001) Massive gene decay in the leprosy bacillus. Nature 409:1007–1011

Danchin A (1989) Homeotopic transformation and the origin of translation. Prog Biophys Mol Biol 54:81–86

Deng ZL, Connor DH (1985) Progressive disseminated penicilliosis caused by *Penicillium marneffei*. Report of eight cases and differentiation of the causative organism from *Histoplasma capsulatum*. Am J Clin Pathol 84:323–327

Espitia C, Laclette JP, Mondragon-Palomino M, Amador A, Campuzano J, Martens A, Singh M, Cicero R, Zhang Y, Moreno C (1999) The PE-PGRS glycine-rich proteins of *Mycobacterium tuberculosis*: a new family of fibronectin-binding proteins? Microbiology 145:3487–3495

Farber P, Geisen R (2000) Karyotype of *Penicillium nalgiovense* and assignment of the penicillin biosynthetic genes to chromosome IV. Int J Food Microbiol 58:59–63

Fierro F, Gutierrez S, Diez B, Martin JF (1993) Resolution of four large chromosomes in penicillin-producin filamentous fungi: the penicillin gene cluster is located on chromosome II (9.6 Mbp) in *Penicillium notatum* and chromosome I (10.4 Mbp) in *Penicillium chrysogenum*. Mol Gen Genet 241:573–578

Frangeul L, Nelson KE, Buchrieser C, Danchin A, Glaser P, Kunst F (1999) Cloning and assembly strategies in microbial genome projects. Microbiology 145:2625–2634

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM (1995) The minimal gene complement of *Mycoplasma genitalium*. Science 270:397–403

Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, Gwinn M, Hickey EK, Clayton R, Ketchum KA, Sodergren E, Hardham JM, McLeod MP, Salzberg S, Peterson J, Khalak H, Richardson D, Howell JK, Chidambaram M, Utterback T, McDonald L, Artiach P, Bowman C, Cotton MD, Venter JC (1998) Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. Science 281:375–388

Glemet E, Codani JJ (1997) LASSAP, a LArge Scale Sequence compArison Package. Comput Appl Biosci 13:137–143

Hamilton AJ, Jeavons L, Youngchim S, Vanittanakom N, Hay RJ (1998) Sialic acid-dependent recognition of laminin by *Penicillium marneffei* conidia. Infect Immun 66:6024–6026

Hamilton AJ, Jeavons L, Youngchim S, Vanittanakom N (1999) Recognition of fibronectin by *Penicillium marneffei* conidia via a sialic acid-dependent process and its relationship to the interaction between conidia and laminin. Infect Immun 67:5200–5205

Himmelreich R., Hilbert H, Plagens H, Pirkl E, Li BC, Herrmann R (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. Nucleic Acids Res 24:4420–4449

Imwidthaya P, Thipsuvan K, Chaiprasert A, Danchaivijitra S, Sutthent R, Jearanaisilavong J (2001) *Penicillium marneffei*: types and drug susceptibility. Mycopathologia 149:109–115

Kao LR, Peterson J, Ji R, Bender L, Bender A (1996) Interactions between the ankyrin repeat-containing protein Akr1p and the pheromone response pathway in *Saccharomyces cerevisiae*. Mol Cell Biol 16:168–178

Kappe R., Fauser C, Okeke CN, Maiwald M (1996) Universal fungus-specific primer systems and group-specific hybridization oligonucleotides for 18S rDNA. Mycoses 39:25–30

Kayser T, Schulz G (1991) Electrophoretic karyotype of cellulolytic *Penicillium janthinellum* strains. Curr Genet 20:289–291

Kondrashov AS (1994) Muller' ratchet under epistatic selection. Genetics 136:1469–1473

Kwan EY, Lau YL, Yuen KY, Jones BM, Low LC (1997) Penicillium marneffei infection in a non-HIV infected child. J Paediatr Child Health 33:267–271

Lo CY, Chan DT, Yuen KY, Li FK, Cheng KP (1995) Penicillium marneffei infection in a patient with SLE. Lupus 4:229–231

LoBuglio KF, Taylor JW (1995) Phylogeny and PCR identification of the human pathogenic fungus *Penicillium marneffei*. J Clin Microbiol 33:85–89

Pryciak PM, Hartwell LH (1996) *AKR1* encodes a candidate effector of the G βγ complex in the *Saccharomyces cerevisiae* pheromone response pathway and contributes to control of both cell shape and signal transduction. Mol Cell Biol 16: 2614–2626

Stachelhaus T, Marahiel MA (1995) Modular structure of peptide synthetases revealed by dissection of the multifunctional enzyme GrsA. J Biol Chem 270:6163–6169

Stoesser G, Baker W, Van den Broek A, Camon E, Garcia-Pastor M, Kanz C, Kulikova T, Lombard V, Lopez R, Parkinson H, Redaschi N, Sterk P, Stoehr P, Tuli MA (2001) The EMBL nucleotide sequence database. Nucleic Acids Res 29:17–21

Supparatpinyo K, Khamwan C, Baosoung V, Nelson KE, Sirisanthana T (1994) Disseminated *Penicillium marneffei* infection in southeast Asia. Lancet 344:110–113

Udagawa S, Furuya K, Horie Y (1973) Mycological reports from New Guinea and the Solomon Islands (compiled by Y. Kobayasi). 19. Notes on some ascomycetous microfungi from soil. Bull. Natl Sci Mus Tokyo 16:503–520

Vanittanakom N, Merz WG, Sittisombut N, Khamwan C, Nelson KE, Sirisanthana T (1998) Specific identification of *Penicil- lium marneffei* by a polymerase chain reaction/hybridization technique. Med Mycol 36:169–175

Verweij PE, Meis JF, Van den Hurk P, Zoll J, Samson RA, Melchers WJ (1995) Phylogenetic relationships of five species of *Aspergillus* and related taxa as deduced by comparison of sequences of small subunit ribosomal RNA. J Med Vet Mycol 33:185–190

Wong KH, Lee SS (1998) Comparing the first and second hundred AIDS cases in Hong Kong. Singapore Med J 39:236–240

Wong LP, Woo PC, Wu, AY, Yuen KY (2002) DNA immunization using a secreted cell wall antigen Mp1p is protective against Penicillium marneffei infection. Vaccine 20:2878–2886

Wong SS, Siau H, Yuen KY (1999) Penicilliosis marneffei—West meets East. J Med Microbiol 48:973–975

Wong SS, Ho TY, Ngan AH, Woo PC, Que TL, Yuen KY (2001a) Biotyping of *Penicillium marneffei* reveals concentration-dependent growth inhibition by galactose. J Clin Microbiol 39: 1416–1421

Wong SS, Wong KH, Hui WT, Lee SS, Lo JY, Cao L, Yuen KY (2001b) Differences in clinical and laboratory diagnostic characteristics of penicilliosis marneffei in human immunodeficiency virus (HIV)- and non-HIV-infected patients. J Clin Microbiol 39:4535–4540

Wong SS, Woo PC, Yuen KY (2001c) *Candida tropicalis* and *Penicillium marneffei* mixed fungaemia in a patient with Waldenstrom's macroglobulinaemia. Eur J Clin Microbiol Infect Dis 20:132–135

Wu S, Guo N, Yin Z, Chai J. (1996) Characterization of pathogenic fungi genomes using pulsed field gel electrophoresis. Chin Med Sci J 11:188–190

Young C, Itoh Y, Johnson R, Garthwaie I, Miles CO, Munday-Finch SC, Scott B (1998) Paxilline-negative mutants of *Penicillium paxilli* generated by heterologous and homologous plasmid integration. Curr Genet 33:368–377

Yuen KY, Wong SS, Tsang DN, Chau PY (1994) Serodiagnosis of *Penicillium marneffei* infection. Lancet 344:444–445

G Conclusion et perspectives

Suite à cette étude des biais compositionnels en acides aminés des protéines procaryotes, des conclusions peuvent être tirées et l'utilisation des analyses multivariées dans de nouveaux cadres, envisagées.

Tout d'abord, malgré de grandes distances phylogéniques, d'importantes différences phénotypiques, génétiques et morphologiques et l'extrême diversité des environnements de croissance, les procaryotes présentent de nombreux points communs au niveau de la composition en acides aminés de leur protéome. On pense bien évidement à l'universalité de composition des protéines intégrées à la membrane interne des cellules, mais également à l'étonnante récurrence des biais dus à l'aromaticité des protéines orphelines ou encore aux résidus issus des codons riches en adénosine (AAN).

Néanmoins, les pressions de sélections intrinsèques et extrinsèques influent tant sur la survie des organismes qu'elles transparaissent parfois de manière spécifique dans la composition en acides aminés des protéines. Ceci a été particulièrement observé lors de l'étude de la bactérie psychrophile *P. haloplanktis* dont le protéome présente un biais en asparagine typique des bactéries du froid. Ce biais révèle en effet que les basses températures autorisent la présence plus fréquente de cet acide aminé sans risquer un vieillissement trop prématuré des protéines, comme peuvent le subir celles d'organismes mésophiles et thermophiles.

Dans la lignée des travaux entrepris, il sera intéressant d'entreprendre des études non plus sur des protéomes entiers mais sur des familles de protéines particulières d'un seul ou de plusieurs procaryotes différents (i.e. protéines liées à la pathogénicité, à la symbiose, à l'aérobie). Ceci permettra de caractériser au plus près les différences au sein d'un même groupe fonctionnel. D'autre part, à l'image de ce qui a été réalisé dans les travaux décrits dans l'article I (suppression de l'acide aminé cystéine dans l'AFC pour mettre en évidence d'autres biais que le biais cystéine des protéines de *M. jannaschii)*, les différents biais pourraient être retirés les uns après les autres (suppression des IIMPs ou encore des résidus aromatiques) pour en laisser apparaître de nouveaux. Même potentiellement moins forts que ceux déjà mis en lumière, la description de ces nouveaux biais permettrait de compléter les résultats de nos investigations. A l'inverse, les groupes de protéines d'ores et déjà identifiés (la famille de protéines PPE, PE, PE-PGRS de *Mycobacterium tuberculosis*, ou les protéines exportées de *Ralstonia solanacearum*) pourraient être analysés individuellement dans une AFC « débarrassée » du reste du protéome. Ce genre d'approche permettrait sans doute une caractérisation plus fine des protéines appartenant à ces groupes. Enfin, une autre voie d'exploration consisterait à coupler l'AFC et des analyses phylogéniques, en étudiant par exemple les bactéries d'une branche phylogénique donnée et tenter de mettre en évidence les caractéristiques communes ou spécifiques de chaque protéome (analyses des α-protéobactéries, des mycoplasmes). Par conséquent, même si un nombre important de propriétés des protéines ont déjà été décelées, de nombreuses autres restent très certainement à découvrir.

H Références

Akashi, H. et T. Gojobori (2002). "Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis." Proc Natl Acad Sci U S A **99**(6): 3695-700.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers et D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-10.

Antonio Yáñez, L. C., Olivier Neyrolles, Encarnación Alonso, Marie-Christine Prévost, Jorge Rojas, Harold L. Watson, Alain Blanchard, and Gail H. Cassell (1996). Mycoplasma penetrans Bacteremia and Primary Antiphospholipid Syndrome. the 11th International Congress of the International Organization for Mycoplasmology., Orlando, FL, USA.

Benzecri, J.-P. (1973). L'analyse des données, L'Analyse des Correspondances. Paris, France, Dunod Edition.

Blake, R. D. et P. W. Hinds (1984). "Analysis of the codon bias in *E. coli* sequences." J Biomol Struct Dyn **2**(3): 593-606.

Bocs, S., S. Cruveiller, D. Vallenet, G. Nuel et C. Medigue (2003). "AMIGene: Annotation of MIcrobial Genes." Nucleic Acids Res **31**(13): 3723-6.

Braun, V. et M. Braun (2002). "Iron transport and signaling in Escherichia coli." FEBS Lett **529**(1): 78-85.

Brennan, M. J. et G. Delogu (2002). "The PE multigene family: a 'molecular mantra' for mycobacteria." Trends Microbiol **10**(5): 246-9.

Brooks, D. J. et J. R. Fresco (2002). "Increased frequency of cysteine, tyrosine, and phenylalanine residues since the last universal ancestor." Mol Cell Proteomics **1**(2): 125-31.

Brooks, D. J. et J. R. Fresco (2003). "Greater GNN pattern bias in sequence elements encoding conserved residues of ancient proteins may be an indicator of amino acid composition of early proteins." Gene **303**: 177-85.

Buchanan, S. K., B. S. Smith, L. Venkatramani, D. Xia, L. Esser, M. Palnitkar, R. Chakraborty, D. van der Helm et J. Deisenhofer (1999). "Crystal structure of the outer membrane active transporter FepA from Escherichia coli." Nat Struct Biol **6**(1): 56-63.

Burley, S. K. et G. A. Petsko (1985). "Aromatic-aromatic interaction: a mechanism of protein structure stabilization." Science **229**(4708): 23-8.

Cao, L., C. M. Chan, C. Lee, S. S. Wong et K. Y. Yuen (1998). "MP1 encodes an abundant and highly antigenic cell wall mannoprotein in the pathogenic fungus Penicillium marneffei." Infect Immun **66**(3): 966-73.

Cao, L., D. L. Chen, C. Lee, C. M. Chan, K. M. Chan, N. Vanittanakom, D. N. Tsang et K. Y. Yuen (1998). "Detection of specific antibodies to an antigenic mannoprotein for diagnosis of Penicillium marneffei penicilliosis." J Clin Microbiol **36**(10): 3028-31.

Carbone, A., F. Kepes et A. Zinovyev (2005). "Codon bias signatures, organization of microorganisms in codon space, and lifestyle." Mol Biol Evol **22**(3): 547-61.

Clark, E. H., J. M. East et A. G. Lee (2003). "The role of tryptophan residues in an integral membrane protein: diacylglycerol kinase." Biochemistry **42**(37): 11065-73.

Costerton, J. W., P. S. Stewart et E. P. Greenberg (1999). "Bacterial biofilms: a common cause of persistent infections." Science **284**(5418): 1318-22.

Daniel, R. M., M. Dines et H. H. Petach (1996). "The denaturation and degradation of stable enzymes at high temperatures." Biochem J **317 (Pt 1)**: 1-11.

Daubin, V. et H. Ochman (2004). "Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli." Genome Res **14**(6): 1036-42.

David, P. S., P. S. Dutt, B. Wathen, Z. Jia et B. C. Hill (2000). "Characterization of a structural model of membrane bound cytochrome c-550 from Bacillus subtilis." Arch Biochem Biophys **377**(1): 22-30.

Decho, A. W. (1990). Microbial exopolymer secretions in ocean environments: their role(s) in food webs and marine processes. Oceanography and marine biology annual Review. M. Barnes. Aberdeen, Aberdeen University Press**:** 73-153.

Delorme, M. O. et A. Hénaut (1988). "Merging of distance matrices and classification by dynamic clustering." Comput Appl Biosci **4**(4): 453-8.

Deng, Z. L. et D. H. Connor (1985). "Progressive disseminated penicilliosis caused by Penicillium marneffei. Report of eight cases and differentiation of the causative organism from Histoplasma capsulatum." Am J Clin Pathol **84**(3): 323-7.

Diday, E. (1971). "Une nouvelle méthode en classification automatique et reconnaissance des formes: la méthode des nuées dynamiques." Rev. Stat. Appliquée **19**(2): 19-33.

Dresios, J., Y. L. Chan et I. G. Wool (2002). "The role of the zinc finger motif and of the residues at the amino terminus in the function of yeast ribosomal protein YL37a." J Mol Biol **316**(3): 475-88.

Fang, G., E. Rocha et A. Danchin (2005). "How Essential Are Nonessential Genes?" Mol Biol Evol **22**(11): 2147-2156.

Farias, S. T. et M. C. Bonato (2003). "Preferred amino acids and thermostability." Genet Mol Res **2**(4): 383-93.

Feller, G. et C. Gerday (2003). "Psychrophilic enzymes: hot topics in cold adaptation." Nat Rev Microbiol **1**(3): 200-8.

ffrench-Constant, R., N. Waterfield, P. Daborn, S. Joyce, H. Bennett, C. Au, A. Dowling, S. Boundy, S. Reynolds et D. Clarke (2003). "Photorhabdus: towards a functional genomic analysis of a symbiont and pathogen." FEMS Microbiol Rev **26**(5): 433-56.

Forst, S. et D. Clarke (2002). Bacteria-nematode symbiosis. Entomopathogenic Nematology. R. Gaugler. London, CAB International**:** 57-77.

Francino, M. P. et H. Ochman (1997). "Strand asymmetries in DNA evolution." Trends Genet **13**(6): 240-5.

Frank, A. C. et J. R. Lobry (1999). "Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms." Gene **238**(1): 65-77.

Fuchs, T. M., B. Schneider, K. Krumbach, L. Eggeling et R. Gross (2000). "Characterization of a bordetella pertussis diaminopimelate (DAP) biosynthesis locus identifies dapC, a novel gene coding for an N-succinyl-L,L-DAP aminotransferase." J Bacteriol **182**(13): 3626-31.

Fukuchi, S. et K. Nishikawa (2001). "Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria." J Mol Biol **309**(4): 835-43.

Georgis, R. et G. J. Poinar (1990). Field effectiveness of entomophilic nematodes neoaplectana and heterorhabditis. Integrated pest management for turfgrass and ornementals. A. R. Leslie et R. L. Metcalf. Washington D.C., United States Environmental Protection Agency.

Gouy, M. et C. Gautier (1982). "Codon usage in bacteria: correlation with gene expressivity." Nucleic Acids Res **10**(22): 7055-74.

Grantham, R., C. Gautier et M. Gouy (1980). "Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type." Nucleic Acids Res **8**(9): 1893-912.

Grantham, R., C. Gautier, M. Gouy, M. Jacobzone et R. Mercier (1981). "Codon catalog usage is a genome strategy modulated for gene expressivity." Nucleic Acids Res **9**(1): r43-74.

Grantham, R., C. Gautier, M. Gouy, R. Mercier et A. Pave (1980). "Codon catalog usage and the genome hypothesis." Nucleic Acids Res **8**(1): r49-r62.

Guerdoux-Jamet, P., A. Henaut, P. Nitschke, J. L. Risler et A. Danchin (1997). "Using codon usage to predict genes origin: is the Escherichia coli outer membrane a patchwork of products from different genomes?" DNA Res **4**(4): 257-65.

Guerinot, M. L. (1994). "Microbial iron transport." Annu Rev Microbiol **48**: 743-72.

Gupta, S. K. et T. C. Ghosh (2001). "Gene expressivity is the main factor in dictating the codon usage variation among the genes in Pseudomonas aeruginosa." Gene **273**(1): 63-70.

Haft, D. H., J. D. Selengut et O. White (2003). "The TIGRFAMs database of protein families." Nucleic Acids Res **31**(1): 371-3.

Handley, P. S., F. F. Correia, K. Russell, B. Rosan et J. M. DiRienzo (2005). "Association of a novel high molecular weight, serine-rich protein (SrpA) with fibril-mediated adhesion of the oral biofilm bacterium Streptococcus cristatus." Oral Microbiol Immunol **20**(3): 131-40.

Hartmann, M., A. Tauch, L. Eggeling, B. Bathe, B. Mockel, A. Puhler et J. Kalinowski (2003). "Identification and characterization of the last two unknown genes, dapC and dapF, in the succinylase branch of the L-lysine biosynthesis of Corynebacterium glutamicum." J Biotechnol **104**(1-3): 199-211.

Hill, M. O. (1974). "Correspondence analysis: a neglected multivariate method." Applied Statistics **23**(3): 340-54.

Huang, S. L., L. C. Wu, H. K. Liang, K. T. Pan, J. T. Horng et M. T. Ko (2004). "PGTdb: a database providing growth temperatures of prokaryotes." Bioinformatics **20**(2): 276-8.

Hunter, C. A., J. Singh et J. M. Thornton (1991). "Pi-pi interactions: the geometry and energetics of phenylalanine-phenylalanine interactions in proteins." J Mol Biol **218**(4): 837-46.

Jordan, I. K., F. A. Kondrashov, I. A. Adzhubei, Y. I. Wolf, E. V. Koonin, A. S. Kondrashov et S. Sunyaev (2005). "A universal trend of amino acid gain and loss in protein evolution." Nature **433**(7026): 633-8.

Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno et M. Hattori (2004). "The KEGG resource for deciphering the genome." Nucleic Acids Res **32**(Database issue): D277-80.

Kanz, C., P. Aldebert, N. Althorpe, W. Baker, A. Baldwin, K. Bates, P. Browne, A. van den Broek, M. Castro, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, J. Gamble, F. G. Diez, N. Harte, T. Kulikova, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, M. McHale, F. Nardone, V.

Silventoinen, S. Sobhany, P. Stoehr, M. A. Tuli, K. Tzouvara, R. Vaughan, D. Wu, W. Zhu et R. Apweiler (2005). "The EMBL Nucleotide Sequence Database." Nucleic Acids Res **33**(Database issue): D29-33.

Karlin, S. et J. Mrazek (2000). "Predicted highly expressed genes of diverse prokaryotic genomes." J Bacteriol **182**(18): 5238-50.

Karlin, S., J. Mrazek et A. M. Campbell (1997). "Compositional biases of bacterial genomes and evolutionary implications." J Bacteriol **179**(12): 3899-913.

Karlin, S., J. Mrazek et A. M. Campbell (1998). "Codon usages in different gene classes of the Escherichia coli genome." Mol Microbiol **29**(6): 1341-55.

Knight, R. D., S. J. Freeland et L. F. Landweber (2001). "A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes." Genome Biol **2**(4): research0010.1 - 0010.13.

Koshi, J. M. et W. J. Bruno (1999). "Major structural determinants of transmembrane proteins identified by principal component analysis." Proteins **34**(3): 333-40.

Kreil, D. P. et C. A. Ouzounis (2001). "Identification of thermophilic species by the amino acid compositions deduced from their genomes." Nucleic Acids Res **29**(7): 1608-15.

Krembs, C., H. Eicken, K. Junge et J. W. Deming (2002). "High concentrarions of exopolymeric substances in Arctic winter sea ice: implications for the polar ocean carbon cycle and cryoprotection of diatoms." Deep sea research Part I: Oceanographic Research Papers **49**: 2163-2181.

Krieger, C. J., P. Zhang, L. A. Mueller, A. Wang, S. Paley, M. Arnaud, J. Pick, S. Y. Rhee et P. D. Karp (2004). "MetaCyc: a multiorganism database of metabolic pathways and enzymes." Nucleic Acids Res **32**(Database issue): D438-42.

Krummenacker, M., S. Paley, L. Mueller, T. Yan et P. D. Karp (2005). "Querying and computing with BioCyc databases." Bioinformatics.

La, D., M. Silver, R. C. Edgar et D. R. Livesay (2003). "Using motif-based methods in multiple genome analyses: a case study comparing orthologous mesophilic and thermophilic proteins." Biochemistry **42**(30): 8988-98.

Lafay, B., A. T. Lloyd, M. J. McLean, K. M. Devine, P. M. Sharp et K. H. Wolfe (1999). "Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases." Nucleic Acids Res **27**(7): 1642-9.

Lawrence, J. G. et H. Ochman (2002). "Reconciling the many faces of lateral gene transfer." Trends Microbiol **10**(1): 1-4.

Liang, H. K., C. M. Huang, M. T. Ko et J. K. Hwang (2005). "Amino acid coupling patterns in thermophilic proteins." Proteins **59**(1): 58-63.

Lin, K., Y. Kuang, J. S. Joseph et P. R. Kolatkar (2002). "Conserved codon composition of ribosomal protein coding genes in Escherichia coli, Mycobacterium tuberculosis and Saccharomyces cerevisiae: lessons from supervised machine learning in functional genomics." Nucleic Acids Res **30**(11): 2599-607.

Lindner, H. et W. Helliger (2001). "Age-dependent deamidation of asparagine residues in proteins." Exp Gerontol **36**(9): 1551-63.

Lobry, J. R. (1996). "Asymmetric substitution patterns in the two DNA strands of bacteria." Mol Biol Evol **13**(5): 660-5.

Lobry, J. R. (1997). "Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species." Gene **205**(1-2): 309-16.

Lobry, J. R. et D. Chessel (2003). "Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria." J Appl Genet **44**(2): 235-61.

Lobry, J. R. et C. Gautier (1994). "Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes." Nucleic Acids Res **22**(15): 3174-80.

Lobry, J. R. et J. M. Louarn (2003). "Polarisation of prokaryotic chromosomes." Curr Opin Microbiol **6**(2): 101-8.

LoBuglio, K. F. et J. W. Taylor (1995). "Phylogeny and PCR identification of the human pathogenic fungus Penicillium marneffei." J Clin Microbiol **33**(1): 85-9.

Lynn, D. J., G. A. Singer et D. A. Hickey (2002). "Synonymous codon usage is subject to selection in thermophilic bacteria." Nucleic Acids Res **30**(19): 4272-7.

Mancuso Nichols, C. A., S. Garon, J. P. Bowman, G. Raguenes et J. Guezennec (2004). "Production of exopolysaccharides by Antarctic marine bacterial isolates." J Appl Microbiol **96**(5): 1057-66.

Martinez, J. S., J. N. Carter-Franklin, E. L. Mann, J. D. Martin, M. G. Haygood et A. Butler (2003). "Structure and membrane affinity of a suite of amphiphilic siderophores produced by a marine bacterium." Proc Natl Acad Sci U S A **100**(7): 3754-9.

Marty, P., S. Brun et M. Gari-Toussaint (2000). "[Systemic tropical mycoses]." Med Trop (Mars) **60**(3): 281-90.

