Métagénomique bactérienne et virale

Nouvelles définitions d'espace microbien et nouveaux défis algorithmiques

Julie Baussand — Alessandra Carbone

Génomique Analytique, Université Pierre et Marie Curie-Paris 6, INSERM U511
91 blv de l'Hôpital, 75012 Paris
baussand@infop6.jussieu.fr et Alessandra.Carbone@lip6.fr

RÉSUMÉ. *Plusieurs problèmatiques de la métagénomique concernant l'assemblage des génomes, la classification et la quantification des espèces environnementales, la reconstruction phylogénétique et la spécificité des communautés sont discutées. Le défi est important étant donné les difficultés liées notamment aux séquences incomplètes détectées pour les communautés microbiennes non-cultivables, le manque d'outils pour la détection des protéines très divergentes, le manque de concepts adaptés à la classification des communautés environnementales pour ne citer que ces quelques problèmes. Une vue d'ensemble est donnée sur les projets en cours et les données disponibles, ainsi que sur des questions algorithmiques concernant la métagénomique.*

ABSTRACT. *Several problems in metagenomics are discussed concerning genome assembly, environmental species classification, phylogenetic reconstruction, communities specificity, communities quantification. The challenge is great because of the difficulties due to incomplete available sequences detected for uncultured microbial communities, the lack of tools for detecting homology between divergent proteins, missing concepts on which to base environmental classification to cite just a few hurdles. The paper gives a concise overview of current projects and available data, and sets some algorithmic questions.*

MOTS-CLÉS : *métagénomique, virus, bactéries, espaces microbiens, classification environnementale, classification phylogénétique.*

KEYWORDS: *metagenomics, virus, bacteria, microbial spaces, environmental classification, phylogenetic classification.*

1. Introduction

Le terme *métagénomique*, ou encore *génomique environnementale* ou *génomique des communautés*, indique l'étude des *métagénomes*, qui lui-même réfère à l'ensemble des séquences d'ADN extraites de communautés multi-espèces prélevées dans l'environnement. Ces communautés sont généralement composées d'organismes non cultivables, soit qu'ils ne sont pas ciblés par les conditions de culture car non-connus, soit qu'ils résistent aux tentatives de culture. Il a été estimé que seul 1% des procaryotes de la plupart des environnements peuvent être cultivés (Amann *et al.*, 1990) nécessitant que le clonage soit effectué directement sur les échantillons prélevés dans l'environnement. Plusieurs lignées de bactéries (archéa et eubactérie) ont pu être ainsi détectées par des approches de phylogénétique moléculaire, mais leur caractérisation reste élusive car aucune souche cultivable en laboratoire n'est disponible. De même, l'étude des phages (virus infectant les procaryotes), qui sont les microbes les plus abondant sur terre, est limitée par la difficulté à identifier et cultiver leurs hôtes.

Ainsi, le défi de la métagénomique consiste à relier les informations génomiques issues des clones à l'organisme ou à l'écosystème duquel l'ADN a été extrait. Voir les articles dédiés (*Nature, 2005; Handelsman, 2004; Rodriguez-Valera, 2004)*. Déjà des stratégies métaboliques ont pu être proposées pour plusieurs espèces de bactéries non-cultivables, laissant envisager une compréhension plus précise des espèces microbiennes, de la variabilité de leur génome, et de la distribution intra- et inter-espèce des gènes et des chemins biochimiques. Désormais, on souhaiterait pouvoir répondre à des questions biologiques telles que : Comment mesurer la diversité d'une communauté ? Quelles sont les relations d'évolution entre l'ADN issu de transferts horizontaux, les éléments génétiques mobiles et les génomes ? Comment identifier les hôtes possibles d'un bactériophage ? Après modélisation, comment déterminer les dynamiques temporelles et géographiques des mycètes, des bactéries et de leurs bactériophages ?

D'un point de vue informatique, la grande quantité de données métagénomiques nouvellement disponible demande de revisiter les algorithmes initialement conçus pour traiter des séquences issues d'un même organisme (Chen *et al.*, 2005). Il est en effet nécessaire d'adapter ces algorithmes au nombre d'espèces présents dans ces populations, et à la fréquence des réarrangements génomiques, et á l'insertion/délétion des gènes affectant ces organismes. Mais également d'en inventer de nouveaux basés sur des notions alternatives d'espaces de séquences microbiennes. Des questions algorithmiques peuvent déjà se poser : Les techniques de co-assemblage peuvent-elles être développées pour assembler des génomes polymorphiques de communautés complexes ? Comment les programmes de prédiction de gènes peuvent être adaptés aux séquences issues de génome à faible taux de couverture ? Quels paramètres permettent de différencier des espèces proches ou de relier des espèces éloignées ? Quelles sont les méthodes de clustering les plus adaptées à la découverte des chemins métaboliques et groupes fonctionels témoin de l'adaptation des microorganismes à leur environnement ? Comment optimiser la construction d'arbres et l'alignement des séquences avec des données métagénomiques incomplétes ?

