

Texte de la conférence du Professeur Georges Cohen (Institut Pasteur)

De même que le décryptage du génome humain promet de révolutionner les sciences médicales, l'application des techniques de la génomique à l'évolution microbienne et à la biologie de l'environnement promet de révolutionner la microbiologie. Les retombées ne portent pas seulement sur une connaissance accrue, mais aussi sur l'économie. Une compréhension plus complète de la diversité microbienne et des processus environnementaux nécessitera non seulement un inventaire, mais une compréhension plus complète des unités fondamentales d'organisation et de leurs interactions.

Ce sont les communautés, non la biomasse totale qui dirigent les cycles géobiochimiques qui entretiennent la biosphère. Donc, les descriptions des dimensions spatio-temporelles de la structure des communautés microbiennes et les patterns complexes de l'expression des gènes qui sous-tendent les interactions trophiques sont fondamentales pour une compréhension totale de notre biosphère.

D'autre part, cette compréhension sera incomplète sans la connaissance des mécanismes fondamentaux qui contribuent à la variation génétique et à la spéciation. La séquence des génomes a révélé une plasticité génétique totalement inattendue à l'intérieur d'une espèce et entre des espèces bien déterminées. On a pu montrer que les échanges

horizontaux d'ADN représentaient un élément déterminant de la construction des génomes et de l'innovation biochimique.

Les progrès rapides de la génomique doivent être complétés par un investissement sérieux dans la systématique, afin de développer une taxonomie mieux adaptée à l'information génétique. En effet, les concepts de la taxonomie traditionnelle, c'est à dire les espèces, genres et familles ne servent pas la systématique microbienne, qui ne tient pas compte des transferts horizontaux, et des mécanismes variés et complexes de la spéciation et de l'évolution.

Les microbes ont dominé la vie sur la terre pendant la plus grande partie des quelques 4,5 milliards d'années de son histoire. Ils sont le fondement de la biosphère, contrôlant les cycles biogéochimiques et affectant la géologie, l'hydrologie et les climats locaux et globaux. Toute la vie sur terre dépend d'eux. L'espèce humaine ne peut survivre sans leur riche diversité mais la plupart des espèces microbiennes peuvent survivre sans les humains

Les progrès extraordinaires de la technologie moléculaire ont permis une explosion de l'information en biologie microbienne. On sait maintenant que les espèces microbiennes en culture pure ne représentent que pauvrement la richesse de leur diversité naturelle. Ceci a été révélé au cours des dix années écoulées par l'utilisation de nouveaux outils moléculaires permettant d'explorer la diversité environnementale et aboutissant à une croissance explosive de l'écologie microbienne.

La séquence complète des génomes a été identifiée comme un élément essentiel du développement de la stratégie et il a été recommandé en 1997 de séquencer le génome d'au moins une espèce de chacune des grandes divisions des microorganismes. Cet objectif, qui était ambitieux à l'époque, est complètement dépassé huit ans plus tard grâce aux progrès remarquables de la technologie de la séquence et de la bioinformatique; en outre, de nouvelles technologies sont apparues comme les puces à ADN, qui permettent d'obtenir des données essentielles sur la physiologie d'un organisme. Les développements ont été si rapides qu'ils peuvent conduire à l'analyse de la complexité extraordinaire des communautés écologiques naturelles.

Par exemple, la séquence d'une espèce anaérobie non isolée en culture pure, mais faisant partie d'une communauté microbienne présente dans le fermenteur expérimental des eaux d'épuration d'Evry a été récemment obtenue au Génoscope. La séquence de son ARN 16S indique qu'il s'agit d'une espèce que l'on ne peut rattacher à aucune des familles connues. Depuis les années 1990 qui ont vu l'apparition de technologies de séquençage à haut débit, les génomes complets de centaines d'organismes ont été séquencés et archivés. Le succès de ces programmes est impressionnant, les génomes d'organismes aussi divers que ceux de bactéries et d'*Homo sapiens* et aussi extraordinaires que ceux du poisson *Tetraodon* et du pin *Pinus taeda* (loblolly pine) ont été catalogués. Des centaines de milliers de gènes ont été

séquencés et nous sommes en plein dans une révolution génomique.

Toutefois, le travail plus difficile d'interprétation de ces séquences a à peine commencé. Il est difficile de définir en science les limites de la connaissance à un instant précis. Cependant, c'est le cas de la génomique.