Medigue, C., E. Krin, G. Pascal, V. Barbe, A. Bernsel, P. N. Bertin, F. Cheung, S. Cruveiller, S. D'Amico, A. Duilio, G. Fang, G. Feller, C. Ho, S. Mangenot, G. Marino, J. Nilsson, E. Parilli, E. P. Rocha, Z. Rouy, A. Sekowska, M. L. Tutino, D. Vallenet, G. von Heijne et A. Danchin (in press). "Coping with cold: the genome of the versatile marine Antarctica bacterium Pseudoalteromonas haloplanktis TAC125." Genome Res.

Medigue, C., T. Rouxel, P. Vigier, A. Hénaut et A. Danchin (1991). "Evidence for horizontal gene transfer in *Escherichia coli* speciation." J Mol Biol **222**(4): 851-6.

Moeck, G. S. et J. W. Coulton (1998). "TonB-dependent iron acquisition: mechanisms of siderophore-mediated active transport." Mol Microbiol **28**(4): 675-81.

Moszer, I., L. M. Jones, S. Moreira, C. Fabry et A. Danchin (2002). "SubtiList: the reference database for the Bacillus subtilis genome." Nucleic Acids Res **30**(1): 62-5.

Musto, H., H. Naya, A. Zavala, H. Romero, F. Alvarez-Valin et G. Bernardi (2004). "Correlations between genomic GC levels and optimal growth temperatures in prokaryotes." FEBS Lett **573**(1-3): 73-7.

Nahlik, M. S., T. J. Brickman, B. A. Ozenberger et M. A. McIntosh (1989). "Nucleotide sequence and transcriptional organization of the Escherichia coli enterobactin biosynthesis cistrons entB and entA." J Bacteriol **171**(2): 784-90.

Nakashima, H., S. Fukuchi et K. Nishikawa (2003). "Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures." J Biochem (Tokyo) **133**(4): 507-13.

Natale, D. A., U. T. Shankavaram, M. Y. Galperin, Y. I. Wolf, L. Aravind et E. V. Koonin (2000). "Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs)." Genome Biol **1**(5): research0009.1-0009.19.

Naya, H., A. Zavala, H. Romero, H. Rodriguez-Maseda et H. Musto (2004). "Correspondence analysis of amino acid usage within the family Bacillaceae." Biochem Biophys Res Commun **325**(4): 1252-7.

Nolling, J., G. Breton, M. V. Omelchenko, K. S. Makarova, Q. Zeng, R. Gibson, H. M. Lee, J. Dubois, D. Qiu, J. Hitti, Y. I. Wolf, R. L. Tatusov, F. Sabathe, L. Doucette-Stamm, P. Soucaille, M. J. Daly, G. N. Bennett, E. V. Koonin et D. R. Smith (2001). "Genome sequence and comparative analysis of the solvent-producing bacterium Clostridium acetobutylicum." J Bacteriol **183**(16): 4823-38.

Pal, C., B. Papp et L. D. Hurst (2001). "Highly expressed genes in yeast evolve slowly." Genetics **158**(2): 927-31.

Palacios, C. et J. J. Wernegreen (2002). "A strong effect of AT mutational bias on amino acid usage in Buchnera is mitigated at high-expression genes." Mol Biol Evol **19**(9): 1575-84.

Pan, A., C. Dutta et J. Das (1998). "Codon usage in highly expressed genes of Haemophillus influenzae and Mycobacterium tuberculosis: translational selection versus mutational bias." Gene **215**(2): 405-13.

Papp, B., C. Pal et L. D. Hurst (2003). "Dosage sensitivity and the evolution of gene families in yeast." Nature **424**(6945): 194-7.

Pascal, G., C. Medigue et A. Danchin (2005). "Universal biases in protein composition of model prokaryotes." Proteins **60**(1): 27-35.

Patel, K. et R. T. Borchardt (1990). "Chemical pathways of peptide degradation. II. Kinetics of deamidation of an asparaginyl residue in a model hexapeptide." Pharm Res **7**(7): 703-11.

Pedulla, M. L., M. E. Ford, J. M. Houtz, T. Karthikeyan, C. Wadsworth, J. A. Lewis, D. Jacobs-Sera, J. Falbo, J. Gross, N. R. Pannunzio, W. Brucker, V. Kumar, J. Kandasamy, L. Keenan, S. Bardarov, J. Kriakov, J. G. Lawrence, W. R. Jacobs, Jr., R. W. Hendrix et G. F. Hatfull (2003). "Origins of highly mosaic mycobacteriophage genomes." Cell **113**(2): 171-82.

Pe'er, I., C. E. Felder, O. Man, I. Silman, J. L. Sussman et J. S. Beckmann (2004). "Proteomic signatures: amino acid and oligopeptide compositions differentiate among phyla." Proteins **54**(1): 20-40.

Pereto, J., P. Lopez-Garcia et D. Moreira (2004). "Ancestral lipid biosynthesis and early membrane evolution." Trends Biochem Sci **29**(9): 469-77.

Perret, S., A. Belaich, H. P. Fierobe, J. P. Belaich et C. Tardif (2004). "Towards designer cellulosomes in Clostridia: mannanase enrichment of the cellulosomes produced by Clostridium cellulolyticum." J Bacteriol **186**(19): 6544-52.

Perriere, G., J. R. Lobry et J. Thioulouse (1996). "Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acid sequences." Comput Appl Biosci **12**(6): 519-24.

Plummer, C., H. Wu, S. W. Kerrigan, G. Meade, D. Cox et C. W. Ian Douglas (2005). "A serine-rich glycoprotein of Streptococcus sanguis mediates adhesion to platelets via GPIb." Br J Haematol **129**(1): 101-9.

Postle, K. et R. F. Good (1983). "DNA sequence of the Escherichia coli tonB gene." Proc Natl Acad Sci U S A **80**(17): 5235-9.

Pupo, G. M., R. Lan et P. R. Reeves (2000). "Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics." Proc Natl Acad Sci U S A **97**(19): 10567-72.

Rahman, S. A., P. Advani, R. Schunk, R. Schrader et D. Schomburg (2005). "Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC)." Bioinformatics **21**(7): 1189-93.

Robinson, N. E. et A. B. Robinson (2001). "Prediction of protein deamidation rates from primary and three-dimensional structure." Proc Natl Acad Sci U S A **98**(8): 4367-72.

Rocha, E. P. et A. Danchin (2001). "Ongoing evolution of strand composition in bacterial genomes." Mol Biol Evol **18**(9): 1789-99.

Rocha, E. P. et A. Danchin (2002). "Base composition bias might result from competition for metabolic resources." Trends Genet **18**(6): 291-4.

Rocha, E. P. et A. Danchin (2004). "An analysis of determinants of amino acids substitution rates in bacterial proteins." Mol Biol Evol **21**(1): 108-16.

Rocha, E. P., A. Danchin et A. Viari (1999). "Translation in *Bacillus subtilis*: roles and trends of initiation and termination, insights from a genome analysis." Nucleic Acids Res **27**(17): 3567-76.

Rocha, E. P., A. Danchin et A. Viari (1999). "Universal replication biases in bacteria." Mol Microbiol **32**(1): 11-6.

Schalk, I. J., W. W. Yue et S. K. Buchanan (2004). "Recognition of iron-free siderophores by TonB-dependent iron transporters." Mol Microbiol **54**(1): 14-22.

Serres, M. H., S. Goswami et M. Riley (2004). "GenProtEC: an updated and improved analysis of functions of Escherichia coli K-12 proteins." Nucleic Acids Res **32**(Database issue): D300-2.

Serres, M. H. et M. Riley (2000). "MultiFun, a multifunctional classification scheme for Escherichia coli K-12 gene products." Microb Comp Genomics **5**(4): 205-22.

Sharp, P. M. et W. H. Li (1987). "The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications." Nucleic Acids Res **15**(3): 1281-95.

Siboo, I. R., H. F. Chambers et P. M. Sullam (2005). "Role of SraP, a Serine-Rich Surface Protein of Staphylococcus aureus, in binding to human platelets." Infect Immun **73**(4): 2273-80.

Singer, G. A. et D. A. Hickey (2003). "Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content." Gene **317**(1-2): 39-47.

Skinner, L. M. et M. P. Jackson (1997). "Investigation of ribosome binding by the Shiga toxin A1 subunit, using competition and site-directed mutagenesis." J Bacteriol **179**(4): 1368-74.

Sorimachi, K. (1999). "Evolutionary changes reflected by the cellular amino acid composition." Amino Acids **17**(2): 207-26.

Tekaia, F., A. Lazcano et B. Dujon (1999). "The genomic tree as revealed from whole proteome comparisons." Genome Res **9**(6): 550-7.

Tekaia, F., E. Yeramian et B. Dujon (2002). "Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis." Gene **297**(1-2): 51-60.

Thomas, A., R. Meurisse et R. Brasseur (2002). "Aromatic side-chain interactions in proteins. II. Near- and far-sequence Phe-X pairs." Proteins **48**(4): 635-44.

Thomas, A., R. Meurisse, B. Charloteaux et R. Brasseur (2002). "Aromatic side-chain interactions in proteins. I. Main structural features." Proteins **48**(4): 628-34.

Tie, J. K., C. Nicchitta, G. von Heijne et D. W. Stafford (2005). "Membrane topology mapping of vitamin K epoxide reductase by in vitro translation/cotranslocation." J Biol Chem **280**(16): 16410-6.

Ulmschneider, M. B. et M. S. Sansom (2001). "Amino acid distributions in integral membrane protein structures." Biochim Biophys Acta **1512**(1): 1-14.

Van Domselaar, G. H., P. Stothard, S. Shrivastava, J. A. Cruz, A. Guo, X. Dong, P. Lu, D. Szafron, R. Greiner et D. S. Wishart (2005). "BASys: a web server for automated bacterial genome annotation." Nucleic Acids Res **33**(Web Server issue): W455-9.

van Geest, M. et J. S. Lolkema (2000). "Membrane topology and insertion of membrane proteins: search for topogenic signals." Microbiol Mol Biol Rev **64**(1): 13-33.

Wallin, E. et G. von Heijne (1998). "Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms." Protein Sci **7**(4): 1029-38.

White, D. C. (1986). Quantitative physiochemical characterization of bacterial habitats. Methods and special applications in bacterial ecology. J. S. Poindexter et E. R. Leadbetter. New York, Plenum Press. **2:** 177-203.

Wilquet, V. et M. Van de Casteele (1999). "The role of the codon first letter in the relationship between genomic GC content and protein amino acid composition." Res Microbiol **150**(1): 21-32.

Woese, C. R. et G. E. Fox (1977). "Phylogenetic structure of the prokaryotic domain: the primary kingdoms." Proc Natl Acad Sci U S A **74**(11): 5088-90.

Woese, C. R., O. Kandler et M. L. Wheelis (1990). "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya." Proc Natl Acad Sci U S A **87**(12): 4576-9.

Wong, K. H. et S. S. Lee (1998). "Comparing the first and second hundred AIDS cases in Hong Kong." Singapore Med J **39**(6): 236-40.

Wright, H. T. (1991). "Nonenzymatic deamidation of asparaginyl and glutaminyl residues in proteins." Crit Rev Biochem Mol Biol **26**(1): 1-52.

Zavala, A., H. Naya, H. Romero et H. Musto (2002). "Trends in codon and amino acid usage in Thermotoga maritima." J Mol Evol **54**(5): 563-8.

Zhu, K., A. Jutila, E. K. Tuominen, S. A. Patkar, A. Svendsen et P. K. Kinnunen (2001). "Impact of the tryptophan residues of Humicola lanuginosa lipase on its thermal stability." Biochim Biophys Acta **1547**(2): 329-38.

I Annexes

I.I Données supplémentaires extraites de l'article 2.

I.I.1 Tableau supplémentaire 1.

Suppl-Tableau I : Occurrence and percentage of Gene Ontology (GO) found in different clusters in the proteome. GO:0000027 (biological process: ribosomal large subunit assembly and maintenance); GO:000003735 (molecular function: structural constituent of ribosome); GO:005840 (cellular component: ribosome); GO:0007046 (biological process: ribosome biogenesis); GO:0008999 (molecular function: ribosomal-protein-alanine N-acetyltransferase activity); GO:0015934 (cellular component: large ribosomal subunit); GO:0015935 (cellular component: small ribosomal subunit); GO:0042254 (biological process: ribosome biogenesis and assembly).

	cluster	0000027		0003735		0005840		0007046		0008999		0015934		0015935		0042254	
		%	nb	%	nb	%	nb	%	nb	%	nb	%	nb	%	nb	%	nb
A. fulgidus	orange																
	yellow																
	blue			64	27/42	56	15/27	100	1/1			100	4/4	71	5/7	50	1/2
	green			36	15/42	44	12/27							29	2/7	50	1/2
B. bacteriovorus	orange																
	yellow			11	4/36	8	2/24							29	2/7		
	blue			3	1/36					100	2/2			14	1/7		
	green			83	30/36	88	21/24					100	4/4	57	4/7	100	1/1
	pink			3	1/36	4	1/24										
B. burgdorferi	orange																
	yellow			94	30/32	90	19/21					100	5/5	100	6/6	100	1/1
	blue																
	green			6	2/32	10	2/21										
B. halodurans	orange	100	1/1	26	9/35	18	4/22							50	3/6		
	yellow			66	23/35	77	17/22					100	4/4	33	2/6	100	1/1
	blue			6	2/35	5	1/22							17	1/6		
	green			3	1/35												
B. japonicum	orange			3	1/35									17	1/6		
	yellow			20	7/35	22	5/23					20	1/5	17	1/6		
	blue	100	1/1	77	27/35	78	18/23					80	4/5	67	4/6	100	1/1
	green																
C. acetobutylicum	orange			8	3/37									13	1/8		
	yellow	100	1/1	92	34/37	100	19/19					100	4/4	88	7/8	100	1/1
	blue																
	green																
C. jejuni	orange																
	yellow																
	blue	100	1/1	100	32/32	100	21/21					100	5/5	100	6/6	100	1/1
	green																
C. trachomatis	orange	100	1/1	100	32/32	100	22/22					100	4/4	100	6/6	100	1/1
	yellow																
D. radiodurans	orange	100	1/1	91	31/34	100	19/19	100	1/1			100	5/5	100	6/6	100	1/1
	yellow			9	3/34												
	blue																
	green																
	pink																
E. coli O157:H7	orange	100	1/1	97	29/30	95	19/20	100	1/1			100	4/4	100	6/6	100	1/1
	yellow									100	2/2						
	blue			3	1/30	5	1/20										
	green																
F. nucleatum	orange	100	1/1	100	30/30	100	18/18			100	3/3	100	4/4	100	6/6	100	1/1
	yellow																

	cluster	0000027		0003735		0005840		0007046		0008999		0015934		0015935		0042254	
		%	nb	%	nb	%	nb	%	nb	%	nb	%	nb	%	nb	%	nb
H. salinarum	orange			100	36/36	100	23/23					100	3/3	100	7/7	100	2/2
	yellow																
M. kandleri	orange			88	38/43	83	25/30					100	4/4	100	7/7	100	2/2
	yellow			12	5/43	17	5/30	100	2/2								
	blue																
M. penetrans	orange			97	32/33	100	20/20					100	5/5	100	6/6	100	1/1
	yellow			3	1/33												
	blue																
	green																
Nostc	orange			12	4/34	9	2/22					20	1/5	17	1/6		
	yellow	100	1/1	82	28/34	82	18/22					80	4/5	83	5/6	100	1/1
	blue			3	1/34	5	1/22										
	green			3	1/34	5	1/22										
P. abyssi	orange			98	40/41	96	27/28	100	2/2			100	4/4	100	7/7	100	2/2
	yellow			2	1/41	4	1/28										
	blue																
R. solanacearum	orange			6	2/33	5	1/22										
	yellow																
	blue	100	1/1	94	31/33	95	21/22					100	4/4	100	6/6	100	1/1
	green																
S. coelicolor	orange																
	yellow																
	blue			9	3/35	8	2/24					20	1/5				
	green	100	1/1	91	32/35	92	22/24					80	4/5	100	6/6		
Synechocystis	orange			17	6/35	13	3/23					20	1/5	33	2/6		
	yellow			54	19/35	52	12/23					80	4/5	33	2/6	100	1/1
	blue	100	1/1	29	10/35	35	8/23							33	2/6		
	green																
T. acidophilum	orange																
	yellow			100	37/37	100	25/25					100	3/3	100	7/7	100	2/2
	blue																
	green																
T. maritima	orange	100	1/1	90	27/30	89	17/19					100	5/5	83	5/6	100	1/1
	yellow			10	3/30	11	2/19							17	1/6		
	blue																
Y. pestis	orange																
	yellow	100	1/1	16	5/31	10	2/21					25	1/4	33	2/6		
	blue			6	2/31	10	2/21			100	1/1						
	green			77	24/31	81	17/21					75	3/4	67	4/6	100	1/1

I.I.2 *Tableau supplémentaire 2.*

Suppl-Tableau II : CA axes built by Cys and Trp biases

Organism	CA axis	Amino acid bias
Archaea		
A. pernix	3	cysteine
A. fulgidus	2 and 4	cysteine
M. kandleri	4	cysteine
P. abyssi	3	cysteine
T. acidophilum	4	cysteine
Gram positive bacteria		
D. radiodurans	4	cysteine & tryptophan
B. halodurans	4	cysteine
Proteobacteria		
C. jejuni	3	cysteine
Others bacteria		
A. aeolicus	3	cysteine
B. burgdorferi	4	tryptophan
F. nucleatum	4	tryptophan
S. coelicolor	4	tryptophan
T. maritima	3	tryptophan

I.I.3 *Tableau supplémentaire 3.*

Suppl-Tableau III : features of amino acids

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Hydrophobicity																				
Hydrophilic			1	1			1		1			1		1	1					
Hydrophobic	1	1			1			1		1	1							1	1	1
Volume																				
Big					1		1								1				1	1
Medium			1									1	1				1			
Small	1	1				1										1				
Atoms																				
Nitrogen							2		1			1		1	3				1	
Oxygen																1	1			1
Sulphur		1									1									
Carbon	3	3	4	5	9	2	6	6	6	6	5	4	5	5	6	3	4	5	11	9
Lateral chain																				
Aromatic					2		1												2	2
Basic							1		2						3					
Acid			1	1																
Metabolic cost																				
ATP	0	3	1	1	4	1	6	3	2	3	3	2	2	2	6	1	3	3	5	4
NAD(P)H	0	3	0	-1	1	0	-2	1	2	0	6	0	1	-1	0	-1	2	1	1	0
Dayhoff classes																				
astpg	1					1							1			1	1			
mliv								1		1	1							1		
denq			1	1								1		1						
rkh							1		1						1					
fyw					1														1	1
Post-translational modifications																				
Phosphorylation			1				1									1	1			1
Carbamylation									1											
Glycosylation												1								
Methylation			1				1		1											
Others																				
Number of codons	4	2	2	2	2	4	2	3	2	6	1	2	4	2	6	6	4	4	1	2
G+C content	0.83	0.50	0.50	0.50	0.17	0.83	0.50	0.11	0.17	0.39	0.33	0.17	0.83	0.50	0.72	0.50	0.50	0.50	0.67	0.17
Role in aging			1									2								
metabolic steps from pyruvate)	1	8	4	6	11	7	20	12	13	8	11	5	10	7	12	6	9	4	14	11

I.II Article 6.

MaGe - a microbial genome annotation system supported by synteny results

David Vallenet, Laurent Labarre, Zoé Rouy, Valérie Barbe[1], Stéphanie Bocs, Stéphane Cruveiller, Aurélie Lajus, Géraldine Pascal, Claude Scarpelli[1] and Claudine Médigue.

Atelier de Génomique Comparative, CNRS-UMR8030 2 rue Gaston Crémieux, 91057 Evry, Cedex, France and [1]Genoscope and CNRS-UMR8030 2 rue Gaston Crémieux, 91057 Evry, Cedex, France.

Keywords: annotation system, microbial genome, gene context, synteny, metabolic pathway

ABSTRACT

Magnifying Genomes (MaGe) is a microbial genome annotation system based on a relational database containing information on bacterial genomes, as well as a web interface to achieve genome annotation projects. Our system allows one to initiate the annotation of a genome at the early stage of the finishing phase. MaGe's main features are: (i) integration of annotation data from bacterial genomes enhanced by a gene coding re-annotation process using accurate gene models, (ii) integration of results obtained with a wide range of bioinformatics methods, among which exploration of gene context by searching for conserved synteny and reconstruction of metabolic pathways, (iii) an advanced web interface allowing multiple users to refine the automatic assignment of gene product functions. MaGe is also linked to numerous well-known biological databases and systems. Our system has been thoroughly tested during the annotation of complete bacterial genomes (*Acinetobacter ADP1*, *Pseudoalteromonas haloplanktis*, *Frankia alni*) and is currently used in the context of several new microbial genome annotation projects. In addition, MaGe allows for annotation curation and exploration of already published genomes from various genera (*e.g.* Yersinia, Bacillus, and Neisseria). MaGe can be accessed at: http://www.genoscope.cns.fr/agc/mage.

INTRODUCTION

During the last few years, the genomes of about two hundred and fifty bacteria have been completely sequenced, leading to an enormous demand for fast and accurate analysis of the resulting biological sequences. The information obtained from a genome depends largely on the quality of the annotation of its complete sequence (mainly, CoDing Sequence (CDS) identification and function prediction). The quality of the annotation itself depends on the implemented bioinformatics tools and on the work and time dedicated to it by the annotators (1). A common annotation process starts with the use of highly automatic prediction of genes and biological functions of their product. Because of the large number of genomes currently annotated, part of this data from automatic methods is often directly stored in public databanks, leading to the propagation of existing annotation errors (2). Actually, validation of the first set of automatic data involves tedious manual work in which an expert performs additional searches and analyses. The first steps of such annotation processes obviously require powerful automatic tools. The recent publication of BaSys, a web software which permits a complete automatic annotation process for a new bacterial genome, highlights this need (Van Domselaar, Stothard et al. 2005). In addition, databases for storage and management of heterogeneous data, together with complex but user-friendly interfaces are also essential to manually annotate a genome efficiently.

To achieve the annotation of a complete genome, a number of annotation tools have been designed, with the first, strictly automatic systems focusing on human readable HTML reports (4-6). Since then, many efforts have been made in terms of project management (*i.e,* complex biological data models and integrated databases), spectrum of bioinformatics tools applied (including multiple genome comparison-based annotation strategies), sophistication of the user interfaces (extensive visualizations, fully interactive graphical interfaces), and the presence of convenient features such as data editors. Examples of commonly used annotation platforms are given by commercial systems such as ERGO (7) or Pedant-Pro (successor of PEDANT), and open source systems such as Artemis (8), GenDB (9) or Manatee (TIGR, unpublished). In addition, some newly-developed tools perform automatic tasks for contig-assembly analysis together with automatic annotation of the successive assembly updates, thus allowing annotation of a genome to start during the finishing phase of the sequencing process (10, 11).

In the study of microbial genomes, the increasing number and the diversity of sequenced genomes have led to the development of novel methods for the contextual analysis of genes and proteins, to detect functional constraints on genome evolution (12-15). Although results from these methods clearly demonstrate the added-value of genomic context analysis in the process of prokaryotic genome annotation (16), no existing annotation systems, except perhaps ERGO, systematically integrates them. To address this problem, we have developed a new microbial genome annotation system, called MaGe (Magnifying Genomes), which shares several functionalities with existing systems, mainly (i) an automatic annotation process including syntactic and

functional annotations together with classification inferences, (ii) a relational database used to store the sequences and the analysis results, (iii) a WEB interface allowing multiple users to simultaneously annotate a genome and to query the database (*e.g.* search for functionalities and/or gene content between related species), and (iv) several connectivities and/or integration of other systems and databases. Since MaGe has been developed by people who are involved in manual expert annotation themselves, it offers original features such as a graphical gene context exploration. In order to detect gene groups that share locally conserved chromosomal organization, the annotated genome is compared to publicly available other ones. Synteny map visualization is then useful to quickly pinpoint genome rearrangements between related bacterial species. In addition, a customizable user-friendly gene editor has been developed to take into account the specificities of each bacterial annotation project (functional classifications, comparisons to reference genome data, etc). In the context of the expert validation of the automatic predictions, the MaGe cartographic representations have already been shown to improve notably the final annotation quality (17, 18). Our system is currently being used for the annotation or re-annotation of more than sixteen microbial genomes (http://www.genoscope.cns.fr/agc/mage).

The MaGe system consists of three main components which are described in the following sections: (i) a set of bioinformatics methods currently implemented in the system, (ii) a relational database which contains sequence data and the results from the set of analysis methods and (iii) a graphical web interface. The setup and the management of a new annotation project are described in the last section of this paper.