2. Complexité des communautés microbiennes et approches expérimentales

Les premières analyses de métagénomes, réalisées par PCR, ont consisté à explorer la diversité des séquences d'ARN ribosomal extraites des communautés microbiennes (Pace *et al.*, 1986; Pace *et al.*, 1985). Ces expériences ont mis en évidence une diversité microbienne bien plus compliquée que pouvaient le faire penser les méthodes par culture. En 2002, une approche expérimentale a permis de montrer la présence de >5000 différents virus dans 200 litres d'eau de mer (Breitbart *et al.*, 2002). D'autres études ont montré la présence de >1000 espèces virales dans les fèces humaines, et probablement un million de virus différents par kilogramme de sédiments marins. Essentiellement tous ces virus étaient de nouvelles espèces. Également, une étude de la mer des Sargasses a mis en évidence environ 2000 différentes espèces microbiennes dont 148 types de génome bactérien jamais observés auparavent (Venter *et al.*, 2004).

Complexité variable des communautés microbiennes. La complexité du problème d'assemblage simultané de fragments d'ADN provenant de communautés hétérogénes semble dépendre du nombre d'espèces contenues dans l'échantillon. Les communautés étudiées jusqu'à maintenant montrent un nombre d'espèces inversement proportionnel au pourcentage des fragments d'ADN qui ont pu être assemblés en contigs (Schloss *et al.*, 2005) : la communauté du drainage minier acide était estimée à 6 espèces et 85% des 100 000 fragments d'ADN ont été assemblés en contigs (Tyson *et al.*, 2004) ; les échantillons de carcasse de baleine comptent entre 20 et 150 espèces et un assemblage en contigs d'environ 40% des fragments d'ADN (sur 40 000 par échantillon) a pu être effectué (Tringe *et al.*, 2005) ; les échantillons de la mer des Sargasses comptant environ 300 espèces (Venter *et al.*, 2004) et le sol d'une ferme du Minnesota comptant >3000 espèces (Tringe *et al.*, 2005) n'ont permis l'assemblage en contigs que de $< 1\%$ des fragments d'ADN (sur 325 000 et 150 000 respectivement).

Plusieurs génomes viraux ont été séquencés à partir de données métagénomiques, dont les 17 premiers phages marins de longueur de génomes variant entre 10 et 800Kb (Paul *et al.*, 2005). Les génomes bactériens de *Kuenenia stuttgartiensis* (Strous *et al.*, 2006), de *Ferroplasma* type II et de *Leptospirillum* groupe II ont été presque complétement assemblés et trois autres ont été partiellement reconstitués (Tyson *et al.*, 2004). C'est le faible nombre d'espèces au sein de la communauté et le taux relativement bas de réarrangements génomiques qui a permis le séquençage de ces organismes. Des techniques expérimentales d'amplification de matériel génétique et de division des communautés ont rendu possible la reconstitution de certains de ces génomes. De nouvelles approches sont recherchées afin d'obtenir un découpage plus fin des populations et de réduire ainsi la complexité du problème de caractérisation des métagénomes. Les méthodes d'enrichissement de génomes existantes permettent de séparer notamment les génomes symbiotiques de leur hôte, les génomes de contenu en GC différents ou encore de sélectionner les composants actifs au sein d'une population microbienne. Les méthodes d'enrichissement de gènes permettent, quant à elles, de sélectionner les gènes d'intérêt impliqués dans un chemin biologique spécifique ou de séquence connues, ou d'identifier des marqueurs génétiques spécifiques (Cowan *et al.*, 2005).