On peut identifier quatre domaines qui laissent à désirer dans le cas des gènes d'organismes procaryotes :

- 1) environ 40% de tous les gènes prédits ne possèdent pas d'annotation fonctionnelle
- 2) de nombreux gènes ont une fonction prédite, mais la prédiction n'a pas été expérimentalement validée
- 3) de 5 à 10% des fonctions prédites peuvent être incorrectes
- 4) de nombreux enzymes connus ne possèdent pas de gènes correspondants dans les bases de données de séquence

Je donnerai quelques exemples à la fin de mon exposé.

Les séquences disponibles actuellement sont d'une grande utilité mais afin d'exploiter leur potentiel, elles doivent être annotées avec précision en utilisant une approche combinée de bioinformatique et d'expérimentation. Comme de nombreux gènes sont trouvés dans une forme voisine dans plus d'une espèce, l'attribution d'une fonction à un gène donné peut accroître notre compréhension de la physiologie de beaucoup d'organismes différents. Dans ces séquences sont cachées des cibles pour de nouveaux médicaments, de nouveaux enzymes pour

la biotechnologie et une abondance de nouveaux éléments de régulation.

La simple annotation n'est qu'une première, bien qu'essentielle, étape de la compréhension de la complexité d'un organisme. L'annotation fonctionnelle doit être plus qu'un catalogue de fonctions protéiques. Elle doit inclure de l'information sur les interactions entre les produits des gènes, ces interactions résultant dans une hiérarchie de fonctions dont nous ne connaissons que très peu de choses. La conséquence de ces interactions est un système qui est d'un autre ordre que la somme de ses constituants et qui résulte en un organisme qui en produit deux. Par conséquent, un des buts de l'annotation fonctionnelle est de fournir les fondations de l'exploration et de la compréhension d'un organisme entier. Ceci est du domaine de la biologie des systèmes, qui cherche à comprendre la physiologie d'un organisme en étudiant les produits du génome, comment ils interagissent en réseaux synergiques pour accomplir les fonctions complexes de la vie. Comme la biologie des systèmes dépend directement des produits du génome, une annotation soigneuse et précise se doit de décrire les rôles individuels de ces produits. Malheureusement, les efforts dans la biologie des systèmes sont ralentis par la lenteur de l'annotation des gènes. Il est donc essentiel de révéler le potentiel biochimique complet des gènes procaryotiques si nous devons un jour comprendre la machinerie de ces formes les plus simples de la vie.

La lenteur de l'annotation a conduit à la situation actuelle où un grand nombre d'annotations putatives sont basées sur un bien plus petit nombre de produits de gènes fonctionnellement caractérisés. Cette pyramide renversée n'est pas une base satisfaisante à partir de laquelle des hypothèses peuvent être formulées. Avec une connaissance limitée sur la diversité des fonctions des gènes, des fonctions peuvent être facilement mal interprétées et une confiance exagérée peut être placée sur des corrélations entre gènes dont les fonctions sont en réalité éloignées.

En conséquence, afin d'utiliser au mieux les données des séquences génomiques, une annotation expérimentale des produits des gènes doit recevoir une priorité élevée.

Sources courantes de l'annotation fonctionnelle et leurs limitations.

Différentes sources d'annotation fonctionnelle sont à la disposition des chercheurs. Certaines sont bien connues ; d'autres sont plus obscures ou en cours de développement. Mais elles présentent toutes des limitations et ne répondent pas aux besoins de la recherche actuelle. La littérature scientifique est la meilleure source d'annotation, mais l'information est souvent dispersée dans de nombreux articles et peut ne pas être largement connue. De plus, un grand nombre d'informations a été publiée avant l'acquisition des séquences génomiques et n'a pas été

systématiquement incorporée dans les annotations actuellement disponibles.

L'apparition d'outils de computation automatique a également créé un certain nombre de problèmes pour l'acquisition et la compilation des données. Comme ces méthodes dominent actuellement l'annotation, des erreurs anciennes se sont propagées. En outre, des fonctions découvertes après les premières annotations ont été négligées et n'ont pas été incorporées dans les bases de données. Pour répondre à ces problèmes, la nécessité se fait sentir de la création d'une source centrale et unique d'information régulièrement mise à jour, distincte des bases de données actuelles.