BIOINFORMATICS METHODS

Prediction of genes and signals

The annotation process begins with the FASTA formatted contig(s) on which appropriate algorithms for identification of coding regions and various genetic elements are first applied. A preliminary and essential step for a new genome annotation (anonymous DNA sequence) consists in construction of appropriate gene models, *i.e.* one or several Markov models that fit well with the input genomic data. For this purpose, the longest ORFs are first retained to derive parameters of the Markov model of the protein-coding region. This pre-matrix is used in the AMIGene program for predicting coding regions in the original genomic sequence(s) (19). AMIGene is a gene finding program highly similar to GeneMark in its sequence parsing step for prediction of CDSs (20). In a second step, the first set of predicted CDSs is used to investigate codon usage differences running the multivariate statistical technique of correspondence analysis (21). Identification of major trends within the data set (*i.e*, several major classes of CDSs differing in their codon bias) requires use of a method that automatically clusters the objects that are located close to one another (22). The CDS classes define the training sets for protein-coding regions and the rest of the sequence is included in the noncoding training set. Corresponding gene models (generally 1 to 4 Markov models) are then generated and subsequently used in the core of AMIGene. This preliminary procedure which consists of the construction of gene models fitting well with the input genomic data, leads to more accurate prediction of small genes and/or atypical gene composition. In order to increase the reliability of the AMIGene results in terms of start codon positions, we have integrated the RBSfinder program into our software, which searches for ribosome-binding sites in the extragenic regions (23). tRNAscan-SE (24) has also been included in the annotation pipeline for the prediction of tRNA-encoding genes. In addition, other RNA structures, such as small RNAs and riboswitches, are identified using the Rfam database (25). We have also integrated the Petrin program (26) to identify putative rho-independent transcription termination sites. Other genetic elements such as intrachromosomal repeats are detected using the method described in Achaz *et al.* (27).

Functional annotation tools

Extracted gene products are subjected to exhaustive bioinformatics analysis, including the gapped blastP algorithm (28) for general-purpose homology searches against the full non-redundant protein sequence databank UniProt (29). Queries are also submitted to more sensitive sequence similarity search tools, using motif/pattern/protein families compiled in the InterPro database (30) and the COG databank (31). In addition, genes coding for enzymes are classified using the PRIAM software (32), the results of which are used for metabolic pathway reconstruction (see below). Finally, functional assignations are also made using the HAMAP (High-quality Automated and Manual Annotation of Microbial Proteomes) web server (http://www.expasy.org/sprot/hamap). Based on the curated proteins from the Swiss-Prot databank, the HAMAP project aims to

develop an automatic process for a high-quality annotation of microbial proteomes (33). For each well-defined (sub)family, a rule system describes the level and extent of annotations that can be assigned by similarity with a prototype manually-annotated entry. In terms of predicted structural features, we search for alpha-helical transmembrane regions with the tmHMM program (34), and signal peptides with SignalP (35). Highly sensitive comparison of each predicted protein with the SCOP database of known structural domains (36,37) is also carried out. To predict probable subcellular localization of the annotated protein in the cell, another original approach has been developed in our group. We used correspondence analysis (21) to identify the major factors that shape variation in amino acid usage among proteins of the organism of interest. The distribution of proteins on the correspondence analysis space, joined with clustering method (22), shows systematically and prominently well-separated groups as individual clusters. It was proven that one of them was exclusively constituted of Integral Inner Membrane Proteins (IIMPs) (38). This method brings about a novel feature of a particular consistent class of inner membrane proteins that have key cellular roles, while they can be distinguished from all others membrane protein types.

Finally, along with the fast growing number of sequenced prokaryotic genomes, an additional method that relies on gene context rather than on sequence similarity only, has been developed in our group: synteny computation, which is undoubtedly one of the most original components of the MaGe system.

Comparative genomics through synteny analysis

For assigning function to novel proteins, gene context approaches can complement the classical homology-based gene annotation. These 'nonhomology-based' inference methods rely on the fact that functionally associated proteins are encoded by genes that share similar selection pressures. In most of the proposed methods (31,39), orthologous pairs of proteins satisfy the *bi-directionnal best hit* (BBH) criterion, based on Smith-Waterman (40) comparisons of complete genomes with one another. An innovative aspect of our approach is that we offer the possibility of retaining more than one homologous gene. Pairwise comparisons between predicted protein sequences of the studied genome and the proteins of another genome, allow computation of ranked hits and BBH (for each protein, the three best hits are kept). Putative orthologous relations between two genomes are defined as gene couples satisfying the BBH criterion or an alignment threshold (generally, a minimum of 30% sequence identity on 80% of the length of the smallest protein). These relations are subsequently used to search for conserved gene clusters, *e.g.* synteny groups among several bacterial genomes(41)**.** Our method, called the Syntonizer (unpublished), allows for multiple correspondences between genes and thus, paralogy relations and/or gene fusions are easily detected. All possible kinds of chromosomal rearrangements are allowed (inversion, insertion/deletion, see Figure 1A). A *"gap"* parameter, representing the maximum number of consecutive genes which are not involved in a synteny group, is generally set to five genes. Comparative annotations of a new bacterial genome involve the computation of these synteny groups across all available microbial proteomes (NCBI databank, RefSeq section (42)).

From these comparison results, we define a species-specific gene as a gene having no ortholog in the compared species (significant similarities were not detected). This allowed us to compute specific regions between the genome under analysis, and a set of genomes selected for their phylogenetic proximity. Such regions are defined by at least two consecutive specific genes (Figure 1B). Insertion of genes which have homologies in the compared species is allowed in a specific region. A *"gap"* parameter, representing the maximum number of consecutive genes with homologies, is generally set to two genes. Finally, the predictive power of chromosomal clustering obtained with the Syntonizer method is used to assign a putative function, even in the absence of relevant sequence similarity.

Automatic functional assignations

The computational methods described above form the core of the MaGe processing pipeline. This fully automated first round of annotation ends with a functional assignation procedure allowing inference, as precisely as possible, of specific function(s) for each individual gene. Our procedure assigns biological function descriptions (gene products), gene names, Enzyme Commission (EC) numbers and functional classes when possible. Assignation of Gene Ontology terms (43) is directly obtained from the InterProScan results (30). The first step of our procedure uses reference annotation data of the *Escherichia coli* genome (44,45). If a significant match is found with this set of data, functional description and functional classes, gene names and synonyms are kept. Results from HAMAP functional assignations are then considered. If no HAMAP family is assigned, our procedure combines evidence from different methods, and defines a priority rank for each method. Pairwise comparisons with

curated annotations of model organisms (such as *Bacillus subtilis*, or other related bacterial genomes) are evaluated, taking synteny results into account. If no orthology relation exists, the program explores results against two protein domain databanks: TIGRFAMs (46) and Pfam (47). A hit is retained if the score is above the cutoff defined for each Hidden Markov Model (HMM). Priority is first given to TIGRFAMs results, and then, to those of Pfam. In case of multiple none-overlapping HMM hit results, a modular protein annotation using the "multifunctional protein" keywords is created, as well as a concatenation of the different domain descriptions. If no valuable HMM hit exists, the blastP results against UNIPROT (29) are evaluated given priority to the curated Swiss-Prot annotations. Only full-length matches with a high percent identity are considered and retained as a definitive or putative assignation. In all cases, PRIAM results (32) are used to assign EC number(s) to genes described as (putative) enzymes. Finally, if the selected UNIPROT match is described as a "(conserved) hypothetical protein", PRIAM results (if any) are checked to assign the description of the putative corresponding enzymatic function. If no PRIAM results exist, the predicted protein is annotated as a "conserved hypothetical protein". A protein with no blastP, HMM, or PRIAM matches remains a protein of unknown function. To complete the annotation, a (conserved) hypothetical protein is considered as "putative membrane protein" if at least three alpha-helical transmembrane regions have been retrieved by the tmHMM program (34), or as a "putative exported protein" if a signal peptide has been predicted by the SignalP program (35).

Sequence and annotation updates

The finishing phase term is highly variable depending on the genome coverage by the DNA libraries, the number of clones sequenced during the random phase and the number of repeated sequences present in the genome. To give researchers a quicker access to genome information, it is therefore important to start the annotation of a genome during the finishing phase of a project. Progression of the finishing phase can involve the alteration of numerous CDSs due to sequence gap closure and the addition, deletion, or modification of one or more bases. We have therefore developed a procedure which maps annotated features from an earlier version to the updated version of the genome sequence assembly. For each new assembly, a syntactic annotation is made (see 'Prediction of genes and functional annotation tools' section) and predicted genes are compared against the previous set of annotated genes. Only corresponding genes which align perfectly are mapped, taking into account a possible modification of the start codon position. In case of multiple correspondences (e.g. duplicated genes), the genomic context is explored to map only genes having a conserved neighborhood. At the end of the process, expert annotations of the mapped genes are transferred to the new version of the database. Then, a report allows one to retrieve locus name (*i.e.*, label) correspondences between mapped genes, genes that no longer exist and newly predicted genes (after the gap closure).

The synteny results can be used in an alternative way to identify one (or several) possible supercontig organization on the final chromosome by comparison with a phylogenetically related complete genome (hereafter, reference genome) (Supplementary Figure 1). This process, that may be very helpful for the progression of the finishing phase, is achieved in two ways: (i) finding the best supercontig order (and orientation) which maximizes a global conservation of the co-linearity between the reference genome and the draft of the sequenced genome (ii) looking for significant synteny groups on the supercontig ends which are neighbors on the reference genome (Supplementary Figure 1A and 1B). All proposed results must be experimentally validated by PCR analysis.

Metabolic pathway reconstruction

The set of annotated EC numbers provides an access to the chemical repertoire of the organism and allows for reconstruction of metabolic pathways. Two sets of reference metabolic pathways are used and linked to the MaGe annotations (Table 1). In KEGG (48), all known chemical reaction interconnections leading to a metabolic pathway are represented. Correspondences between EC numbers, reactions and pathways are modeled in the LIGAND database (49) and, from the MaGe interface, a dynamic request to the KEGG server allows one to visualize colored EC numbers on the KEGG diagrams (see 'Metabolic pathway visualization' section). The main advantage of this dataset relies on the completeness of the KEGG metabolic networks, and the automatic updating of the metabolic pathway reconstruction (Table 1). The BioCyc system (50,51) is a collection of Pathway/Genome DataBases (PGDBs). Each database describes the metabolic pathways of a single organism. The MetaCyc PGDB (52) is somewhat different in that it provides a reference collection of more than 600 experimentally elucidated metabolic pathways from many organisms (Table 1). For each prokaryotic genome being annotated in MaGe, an instance of the BioCyc scheme (built on an object database system, Ocelot) is created. The Pathway Tools software (50) starts with the organism annotations and analyzes the list of predicted EC numbers and the product name of the CDSs, to identify a

set of possible reactions which are subsequently matched against all pathways from MetaCyc. Each pathway is then evaluated and retained or not for the studied organism. At the end of the process, a PGDB is built (this new database is usually named *organism*Cyc, *i.e.* AcinetoCyc, FrankiaCyc, etc). In a second step, the Pathway Hole Filler program (53) is executed in order to find putative gene candidates for missing enzymes in the previously predicted metabolic pathways. KEGG and BioCyc metabolic network tools are clearly complementary, both in terms of metabolic datasets and of metabolic pathways graphical representation. However, these two homology-based metabolic pathway reconstruction systems cannot predict novel pathways. For this purpose, the MaGe system is connected to the Pathway Hunter Tool (PHT) web server (54). Starting from the set of MaGe annotated EC numbers, and a source/destination metabolite pair selected by the user, the shortest metabolic pathways (k-shortest pathways) are computed by PHT (Table 1). Alternative routes can then be evaluated for biological significance. Used together, these three methods are helpful to infer functional coupling of genes participating in the same cellular process.

THE RELATIONAL DATABASE

The MaGe system uses a relational database called PkGDB (Prokaryotic Genome DataBase) for storing, modifying and accessing very large datasets. A simplified view of the PkGDB data model is depicted in Figure 2. PkGDB supports several tables which model the components of the database shared by the different annotation projects: sequence and annotation data, functional predictions and annotation management (green, purple and blue colors in Figure 2A). Moreover, these three main components can be supplemented by other relational tables which take the specificities of each annotation project into account (tables surrounded by a red rectangle in Figure 2A).

Sequence and annotation data

PkGDB core tables store information on organisms, sequences and genomic objects (RNA genes, CDSs, etc). These tables merge annotations coming from three main origins. Firstly, in the case of a newly sequenced genome, gene prediction tools are run and compiled as new objects to be validated (see 'Bioinformatics Methods' section). Secondly, all public complete bacterial proteomes are available in the MaGe system. Sequences and annotations are extracted from the NCBI RefSeq (42) and EBI Genome Reviews (55) databanks, and stored in PkGDB. Annotations of several interesting complete bacterial genomes (*i.e.*, which could be improved in the context of a new MaGe genome project), are submitted to a human computer-assisted process, in order to improve their qualities. Some inconsistencies are corrected (erroneous reading frame, missing stop codon, in-frame stop codon, etc), and DNA regions suspected to contain pseudogenes are carefully re-annotated in the same way (the positions, on the DNA sequence, of the corresponding gene fragments are properly stored in PkGDB). This kind of information, which is not available in public protein databanks, can then be easily retrieved for further analysis. Moreover, the original sets of annotations are also submitted to a gene content re-annotation process which detect putative missing genes or wrongly annotated genes (56). These enriched sets of annotation data will be subsequently used to search for synteny groups in the genome(s) to be annotated. The third data origin consists of model bacteria annotations. Many specialized databases (*e.g*. GenprotEC (Serres, Goswami et al. 2004) and Ecogene (44) for *Escherichia coli*, PseudoCAP for *Pseudomonas aeruginosa* (57) and Subtilist for *Bacillus subtilis* (58)) focus their information on one or a few organisms. Annotations are continuously updated using experimental evidence. For each MaGe project, a set of reference organism annotations can be defined and integrated in the automatic and manual annotation process.

Functional predictions

Around this core structure, additional tables store functional prediction results (see 'Bioinformatics Methods' section). To retrieve and query results, each databank (*e.g.* UNIPROT, InterPro, COG, BioCyc, KEGG/LIGAND) used by a method is indexed in one or several tables. This integrative strategy improves efficiency in data access and querying. The system architecture also permits easy integration of new method results. To compute orthologs, paralogs, syntenies and specific regions, genome pair comparisons are defined for each project. This project customization reduces the number of pairwise comparisons and allows for more efficiency in query time execution. In the same way, blastP thresholds are evaluated depending on the studied organism (*i.e.*, the chosen values take into account the phylogenetic proximity of the newly annotated genome to the available bacterial proteomes). The corresponding results are stored in the database (Figure 2B) and used in the MaGe graphical representation of the synteny maps (see 'Genome browser and synteny maps' section).

Annotation management

The database architecture supports integration of automatic and manual annotations and records a history of all the modifications. Automatic annotation can be updated at any stage of the project. Furthermore, sequence updates and annotation transfer are stored in the database, allowing users to check mapped genes, new genes and genes that do not exist anymore. Three user groups are defined: 'curator', 'annotator' and 'guest'. 'Curators' and 'annotators' receive a username with password for data modification. The 'curator' status allows a user to save new annotations directly in the gene editor. Users having an 'annotator' status cannot directly save a novel annotation but instead, a mail is automatically sent to the 'curators' for a final review (acceptance or rejection of the submitted annotation). In case of a public project, a 'guest' login status is activated and annotations are immediately made available. Anonymous users can then query and browse the data using the MaGe's functionalities.

Depending on the organism properties, various functional classifications can be integrated in the thematic database: either an already defined classification (e.g. MultiFun (59)(Serres et Riley 2000), *B. subtilis* (58)(Moszer, Jones et al. 2002), TIGR (46)(Haft, Selengut et al. 2003), COG (31), or FunCat (60) classifications) or a completely new one.

FUNCTIONALITIES OF THE WEB INTERFACE

The MaGe web interface consists of numerous dynamic web pages containing textual and graphical representations for accessing and querying data (Supplementary Figure 2). A specific effort has been made in terms of graphical representations of available analysis results, to make the manual expert annotation easier and more efficient.

Genome browser and synteny maps

MaGe's main innovative functionality is a cartographic gene context exploration of the studied genome compared against all the available microbial genomes. This comparative genomics environment provides quality checks for both the automatic annotations and manual analysis. In Figure 3A, the first graphic map (genome browser), contains the complete *Acinetobacter sp*. ADP1 chromosome, over which the user can navigate with complete freedom (moving and zooming functionalities). The predicted coding genes are drawn, on the six reading frames, in red rectangles together with the coding prediction curves which are computed with the selected gene model (here, matrix 1).

The two following maps are representations of the synteny results (Figure 3A): each line shows the similarity results between the genome being annotated (here, *Acinetobacter* ADP1) and a given genome (*i.e.*, the first three lines of the second synteny map are with three Pseudomonas species). The first synteny map is a selection of the hundred curated genomes in PkGDB to date (see 'The relational database' section), and the second one is a selection of the two hundred and forty complete prokaryotic proteomes available in public databanks. On these maps, a rectangle flags the existence of a gene in a compared organism which is similar to the opposite gene in the annotated genome (*Acinetobacter* ADP1). If, for several co-localized CDSs on the annotated genome, there are several co-localized homologs on the compared genome, the rectangles will all be of the same color; otherwise, the rectangle is white. A group of rectangles of the same color thus indicates a synteny group. This graphical representation allows the user to quickly see if the part of the genome being annotated shares similarities and locally conserved organization with the selected bacterial sequences ('Options' functionality). As shown in Figure 3A, this is the case with *A. baumannii* (unpublished), and the two selected Pseudomonas species, with the *P. aeruginosa* genome sharing the most important number of synteny groups in this part of the ADP1 genome.

In contrast with the genome browser, there is no notion of scale on the synteny maps: to see how homologous genes are organized in a synteny group, the user can simply interact on one gene in a given synteny group. For example, by clicking on one rectangle of the green synteny group between *P. aeruginosa* and *Acinetobacter* ADP1, both corresponding genome regions of the compared organisms are shown and orthologs are linked, allowing the user to explore fusion/fission events, duplication, inversion and insertion/deletion of genes (Figure 3B). In our example, one interesting rearrangement appears clearly: the two *P. aeruginosa* homologs of the ADP1 CDS named ACIAD1137 are co-localized and transcribed on the same strand, showing that

the corresponding biological functions (*i.e.*, ribonuclease H and epsilon subunit of the DNA polymerase III) have been fused in the genome of *Acinetobacter* ADP1. Actually, the graphical representation of the synteny maps itself is also useful for detecting this kind of interesting feature: on each line, a rectangle has the same size as the corresponding annotated CDS in the studied genome. In addition, rectangles are colored depending on the part of the protein which aligns with the corresponding ADP1 protein (Figure 3A). It then becomes easy to see that ACIAD1137 has always two homologous genes in all the selected compared genomes (except with *A. baumannii*). However, the corresponding ADP1 protein aligns only on its N-terminal part with the first corresponding genes (*rnhA* gene), and on its C-terminal part with the second corresponding genes (*dnaQ* gene). Finally, these two homologous genes are involved in a synteny group containing 8 genes in Pseudomonas species, 6 genes in *Ralstonia solanacearum*, 3 genes in *Escherichia coli*, *Pseudoalteromonas haloplanktis*, and *Xanthomonas axonopodis*, and only 2 genes in *Shewanella oneidensis*. In these last four bacteria, *dnaQ* and *rnhA* genes are transcribed anti-clockwise and in *R. solanacearum, dnaQ* gene is not co-localized with the *rnhA* gene (white rectangle). This raises interesting evolutionary questions concerning the fusion of these two biological functions (involved in DNA replication).

Just below the three maps, several functionalities are available, such as the exploration of synteny results or annotated data using keywords ('Explore'), the search for similarities using blast functionalities (28), or for patterns in DNA or protein sequences ('Search'). At any time the user can download data in different common file formats (FASTA, EMBL, GenBank, etc) or extract part of its DNA sequence ('Export Data'). He/she can work with Artemis software (8) which is very useful for modifying erroneous start codon positions, for example, or explore KEGG (48), BioCyc (51), or PHT (54) metabolic pathways with MaGe annotations as input (see 'Metabolic pathway reconstruction' and 'Metabolic pathway visualization' sections).

Automatic versus manual annotation

In spite of the continuous improvement in the overall quality of bioinformatic methods, some difficulties in gene functional assignment can hardly be addressed in a completely automatic way. Most notably, the problem of error propagation in databases (61), which is today very strong in the context of common 'industrial' production of genome data, can only be solved with human intervention. Thus, the set of automatic annotations produced by any system should be considered only as a useful first approximation.

In MaGe, automatic annotation is always available in the gene editor ('Automatic annotation' section; Figure 4). This information is updated each time a new version of the complete genome sequence becomes available. Improvement of the annotation data quality can be made in the 'Gene Validation' section of the gene editor, which allows the user to modify, delete and add information. Several fields are mandatory such as 'Product', 'ProductType' (59), 'ECnumber', 'Roles' (*i.e.*, functional categories which have been chosen by the group of annotators), 'Localization' (cellular localization) and 'Class' (*i.e.*, known protein, strong similarity with known protein, no significant database hit, etc). Other fields are optional such as 'Comments' (free text), 'BioProcess' (biological processes), and 'PubmedID' (this field may contain the PubMed identification number(s) of any publication describing a biological function experimentally verified). Most of these fields are constrained by controlled vocabulary in order to provide annotation consistency and interoperability between genome annotation projects. In addition, annotation homogenization is also achieved *via* a procedure which is automatically launched when gene annotations are saved in the database. This allows for a minimal checking of the annotation coherence. For instance, 'ProductType' field must be equal to *enzyme* if an EC number is given ; 'Gene' and 'Synonyms' fields should be empty if 'Product' field contains *putative* or if 'Class' field is equal to 3 (*Function proposed based on presence of conserved amino acid motif, structural feature or limited homology*)'. A further advantage of MaGe's manual annotation system is that it enables a group of users, possibly at different locations, to easily co-operate on specific annotations: e-mail addresses of either the last annotator (in the gene editor, Figure 4) or all the different annotators for a specific gene (in the 'History' functionality, not shown) are available.

To help the user in the manual annotation of a gene, a summary of available method results are visualized in a completely customizable list (Figure 4). This part of the gene editor is essentially a workbench for curation and analysis of a single gene or its protein family. Primary information for the ORF shown in Figure 4 (CENIA1328, *Cenibacterium arsenoxidans ligA* gene) is presented in separate tables. This includes gene prediction (AMIGene) and duplication results, similarity results against (i) annotation data from reference genomes (*E. coli*, *B. subtilis* and *Acinetobacter* ADP1 in Figure 4), (ii) Swiss-Prot curated annotations and TrEMBL databank (only the ten best hits are kept), (iii) synteny results using PkGDB curated proteomes (about

100 to date) and complete prokaryotic genomes stored in the NCBI RefSeq section (about 240 to date). These comprehensive synteny results are useful to update, if necessary, the list of currently selected genomes which are visualized in the synteny maps. Other tables include enzymatic function predictions (PRIAM results), similarity results against COG (COGnitor), protein domain databanks (InterProScan), and HAMAP families. Finally, clues on the probable protein localization are given by the SignalP, and tmHMM results (Figure 4). For each set of results, external links, if any, are provided (NiceProt, NiceEnzyme, InterPro and COG databases, HAMAP families). In addition, direct interaction with PubMed (only if the field 'PubMedID' is filled), and with KEGG (external link) or BioCyc (internal link) metabolic pathway(s) involving the encoded enzyme (EC 6.5.1.2, Figure 4), is available.

This integrative strategy allows annotators to quickly browse functional evidence, tracking the history of a function and checking the gene context conservation with an orthologous gene having an experimentally demonstrated biological function.

A specific annotation can be saved using several status: 'in progress' (*i.e.*, the first step of expert work is not finished), 'finished' (*i.e.*, the first check of the automatic annotation is now complete), 'curated' (*i.e.*, the annotation has been modified during an expert analysis dedicated to biological process annotation). When a gene seems to be wrongly predicted, the user can select the 'Artefact' status (these genes are removed from the set of annotations before submission to public databanks). Finally, the 'CheckSeq' status is used when a sequence error is suspected (reads corresponding to these genes have to be checked for errors in the assembly).

Metabolic pathway visualization

Using MaGe, metabolic pathway exploration is accessible through three different tools: KEGG, BioCyc and PHT (see 'Metabolic pathway reconstruction' section). For a selected KEGG metabolic pathway, the user can obtain two lists of genes which are involved in this pathway. The first one gives the current annotated EC numbers, and the second one, the primary PRIAM predictions. Starting from these two EC number lists, metabolic maps are dynamically drawn via a request to the KEGG web server. A color-based code enables comparison of the studied organism enzyme content against a selected related organism (Supplementary Figure 3C). To easily retrieve co-localized genes involved in the same metabolic pathway, enzymes encoded by genes localized on the current MaGe genome region are highlighted in yellow. The useful representation of KEGG interconnected metabolic pathways is supplemented by the organism-specific PGDB built with the BioCyc system. This other set of metabolic data is made available to the expert through the MaGe interface via a web server (for each genome being annotated, we add a link to the corresponding *organism*Cyc database). Finally, to give the user the opportunity to find putative novel metabolic pathways, access to a PHT web form, which implements an *ab-initio* prediction method, is also available in MaGe.