3. Projets métagénomiques, bibliothèques de clones et base de données

Le *Genome Online Database* (Liolios *et al.*, 2006; Bernal *et al.*, 2001) liste 56 projets métagénomiques différents tels que : le métagénome du symbionte bactérien de la surface dorsale d'un ver (134 million bp) ; un échantillon de l'environnement d'un mammouth laineux (28 million bp) ; la station océanographique de "Hawai Ocean Times Series" ($2e^{+06}$Kbp) ; l'air de la ville de New York (en cours). La Diversa Corporation projette de recouvrir presque 100% de l'ADN des microorganismes résidant dans le sol, l'air, l'eau, les plantes et les animaux à travers le monde (Short, 1997; Tringe *et al.*, 2005). Un autre projet concerne la cartographie des 10^{12} microbes formant la microflore du corps humain (Relman *et al.*, 2001). Ces projets nécessitent la construction de bibliothèques métagénomiques et de base de données dédiées.

Les bibliothèques métagénomiques. Depuis la construction de la première bibliothèque métagénomique (Schmidt *et al.*, 1991), des centaines d'autres ont été mises en place : 2 500 bibliothèques contenant des millions de génomes ont été créées par la Diversa Corporation. Cinq bibliothèques métagénomiques virales ont été publiées : 2 d'eau de mer de rivage, 1 de sédiment marin, 1 de féces humaines et 1 de féces équines. Plus de 60% des ORFs dans chacun de ces 5 groupes (mesuré en décembre 2004) sont nouvelles. On constate que 65% des gènes phagiques n'ont pas d'homologues connus tandis que seulement 10% des gènes bactériens, issus de métagénomes et de cultures microbiennes, sont nouveaux. Cela suggère que la plupart du métagénome bacterien global est échantillonné alors que le métagénome viral global reste relativement mal caractérisé. Voir (Daniel, 2005) pour les méthodes d'analyses des bibliothèques métagénomiques.

Les bases de données. En plus des bases de données de dépot de séquences (voir *Genome Online Database*), des bases de données dédiées aux métagénomes ont été construites telles que : *megx.net* qui se restreint aux données métagénomiques des organismes marins et propose 25 génomes complets ; *CAMERA* qui se dédie à la maintenance et à la mise à disposition des séquences environnementales brutes et aux métadonnées associées. De nouvelles bases de données émergent qui intégrent aux séquences métagénomiques des données géographiques et environnementales : *Micro-Mar* (Pushker *et al.*, 2005) est une base de données qui stocke 8187 entrées (959 archea, 6351 bactéries, 877 non-classées) dont 93% disposent d'informations géographiques complétes ; *MetaFunctions* développe un système de data mining qui corrèle les motifs génétiques des génomes et des métagénomes aux données environnementales contextuelles.

4. Analyses des séquences métagénomiques

Assemblage de séquences métagénomiques. Le problème d'assemblage des fragments d'ADN issus d'un seul génome n'est pas complètement résolu, et en métagénomique le problème est encore plus difficile puisqu'il est nécessaire de pouvoir associer le

fragment de séquence à l'espèce duquel il provient. Les programmes d'assemblage comme Phred/Phrap et Sequencher ont été mis au point pour combiner des séquences issues d'un même génome et sont fondés sur des hypothèses fausses lorsqu'elles sont appliquées à un métagénome. Par exemple, une différence de base entre 2 séquences est considérée comme une erreur de séquençage par ces programmes, alors que cela pourrait être le signe d'une différence d'origine en métagénomique. Le grand nombre de génomes viraux (jusqu'à plusieurs millions) prélevés dans l'environnement ainsi que les répétitions de séquences de type transposons rendent la discrimination entre les organismes et l'assemblage des séquences d'autant plus difficile.

Ces problèmes nécessitent des approches algorithmiques fines, des fragments d'ADN plus longs, ainsi que des taux de couverture plus importants. Voir (Chen *et al.*, 2005) pour une vue d'ensemble. Récemment, la différence de nombre de répétitions de tétranucléotides entre génomes a pu être utilisée pour discriminer ces génomes, mais cette approche nécessite un échantillon de faible complexité et une répartition équitable des génomes (Teeling *et al.*, 2004). Le polymorphisme des séquences nucléotidiques, la duplication des gènes et le transfert horizontal de gène sont tous des obstacles à la fiabilité de l'assemblage des génomes.