Les sources d'annotation actuelles appliquent fréquemment des seuils d'évidence trop incertains et fournissent rarement une estimation quantitative de la confiance que l'on peut accorder à l'évidence utilisée dans l'annotation. En outre, les méthodes expérimentales utilisées sont souvent différentes d'un groupe de chercheurs à un autre. En conséquence, des annotations effectuées à l'aide de techniques inadéquates peuvent ne pas être détectées et être propagées d'un génome à un autre.

La mise à jour régulière des annotations est également difficile. De nouvelles informations et des caractérisations de produits de gènes qui apparaissent dans la littérature ne sont pas toujours transmis aux sources d'annotation électronique faute d'avoir établi des procédures de mise à jour. D'autre part, la transmission à la communauté scientifique des erreurs d'annotation connues est un problème

souvent négligé..Des exceptions notables sont les bases de données modèles telles que EcoCyc,Flybase,et les bases de données de *Saccharomyces cerevisiae* qui effectuent régulièrement des efforts de «curation» de la littérature et mettent leurs bases de données régulièrement à jour. Toutefois, ces efforts sont limités à moins de 10% des génomes séquencés;de plus,ces annotations ne sont pas automatiquement transférés aux gènes correspondants éventuellement présents dans d'autres organismes.

Enfin,les noms employés pour les gènes et leurs produits entretiennent une confusion extrême chez les chercheurs, car peu de résumés (abstracts) utilisent des descriptions fonctionnelles standard ou des identifiants systématiques des gènes.Par exemple le sigle *ilvA* représente la thréonine déshydratase chez *Saccharomyces cerevisiae* et l'alpha acétolactate synthétase chez *Schizosaccharomyces pombe*.

Avancées récentes en bioinformatique.

La majorité des algorithmes déduisent la fonction d'un gène par la similarité de sa séquence avec celle de gènes précédemment caractérisés.Mais qu'en est-il quand aucune fonction n'a été attribuée aux homologues d'un gène donné? Ou bien quand une phase ouverte n'a aucun homologue dans les bases de données publiques? Ces situations constituent des difficultés importantes pour les prédictions fondées sur les méthodes utilisant les séquences.

Des suppositions sur la fonction peuvent alors être utiles car elles peuvent indiquer quels essais biochimiques peuvent être appliqués pour confirmer ou infirmer la fonction hypothétique.

Trois méthodes existent qui augmentent les chances d'avoir un point de départ pour les recherches expérimentales des biochimistes.

1) la première est fondée sur des techniques de génomique comparée qui infèrent des associations fonctionnelles entre des gènes à partir des patterns observés dans de nombreux génomes. Par exemple, supposons qu'un gène A soit adjacent à un gène B dont on connaît la fonction. En elle-même, cette observation est de peu de conséquence, car les voisins d'un gène ont fréquemment des fonctions complètement différentes de celle de ce gène. Mais si une analyse plus poussée montre que cette proximité se retrouve chez 5 ou plus organismes divers, on peut supposer raisonnablement que les deux gènes A et B ont des fonctions apparentées, comme par exemple l'appartenance à une même chaîne métabolique.

2) la seconde classe de méthodes est fondée sur l'application de techniques informatiques qui présupposent que les protéines ayant des fonctions similaires tendent à avoir des propriétés physico-chimiques voisines. Ce type de méthode a été utilisé pour inférer si une protéine est un enzyme ou non et si oui, à quelle classe des six classes principales de l'Enzyme Commission il appartient, en se fondant sur des informations concernant par exemple, la composition en amino acides, le poids moléculaire, le

point isoélectrique, ou encore une signature particulière à une fonction.

La troisième classe de méthodes est fondée sur des technologies intégratives telles que le screening systématique des interactions protéine-protéine, le screening de la génomique fonctionnelle, l'information sur les voies métaboliques, qui peuvent fournir une vision globale du protéome d'un organisme donné.

Il faudrait trouver une séquence génétique pour chaque fonction enzymatique connue. En se servant des EC numbers disponibles avec l'information sur des gènes correspondant à des enzymes connus provenant de diverses banques de données, y compris Swiss-Prot et TrEMBL, on voit qu'il y a au moins 1400 enzymes pour lesquels on ne dispose pas d'une séquence correspondante. Ces enzymes sont appelés orphelins, un terme correspondant à leur héritage génétique non défini.