Exploration of metabolic pathways could be enhanced through gene context analysis. For example, in the case of lysine biosynthesis, three alternative routes are described in the literature: the succinylase, dehydrogenase and acetylase branches (62). During the study of *Frankia alni* genome, MaGe annotations combined with the FrankiaCyc PGDB revealed only one possible pathway involving the succinylase branch (see Supplementary Figure 3A). All of the genes coding for the enzymes of this pathway (*ask*, *asd*, *dapA*, *dapB*, *dapD*, *dapE*, *dapF* and *lysA*) have been found, except for the *dapC* gene which encodes a succinyldiaminopimelate amino transferase activity. In *Escherichia coli*, the *dapC* gene does not exist, but the ArgD protein possesses both an acetylornithine and a succinyldiaminopimelate aminotransferase activity for arginine and lysine biosynthesis respectively (63). In *F. alni*, the *argD* gene has been identified, its presence could explain the absence of *dapC*. Actually, studying the *Frankia alni* genomic context of the genes involved in lysine biosynthesis, we found a gene (FRAAL6125) described as a putative aminotransferase. This gene is co-localized with the characterized *dapE* and *dapD* genes which encode two of the three steps of the succinylase branch (Supplementary Figure 3B). In addition, the corresponding KEGG map reveals the apparent lack of DapC activity and a co-localization of *dapE* and *dapD* genes (Supplementary Figure 3C). Furthermore, the synteny results among thirty organisms show a chromosomal conservation of this three-gene organization. All these evidence leads us to assume that FRAAL6125 is a good candidate for *dapC*. These assumptions were confirmed by sequence comparison with experimentally demonstrated *dapC* genes in *Corynebacterium glutamicum* (64)(Hartmann, Tauch et al. 2003) and *Bordetella pertussis* (65)(Fuchs, Schneider et al. 2000) (respectively, 52% and 32% amino acid identity). In contrast to the *dapC* homolog in other organims, in *Frankia alni* the protein encoded by FRAAL6125 possesses an additional C-terminal

domain of unknown function which is characterized by a glutamine and glycine rich content. Two other strains of the Frankia genus (Cci3 and EAN1pec), sequenced by the United States Department Of Energy, show a similar genomic organization of the *dapCDE* gene cluster. But only the strain EAN1pec possesses this C-terminal domain. This Frankia-specific C-terminal domain of DapC calls for more experimental investigation. This example shows that MaGe integration of gene context methods is a powerful tool for experts in metabolic analysis.

Data Exploration

Although the notion of multigenome comparisons is omnipresent in the graphical interface of our system, the exploration functionality developed in MaGe is linked to the genome being selected for expert annotation only ('Display organism' in the 'Options' functionality). A simple keyword search enables the user to quickly retrieve genes of the annotated genome having a particular function. Several sets of data can be queried, such as automatic and validated annotations (expert work), or a specific set of annotated CDSs corresponding, for example, to conserved hypothetical proteins which are in synteny with other organisms. In addition, each kind of computed result (PRIAM, InterPro, blast similarities in reference genome annotation data, and in Swiss-Prot or TrEMBL databanks) can be retrieved. The result output is a list of candidate genes. The genomic contexts of which can be easily visualized (automatic displacement of the genome browser centered on a gene of interest).

In a second section, called 'PhyloProfile and Synteny' (Supplementary Figure 4), the user can search for genes of the studied organism which are homologs of genes in certain organisms, and exclude those that are homologs of genes in other organisms. The phylogenetic profile method is designed to infer functional relationships between genes: proteins involved in the same biological process are likely to evolve in a correlated fashion (15). This method, combined with the integration of synteny results, allows one to detect a coevolution of gene groups which have a similar chromosomal organization. Integration of chromosomal proximity and gene content information has been reported to be more accurate than the single-gene phylogenetic profiles (66).

Using the synteny results stored in our database (see 'The relational database' section) the fusion/fission events can easily be computed. Our procedure detects synteny groups having two genes from a compared genome corresponding to a single annotated CDS in the target genome (Figure 1A). BlastP correspondences are evaluated to exclude the detection of tandem duplications by keeping only non-overlapping side-by-side alignments. These events are listed in the 'Fusion/Fission' item of the 'Explore' functionality (Supplementary Figure 4), and split into two tables: one containing the list of putative fused genes, and the other for fission events. Annotators can then browse results by checking for possible pseudogenes or for true functional evidence leading to the annotation of a multifunctional protein (see above, the case of *rnhA* and *dnaQ* gene fusion in *Acinetobacter ADP1*).

In a fourth section of the 'Explore' functionality, specific regions between the genome under analysis and a set of genomes selected for their phylogenetic proximity can be browsed (Supplementary Figure 4). Data is represented in a table listing gene clusters that have no correspondences in one or more compared organisms. One application of this comparative genomic analysis is the detection of genomic islands. A comparative study between two *Acinetobacter baumannii* strains, AYE a multi-drug resistant strain and SDF a fully susceptible one, led us to decipher a 86-kb AYE specific region where more than forty resistance genes are clustered ('Genomics of Multi-Drug Resistance in *Acinetobacter baumannii*', P.-E. Fournier et al., submitted).

SETTING UP A NEW ANNOTATION PROJECT

The MaGe system can be used either for the annotation of novel genomes, or for curation of already annotated genomes available in public databanks (re-annotation projects). To start a new project, we first work on the integration, in PkGDB, of available bacterial genomes which are of interest in the context of the new thematic database (Supplementary Figure 5). Both complete and unfinished bacterial genomes are integrated in our database. The sequence(s) of the novel genome(s) are then submitted to the complete annotation pipeline analysis, including computation of synteny results with all the available proteomes in PkGDB and in the NCBI RefSeq databank. As explained in the 'Metabolic pathway reconstruction' section, a Pathway/Genome DataBase (PGDB) is built using the BioCyc software (51), and the corresponding database is made available

from the MaGe interface (Supplementary Figure 5). Some changes in the gene editor are made to take into account the specificity of each project. For example, the *E. coli* functional classification which is the default can be changed, or additional 'BioProcess' classes can be added (for the RhizoScope project shown in Supplementary Figure 5, three additional processes were added: Nitrogen fixation, Photosynthesis and Symbiosis). Finally, the new thematic database is made available to the research teams involved in the project (via a secure connection which requires authenticated login). In addition, the portion of the database information corresponding to bacterial genomes available in public databanks is made freely accessible via the MaGe interface (Supplementary Figure 5).

The MaGe system, initially developed and used in the context of the *Acinetobacter* ADP1 genome annotation (17), has also been used for the analysis of *Pseudoalteromonas haloplanktis* (18), *Frankia alni* and *Pseudomonas entomophila* (in preparation). In the context of the MicroScope project which aims to build thematic databases for the (re)-annotation of prokaryotic genomes (http://www.genoscope.cns.fr/agc/microscope), a number of microbial genomes are currently being annotated using the MaGe system (sixteen projects to date): this includes pathogenic species (such as *Leptospira biflexa*, *Neisseria meningitidis* NEM8013, and *Escherichia coli* strains), or environmental bacteria (such as *Cenibacterium arsenoxidans*, and *Bradyrhizobium* sp. ORS278). In addition, our group is involved in a metagenomic project which aims to produce an inventory of the microorganisms present at two main stages of waste water treatment. Several large genomic regions from yet uncultured microorganisms have already been annotated and analyzed, giving us the opportunity to propose specific culture media for enrichment cultures for the corresponding bacteria.

IMPLEMENTATION OF MaGe

UNIX shell and perl scripts manage data integration and computations. Program executions are dispatched on a multi-processor computer system (40 Alpha 1 GHz CPUs) by the Platform LSF software (a batch application workload processing). Pattern search and sequence alignments are performed with the Biofacet package (67). The free MySQL database management system which is used by PkGDB provides a fast and a reliable access to data. For the MaGe web server, the Apache system and the PHP (Hypertext Preprocessor) language are used. PHP is a HTML-embedded scripting language allowing dynamic generation of the HTML page contents. Associated with the GD graphics library, web interface images are dynamically generated in PNG (Portable Network Graphics) format. The PkGDB database scheme and the MaGe web frontend are available upon request for a local installation. Furthermore, on demand, we can customize the MaGe system for a specific genome project (http://www.genoscope.cns.fr/agc/microscope).

CONCLUSION

The MaGe annotation platform (*i.e.*, a software suite with a multigenomes relational database and a web graphical interface) has proved to be a useful tool for expert annotation, mainly because it avoids most of the main automatic sequence annotation pitfalls. In the process of the expert annotation, our graphical representation of synteny results is obviously invaluable to highlight interesting features. Due to the dynamic nature of the bioinformatics field, constant efforts are made to keep the set of computational techniques (*i.e.*, additional methods and/or links to useful web sites are regularly added) and the integrated databases up-to-date. Our automatic annotation procedure takes into account the spurious function assignments caused by multidomain proteins, and exploits functional coupling between genes located in adjacent positions on the chromosome. However, we plan to improve some decision rules, mainly by introducing data from predicted metabolic pathways obtained with the BioCyc software (50,51). In addition, MaGe is often used to annotate several closely related genomes, and a novel functionality is clearly required, which will permit a manual refinement of annotation on several related species at the same time. Other planned developments include: new features in the genome browser (*i.e.*, representation of global DNA and protein statistical tendencies), new features in the gene editor (*i.e.*, graphical representation of functional annotations on the corresponding protein), and an improved interface for queries in the PkGDB database.

The growing availability of expression profiles (from microarray data and proteomics), supplemented with gene essentiality and regulation, protein-protein interaction and metabolomics data, brings a major source of clues for the clarification of gene

function. In that sense, the Genostar exploratory genomics platform (http://www.genostar.org), which has already been connected to PkGDB, is a promising tool: it offers a unified way of representing and managing data of various types and origins through a set of software modules which can exchange information. In the context of some annotation projects, MaGe high quality annotations are imported into Genostar and linked to various types of experimental data modeled in the GenoLink module (68). The MaGe functionalities combined with the advanced query interface of this module should contribute to the characterization of the functions of orphan genes.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank all MaGe users for their feedback that helped greatly in optimizing and improving many functionalities of the system. This work was supported by the French Centre National de la Recherche Scientifique (CNRS-URA8030), the GENOPOLE of Evry, and the French Ministry of Research (funds allocated by the ACI IMPBio2004). We thank Susan Cure and Denis Bayada for their help in writing the manuscript. A particular thanks for François Le Fèvre for his help in setting up the BioCyc system. We thank the entire system network team of Genoscope for its essential contribution to the efficiency of the MaGe web interface.

REFERENCES

1. Galperin, M.Y. and Koonin, E.V. (1998) *In Silico Biol*, **1,** 55-67.
2. Brenner, S.E. (1999) *Trends Genet*, **15,** 132-133.
3. Van Domselaar, G.H., Stothard, P., Shrivastava, S., Cruz, J.A., Guo, A., Dong, X., Lu, P., Szafron, D., Greiner, R. and Wishart, D.S. (2005) *Nucleic Acids Res*, **33,** W455-459.
4. Riley, M.L., Schmidt, T., Wagner, C., Mewes, H.W. and Frishman, D. (2005) *Nucleic Acids Res*, **33,** D308-310.
5. Gaasterland, T. and Sensen, C.W. (1996) *Trends Genet*, **12,** 76-78.
6. Hoersch, S., Leroy, C., Brown, N.P., Andrade, M.A. and Sander, C. (2000) *Trends Biochem Sci*, **25,** 33-35.
7. Overbeek, R., Larsen, N., Walunas, T., D'Souza, M., Pusch, G., Selkov, E., Jr., Liolios, K., Joukov, V., Kaznadzey, D., Anderson, I. *et al.* (2003) *Nucleic Acids Res*, **31,** 164-171.
8. Berriman, M. and Rutherford, K. (2003) *Brief Bioinform*, **4,** 124-132.
9. Meyer, F., Goesmann, A., McHardy, A.C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R. *et al.* (2003) *Nucleic Acids Res*, **31,** 2187-2195.
10. Frangeul, L., Glaser, P., Rusniok, C., Buchrieser, C., Duchaud, E., Dehoux, P. and Kunst, F. (2004) *Bioinformatics*, **20,** 790-797.
11. Almeida, L.G., Paixao, R., Souza, R.C., Costa, G.C., Barrientos, F.J., Santos, M.T., Almeida, D.F. and Vasconcelos, A.T. (2004) *Bioinformatics*, **20,** 2832-2833.
12. Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998) *Trends Biochem Sci*, **23,** 324-328.
13. Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999) *Science*, **285,** 751-753.
14. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. and Maltsev, N. (1999) *In Silico Biol*, **1,** 93-108.
15. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999) *Proc Natl Acad Sci U S A*, **96,** 4285-4288.
16. Doerks, T., von Mering, C. and Bork, P. (2004) *Nucleic Acids Res*, **32,** 6321-6326.
17. Barbe, V., Vallenet, D., Fonknechten, N., Kreimeyer, A., Oztas, S., Labarre, L., Cruveiller, S., Robert, C., Duprat, S., Wincker, P. *et al.* (2004) *Nucleic Acids Res*, **32,** 5766-5779.
18. Medigue, C., Krin, E., Pascal, G., Barbe, V., Bernsel, A., Bertin, P.N., Cheung, F., Cruveiller, S., D'Amico, S., Duilio, A. *et al.* (in press) *Genome Res*.
19. Bocs, S., Cruveiller, S., Vallenet, D., Nuel, G. and Medigue, C. (2003) *Nucleic Acids Res*, **31,** 3723-3726.
20. Besemer, J. and Borodovsky, M. (2005) *Nucleic Acids Res*, **33,** W451-454.
21. Benzecri, J.-P. (1973) *L'analyse des données, L'Analyse des Correspondances*. Dunod Edition, Paris, France.
22. Delorme, M.O. and Hénaut, A. (1988) *Comput Appl Biosci*, **4,** 453-458.
23. Suzek, B.E., Ermolaeva, M.D., Schreiber, M. and Salzberg, S.L. (2001) *Bioinformatics*, **17,** 1123-1130.
24. Lowe, T.M. and Eddy, S.R. (1997) *Nucleic Acids Res*, **25,** 955-964.
25. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R. and Bateman, A. (2005) *Nucleic Acids Res*, **33,** D121-124.
26. d'Aubenton Carafa, Y., Brody, E. and Thermes, C. (1990) *J Mol Biol*, **216,** 835-858.
27. Achaz, G., Coissac, E., Viari, A. and Netter, P. (2000) *Mol Biol Evol*, **17,** 1268-1275.
28. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) *Nucleic Acids Res*, **25,** 3389-3402.
29. Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. *et al.* (2005) *Nucleic Acids Res*, **33,** D154-159.
30. Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L. *et*

al. (2005) *Nucleic Acids Res*, **33,** D201-205.
31. Natale, D.A., Galperin, M.Y., Tatusov, R.L. and Koonin, E.V. (2000) *Genetica*, **108,** 9-17.
32. Claudel-Renard, C., Chevalet, C., Faraut, T. and Kahn, D. (2003) *Nucleic Acids Res*, **31,** 6633-6639.
33. Gattiker, A., Michoud, K., Rivoire, C., Auchincloss, A.H., Coudert, E., Lima, T., Kersey, P., Pagni, M., Sigrist, C.J., Lachaize, C. *et al.* (2003) *Comput Biol Chem*, **27,** 49-58.
34. Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) *J Mol Biol*, **305,** 567-580.
35. Bendtsen, J.D., Nielsen, H., von Heijne, G. and Brunak, S. (2004) *J Mol Biol*, **340,** 783-795.
36. Gough, J., Karplus, K., Hughey, R. and Chothia, C. (2001) *J Mol Biol*, **313,** 903-919.
37. Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C. and Murzin, A.G. (2004) *Nucleic Acids Res*, **32,** D226-229.
38. Pascal, G., Medigue, C. and Danchin, A. (2005) *Proteins*, **60,** 27-35.
39. von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A. and Bork, P. (2005) *Nucleic Acids Res*, **33,** D433-437.
40. Smith, T.F. and Waterman, M.S. (1981) *J Mol Biol*, **147,** 195-197.
41. Boyer, F., Morgat, A., Labarre, L., Pothier, J. and Viari, A. (in press) *Bioinformatics*.
42. Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2005) *Nucleic Acids Res*, **33,** D501-504.
43. Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C. *et al.* (2004) *Nucleic Acids Res*, **32,** D258-261.
44. Rudd, K.E. (2000) *Nucleic Acids Res*, **28,** 60-64.
45. Serres, M.H., Goswami, S. and Riley, M. (2004) *Nucleic Acids Res*, **32,** D300-302.
46. Haft, D.H., Selengut, J.D. and White, O. (2003) *Nucleic Acids Res*, **31,** 371-373.
47. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L. *et al.* (2004) *Nucleic Acids Res*, **32,** D138-141.
48. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) *Nucleic Acids Res*, **32,** D277-280.
49. Goto, S., Okuno, Y., Hattori, M., Nishioka, T. and Kanehisa, M. (2002) *Nucleic Acids Res*, **30,** 402-404.
50. Karp, P.D., Paley, S. and Romero, P. (2002) *Bioinformatics*, **18 Suppl 1,** S225-232.
51. Krummenacker, M., Paley, S., Mueller, L., Yan, T. and Karp, P.D. (2005) *Bioinformatics*.
52. Krieger, C.J., Zhang, P., Mueller, L.A., Wang, A., Paley, S., Arnaud, M., Pick, J., Rhee, S.Y. and Karp, P.D. (2004) *Nucleic Acids Res*, **32,** D438-442.
53. Green, M.L. and Karp, P.D. (2004) *BMC Bioinformatics*, **5,** 76.
54. Rahman, S.A., Advani, P., Schunk, R., Schrader, R. and Schomburg, D. (2005) *Bioinformatics*, **21,** 1189-1193.
55. Kersey, P., Bower, L., Morris, L., Horne, A., Petryszak, R., Kanz, C., Kanapin, A., Das, U., Michoud, K., Phan, I. *et al.* (2005) *Nucleic Acids Res*, **33,** D297-302.
56. Cruveiller, S., Le Saux, J., Vallenet, D., Lajus, A., Bocs, S. and Medigue, C. (2005) *Nucleic Acids Res*, **33,** W471-479.
57. Winsor, G.L., Lo, R., Sui, S.J., Ung, K.S., Huang, S., Cheng, D., Ching, W.K., Hancock, R.E. and Brinkman, F.S. (2005) *Nucleic Acids Res*, **33,** D338-343.
58. Moszer, I., Jones, L.M., Moreira, S., Fabry, C. and Danchin, A. (2002) *Nucleic Acids Res*, **30,** 62-65.
59. Serres, M.H. and Riley, M. (2000) *Microb Comp Genomics*, **5,** 205-222.
60. Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Guldener, U., Mannhaupt, G., Munsterkotter, M. *et al.* (2004) *Nucleic Acids Res*, **32,** 5539-5545.
61. Bork, P. and Bairoch, A. (1996) *Trends Genet*, **12,** 425-427.
62. Velasco, A.M., Leguina, J.I. and Lazcano, A. (2002) *J Mol Evol*, **55,** 445-459.
63. Ledwidge, R. and Blanchard, J.S. (1999) *Biochemistry*, **38,** 3019-3024.
64. Hartmann, M., Tauch, A., Eggeling, L., Bathe, B., Mockel, B., Puhler, A. and Kalinowski, J. (2003) *J Biotechnol*, **104,** 199-211.
65. Fuchs, T.M., Schneider, B., Krumbach, K., Eggeling, L. and Gross, R. (2000) *J Bacteriol*, **182,** 3626-3631.
66. Zheng, Y., Roberts, R.J. and Kasif, S. (2002) *Genome Biol*, **3,** RESEARCH0060.
67. Glemet, E. and Codani, J.J. (1997) *Comput Appl Biosci*, **13,** 137-143.
68. Durand, P., Labarre, L., Meil, A. and Wojcik, J. (2005) *ERCIM*, **60,** 31-32.

Table 1: Main features of the metabolic data sets used in MaGe

	KEGG	BioCyc	PHT
Enzyme, reaction data	Ligand	Enzyme database + in-house curation	Ligand and Brenda databases
Pathway data	Multi-organisms, generic representation	Organism specific, experimentally validated (MetaCyc)	No
Gene/reaction correspondences	EC numbers	EC numbers + product names	EC numbers
Pathway reconstruction	Homology based (EC number mapping)	Homology based (pathway selection algorithm)	*Ab-initio* reconstruction (k-shortest pathways)
Hole detection	No	Yes + Pathway Hole Filler	No
Data management	Flat files	Object Database, Ocelot	?
MaGe integration	Web service	Local installation	Web service
MaGe annotation updates	Dynamic	Re-execution of Pathway Tools	Dynamic

This table shows the main features of the three metabolic pathway reconstruction systems integrated in MaGe: KEGG (Kanehisa, Goto et al. 2004), BioCyc (Krummenacker, Paley et al. 2005) and PHT (Pathway Hunter Tool) (Rahman, Advani et al. 2005). KEGG and BioCyc use the homology method to reconstruct metabolic pathways. PHT, which uses an *ab-initio* algorithm to compute the shortest pathways between two metabolites, helps the user to find alternative routes.

Figure legends

Figure 1: Synteny groups and specific regions detection

A. Example of synteny groups (rectangles with green borders) between two genomes A and B. Syntonizer software allows multiple correspondences between genes (red arrows, e.g. blastP similarity results) to detect duplications and gene fusion/fission events. Local rearrangements (inversion; insertion/deletion) are allowed in our method. The gap parameter defines the number of consecutive genes not involved in synteny. The first synteny group shows a gene fusion event in genome A. The second synteny group shows a perfect gene order conservation in the two compared genomes. The third one is the result of a duplication in genome B together with the insertion of two genes (the gap parameter is then equal to 2).

B. Example of a specific region (rectangle with green border) in the genome A. Co-localized genes (plain green rectangles in genome A) have no ortholog in the compared genome B. Lack of correspondence relations (green arrows) are explicitly represented. A gap parameter represents the maximum number of consecutive genes with homologies in the compared genome. In this example, two genes are inserted (the gap parameter is then equal to 2).

Figure 2: Simplified PkGDB relational model.

A. PkGDB is made of three main components: sequence and annotation data (in green), annotation management (in blue) and functional predictions (in purple). Sequences and annotations come from three sources namely public databanks, sequencing centers and specialized databases focused on model organisms. For genomes of interest, a (re)-annotation process is performed using AMIGene (Bocs, Cruveiller et al. 2003) and leads to the creation of new 'Genomic Objects'. Each 'Genomic Object' and associated functional prediction results are stored in PkGDB. The database architecture supports integration of automatic and manual annotations, and management of a history of annotations and sequence updates. The core of PkGDB can be supplemented by other tables to take into account genome project specificities ('Project customization', red rectangle).

B. Part of the PkGDB relational structure for storage of synteny results. Correspondences between 'Genomic Objects' (identified by the primary key GO_id) are modeled by the table 'GO_GO_CPD' which links pairs of 'Genomic Objects' (GO_id_1 and GO_id_2 foreign keys). A synteny group is defined by a group of correspondences having the same 'SYNT_id' (identifier of a synteny group). The third table, 'Synteny_group', contains a description of synteny groups (e.g. identifier of compared genomes, location of the groups and number of genes involved for each genome)

Figure 3: MaGe's genome browser and synteny maps

A. The *Acinetobacter* ADP1 chromosomal segment, extending between positions 1117,700 and 1137,700 bp, is represented on this graphical map of the MaGe interface developed on our database. Annotated CDSs are represented in the 6 reading frames of the sequence by red rectangles, and coding prediction curves are superimposed on the predicted CDSs (blue curves). The synteny maps, calculated on a set of selected genomes (3 from PkGDB database and 5 from NCBI databank), are displayed below. In contrast with the graphic interface of the *Acinetobacter* ADP1 genome, there is no notion of scale on the synteny map: a rectangle has the same size of the CDS which is exactly opposite in the ADP1 genome, and it represents a putative ortholog between one CDS of the compared genome and one CDS of the *Acinetobacter* ADP1 genome. In addition, rectangles are colored depending on the part of the protein which aligns with the corresponding ADP1 protein. If, for several CDSs co-localized on the ADP1 genome, there are several co-localized orthologs in the compared genome, the rectangles will all be of the same color; otherwise, the rectangle is white. A group of rectangles of the same color thus indicates synteny between *Acinetobacter* ADP1 and the compared genome.

B. This second graphical representation of synteny has been obtained by clicking on one rectangle of the synteny maps (here one of the eight *P. aeruginosa* green genes). It allows the user to see how homologous genes, in a synteny group, are organized: here, one fusion event in *Acinetobacter* ADP1 (ACIAD1137: rnhA+dnaQ), a duplication of two genes (PA1810, PA1811) and an insertion of two genes (PA1814, PA1813) in *P. aeruginosa*. In addition, ACIAD1138 is similar to the *mtlD* gene of *P. aeruginosa* only in its N-terminal part, the second part of the protein sharing similarity with a COG family annotated as 'LysM-repeat proteins and domains' (COG1388).

Figure 4: MaGe's gene editor

The MaGe's gene editor is used in the context of expert annotation. It is made of three main sections: 1. the 'Gene Validation' section allows the user to modify, delete and add information (see text); 2. the 'Automatic Annotation' section contains the results from the automatic procedure described in the 'Automatic functional assignations' section; 3. the last section gives access to a summary of available tool results, including Blast alignments (see text). External links to useful Websites are provided, together with links to PubMed, KEGG, and the CeniCyc database ('BioCyc' link).