Recherche de gène dans le métagénome. Les algorithmes de recherche de gènes sont fondamentalement basés sur la similarité de séquences, ce qui constitue une réelle limitation à la découverte de nouveaux gènes dans le métagénome. Des nouvelles approches devraient être développées afin de détecter les gènes tronqués aux extrémités des contigs et les génes sans homologues connus comme ceux des virus. Une version prochaine du programme d'annotation EXOGEAN devrait prendre en compte la première de ces limitations (S. Djebali, communication personelle, 2006). Pour les gènes sans homologues, des méthodes basées sur un système de score des régions codantes ont été développées. GLIMMER (Salzberg *et al.*, 1998; Delcher *et al.*, 1999) utilise un modèle de Markov interpolé. Cet outil effectue des prédictions basées sur un contexte variable évoluant selon la composition locale de la séquence, et a déjà été utilisé dans un contexte métagénomique (Lilles *et al.*, 2003).

5. De l'analyse statistique des séquences aux espaces métagénomiques

La classification des séquences génomiques de bactéries et virus/phages, et ultimement des couples phage/hôte, ainsi que la mise en corrélation de ces informations avec l'environnement et la physiologie de ces organismes, demande de développer des outils mathématiques et algorithmiques sophistiqués pour l'analyse des séquences.

Certaines propriétés statistiques des séquences, comme le contenu en GC/AT/GT ou la distribution de tétranucleotides, peuvent constituter une première approche pour regrouper ces séquences. Ces "signatures" semblent être représentatives pour des séquences de longueur minimum de 50Kbp et semblent être corrélées aux classifications phylogénétiques, mais pas aux classifications environnementales (Campbell *et al.*, 1999). Également, l'analyse de séquences nucléotidiques permet de définir une

signature pour les génomes bactériens basée sur la fréquence de mots d'une longueur maximale de 10 lettres (Blaisdell *et al.*, 1993; Deschavanne *et al.*, 1991; Karlin *et al.*, 1992; Gelfand *et al.*, 1997; Deschavanne *et al.*, 1999; Campbell *et al.*, 1999). Ces signatures ont démontré leur utilité pour la classification phylogénétique bactérienne. Un outil fournissant des analyses de séquences et de leurs signatures est décrit dans (Fertil *et al.*, 2005). Cette approche a également été appliquée à un grand nombre de phages (G. Poncelin et P. Deschavanne, communication personelle, 2006).

Classification phylogénétique. La construction d'arbres phylogénétiques à partir de séquences partielles, chevauchantes ou pas, nécessite de nouvelles approches algorithmiques qui pourraient mener à la construction d'arbres "bruts" mais suffisant pour certaines applications comme l'apprentissage de modèles de Markov cachés sur des séquences génomiques courtes. Les méthodes des super-arbres (Bininda-Emonds, 2004), basées sur la construction des arbres à partir de multiples sous-arbres, représentent une direction intéressante à étudier.

Environ 80% des protéines virales sont identifiées dans les POGs (Phage Orthologous Groups of proteins) comme étant très spécifiques aux génomes des phages et peu retrouvés dans les génomes microbiens (Liu *et al.*, 2006). Par conséquent, la classification phylogénétiques des phages pourrait être indépendante de la classification bactérienne : un grand nombre de séquences homologues de divergence variée pourrait s'avérer être un point essentiel pour la compréhension de la classification phylogénétique virale.

Actuellement, la *Ribosomal Database Project II* contient 253 813 séquences d'ARNr de bactéries non-cultivables mais de telles séquences ne fournissent que très peu d'informations sur les capacités physiologiques de ces organismes.

Classification environnementale et physiologique . L'analyse du biais de codon des séquences codantes virales et bactériennes propose une approche alternative pour la reconstruction d'un espace de séquences métagénomiques. Une étude automatisée des biais de codon effectuée sur des génomes complets bactériens (Carbone *et al.*, 2003; Carbone *et al.*, 2004; Carbone *et al.*, 2005) a permis de définir un espace formel où les bactéries peuvent être comparées par rapport à leur %GC, la force du biais traductionnel, la température de croissance optimale, l'aérobisme/anaérobisme et d'autres critères. Ces espaces semblent être gouvernés par l'habitat et la physiologie, ils représentent une classification alternative à la phylogénie et sont plus proches des conditions de vie de l'organisme. L'idée algorithmique utilisée pour l'étude des bactéries devrait être revisitée pour les séquences métagénomiques et adaptée aux génomes viraux. Sur de tels espaces, les hypothèses comme l'adaptabilité des virus aux biais de codons de leurs hôtes peuvent être vérifiées. Des analyses préliminaires supportent cette idée. On envisage d'appliquer cette approche aux bactéries, bactériophages, et au mélange des deux populations.