Une initiative a été récemment proposée lors d'une réunion de la Société Américaine de Microbiologie qui s'est tenue en juillet 2004 à Washington. Les participants à ce colloque comprenaient des bioinformaticiens, des microbiologistes et des biochimistes. Étaient également présents des observateurs du NIH, de la NSF, du Département de l'Énergie, de l'Administration de l'Aéronautique et de l'Espace, et du Département de l'Agriculture. Il y a été convenu qu'une nouvelle base soit développée dont le composant central devrait contenir

- 1) des prédictions concernant les fonctions des gènes de fonction inconnue, déposées par les bioinformaticiens, fondées sur des annotations automatiques, qui serviraient de point de départ pour des recherches expérimentales
- 2) les résultats, positifs ou négatifs de ces recherches, conduisant éventuellement à de nouvelles annotations reposant sur du travail expérimental rigoureux.
- 3) une liste prioritaire des gènes séquencés pour lesquels n'existe actuellement aucune information fonctionnelle
- 4) une liste de fonctions biochimiquement caractérisées auxquelles aucun gène n'a été trouvé correspondre(appelés fonctions orphelines)
- 5) confirmation ou infirmation des protéines caractérisées se trouvant dans les bases de données actuelles.

Les chercheurs de l'Enzyme Genomics Initiative ont produit un système qui prédit la facilité avec laquelle un gène correspondant à un orphelin donné peut être déterminé. La meilleure note est attribuée aux enzymes qui proviennent d'un organisme totalement séquencé ou pour lequel on possède une partie de la séquence protéique. Dans ce cas, l'identification du gène requiert la comparaison automatique des propriétés biochimiques aux ORFs de la séquence génomique et la vérification expérimentale de la prédiction. Une deuxième priorité est donnée aux enzymes qui ont été récemment caractérisés

expérimentalement. Une moins bonne note est donnée aux enzymes pour lesquels on ne possède de l'information que sur certaines caractéristiques physico-chimiques telles que le poids moléculaire, le point isoélectrique, ou encore les résultats d'une digestion protéolytique.

Pour terminer, je voudrais donner ici quelques exemples d'un début d'annotation fonctionnelle qui a été entrepris au Génoscope d'Evry, et à laquelle je collabore. L'organisme étudié est *Acinetobacter baylii* (diapos 1 et 2) et les gènes choisis sont relatifs à l'utilisation pour sa croissance de sources carbonées ou azotées. La diapositive montre en rouge cinq régions que nous avons appelées îlots cataboliques correspondant à la dégradation de certaines de ces sources. La dégradation de certains de ces composés (diapo 3), esters salicyliques, benzoate, catéchol, protocatéchuate, quinate, caffeate, signe la niche écologique d'*Acinetobacter*, bactérie du sol, qui dégrade les produits d'origine végétale qu'il y trouve. Nous avons choisi d'étudier certains de ces îlots pour confirmer ou infirmer leur annotation automatique.

Dégradation de l'urée.

Notre choix s'est d'abord porté sur l'amas de gènes que l'annotation automatique avait désigné comme codant, pour les 9 sous-unités de l'uréase (diapos 4 et 5). Les gènes 1088 à 1096, codant pour ces sous-unités, sont adjacents et transcrits dans la même direction; ils ont été validés par comparaison avec

des gènes correspondant d'autres espèces, dont l'attribution a été dans certains cas ,confirmée biochimiquement(diapo 6).On constate qu' entre les gènes 1091 et 1093, se trouve un gène qui a été annoté automatiquement comme «putative alpha ribazole 5'-phosphatase», enzyme appartenant à la voie de biosynthèse de la vitamine B12.Or,seuls 5 gènes de cette voie ont été annotés chez *Acinetobacter*, sur la trentaine qu'elle comporte.On peut faire raisonnablement l'hypothèse que le gène 1092, dont des homologues existent dans d'autres espèces ,mais sont dans d'autres régions du chromosome, code pour une protéine dont la fonction est d'une manière ou d'une autre, reliée au catabolisme de l'urée.

Au Génoscope, a été initié un programme de disruption(*knock-outs*) de tous les gènes d'*Acinetobacter* et de l'étude des phénotypes consécutifs à ces disruptions. En particulier, les disruptions individuelles de chacun des gènes de l'amas uréase ont été effectuées.

La diapositive suivante (diapo 7) montre que toutes ces disruptions,sauf celle portant sur le gène 1092 ou sur le gène *ureJ* conduisent à des bactéries incapables d'utiliser l'urée comme source d'azote, ce qui valide les annotations automatiques.