Figure 1:

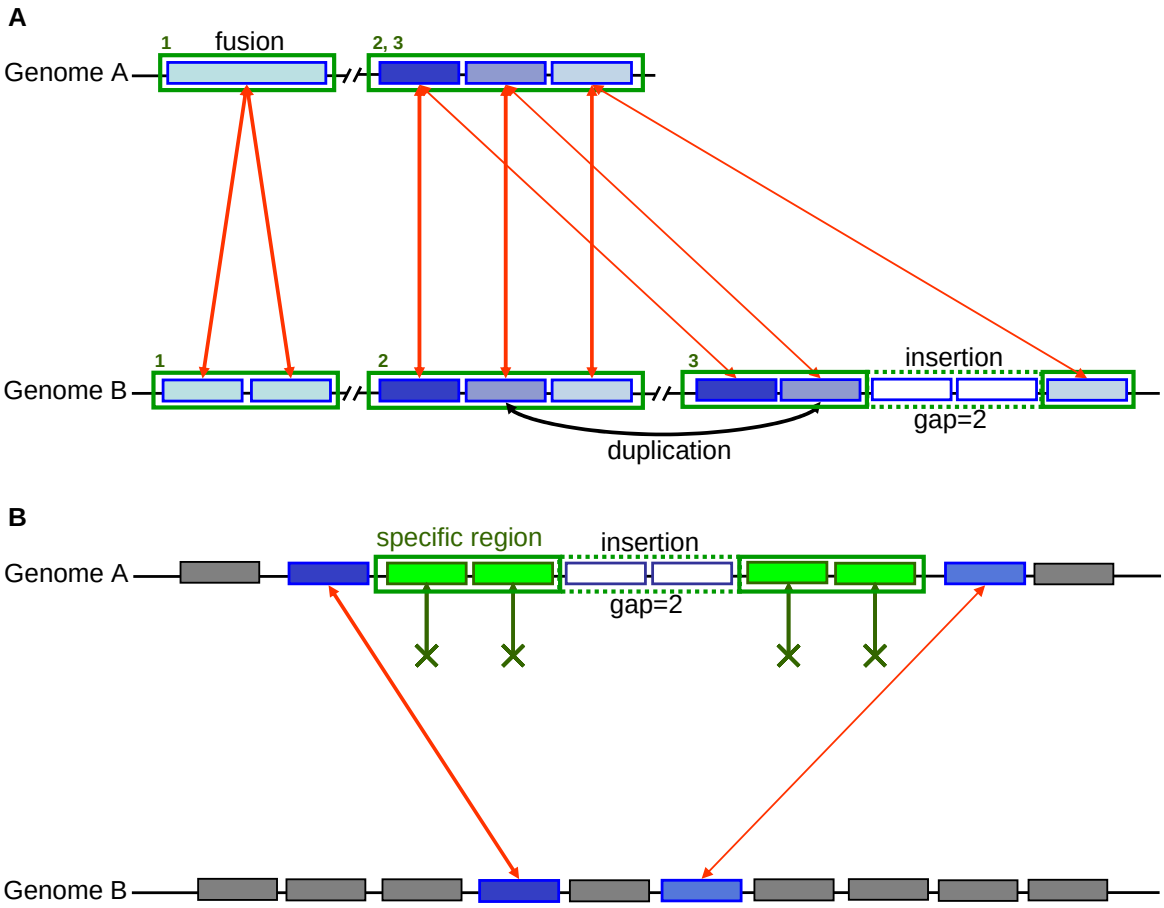
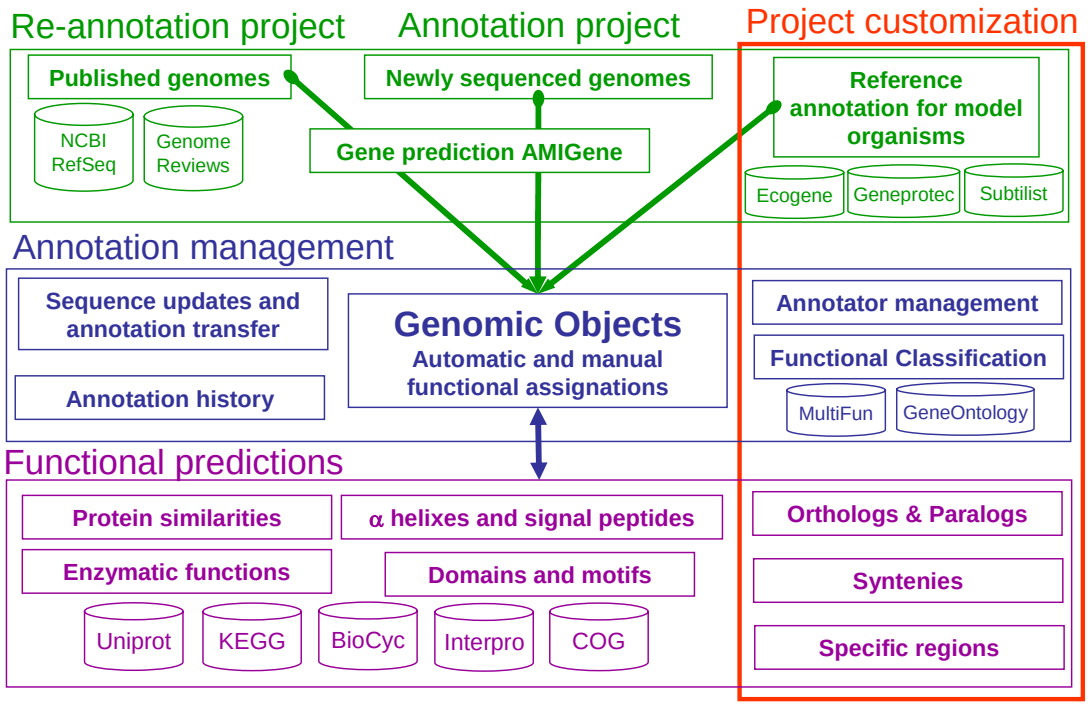

A
1
fusion
2, 3
Genome A
1
2
3
insertion
Genome B
gap=2
duplication
B
specific region
insertion
Genome A
gap=2
Genome B

Figure 2:

A


Re-annotation project
Annotation project
Project customization
Published genomes
Newly sequenced genomes
Reference annotation for model organisms
NCBI RefSeq
Genome Reviews
Gene prediction AMIGene
Ecogene
Geneprotec
Subtilist
Annotation management
Sequence updates and annotation transfer
Genomic Objects
Automatic and manual functional assignations
Annotator management
Functional Classification
Annotation history
MultiFun
GeneOntology
Functional predictions
Protein similarities
α helixes and signal peptides
Orthologs & Paralogs
Enzymatic functions
Domains and motifs
Syntenies
Uniprot
KEGG
BioCyc
Interpro
COG
Specific regions

B

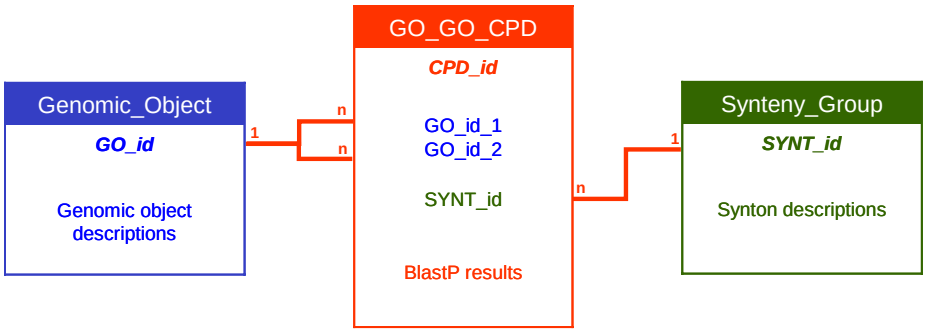
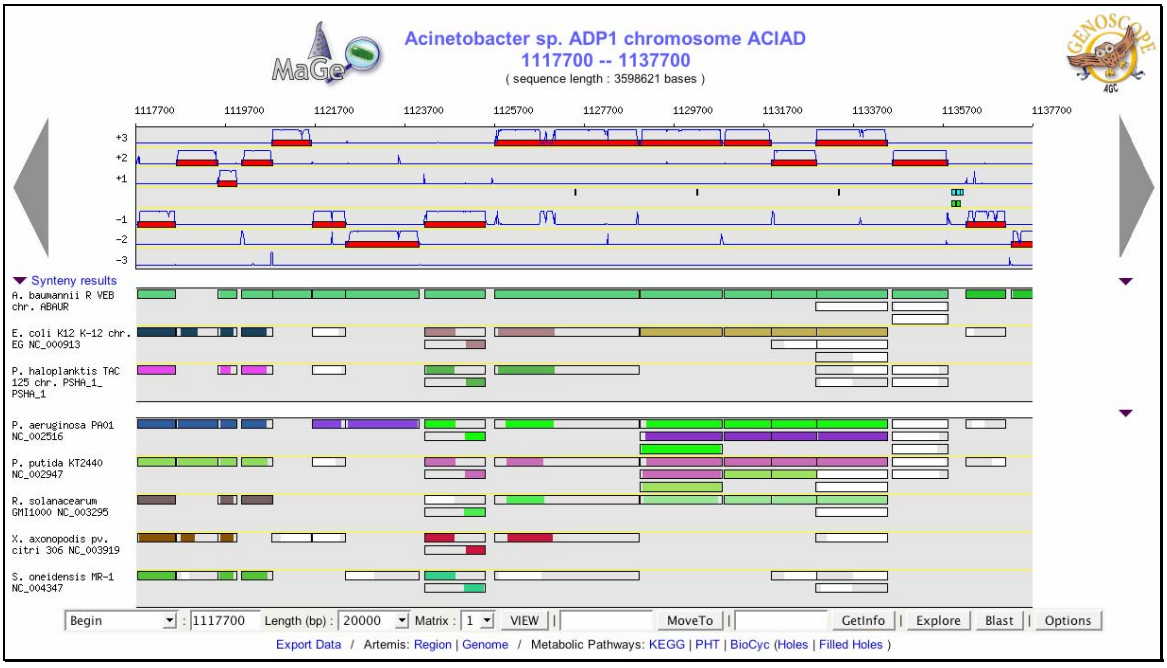
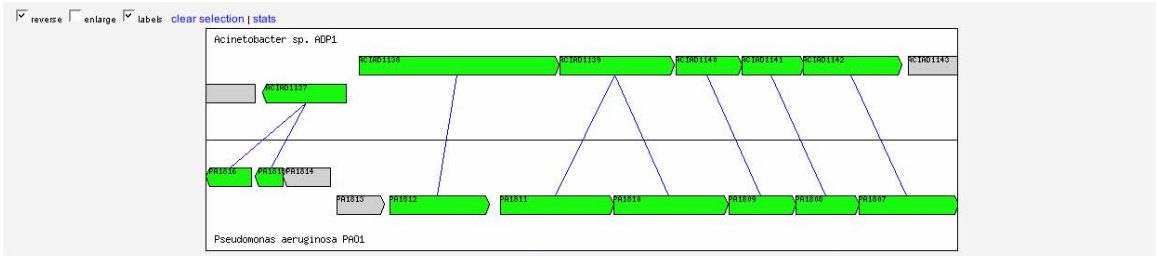

GO_GO_CPD
CPD_id
Genomic_Object
GO_id
GO_id_1
GO_id_2
Genomic object descriptions
SYNT_id
BlastP results
Synteny_Group
SYNT_id
Synton descriptions
1
n
n
1
n

Figure 3:

A



B



Acinetobacter sp. ADP1

Label	Gene	Type	Product
ACIAD1136		CDS	conserved hypothetical protein
ACIAD1137	mhA-dnaQ	CDS	bifunctional protein [Includes: ribonuclease HI; DNA polymerase III ,epsilon subunit, 3-5 exonucleolytic proofreading function]
ACIAD1138		CDS	bifunctional protein [Includes: lytic murein transglycosylase C, membrane-bound (MltD); putative cell wall hydrolase]
ACIAD1139		CDS	putative oligopeptide transport protein (ABC superfamily, peri_bind)
ACIAD1140		CDS	putative oligopeptide transport protein (ABC superfamily, membrane)
ACIAD1141		CDS	putative oligopeptide transport protein (ABC superfamily, membrane)
ACIAD1142		CDS	putative oligopeptide transport protein (ABC superfamily, atp_bind)
ACIAD1143		CDS	putative FMN oxidoreductase

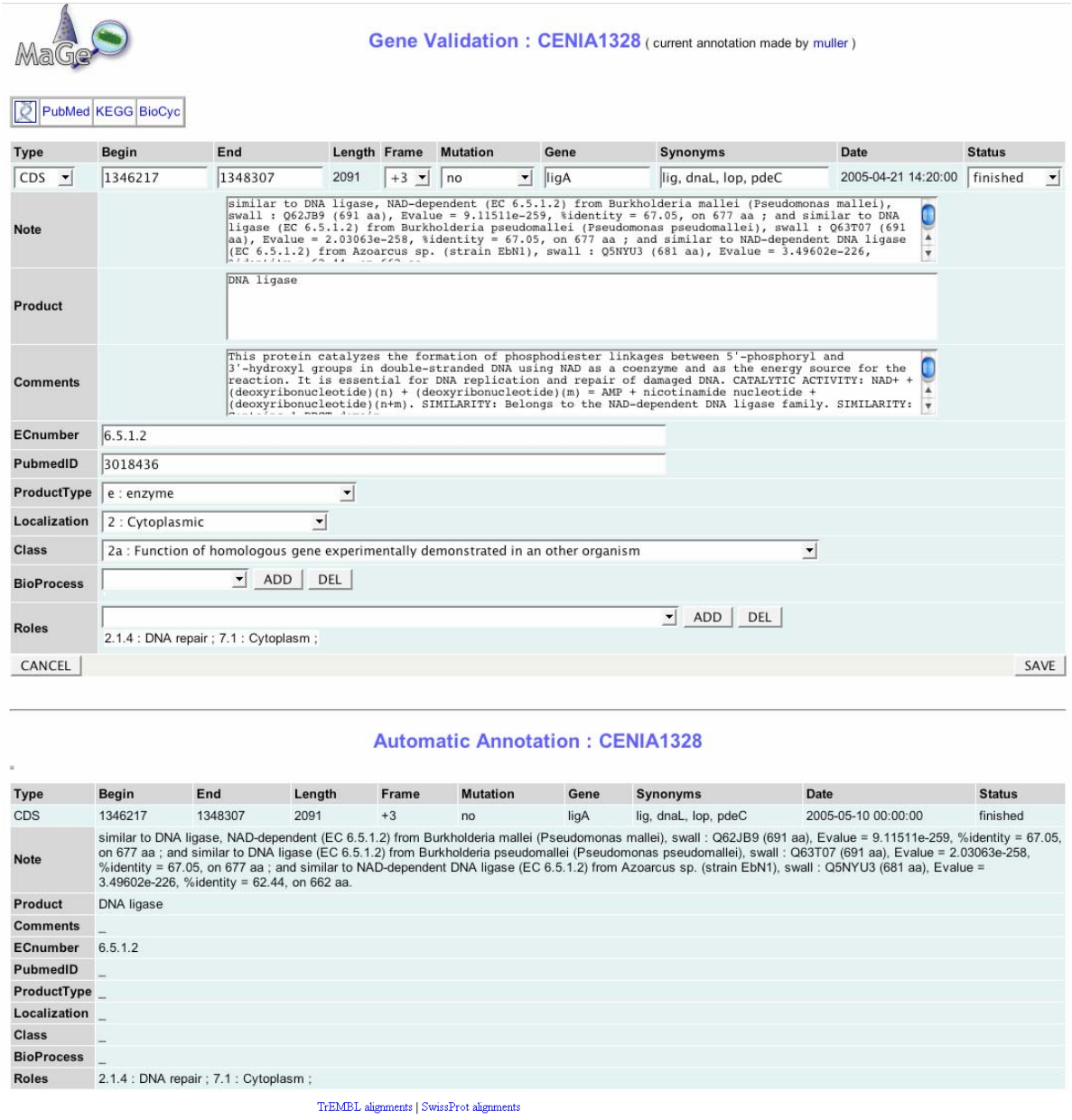
Correspondences

Ident	MatchL	minLrap	LengthQ	LengthB	OrderQ	OrderB
51.72	524	0.975791	537	537	1	1
62.02	337	1.00298	336	340	1	1
61.94	360	1.01408	355	361	1	1
44.5	555	0.900974	616	616	1	1
37.1	601	0.985246	616	610	3	1
41.44	362	0.676636	1077	535	1	1
50.35	141	0.946309	456	149	2	1
43.75	240	0.97166	456	247	1	1

Pseudomonas aeruginosa PAO1

Label	Gene	Type	Product
PA1807			probable ATP-binding component of ABC transporter
PA1808			probable permease of ABC transporter
PA1809			probable permease of ABC transporter
PA1810			probable binding protein component of ABC transporter
PA1811			probable solute-binding protein
PA1812	mltD		membrane-bound lytic murein transglycosylase D precursor
PA1813			probable hydroxyacylglutathione hydrolase
PA1814			hypothetical protein
PA1815	mhA		ribonuclease H
PA1816	dnaQ		DNA polymerase III, epsilon chain

Figure 4:



Gene Validation : CENIA1328 (current annotation made by muller)

PubMed | KEGG | BioCyc

Type	Begin	End	Length	Frame	Mutation	Gene	Synonyms	Date	Status
CDS	1346217	1348307	2091	+3	no	ligA	lig, dnaL, lop, pdeC	2005-04-21 14:20:00	finished

Note	similar to DNA ligase, NAD-dependent (EC 6.5.1.2) from Burkholderia mallei (Pseudomonas mallei), swall : Q62JB9 (691 aa), Evalue = 9.11511e-259, %identity = 67.05, on 677 aa ; and similar to DNA ligase (EC 6.5.1.2) from Burkholderia pseudomallei (Pseudomonas pseudomallei), swall : Q63T07 (691 aa), Evalue = 2.03063e-258, %identity = 67.05, on 677 aa ; and similar to NAD-dependent DNA ligase (EC 6.5.1.2) from Azoarcus sp. (strain EbN1), swall : Q5NYU3 (681 aa), Evalue = 3.49602e-226,
Product	DNA ligase
Comments	This protein catalyzes the formation of phosphodiester linkages between 5'-phosphoryl and 3'-hydroxyl groups in double-stranded DNA using NAD as a coenzyme and as the energy source for the reaction. It is essential for DNA replication and repair of damaged DNA. CATALYTIC ACTIVITY: NAD+ + (deoxyribonucleotide)(n) + (deoxyribonucleotide)(m) = AMP + nicotinamide nucleotide + (deoxyribonucleotide)(n+m). SIMILARITY: Belongs to the NAD-dependent DNA ligase family. SIMILARITY:
ECnumber	6.5.1.2
PubmedID	3018436
ProductType	e : enzyme
Localization	2 : Cytoplasmic
Class	2a : Function of homologous gene experimentally demonstrated in an other organism
BioProcess	ADD DEL
Roles	ADD DEL 2.1.4 : DNA repair ; 7.1 : Cytoplasm ;

CANCEL SAVE

Automatic Annotation : CENIA1328

Type	Begin	End	Length	Frame	Mutation	Gene	Synonyms	Date	Status
CDS	1346217	1348307	2091	+3	no	ligA	lig, dnaL, lop, pdeC	2005-05-10 00:00:00	finished

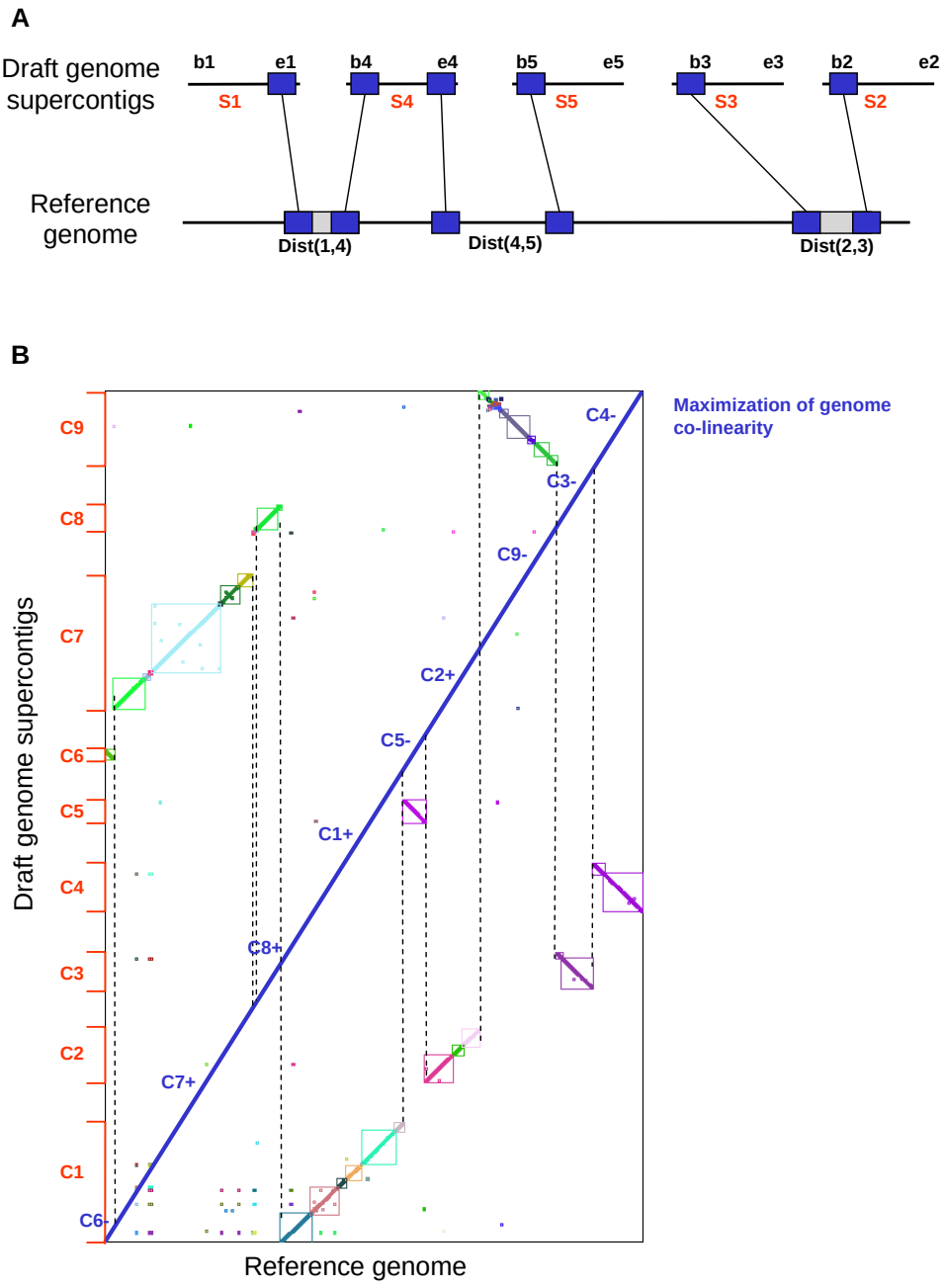
Note	similar to DNA ligase, NAD-dependent (EC 6.5.1.2) from Burkholderia mallei (Pseudomonas mallei), swall : Q62JB9 (691 aa), Evalue = 9.11511e-259, %identity = 67.05, on 677 aa ; and similar to DNA ligase (EC 6.5.1.2) from Burkholderia pseudomallei (Pseudomonas pseudomallei), swall : Q63T07 (691 aa), Evalue = 2.03063e-258, %identity = 67.05, on 677 aa ; and similar to NAD-dependent DNA ligase (EC 6.5.1.2) from Azoarcus sp. (strain EbN1), swall : Q5NYU3 (681 aa), Evalue = 3.49602e-226, %identity = 62.44, on 662 aa.
Product	DNA ligase
Comments	_
ECnumber	6.5.1.2
PubmedID	_
ProductType	_
Localization	_
Class	_
BioProcess	_
Roles	2.1.4 : DNA repair ; 7.1 : Cytoplasm ;

TrEMBL alignments | SwissProt alignments

- **All**
- **AMIGene** (1 Results ordered by **Begin**)
- **Duplications** (0 Results ordered by **Eval**)
- **E. coli Ecogene** (1 Results ordered by **Eval**)
- **B. subtilis** (1 Results ordered by **Eval**)
- **Acinetobacter ADP1** (1 Results ordered by **Eval**)
- **Syntonome** (20 Results ordered by **NbGeneQ**)
- **Syntonome RefSeq** (33 Results ordered by **NbGeneQ**)
- **HAMAP** (0 Results ordered by **H_id**)
- **Similarities SwissProt** (10 Results ordered by **Eval**)
- **Similarities TrEMBL** (10 Results ordered by **Eval**)
- **PRIAM EC number** (1 Results ordered by **Evidence**)
- **COGnitor** (4 Results ordered by **Score**)
- **InterProScan** (20 Results ordered by **IP_id**)
- **SignalP** (0 Results ordered by **SP_proba**)
- **TMhmm** (0 Results ordered by **TM_begin**)
- **All**

I.III Données supplémentaires de l'article 6.

Supplementary figure 1: Ordering supercontigs with synteny results.

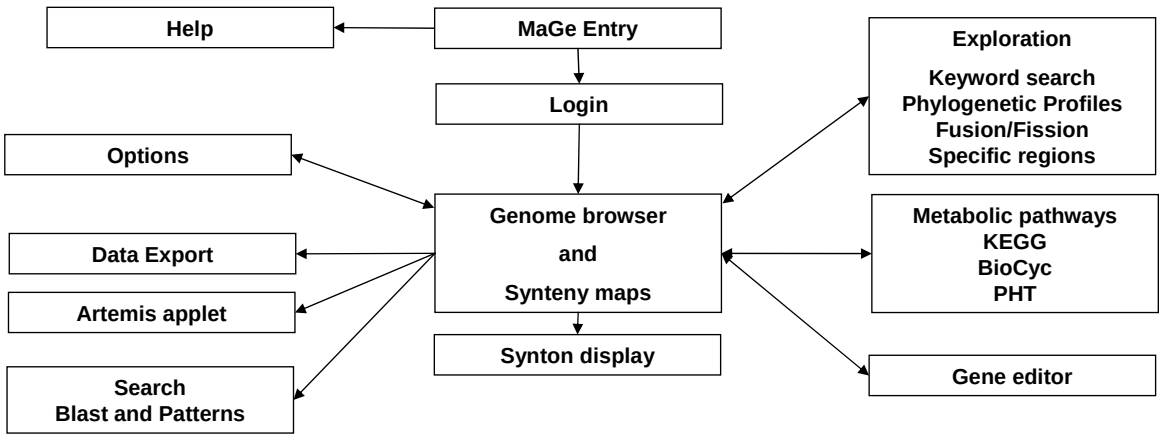


Two strategies relying on synteny results are used in MaGe to find one (or several) possible supercontig organizations of a draft genome.