Une autre espace expérimental qui propose un lien entre l'écologie et l'organisation bactérienne a été défini. La biodiversité des communautés virales non-cultivables peut être évaluée en construisant des modèles de la structure des communautés bac-

teriennes/virales en utilisant un algorithme modifié de Lander-Waterman qui prédit le spectre de contigs. PHACCS (Phage Communities from Contig Spectrum) (Angly *et al.*, 2005), un outil implémentant cette idée, trouve le modèle le plus approprié en optimisant les paramètres du modèle jusqu'à ce que le spectre de contigs prédit soit aussi près que possible du spectre expérimental. Ce modèle est la base pour estimer la richesse d'une communauté virale non-cultivable (à travers la définition d'un indice de similarité et de diversité) et l'abondance du génotype le plus représenté.

Distribution sur l'espace des mots. Une étude à large échelle de l'espace des mots apparaissant dans les séquences microbiennes devrait pouvoir dire si des parties de l'espace sont beaucoup, peu ou pas occupées par ces mots. De nouvelles mesures basées sur les mots devraient alors mener à une comparaison des métagénomes dans un espace formel et leur localisation dans l'espace pourrait suggérer des critères pour regrouper les séquences microbiennes proches et discriminer les plus éloignées. Il faudrait alors vérifier si ces regroupements apportent des informations supplémentaires sur les classifications phylogénétiques ou environnementales.

6. Comparaison de séquences métagénomiques

Le faible taux de couverture des séquences métagénomiques ne permet pas aux algorithmes actuels d'identifier des ORFs de façon fiable, beaucoup peuvent être manquées pour des raisons d'erreurs de séquençage ou de séquences trop courtes pour contenir des informations suffisantes à la détection des gènes. Ces limitations seront surmontées avec des séquences plus longues et des taux de couverture plus forts, mais le principal problème algorithmique reste. Les outils comme tBLASTx comparent les 6 cadres de lecture d'une séquence nucléotidique métagénomique avec les 6 cadres de lecture des séquences nucléotidiques des bases de données, ce qui demande une puissance de calcul substantielle et prend beaucoup plus de temps qu'une comparaison de séquences classiques.

Le problème des graines de similarité. Presque toutes les comparaisons de séquences métagénomiques sont effectuées par des algorithmes de recherche de similarité de séquence tels que BLAST et FASTA. L'observation, selon laquelle la plupart des séquences virales métagénomiques ne présentent pas de similarité avec les séquences de la base de données GenBank, indique que des méthodes algorithmiques plus puissantes sont nécessaires. Le critère du "hit" est l'élément clé des algorithmes heuristiques d'alignement local. Il détermine un ensemble de motifs sur lequel la recherche de similarité est basée, et le choix de cet ensemble est décisif quant aux résultats de la méthode. Les séquences métagénomiques nécessitent un affinement du critère du hit via l'introduction de nouvelles idées algorithmiques telles que la spécification de graines uniques (Noé *et al.*, 2004; Noé *et al.*, 2005) ou multiples. La conception et l'emploi de graines est à la base des nouveaux algorithmes de fouille de base de données (Nicolas *et al.*, 2005).

Le problème d'alignement des protéines divergentes. La classification microbiennes pourrait être suggérée par les protéines, en particulier par les POGs (Phage Orthologous Groups of proteins) (Liu *et al.*, 2006) pour les phages, et par les COGs (Clusters of Orthologous Groups of proteins) pour les bactéries (Tatusov *et al.*, 2001). Le problème de cette approche est que la plupart des protéines virales ne sont pas annotées du fait de leur grande divergence par rapport aux séquences des autres protéines microbiennes connues. L'annotation des séquences très divergentes est un problème fondamental : des génomes déjà séquencés pouvent présenter jusqu'à 60% de gènes non-annotés. Des outils pour détecter les protéines homologues présentant <20% d'identité de séquence sont nécessaires. PHYBAL (Proteins with HYdrophobic Blocks ALignment), un outils dédié à l'alignement des protéines homologues très divergentes, a été développé et sera bientôt disponible (Baussand *et al.*, n.d.). La paramétrisation de nouveaux critères de sélection est en cours.

Le problème de la quantification des communautés. Les familles microbiennes environnementalement et physiologiquement proches ne peuvent pas être facilement discriminées à partir de fragments de séquence, le seuil de pourcentage d'identité en dessous duquel on peut considérer des séquences comme étant issues de 2 familles différentes est difficilement déterminable. Ce problème est intimement lié au problème de la détection des protéines homologues divergentes.