En ce qui concerne la disruption du gène 1092,le dosage de l'uréase qui a été pratiqué montre une activité spécifique bien inférieure à celle du type sauvage, ce qui conduit à l'hypothèse que ce gène code pour un activateur spécifique de l'uréase. Pour vérifier cette hypothèse, le gène sera surexprimé chez

E.coli et des essais d'interaction de la protéine correspondante avec la région promoteur de l'«opéron uréase» devront être effectués.

D'autre part, des expériences sur le transcriptome correspondant sont en cours avec des puces à ADN pour vérifier qu'il s'agit bien d'un opéron et que le gène 1092 en fait partie.

L'annotation automatique indique que la protéine UreJ possède 6 hélices transmembranaires, suggérant qu'il s'agit d'un transporteur impliqué dans le transport soit de l'urée soit du nickel, métal indispensable à l'activité de l'uréase. En fait, il semble s'agir d'une perméase à nickel: alors que la disruption permet la croissance sur urée dans un milieu contenant un excès de nickel, une très faible croissance est observée sur un milieu sans nickel ajouté. Si le nickel est ajouté à ce milieu à la concentration de 5 micromolaire, la croissance est rétablie. Cette expérience devra être confirmée en étudiant directement la perméation du nickel à l'aide de nickel radioactif.

Dégradation du malonate.

La malonate décarboxylase transforme le malonate en acétate et CO₂. Comme l'uréase, c'est un enzyme complexe (diapo 8). L'annotation automatique indique que chez *Acinetobacter*, ses constituants sont codés par cinq protéines distinctes (MdcA, MdcC, MdcD, MdcE, MdcH). En outre, il a été mis en évidence un coenzyme dérivé du coenzyme A, le 2-(5'-

triphosphoribosyl)3'-dephosphocoenzyme A dont la synthèse est assurée par une protéine codée par le gène *mdceB*. Ce coenzyme se lie à une synthase spécifique grâce à une protéine codée par *mcdG*. De plus, deux gènes *mcdL* et *mcdM* participent au transport actif du malonate et un gène *mcdR* code pour un régulateur transcriptionnel spécifique. Les dix gènes impliqués sont adjacents chez *Acinetobacter*. Neuf d'entre eux sont transcrits dans la même direction. Seul celui codant pour le régulateur est transcrit sur le brin opposé.

Ici encore, chacun de ces gènes a été disrupté. Aucun des mutants « knock-out » correspondants n'est capable de croître sur malonate comme seule source de carbone, ce qui confirme le bien-fondé de l'annotation automatique. En outre, une information importante est obtenue : le dérivé du coenzyme A cité plus haut est spécifique de la malonate décarboxylase : en effet, la disruption du gène responsable de sa synthèse n'affecte en rien la croissance sur toutes les autres sources de carbone testées.

Synthèse du pantothénate et du coenzyme A

Le coenzyme A intervient dans la synthèse des acides gras et à divers niveaux du cycle respiratoire, etc... Il est donc indispensable. La diapo suivante (diapo 9) montre que, contrairement aux deux exemples précédents, les gènes qui codent pour les protéines nécessaires à sa synthèse sont répartis tout au long du chromosome circulaire d'*Acinetobacter*.

Un précurseur du coenzyme A est l'acide pantothénique qui résulte de la fusion de la beta-alanine, produit du gène *panD*, et du pantoate, produit du gène *panE*.

L'annotation automatique indique que le gène *panD* est présent chez *Acinetobacter*. Toutefois sa disruption (dûment vérifiée) ne mène pas, comme chez *Escherichia coli*, à une auxotrophie pour la beta-alanine. Nous faisons l'hypothèse qu'il existe chez *Acinetobacter* une voie alternative de la synthèse de beta-alanine et les recherches s'orientent dans ce sens. Quand au gène *panE*, il n'a jamais été identifié avec précision, mais l'enzyme correspondant a été purifié. C'est un exemple de ce que l'Enzyme Genomics Initiative appelle un enzyme orphelin, pour suggérer son héritage génétique non défini.

Si nous progressons dans la voie de biosynthèse, nous notons que le gène codant pour la pantothénate kinase n'a pas été trouvé par l'annotation automatique. La majorité des enzymes conduisant au coenzyme A ayant été détectés, il est vraisemblable que l'activité de phosphorylation du pantothénate est présente. Sa recherche est donc l'une de nos priorités.

Les exemples précédents montrent l'ampleur du travail à accomplir qui nécessitera la collaboration de nombreux chercheurs.

Je vous remercie de votre attention.