A. A distance in bases is determined between two supercontigs in comparison with a reference genome. Synteny groups on the supercontig ends are mapped on the reference genome and the minimal distance between pairs of supercontigs is then computed. If the distance is lower than a defined threshold, a link between the two supercontigs is retained. In this example, the link between supercontigs S1 and S4 and the one between S3 and S2 are kept. The S3 begin (b3) matches with the S2 begin (b2), so the reverse sequence of S3 can be associated with S2.

B. The second method also uses synteny results. The draft genome is made of 9 supercontigs (C1 to C9) and it is compared to a reference genome. Dotplot points represent gene correspondences between the two genomes (e.g. blastP similarity results). Points inside a rectangle which have the same color, symbolize a synteny group. Guided by this representation, the user can then order the supercontigs and assign their relative orientation. In this example, the proposed order is the following: C6-/C7+/C8+/C1+/C5-/C2+/C9-/C3-/C4- (plus and minus symbols refer to direct and reverse orientations of the supercontig).

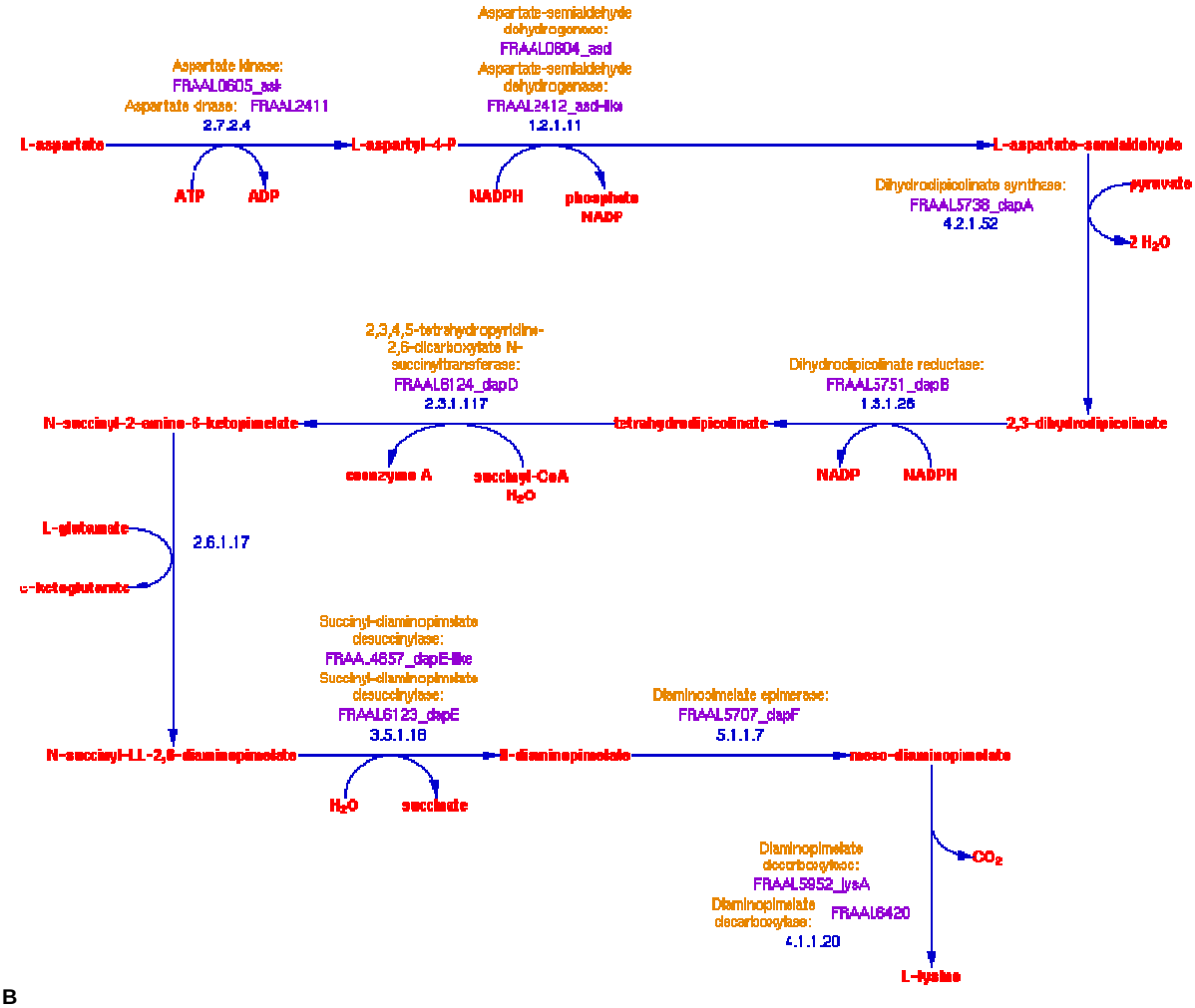
Supplementary figure 2: Structure of the MaGe web server



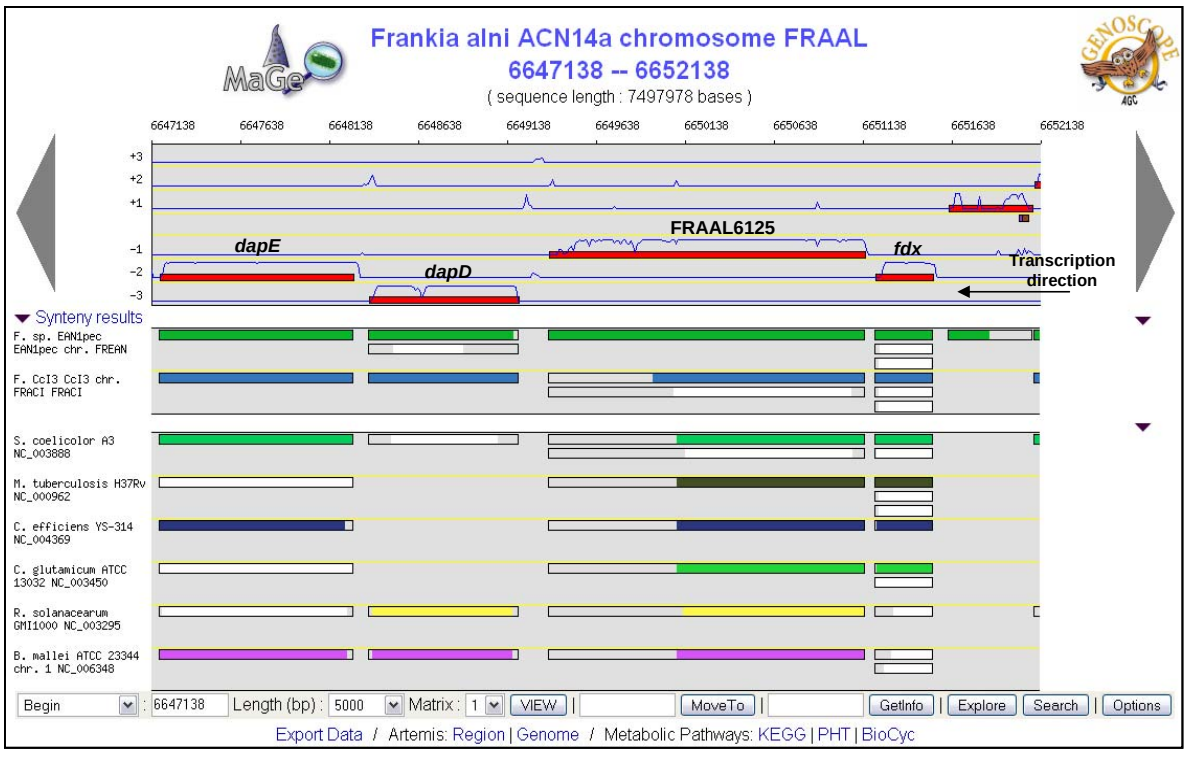
Starting from the 'Genome browser', users can navigate through web pages dealing with several functionalities and various aspects of annotations.

Supplementary figure 3: Lysine biosynthesis in *Frankia alni* genome through the MaGe interfaces.

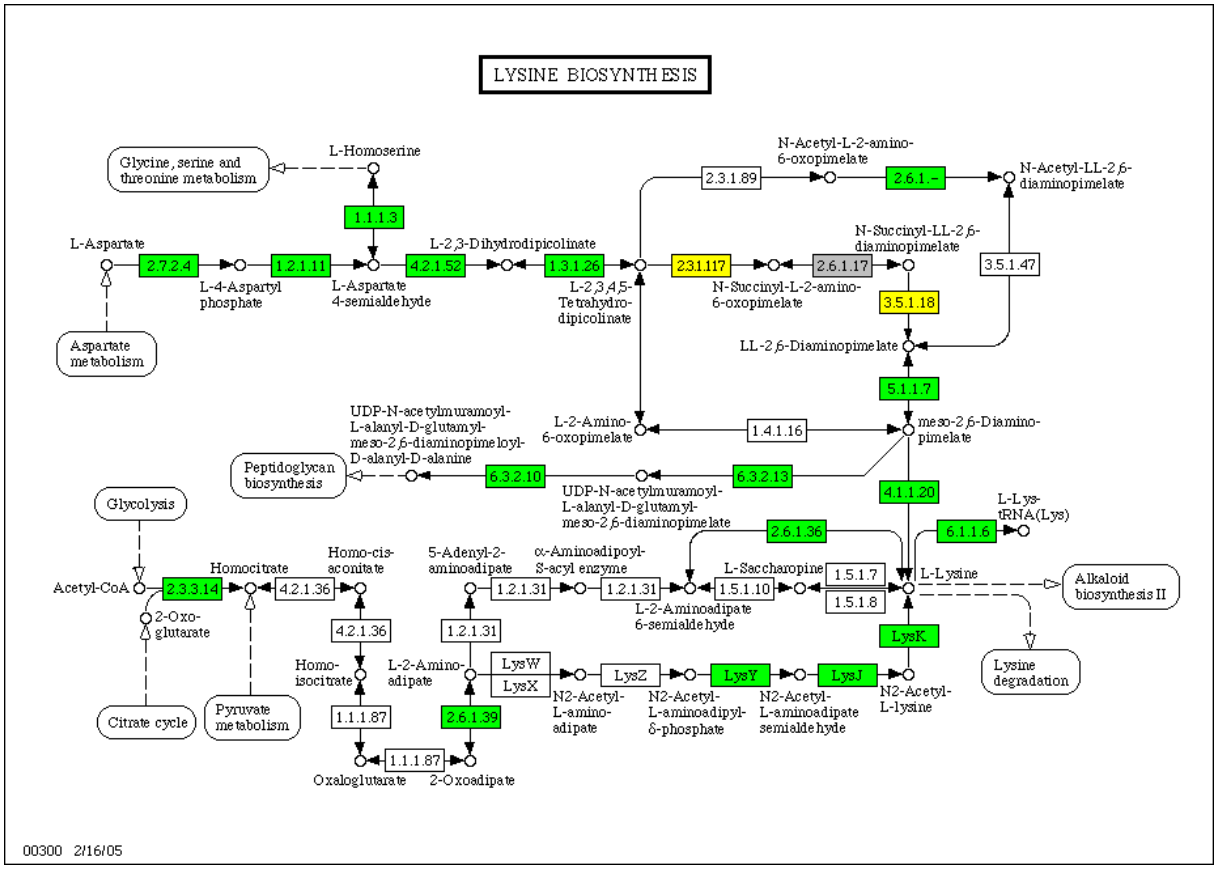
A



B



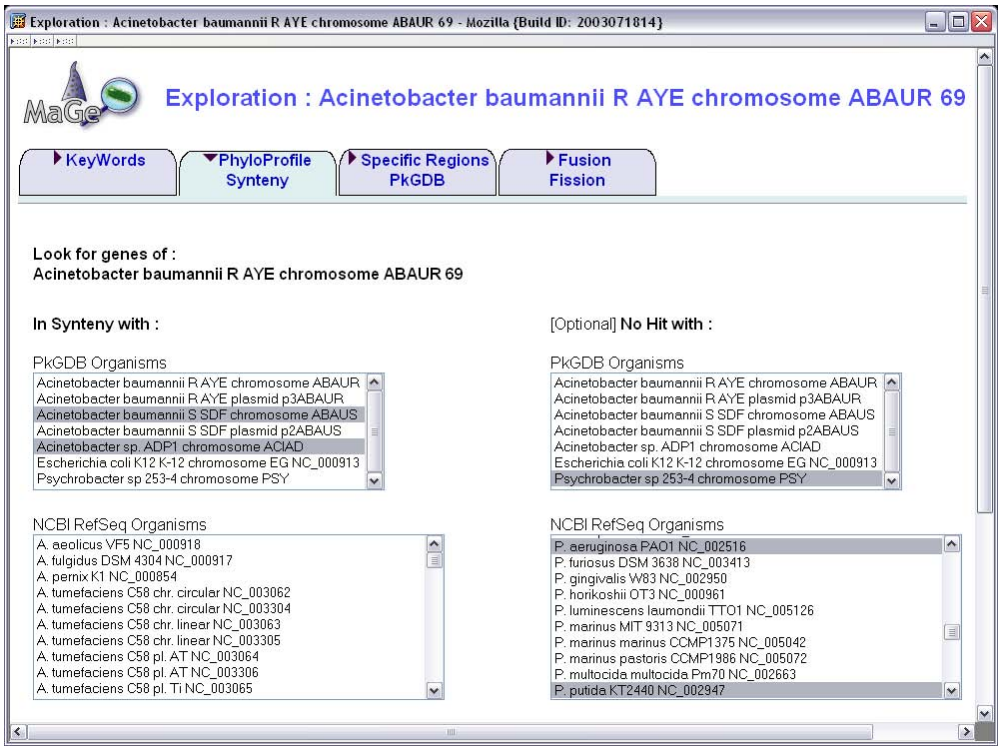
C



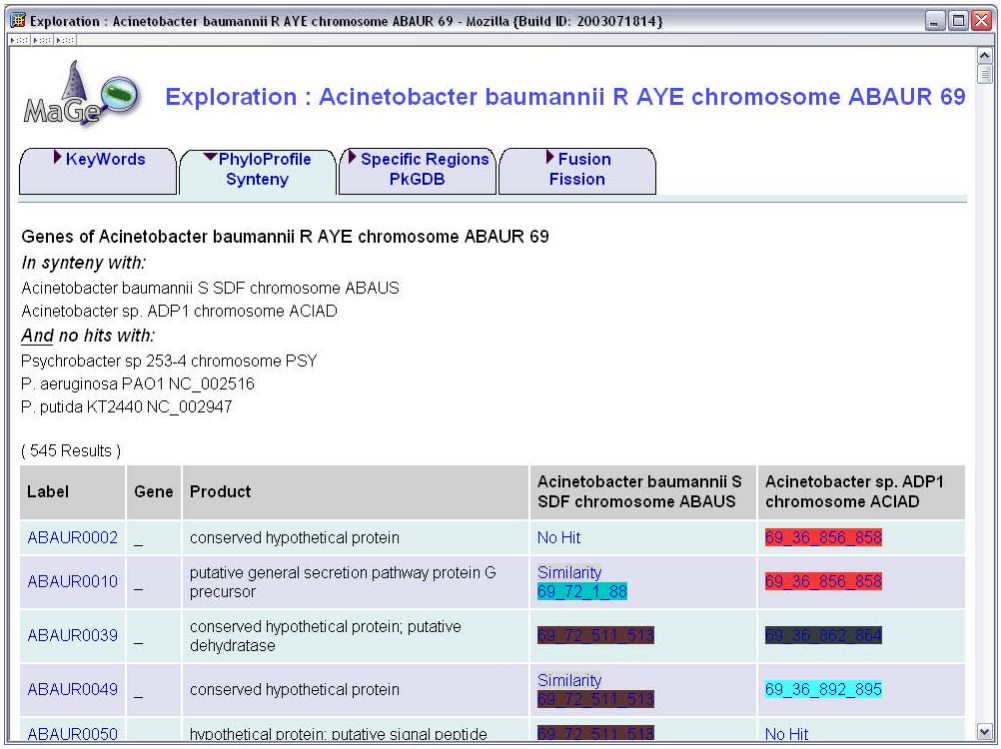
Three screenshots showing lysine biosynthesis in *Frankia alni*. The FrankiaCyc Pathway/Genome DataBase (PGDB) is available through MaGe via a BioCyc web server (A). In addition, the user can obtain KEGG maps by comparison with *Escherichia coli* (C). Yellow rectangles symbolise enzymes encoded by genes in the selected MaGe region (B) while green rectangles represent enzymes encoded by genes localized elsewhere in the studied genome. Grey boxes correspond to known enzymes in *E. coli* that are not present in the genome under study. Lastly, white boxes are enzymatic activities missing in both organisms.
The BioCyc pathway selection algorithm reports only one possible pathway for lysine biosynthesis (A) in *F. alni*. The reported pathway apparently lacks the gene(s) encoding the succinyldiaminopimelate amino transferase activity (EC number 2.6.1.17). The lysine biosynthesis map from KEGG (C) also reports the lack of succinyldiaminopimelate amino transferase activity which has been detected in *E. coli*. Furthermore, genomic context exploration of the genes involved in this pathway, via the MaGe genome browser (B), reveals that the gene FRAAL6125 is co-localized with the characterized *dapE* and *dapD* genes. FRAAL6125 is a good candidate for *dapC*, a gene coding the missing activity and experimentally described in other species (for further details see ‘Metabolic pathway visualization’ section).

Supplementary figure 4: MaGe data exploration.

A

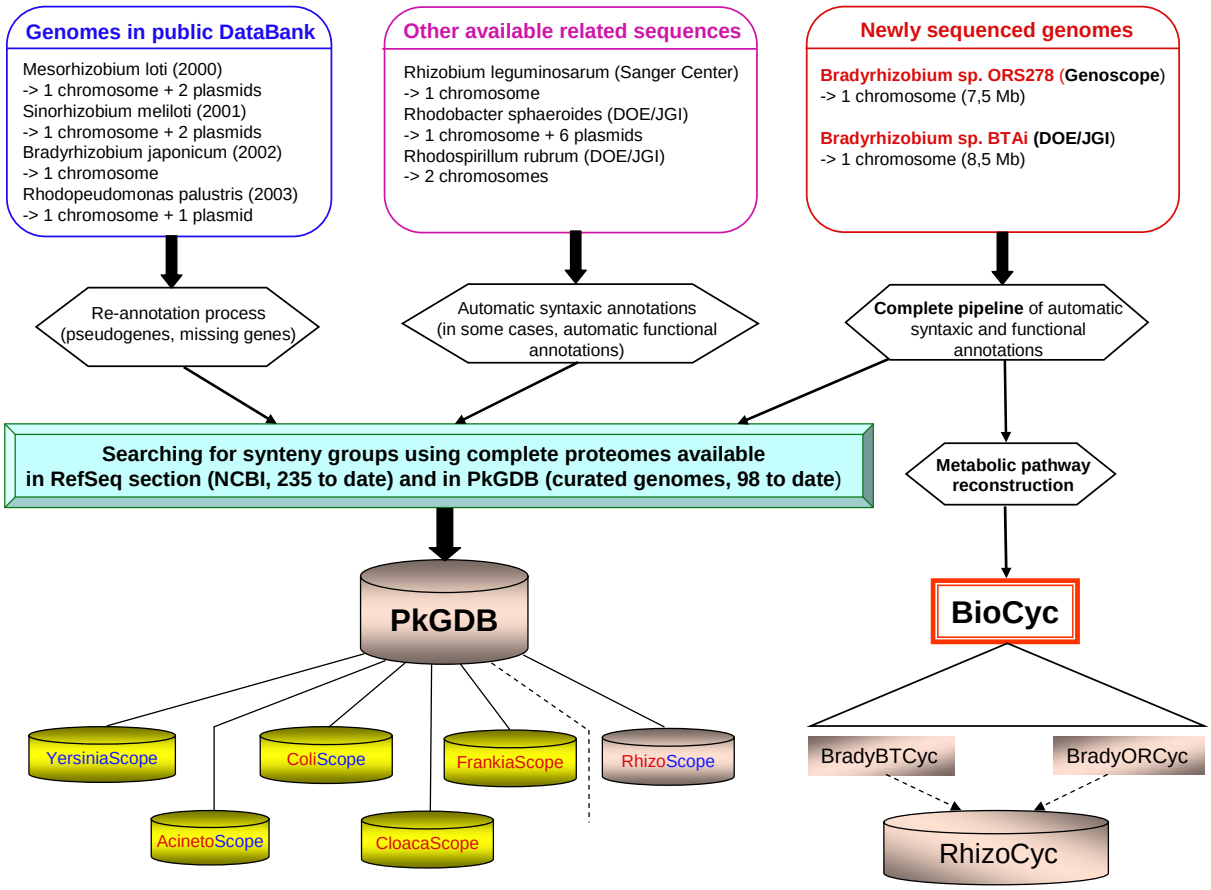


B



Two screenshots of MaGe 'Exploration' functionality are shown as examples of the use of 'PhyloProfile/Synteny' search.
A. Selecting the 'PhyloProfile/Synteny' section, the user can search for genes of *Acinetobacter baumannii* AYE which are homologs to genes in certain organisms (*Acinetobacter* ADP1 and *A. baumannii* SDF) and exclude those that are homologs to genes in other organisms (*Psychrobacter* sp. 253-4, *Pseudomonas aeruginosa* and *P. putida*).
B. The query output is a list of 545 *A. baummannii* AYE genes. The user can then explore gene groups which are specific to the Acinetobacter genus and have a same chromosomal organization (colored rectangles symbolize synteny groups)

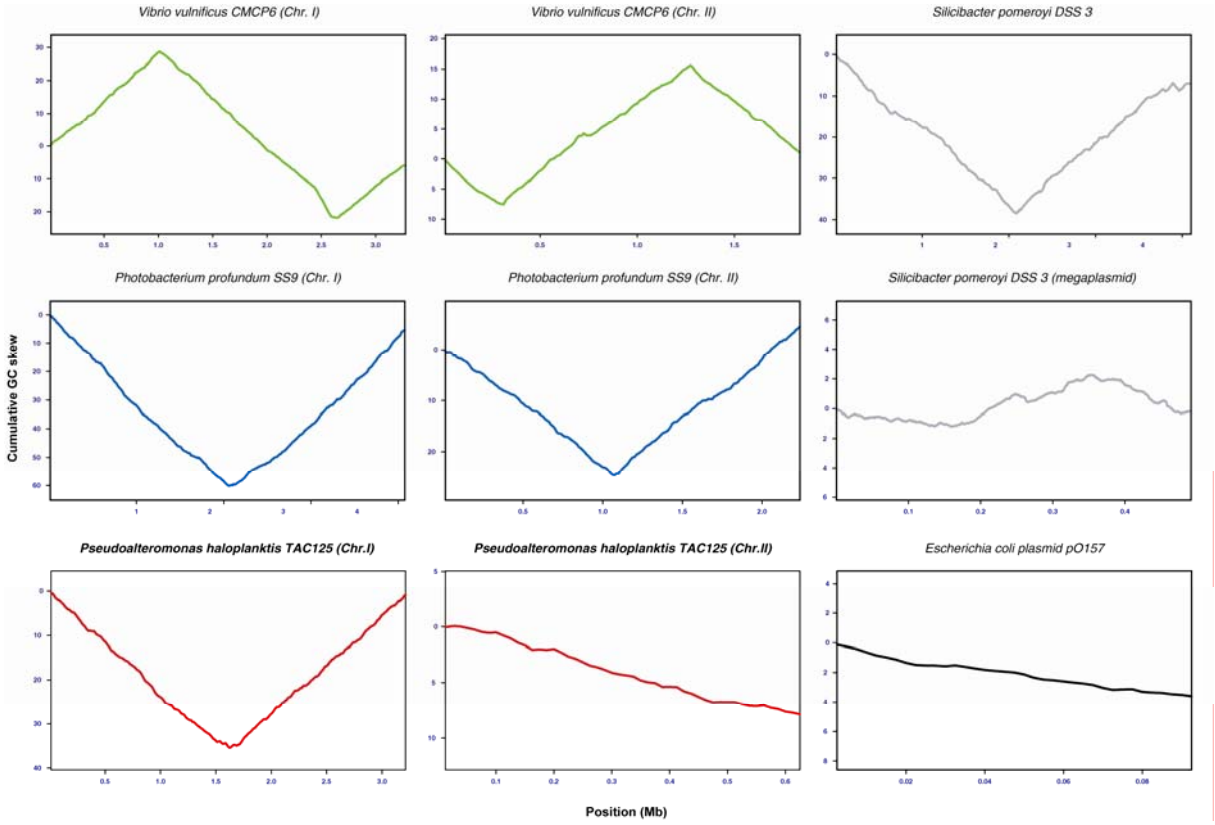
Supplementary figure 5: Setting up a new annotation project: an example.