Traits particuliers caractérisant les communautés. Une façon de caractériser une communauté environnementale est de rechercher les motifs d'ADN partagés entre les individus de la même communauté. On peut rechercher des régularités et irrégularités des séquences métagénomiques telles que a. les séquences courtes inverses complémentaires (de longueur variant entre 3bp et 5000bp) pour lesquelles une méthode originale utilisant les modèles de Markov cachés a été proposée (Robelin *et al.*, 2003) ; b. les séquences codantes qui ne sont que partiellement séquencés.

Une autre approche est de rechercher un chemin biologique spécifique partagé par la communauté. La grande diversité des séquences métagénomiques rende difficile la reconstruction des chemins biologiques et la compréhension du réseau biologique au sein de la communauté. Il existe cependant des exceptions. Un groupe de gènes spécifiques à la réduction du sulfate a été trouvé en comparant des séquences métagénomiques issues de plusieurs bactéries phylogénétiquement très éloignées (Mussmann *et al.*, 2005). L'explication proposée est le transfert horizontal des gènes. Un autre exemple concerne la découverte, chez *Leptospirillum* groupe II, d'une enzyme fondamentale pour l'oxidation du fer dans la communauté microbienne étudiée ainsi qu'un nouveau processus métabolique qui pourrait être une pierre angulaire pour l'ensemble de l'écosystème (Tyson *et al.*, 2004). La reconstruction du génome de *K. stuttgartiensis* a amené à l'identification des gènes responsables de la biosynthèse des ladderanes et du processus d'oxydation anaérobique de l'ammoniac. Le groupe de bactéries anammox est le responsable clé de la génération du nitrogène atmospherique.

En conclusion, la comparaison des fragments de génomes de microbes non-cultivables pourrait mener à la construction d'une cartographie de la diversification et de la spécification des organismes microbiens dans leur environnement. La compa-

raison de larges bibliothèques métagénomiques, témoin de la variété spatiale et temporelle des communautés, a déjà permis de fournir des résultats significatifs sur le style de vie de populations microbiennes (Tringe *et al.*, 2005; Schloss *et al.*, 2005), mais l'organisation á large échelle de ces communautés et leurs interactions restent largement méconnues.

7. Bibliographie

Amann R., Binder B., Olson R., Chisholm S., Devereux R., Stahl D., « Combinationof 16S rRNA targeted oligonucleotide probes with flow-cytometry for analyzing mixed microbial populations », *Appl Environ Microbiol*, vol. 56, p. 1919-1925, 1990.

Angly F., Rodriguez-Brito B., Bangor D., McNairnie P., Breitbart M., Salamon P., Felts B., Nulton J., Mahaffy J., Rohwer F., « PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information », *BMC Bioinformatics*, 2005.

Baussand J., Deremble C., Carbone A., Periodic distributions of hydrophobic amino acids allows to define fundamental building blocks to align distantly related proteins, Submitted manuscript, n.d.

Bernal A., Ear U., Kyrpides N., « Genomes On Line Database (GOLD) : a monitor of genome project world-wide », *Nuc. Ac. Res.*, vol. 29, p. 126-127, 2001.

Bininda-Emonds O. (ed.), *Phylogenetic Supertrees : Combining Information to Reveal the Tree of Life*, Kluwer Academic Publishers, Dordrecht, the Netherlands, 2004.

Blaisdell B., Rudd K., Matin A., Karlin S., « Significant dispersed recurrent DNA sequences in the Escherichia coli genome. Several new groups. », *J Mol Biol.*, vol. 229, n° 4, p. 833-848, 1993.

Breitbart M., *et al.*, « Genomic analysis of uncultured marine viral communities », *Proceedings of the National Academy of Sciences USA*, vol. 99, p. 14250-14255, 2002.

Campbell A., Mrazek J., Karlin S., « Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA », *Proc Natl Acad Sci U S A.*, vol. 96, n° 16, p. 9184-9189, 1999.

Carbone A., Képès F., Zinovyev A., « Codon bias signatures, organisation of microorganisms in codon space and lifestyle », *Molecular Biology and Evolution*, vol. 22, n° 3, p. 547-561, 2004.

Carbone A., Madden R., « Insights on the evolution of metabolic networks of unicellular translationally biased organisms from transcriptomic data and sequence analysis », *Journal of Molecular Evolution*, vol. 61, p. 456-469, 2005.

Carbone A., Zinovyev A., Képès F., « Codon Adaptation Index as a measure of dominating codon bias », *Bioinformatics*, vol. 19, p. 2005-2015, 2003.