To set up a new annotation project (here the annotation of two new Bradyrhizobium species) the first step consists in gathering the available genomic sequences from organisms of interest in PkGDB. These sequences are submitted to various procedures (lozenges), which end with the computation of synteny groups with the set of complete prokaryotic proteomes. A new thematic database is then created (here RhizoScope), the data of which are partly publicly available (*i.e.*, only data corresponding to genomes already stored in public DataBanks; blue color of the word 'Scope'). As shown in this figure, some thematic databases are only accessible by the group of experts (*i.e.*, FrankiaScope, CloacaScope in red), and others are freely available (*i.e*, YersiniaScope in blue). The RhizoScope database contains links to the BradyBTCyc and BradyORCyc metabolic databases which have been built using the BioCyc software. In addition we have recently integrated these metabolic data in the relational scheme of BioWareHouse (MySQL database system; http://bioinformatics.ai.sri.com/biowarehouse). The corresponding database (here RhizoCyc) is very useful for analysis of metabolic content of the compared genomes.

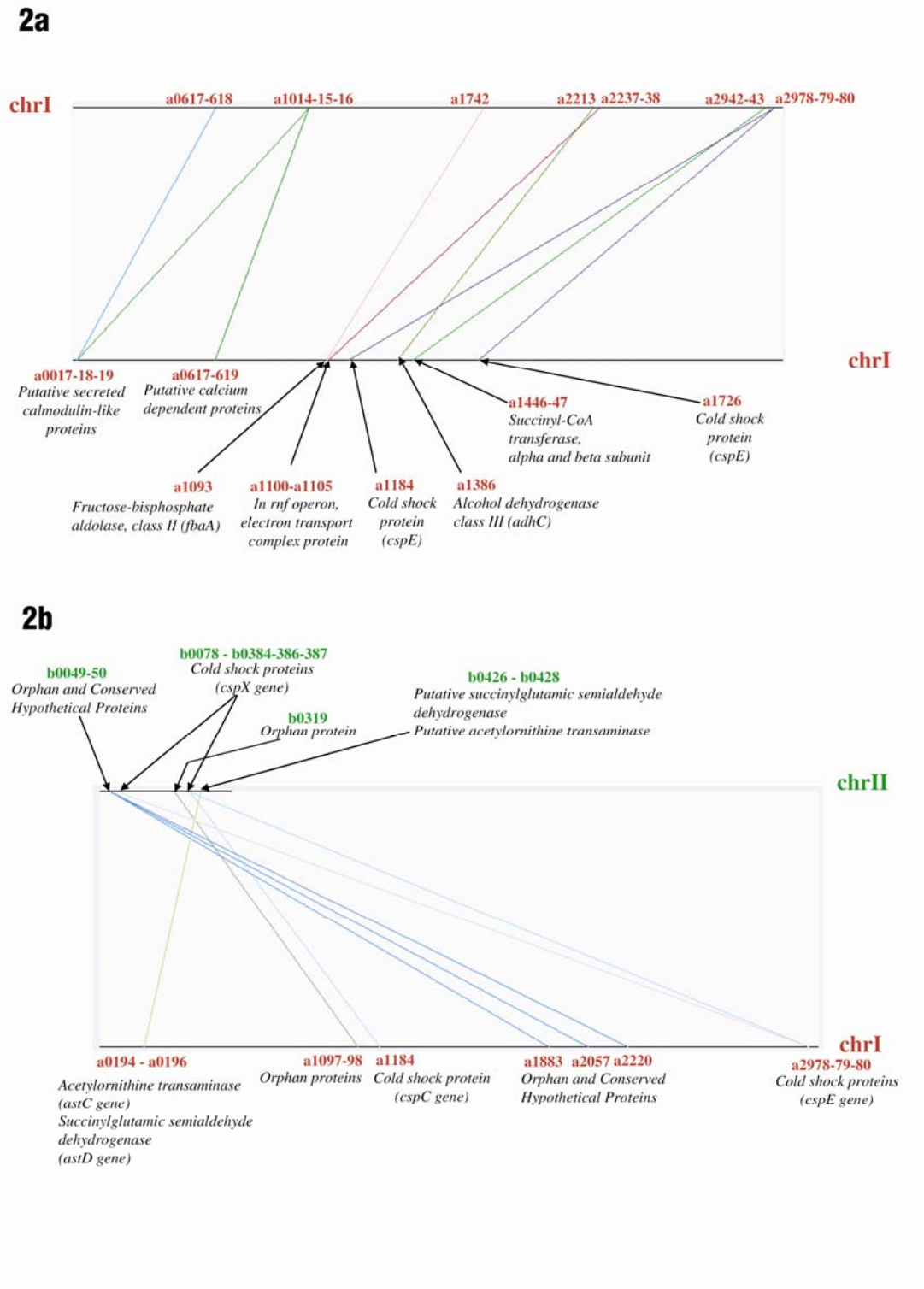
I.IV Données supplémentaires de l'article 3.

I.IV.1 Supplementary Figure 1.



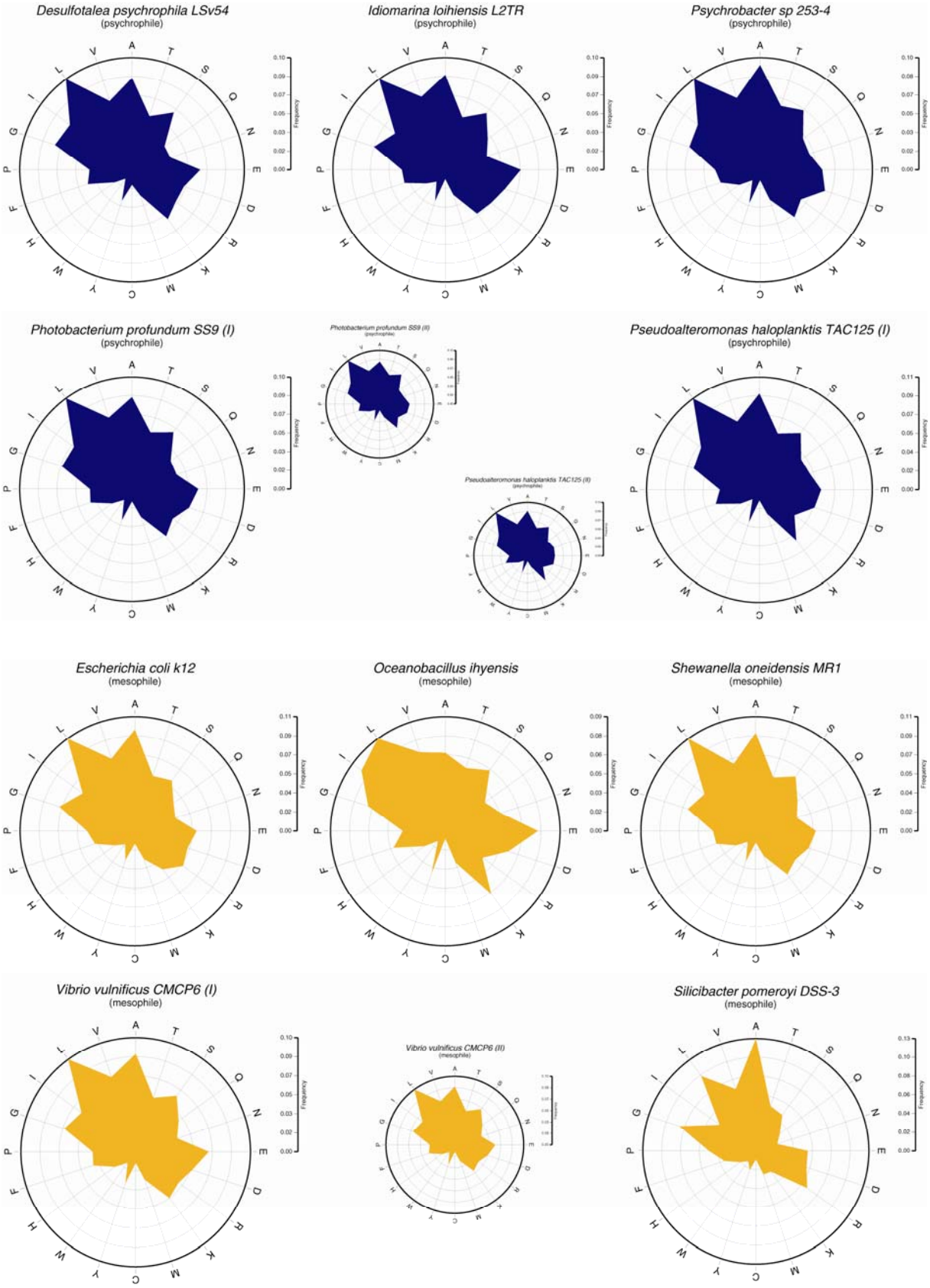
Cumulative GC skew computed on whole chromosome/plasmid of various species. The maximum value of the composite skew index is located at *oriC*, the origin of replication, and the minimum value at *terC*, where termination occurs preferentially. All chromosomes display a typical pattern for bidirectional replication, except for *P. haloplanktis* TAC125 chrII.

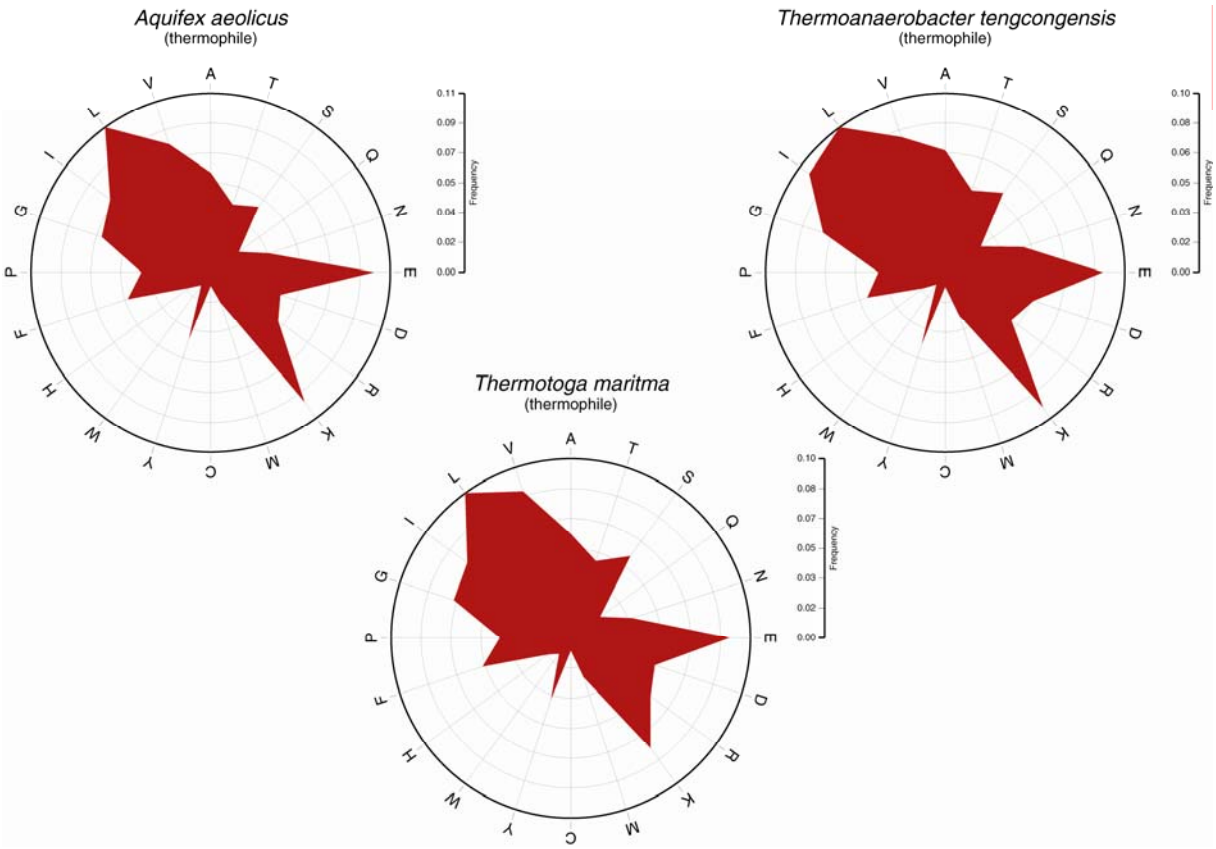
I.IV.2 Supplementary Figure 2.



Duplication events found in chrI (**2a**), and between chrI and chrII (**2b**) of *P. haloplanktis*. The similarity identity threshold between proteins encoded by chrI genes has been set up to 40%, and that between proteins encoded by chrI and chrII genes, has been set up to 70%. In both case, a ratio of 0.8 of the length of the smallest protein is required to establish duplication.

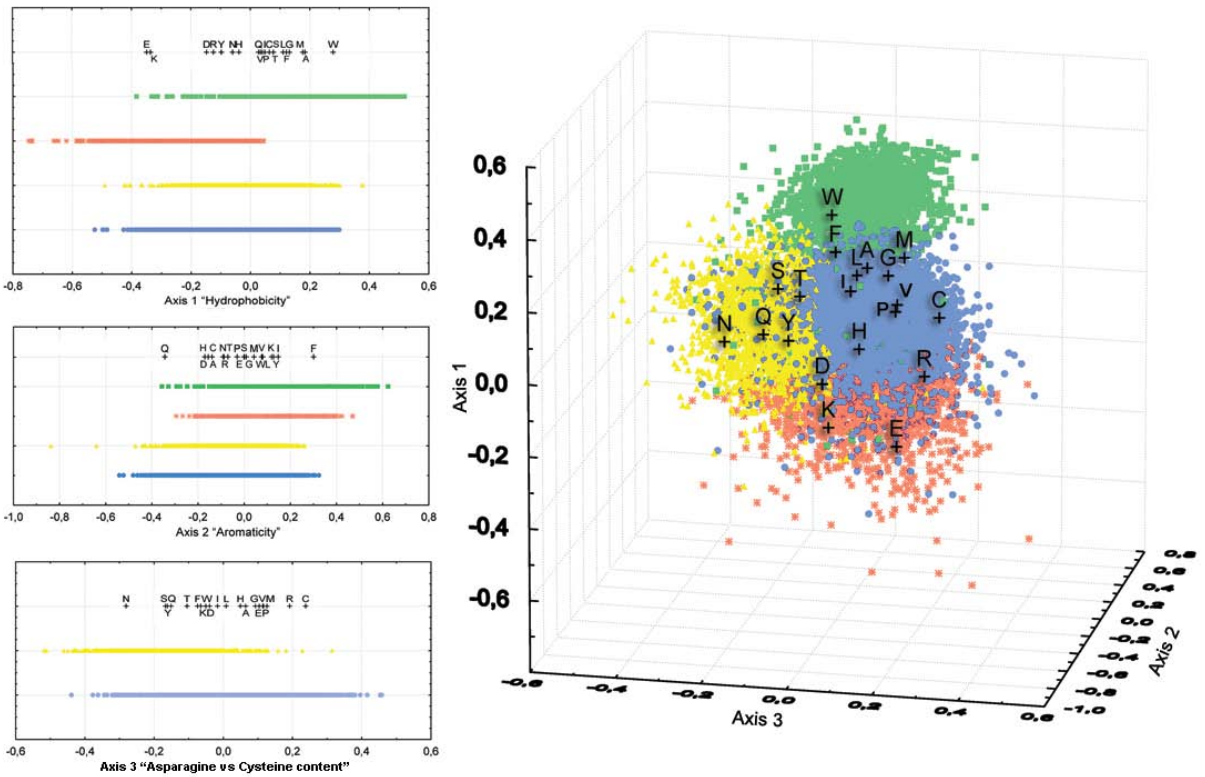
I.IV.3 Supplementary Figure 3.





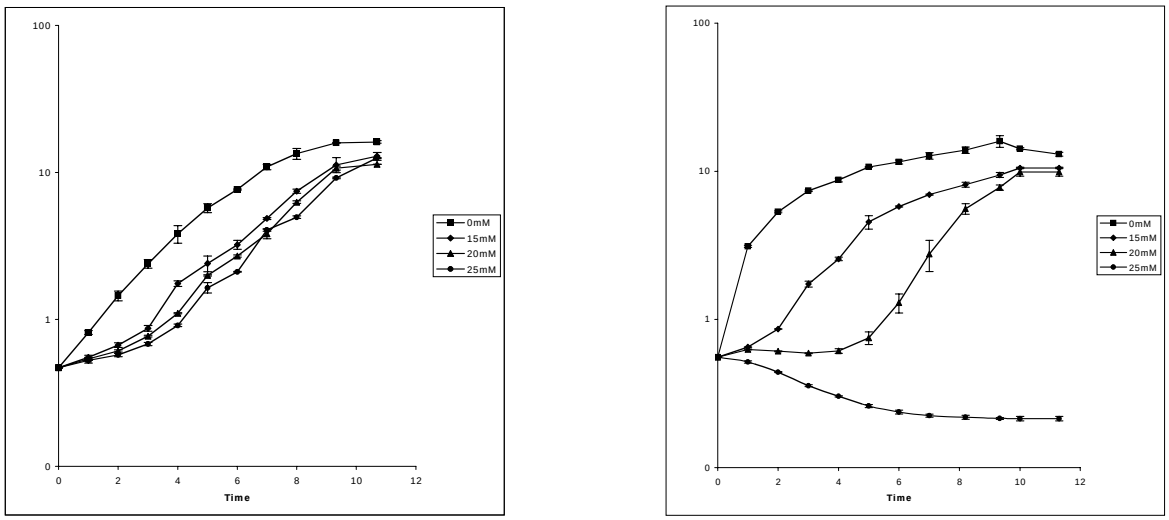
Radar plots showing amino acids frequencies for various bacterial species. Psychrophile, mesophile and thermophile species are reported in blue, orange and red respectively. These plots were made using a PERL script kindly provided by P. F. Hallin (Ussery, 2004).

I.IV.4 *Supplementary Figure 4.*



The 3D graph represents the results of a Correspondence Analysis performed with the pool of two psychrophilic (*D. psychrophila* and *P. haloplanktis*), two mesophilic (*E. coli K-12* and *B. subtilis*) and two thermophilic (*A. aeolicus* and *T. maritima*) bacteria. Axis 1 discriminates proteins by their hydrophobicity, axis 2 by their aromaticity, axis 3 by their asparagine content versus cysteine content. Clustering method gave four groups of proteins. Most of proteins in the orange stars cluster are thermophilic proteins (62% of thermophiles). The group of green squares is composed by integral inner membrane proteins. The blue circles cluster is composed in majority by mesophilic proteins (53% of mesophiles) and most of proteins in the yellow triangles group are psychrophilic (55% of psychrophiles). The amino acid position in this factorial space is shown with a black plus sign.

I.IV.5 Supplementary Figure 5.

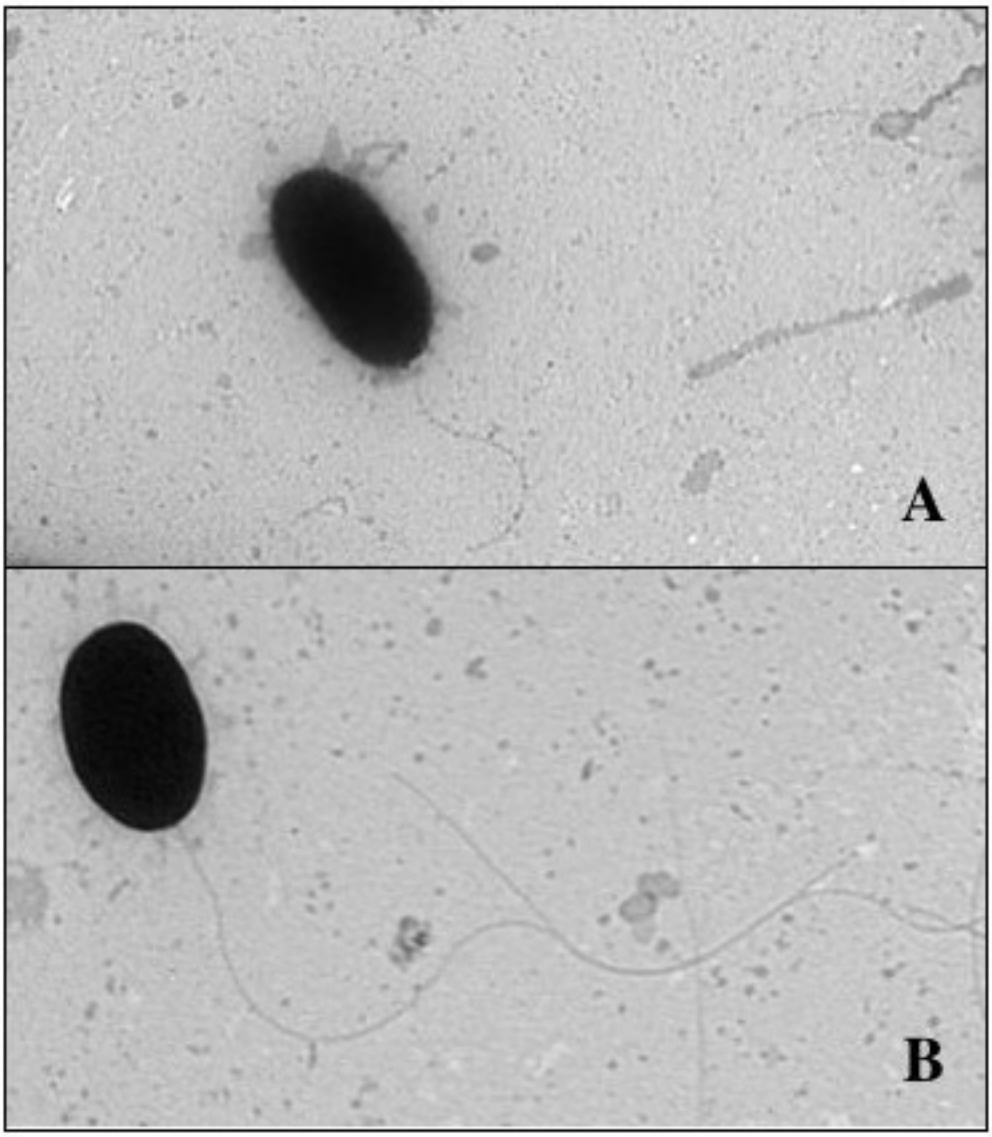


Pseudoalteromonas haloplanktis TAC 125 bacteria (left panel) were grown in parallel with *Escherichia coli* MG1655 cells (right panel) in rich medium supplemented with H_2O_2 at the onset of the growth culture. As can been seen in the figures *P. haloplanktis* is remarkably resistant to H_2O_2: at 25 mM concentration of the toxic molecules the bacteria grow after a slight delay almost as fast as the counterpart without the toxic, while *E. coli* cells are killed and lyse immediately.

I.IV.6 Supplementary Figure 6.

Biofilm assay: Overnight cultures in TYP medium were resuspended in 300 µl of the same medium at an OD_{600} of 0.5 and incubated for 2 days at 4°C in borosilicate tubes.

I.IV.7 *Supplementary Figure 7.*



Effect of salt concentration on polar flagellum synthesis in *P. haloplanktis* TAC125.

Bacterial cells were grown at 4°C under moderate shaking in rich medium with NaCl 0.5% (A) or 2.5% (B) and stained with 0,1% (v/v) osmium tetroxyde prepared in water.