Chen K., Pachter L., « Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities », *PLOS Computational Biology*, vol. 1, p. 106-112, 2005.

Cowan D., Meyer Q., Stafford W., Muyanga S., Cameron R., Wittwer P., « Metagenomic gene discovery : past, present and future », *Trends in Biotechnology (review)*, vol. 23, n° 6, p. 321-329, 2005.

Daniel R., « The metagenomics of soil », *Nature (review)*, vol. 3, p. 470-478, 2005.

Delcher A., Harmon D., Kasif S., White O., Salzberg S., « Improved microbial gene identification with GLIMMER. », *Nuc. Ac. Res.*, vol. 27, p. 4636-4641, 1999.

Deschavanne F., Giron A., Vilain J., Fagot G., Fertil B., « Genomic Signature : Characterization and Classification of Species Assessed by Chaos Game Representation of Sequences », *Molecular Biology and Evolution*, vol. 16, p. 1391-1399, 1999.

Deschavanne P., Radman M., « Counterselection of GATC sequences in enterobacteriophages by the components of the methyl-directed mismatch repair system », *J Mol Evol*, vol. 33, n° 2, p. 125-132, 1991.

Fertil B., Massin M., Lespinats S., Devic C., Dumee P., Giron A., « GENSTYLE : exploration and analysis of DNA sequences with genomic signature », *Nuc. Ac. Res. (Web Server Issue)*, vol. 33, p. W512-W515, 2005.

Gelfand M., Koonin E., « Avoidance of palindromic words in bacterial and archaeal genomes : a close connection with restriction enzymes », *Nuc. Ac. Res.*, vol. 25, n° 12, p. 2430-2439, 1997.

Handelsman J., « Metagenomics : application of genomics to uncultured microorganisms », *Microbiology and Molecular Biology Reviews*, vol. 68, p. 669-685, 2004.

Karlin S., Brendel V., « Chance and statistical significance in protein and DNA sequence analysis », *Science*, vol. 257, n° 5066, p. 39-49, 1992.

Lilles M., Manske B., Bintrim S., Handelsman J., Goodman R., « A Census of rRNA Genes and Linked Genomic Sequences within a Soil Metagenomic Library », *Applied and Environmental Microbiology*, vol. 69 :5, p. 2684-2691, 2003.

Liolios K., Tavernarakis N., Hugenholtz P., Kyrpides N., « The Genomes On Line Database (GOLD) v.2 : a monitor of genome projects worldwide. », *Nuc. Ac. Res.*, vol. 34, p. 332-334, 2006.

Liu J., Glazko G., Mushegian A., « Protein repertoire of double-stranded DNA bacteriophages », *Virus Research*, vol. 117, p. 68-80, 2006.

Mussmann M., *et. al.*, « Clustered Genes Related to Sulfate Respiration in Uncultured Prokaryotes Support the Theory of Their Concomitant Horizontal Transfer », *Journal of Bacteriology*, vol. 187, n° 20, p. 7126-7137, 2005.

Nature, « Focus on metagenomics », *Nature Reviews in Microbiology*, vol. 3, Nature Publishing Group, p. ..., June, 2005.

Nicolas F., Rivals E., « Hardness of optimal spaced seed design », *Combinatorial Pattern Matching*, vol. 3537, Springer, p. 144-155, June, 2005.

Noé L., Kucherov G., « Improved hit criteria for DNA local alignment », *BMC Bioinformatics*, vol. 5, p. 149-156, 2004.

Noé L., Kucherov G., « YASS : enhancing the sensitivity of DNA similarity search », *Nuc. Ac. Res. (Webserver Issue)*, vol. 33, p. W540-W543, 2005.

Pace N., Stahl D., Lane D., Olsen G., « Analysing the natural microbial populations by rRNA sequences », *ASM News*, vol. 5, p. 4-12, 1985.

Pace N., Stahl D., Lane D., Olsen G., « The analysis of natural microbial populations by ribosomal RNA sequences », *Adv. Microb. Ecol.*, vol. 9, p. 1-55, 1986.

Paul J., Sullivan M., « Marine phage genomics : What have we learned ? », *Comp Biochem Physiol B Biochem Mol Biol*, vol. 16, p. 299-307, 2005.

Pushker R., D'Auria G., Alba-Casado J., Rodriguez-Valera F., « Micro-Mar : a database for dynamic representation of marine microbial biodiversity. », *BMC Bioinformatics*, vol. 6, p. 222, 2005.

Relman D., Falkow S., « The meaning and impact of the human genome sequence for microbiology », *Trends Microbiol*, vol. 9, n° 5, p. 206-208, 2001.

Robelin D., Richard H., Prum B., « SIC : a tool to detect short inverted segments in a biological sequence », *Nuc. Ac. Res.*, vol. 31, n° 1, p. 3669-3671, 2003.

Rodriguez-Valera F., « Environmental genomics, the big picture ? », *FEMS Microbiology Letters*, vol. 231, p. 153-158, 2004.

Salzberg S., Delcher A., Kasif S., White O., « Microbial gene identification using interpolated Markov models », *Nuc. Ac. Res.*, vol. 26 :2, p. 544-548, 1998.

Schloss P., Haldelsman J., « Metagenomics for studying unculturable microorganisms : cutting the Gordian knot », *Genome Biology (minireview)*, vol. 6, p. 229, 2005.

Schmidt T., Delong E., Pace N., « Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. », *J. Bacteriol.*, vol. 14, p. 4371-4378, 1991.

Short J., « Recombinant Approaches for Accessing Biodiversity », *Nature Biotechnology*, vol. 15, p. 1322-1323, 1997.

Strous M., Pelletier E., et al, « Deciphering the evolution and metabolism of an anammox bacterium from a community genome. », *Nature*, vol. 440, n° 7085, p. 750-754, 2006.

Tatusov R., Natale D., Garkavtsev I., Tatusova T., Shankavaram U., Rao B., Kiryutin B., Galperin M., Fedorova N., Koonin E., « The COG database : new developments in phylogenetic classification of proteins from complete genomes », *Nuc. Ac. Res.*, vol. 29, p. 22-28, 2001.

Teeling H., Meyerdierks A., Bauer M., Amann R., Glockner F., « Application of tetranucleotide frequencies for the assignment of genomic fragments », *Environ. Microbiol.*, vol. 6, p. 938-947, 2004.

Tringe S., *et al.*, « Comparative metagenomics of microbial communities », *Science*, vol. 308, p. 554-557, 2005.

Tyson G., Chapman J., Hugenholtz P., Allen E., Ram R., Richardson P., Solovyev V., Rubin E., Rokhsar D., Banfield J., « Insight into community structure and metabolism through reconstruction of microbial genomes from the environment », *Nature*, vol. 428, p. 37-43, 2004.

Venter C., *et al.*, « Environmental Genome Shotgun Sequencing of the Sargasso Sea », *Science*, vol. 304, p. 66-74, 2004.

ANNEXE POUR LE SERVICE FABRICATION
A FOURNIR PAR LES AUTEURS AVEC UN EXEMPLAIRE PAPIER
DE LEUR ARTICLE ET LE COPYRIGHT SIGNE PAR COURRIER
LE FICHIER PDF CORRESPONDANT SERA ENVOYE PAR E-MAIL

1. ARTICLE POUR LA REVUE :
   Technique et science informatiques. Volume ...– n.../2006

2. AUTEURS :
   Julie Baussand — Alessandra Carbone

3. TITRE DE L'ARTICLE :
   Métagénomique bactérienne et virale

4. TITRE ABRÉGÉ POUR LE HAUT DE PAGE MOINS DE 40 SIGNES :
   Métagénomique bactérienne et virale

5. DATE DE CETTE VERSION :
   4 septembre 2006

6. COORDONNÉES DES AUTEURS :

   – adresse postale :

   Génomique Analytique, Université Pierre et Marie Curie-Paris 6, INSERM U511
   91 blv de l'Hôpital, 75012 Paris
   baussand@infop6.jussieu.fr et Alessandra.Carbone@lip6.fr

   – téléphone : 01.40.77.97.98

   – télécopie : 01.1.45.83.88.58

   – e-mail : Alessandra.Carbone@lip6.fr

7. LOGICIEL UTILISÉ POUR LA PRÉPARATION DE CET ARTICLE :
   LaTeX, avec le fichier de style `article-hermes2.cls`,
   version 1.23 du 17/11/2005.

8. FORMULAIRE DE COPYRIGHT :
   Retourner le formulaire de copyright signé par les auteurs, téléchargé sur :
   `http://www.revuesonline.com`

SERVICE ÉDITORIAL – HERMES-LAVOISIER
14 rue de Provigny, F-94236 Cachan cedex
Tél. : 01-47-40-67-67
E-mail : revues@lavoisier.fr
Serveur web : http://www.revuesonline.com