I.V Classes fonctionnelles utilisées lors de l'annotation de *P. haloplanktis*

1 Cell envelope and membrane-associated cellular processes
1.1 Cell wall
1.1.1 Biosynthesis of murein sacculus and peptidoglycan
1.1.2 Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides
1.1.3 Surface structures (S-layers?)
1.1.4 Other
1.2 Transport/binding proteins and lipoproteins
1.2.1 Amino acids, peptides and amines
1.2.2 Anions
1.2.3 Carbohydrates, organic alcohols, and acids
1.2.4 Cations
1.2.5 Nucleosides, purines and pyrimidines
1.2.6 Porins
1.2.7 Unknown substrates
1.2.8 Other
1.3 Sensors (signal transduction)
1.3.1 Sensors for chemicals
1.3.2 Mechanosensors
1.3.4 Other
1.4 Membrane bioenergetics (electron transport chain and ATP synthase)
1.4.1 ATP-proton motive force interconversion
1.4.2 Electron transport
1.5 Motility and chemotaxis
1.5.1 Chemotaxis
1.5.3 Motility
1.6 Protein secretion
1.6.1 Protein and peptide secretion and trafficking
1.6.2 Toxin production and resistance
1.7 Cell division
1.7.2 Cell division
1.8 Horizontal gene transfer
1.8.1 Competence
1.8.2 Conjugation
1.9 Collective behaviour
1.9.1 Biofilm formation
2 Intermediary metabolism
2.1 Metabolism of carbohydrates and related molecules
2.1.1 Main glycolytic pathways
2.1.1.1 Glycolysis/gluconeogenesis
2.1.1.2 Pentose-phosphate pathway
2.1.1.3 Entner-Doudoroff pathway
2.1.1.4 Fermentation
2.1.1.5 Pyruvate dehydrogenase and input/output into the TCA cycle
2.1.2 TCA and related cycles
2.1.2.1 TCA cycle
2.1.2.2 Glyoxylate cycle
2.1.2.3 Others (oxaloacetate?)
2.1.3 Metabolism of carbohydrates and related compounds
2.1.3.1 Carbohydrates (other than glucose)
2.1.3.2 Sugar alcohols
2.1.3.3 Sugar acids
2.1.3.4 Amino sugars
2.1.3.5 Biosynthesis and degradation of polysaccharides
2.1.3.6 Others
2.2 Metabolism of amino acids and related molecules
2.2.1 Amino acid biosynthesis and salvage
2.2.1.1 Aromatic amino acid family
2.2.1.2 Aspartate family
2.2.1.3 Glutamate family
2.2.1.4 Pyruvate family
2.2.1.5 Serine family
2.2.1.6 Histidine family
2.2.1.7 Other
2.2.2 Polyamines biosynthesis
2.2.3 Degradation of aminoacids and related molecules
2.2.3.1 Degradation of proteins, peptides, and glycopeptides
2.2.3.2 Degradation of amino acids
2.2.3.3 Degradation of polyamines
2.3 Metabolism of nucleotides and nucleic acids
2.3.1 Biosynthesis and salvage of purines, pyrimidines, nucleosides, and nucleotides
2.3.1.1 Purine ribonucleotide biosynthesis
2.3.1.2 Pyrimidine ribonucleotide biosynthesis
2.3.1.3 Deoxyribonucleotides biosynthesis
2.3.1.4 Salvage of nucleosides and nucleotides
2.3.2 Degradation of purine and pyrimidines
2.3.3 Sugar-nucleotide biosynthesis and conversions
2.3.4 Other
2.4 Metabolism of lipids
2.4.1 Fatty acid and phospholipid metabolism
2.4.1.1 Biosynthesis
2.4.1.2 Degradation
2.4.2 Terpenes and related-molecules
2.4.3 Other
2.5 Metabolism of coenzymes and prosthetic groups
2.5.1 Biotin
2.5.2 Folic acid
2.5.3 Lipoate
2.5.4 Menaquinone and ubiquinone
2.5.5 Molybdopterin
2.5.6 Pantothenate and coenzyme A
2.5.7 Pyridoxine and derivatives
2.5.8 Riboflavin, FMN, and FAD
2.5.9 Glutathione
2.5.10 Thiamine
2.5.11 Pyridine nucleotides
2.5.12 Heme, porphyrin, and cobalamin
2.5.14 Other
2.6 Specific pathways
2.6.4 Other
2.7 Metabolism of phosphate
2.8 Metabolism of sulfur
3 Information transfer pathways
3.1 DNA metabolism
3.1.1 DNA replication
3.1.2 DNA recombination
3.1.3 DNA packaging and segregation
3.1.4 DNA proof-reading and repair
3.1.5 DNA restriction/modification
3.1.6 Degradation of DNA
3.2 RNA metabolism
3.2.1 Transcription
3.2.1.1 Regulation
3.2.1.1.1 Regulatory proteins (DNA interaction)
3.2.1.1.2 Regulatory proteins (RNA interaction)
3.2.1.1.3 Regulatory RNAs
3.2.1.2 Initiation
3.2.1.2.1 Sigma factors
3.2.1.2.2 Other
3.2.1.3 Elongation
3.2.1.3.1 RNA polymerase (core enzyme)
3.2.1.3.2 Transcription / translation coupling
3.2.1.3.3 Other
3.2.1.4 Termination
3.2.2 Processing and degradation of RNA
3.2.2.1 RNA chaperones
3.2.2.2 Ribonucleases
3.2.2.3 Other
3.2.3 RNA modification
3.2.3.1 tRNA modifications
3.2.3.2 rRNA modifications
3.2.3.3 Other modifications
3.3 Protein synthesis, maturation and folding
3.3.1 Ribosomal proteins
3.3.2 Aminoacyl-tRNA synthetases
3.3.3 Initiation
3.3.4 Elongation
3.3.5 Termination

3.3.6 Protein modifications
3.3.6.1 Maturation
3.3.6.2 Chemical modifications
3.3.7 Protein folding
3.3.8 Protein repair
4 Other functions
4.1 Adaptation to atypical conditions
4.2 Detoxification
4.3 Antibiotic production
4.4 Phage-related functions
4.5 Transposons and ISs
4.6 Miscellaneous
4.6.1 Plasmid-related functions
4.6.2 Pathogenesis
5 Similar to unknown proteins
5.1 From P. haloplanktis
5.2 From other organisms
6 No similarity
6.2 Hypothetica

I.VI Proportions des clusters de l'AFC de *P. haloplanktis* selon les classes fonctionnelles de l'annotation.

Fonction	Cluster bleu	Cluster jaune	Cluster rouge	Cluster rose	Cluster vert
1 : cell envelope and membrane-associated cellular processes	8/115 6%	24/115 20%	8/115 6%	56/115 48%	19/115 16%
1.1 : cell wall	2/6 33%	2/6 33%	0/6 0%	2/6 33%	0/6 0%
1.1.1 : biosynthesis of murein sacculus and peptidoglycan	15/35 42%	7/35 20%	4/35 11%	2/35 5%	7/35 20%
1.1.2 : biosynthesis and degradation of surface polysaccharides and lipopolysaccharides	11/30 36%	9/30 30%	4/30 13%	1/30 3%	5/30 16%
1.1.3 : surface structures (s-layers?)	1/5 20%	4/5 80%	0/5 0%	0/5 0%	0/5 0%
1.1.4 : other	1/2 50%	0/2 0%	0/2 0%	0/2 0%	1/2 50%
1.2 : transport/binding proteins and lipoproteins	15/90 16%	14/90 15%	8/90 8%	33/90 36%	20/90 22%
1.2.1 : amino acids, peptides and amines	0/14 0%	0/14 0%	1/14 7%	13/14 92%	0/14 0%
1.2.2 : anions	0/6 0%	0/6 0%	1/6 16%	3/6 50%	2/6 33%
1.2.3 : carbohydrates, organic alcohols, and acids	0/12 0%	1/12 8%	2/12 16%	8/12 66%	1/12 8%
1.2.4 : cations	12/60 20%	10/60 16%	5/60 8%	28/60 46%	5/60 8%
1.2.5 : nucleosides, purines and pyrimidines	0/2 0%	0/2 0%	0/2 0%	2/2 100%	0/2 0%
1.2.6 : porins	0/1 0%	1/1 100%	0/1 0%	0/1 0%	0/1 0%
1.2.7 : unknown substrates	8/57 14%	4/57 7%	11/57 19%	27/57 47%	7/57 12%
1.2.8 : other	0/14 0%	1/14 7%	2/14 14%	6/14 42%	5/14 35%
1.3 : sensors (signal transduction)	4/31 12%	7/31 22%	5/31 16%	0/31 0%	15/31 48%
1.3.1 : sensors for chemicals	0/11 0%	2/11 18%	2/11 18%	0/11 0%	7/11 63%
1.3.2 : mechanosensors	1/3 33%	0/3 0%	0/3 0%	1/3 33%	1/3 33%
1.3.4 : other	0/3 0%	1/3 33%	0/3 0%	0/3 0%	2/3 66%
1.4 : membrane bioenergetics (electron transport chain and atp synthase)	6/15 40%	0/15 0%	1/15 6%	4/15 26%	4/15 26%
1.4.1 : atp-proton motive force interconversion	4/9 44%	0/9 0%	2/9 22%	3/9 33%	0/9 0%
1.4.2 : electron transport	13/39 33%	4/39 10%	2/39 5%	13/39 33%	7/39 17%
1.5 : motility and chemotaxis	6/16 37%	1/16 6%	7/16 43%	1/16 6%	1/16 6%
1.5.1 : chemotaxis	3/10 30%	0/10 0%	0/10 0%	0/10 0%	7/10 70%
1.5.3 : motility	13/35 37%	2/35 5%	10/35 28%	2/35 5%	8/35 22%

Fonction	Cluster bleu	Cluster jaune	Cluster rouge	Cluster rose	Cluster vert
1.6 : protein secretion	8/32 25%	2/32 6%	9/32 28%	0/32 0%	13/32 40%
1.6.1 : protein and peptide secretion and trafficking	1/12 8%	1/12 8%	5/12 41%	3/12 25%	2/12 16%
1.6.2 : toxin production and resistance	0/7 0%	1/7 14%	1/7 14%	3/7 42%	2/7 28%
1.7 : cell division	3/18 16%	1/18 5%	8/18 44%	2/18 11%	4/18 22%
1.7.2 : cell division	4/10 40%	0/10 0%	6/10 60%	0/10 0%	0/10 0%
1.8.1 : competence	1/1 100%	0/1 0%	0/1 0%	0/1 0%	0/1 0%
1.8.2 : conjugation	1/2 50%	1/2 50%	0/2 0%	0/2 0%	0/2 0%
1.9 : collective behaviour	1/1 100%	0/1 0%	0/1 0%	0/1 0%	0/1 0%
1.9.1 : biofilm formation	0/3 0%	3/3 100%	0/3 0%	0/3 0%	0/3 0%
2 : intermediary metabolism	5/7 71%	0/7 0%	0/7 0%	0/7 0%	2/7 28%
2.1 : metabolism of carbohydrates and related molecules	7/14 50%	2/14 14%	2/14 14%	0/14 0%	3/14 21%
2.1.1 : main glycolytic pathways	2/3 66%	1/3 33%	0/3 0%	0/3 0%	0/3 0%
2.1.1.1 : glycolysis/gluconeogenesis	15/17 88%	1/17 5%	1/17 5%	0/17 0%	0/17 0%
2.1.1.2 : pentose-phosphate pathway	4/6 66%	0/6 0%	2/6 33%	0/6 0%	0/6 0%
2.1.1.3 : entner-doudoroff pathway	2/2 100%	0/2 0%	0/2 0%	0/2 0%	0/2 0%
2.1.1.4 : fermentation	3/4 75%	0/4 0%	1/4 25%	0/4 0%	0/4 0%
2.1.1.5 : pyruvate dehydrogenase and input/output into the tca cycle	1/2 50%	0/2 0%	1/2 50%	0/2 0%	0/2 0%
2.1.2 : tca and related cycles	3/3 100%	0/3 0%	0/3 0%	0/3 0%	0/3 0%
2.1.2.1 : tca cycle	10/13 76%	0/13 0%	1/13 7%	1/13 7%	1/13 7%
2.1.2.2 : glyoxylate cycle	1/2 50%	1/2 50%	0/2 0%	0/2 0%	0/2 0%
2.1.2.3 : others (oxaloacetate?)	2/2 100%	0/2 0%	0/2 0%	0/2 0%	0/2 0%
2.1.3 : metabolism of carbohydrates and related compounds	10/13 76%	1/13 7%	1/13 7%	0/13 0%	1/13 7%
2.1.3.1 : carbohydrates (other than glucose)	5/6 83%	1/6 16%	0/6 0%	0/6 0%	0/6 0%
2.1.3.2 : sugar alcohols	2/4 50%	2/4 50%	0/4 0%	0/4 0%	0/4 0%
2.1.3.3 : sugar acids	0/2 0%	1/2 50%	1/2 50%	0/2 0%	0/2 0%
2.1.3.4 : amino sugars	5/7 71%	1/7 14%	0/7 0%	0/7 0%	1/7 14%
2.1.3.5 : biosynthesis and degradation of polysaccharides	2/11 18%	8/11 72%	1/11 9%	0/11 0%	0/11 0%
2.1.3.6 : others	2/2 100%	0/2 0%	0/2 0%	0/2 0%	0/2 0%
2.2 : metabolism of amino acids and related molecules	5/7 71%	1/7 14%	1/7 14%	0/7 0%	0/7 0%

Fonction	Cluster bleu	Cluster jaune	Cluster rouge	Cluster rose	Cluster vert
2.2.1 : amino acid biosynthesis and salvage	16/24 66%	2/24 8%	2/24 8%	1/24 4%	3/24 12%
2.2.1.1 : aromatic amino acid family	8/17 47%	0/17 0%	4/17 23%	0/17 0%	5/17 29%
2.2.1.2 : aspartate family	24/37 64%	3/37 8%	2/37 5%	0/37 0%	8/37 21%
2.2.1.3 : glutamate family	13/16 81%	0/16 0%	1/16 6%	0/16 0%	2/16 12%
2.2.1.4 : pyruvate family	4/6 66%	2/6 33%	0/6 0%	0/6 0%	0/6 0%
2.2.1.5 : serine family	10/22 45%	6/22 27%	3/22 13%	0/22 0%	3/22 13%
2.2.1.6 : histidine family	5/8 62%	0/8 0%	2/8 25%	0/8 0%	1/8 12%
2.2.1.7 : other	7/7 100%	0/7 0%	0/7 0%	0/7 0%	0/7 0%
2.2.2 : polyamines biosynthesis	5/10 50%	4/10 40%	0/10 0%	1/10 10%	0/10 0%
2.2.3 : degradation of amino acids and related molecules	3/7 42%	2/7 28%	0/7 0%	0/7 0%	2/7 28%
2.2.3.1 : degradation of proteins, peptides, and glycopeptides	16/54 29%	20/54 37%	11/54 20%	0/54 0%	7/54 12%
2.2.3.2 : degradation of amino acids	18/31 58%	6/31 19%	2/31 6%	0/31 0%	5/31 16%
2.2.3.3 : degradation of polyamines	1/1 100%	0/1 0%	0/1 0%	0/1 0%	0/1 0%
2.3 : metabolism of nucleotides and nucleic acids	1/4 25%	1/4 25%	2/4 50%	0/4 0%	0/4 0%
2.3.1 : biosynthesis and salvage of purines, pyrimidines, nucleosides, and nucleotides	8/11 72%	0/11 0%	2/11 18%	0/11 0%	1/11 9%
2.3.1.1 : purine ribonucleotide biosynthesis	14/17 82%	0/17 0%	2/17 11%	0/17 0%	1/17 5%
2.3.1.2 : pyrimidine ribonucleotide biosynthesis	4/11 36%	1/11 9%	5/11 45%	0/11 0%	1/11 9%
2.3.1.3 : deoxyribonucleotides biosynthesis	3/6 50%	2/6 33%	1/6 16%	0/6 0%	0/6 0%
2.3.1.4 : salvage of nucleosides and nucleotides	4/7 57%	1/7 14%	1/7 14%	0/7 0%	1/7 14%
2.3.2 : degradation of purine and pyrimidines	2/5 40%	1/5 20%	1/5 20%	0/5 0%	1/5 20%
2.3.3 : sugar-nucleotide biosynthesis and conversions	7/14 50%	3/14 21%	3/14 21%	0/14 0%	1/14 7%
2.3.4 : other	0/1 0%	0/1 0%	1/1 100%	0/1 0%	0/1 0%
2.4 : metabolism of lipids	7/15 46%	3/15 20%	3/15 20%	0/15 0%	2/15 13%
2.4.1 : fatty acid and phospholipid metabolism	20/30 66%	3/30 10%	1/30 3%	5/30 16%	1/30 3%
2.4.1.1 : biosynthesis	14/28 50%	5/28 17%	6/28 21%	2/28 7%	1/28 3%
2.4.1.2 : degradation	5/6 83%	0/6 0%	0/6 0%	0/6 0%	1/6 16%
2.4.2 : terpenes and related-molecules	3/7 42%	0/7 0%	0/7 0%	1/7 14%	3/7 42%
2.4.3 : other	0/1 0%	0/1 0%	0/1 0%	0/1 0%	1/1 100%
2.5 : metabolism of coenzymes and prosthetic groups	2/2 100%	0/2 0%	0/2 0%	0/2 0%	0/2 0%

Fonction	Cluster bleu	Cluster jaune	Cluster rouge	Cluster rose	Cluster vert
2.5.1 : biotin	1/8 12%	0/8 0%	0/8 0%	0/8 0%	7/8 87%
2.5.10 : thiamine	6/10 60%	0/10 0%	3/10 30%	0/10 0%	1/10 10%
2.5.11 : pyridine nucleotides	2/4 50%	0/4 0%	1/4 25%	0/4 0%	1/4 25%
2.5.12 : heme, porphyrin, and cobalamin	6/20 30%	1/20 5%	4/20 20%	3/20 15%	6/20 30%
2.5.14 : other	1/2 50%	0/2 0%	1/2 50%	0/2 0%	0/2 0%
2.5.2 : folic acid	2/10 20%	1/10 10%	3/10 30%	0/10 0%	4/10 40%
2.5.3 : lipoate	0/1 0%	0/1 0%	1/1 100%	0/1 0%	0/1 0%
2.5.4 : menaquinone and ubiquinone	8/14 57%	0/14 0%	2/14 14%	1/14 7%	3/14 21%
2.5.5 : molybdopterin	1/1 100%	0/1 0%	0/1 0%	0/1 0%	0/1 0%
2.5.6 : pantothenate and coenzyme a	4/6 66%	0/6 0%	1/6 16%	0/6 0%	1/6 16%
2.5.7 : pyridoxine and derivatives	4/5 80%	1/5 20%	0/5 0%	0/5 0%	0/5 0%
2.5.8 : riboflavin, fmn, and fad	7/9 77%	1/9 11%	0/9 0%	0/9 0%	1/9 11%
2.5.9 : glutathione	1/9 11%	4/9 44%	0/9 0%	0/9 0%	4/9 44%
2.6.4 : other	0/3 0%	1/3 33%	0/3 0%	0/3 0%	2/3 66%
2.7 : metabolism of phosphate	5/14 35%	3/14 21%	3/14 21%	2/14 14%	1/14 7%
2.8 : metabolism of sulfur	14/37 37%	6/37 16%	8/37 21%	1/37 2%	8/37 21%
3 : information transfer pathways	3/6 50%	1/6 16%	2/6 33%	0/6 0%	0/6 0%
3.1 : dna metabolism	1/10 10%	4/10 40%	0/10 0%	0/10 0%	5/10 50%
3.1.1 : dna replication	2/32 6%	1/32 3%	19/32 59%	0/32 0%	10/32 31%
3.1.2 : dna recombination	4/16 25%	1/16 6%	7/16 43%	0/16 0%	4/16 25%
3.1.3 : dna packaging and segregation	0/7 0%	2/7 28%	3/7 42%	0/7 0%	2/7 28%
3.1.4 : dna proof-reading and repair	8/36 22%	9/36 25%	10/36 27%	0/36 0%	9/36 25%
3.1.5 : dna restriction/modification	0/4 0%	1/4 25%	2/4 50%	0/4 0%	1/4 25%
3.1.6 : degradation of dna	1/6 16%	2/6 33%	1/6 16%	0/6 0%	2/6 33%
3.2 : rna metabolism	1/3 33%	0/3 0%	2/3 66%	0/3 0%	0/3 0%
3.2.1 : transcription	1/9 11%	2/9 22%	5/9 55%	0/9 0%	1/9 11%
3.2.1.1 : regulation	0/34 0%	6/34 17%	17/34 50%	0/34 0%	11/34 32%
3.2.1.1.1 : regulatory proteins (dna interaction)	5/112 4%	7/112 6%	55/112 49%	0/112 0%	45/112 40%
3.2.1.1.2 : regulatory proteins (rna interaction)	1/4 25%	0/4 0%	3/4 75%	0/4 0%	0/4 0%

Fonction	Cluster bleu	Cluster jaune	Cluster rouge	Cluster rose	Cluster vert
3.2.1.1.3 : regulatory rnas	0/1 0%	0/1 0%	0/1 0%	0/1 0%	1/1 100%
3.2.1.2 : initiation	0/1 0%	0/1 0%	1/1 100%	0/1 0%	0/1 0%
3.2.1.2.1 : sigma factors	0/17 0%	1/17 5%	13/17 76%	1/17 5%	2/17 11%
3.2.1.2.2 : other	0/1 0%	1/1 100%	0/1 0%	0/1 0%	0/1 0%
3.2.1.3 : elongation	0/1 0%	0/1 0%	1/1 100%	0/1 0%	0/1 0%
3.2.1.3.1 : rna polymerase (core enzyme)	0/3 0%	0/3 0%	3/3 100%	0/3 0%	0/3 0%
3.2.1.3.2 : transcription / translation coupling	1/2 50%	0/2 0%	1/2 50%	0/2 0%	0/2 0%
3.2.1.3.3 : other	1/3 33%	0/3 0%	2/3 66%	0/3 0%	0/3 0%
3.2.1.4 : termination	0/2 0%	0/2 0%	2/2 100%	0/2 0%	0/2 0%
3.2.2 : processing and degradation of rna	2/8 25%	0/8 0%	5/8 62%	0/8 0%	1/8 12%
3.2.2.1 : rna chaperones	0/1 0%	0/1 0%	1/1 100%	0/1 0%	0/1 0%
3.2.2.2 : ribonucleases	4/13 30%	0/13 0%	6/13 46%	1/13 7%	2/13 15%
3.2.3 : rna modification	5/16 31%	1/16 6%	9/16 56%	0/16 0%	1/16 6%
3.2.3.1 : trna modifications	10/28 35%	2/28 7%	12/28 42%	0/28 0%	4/28 14%
3.2.3.2 : rrna modifications	2/13 15%	2/13 15%	5/13 38%	0/13 0%	4/13 30%
3.2.3.3 : other modifications	0/1 0%	0/1 0%	1/1 100%	0/1 0%	0/1 0%
3.3 : protein synthesis, maturation and folding	4/8 50%	0/8 0%	1/8 12%	0/8 0%	3/8 37%
3.3.1 : ribosomal proteins	14/29 48%	0/29 0%	15/29 51%	0/29 0%	0/29 0%
3.3.2 : aminoacyl-trna synthetases	3/24 12%	3/24 12%	16/24 66%	0/24 0%	2/24 8%
3.3.3 : initiation	0/2 0%	0/2 0%	2/2 100%	0/2 0%	0/2 0%
3.3.4 : elongation	7/9 77%	1/9 11%	1/9 11%	0/9 0%	0/9 0%
3.3.5 : termination	0/3 0%	0/3 0%	3/3 100%	0/3 0%	0/3 0%
3.3.6 : protein modifications	2/14 14%	4/14 28%	2/14 14%	2/14 14%	4/14 28%
3.3.6.1 : maturation	2/4 50%	0/4 0%	1/4 25%	0/4 0%	1/4 25%
3.3.6.2 : chemical modifications	4/8 50%	1/8 12%	1/8 12%	0/8 0%	2/8 25%
3.3.7 : protein folding	10/38 26%	3/38 7%	17/38 44%	4/38 10%	4/38 10%
3.3.8 : protein repair	1/9 11%	6/9 66%	0/9 0%	1/9 11%	1/9 11%
4 : other functions	0/1 0%	0/1 0%	0/1 0%	1/1 100%	0/1 0%
4.1 : adaptation to atypical conditions	19/73 26%	10/73 13%	23/73 31%	8/73 10%	13/73 17%

Fonction	Cluster bleu	Cluster jaune	Cluster rouge	Cluster rose	Cluster vert
4.2 : detoxification	12/33 36%	7/33 21%	3/33 9%	4/33 12%	7/33 21%
4.3 : antibiotic production	2/3 66%	1/3 33%	0/3 0%	0/3 0%	0/3 0%
4.4 : phage-related functions	3/17 17%	7/17 41%	5/17 29%	0/17 0%	2/17 11%
4.5 : transposons and ISs	0/6 0%	3/6 50%	2/6 33%	0/6 0%	1/6 16%
4.6 : miscellaneous	1/2 50%	0/2 0%	1/2 50%	0/2 0%	0/2 0%
4.6.1 : plasmid-related functions	0/4 0%	2/4 50%	1/4 25%	0/4 0%	1/4 25%
4.6.2 : pathogenesis	0/1 0%	0/1 0%	0/1 0%	1/1 100%	0/1 0%
5 : similar to unknown proteins	6/37 16%	13/37 35%	7/37 18%	7/37 18%	4/37 10%
5.1 : from p. haloplanktis	6/40 15%	11/40 27%	6/40 15%	3/40 7%	14/40 35%
5.2 : from other organisms	78/597 13%	189/597 31%	101/597 16%	76/597 12%	153/597 25%
6 : no similarity	4/61 6%	20/61 32%	9/61 14%	4/61 6%	24/61 39%
6.2 : hypothetical	12/171 7%	65/171 38%	26/171 15%	14/171 8%	54/171 31%

Résumé

Les organismes vivants sont soumis à diverses pressions de sélection qui n'agissent pas seulement au niveau du phénotype global mais à chaque niveau de l'organisation de la cellule.

Nous avons analysé la composition globale en acides aminés de l'ensemble les protéines de chaque protéomes étudiés à l'aide, entres autres, de l'Analyse Factorielle des Correspondances et d'un outil de partitionnement, les nuées dynamiques.

Ont été étudiés

(i) les modèles microbiens les mieux connus *E. coli*, *B. subtilis* et *M. jannaschii*,

(ii) un échantillon représentatif du monde procaryote de 28 organismes aux caractéristiques les plus diverses,

(iii) la pathogénicité de *P. luminescens*,

(iv) la qualité psychrophile de la bactérie *P. haloplanktis*, dont la vie à basse température est caractérisée par des protéines fortement biaisées en asparagine et

(v) une perspective d'application aux eucaryotes simples est évoquée au regard des travaux préliminaires sur l'usage des codons de *P. marneffei*, un champignon dimorphique et pathogène.

Abstract

Living organisms are subjected to diverse selection pressures not only on global phenotype but at each level of cellular organisation.

We analysed the global amino acid composition of all proteins of each proteomes studied using multivariate analysis, Correspondence Analysis and a clustering method called dynamic clouds.

We studied:

(i) prokaryote models *E. coli*, *B. subtilis* and *M. jannaschii*,

(ii) a representative sample of the prokaryote world including 28 organisms of with very diverse characteristics,

(iii) the pathogenicity of *P. luminescens*,

(iv) a psychrophile feature of the bacterium *P. haloplanktis*, for which life in the cold is characterised by proteins with a high asparagine bias, and

(v) a perspective will be to apply these analyses to simple eukaryotes through preliminary analysis of codon usage by *P. marneffei*, a dimorphic pathogenic fungus.