Mathématiques et post−génomique

Jean−Philippe Vert
Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611−011, Japan
Tel. : +81−774−38−3296 − Fax. : +81−774−38−3269
E−mail : Jean−Philippe.Vert@mines.org

Résumé
La post−génomique vise à analyser les quantités de données biologiques produites par les divers programmes de séquençage. A travers deux exemples nous montrons comment les mathématiques peuvent contribuer à cette nouvelle discipline, et en bénéficier simultanément.

Abstract
Post−genomics aims at analyzing the wealth of biological data generated by genome−sequencing projects. We show through two examples how mathematics can contribute to this new field and benefit from it in the same time.

Introduction
Depuis les années 1980 et le lancement de programmes de séquençage à grande échelle, les principaux projets de recherche en "bioinformatique" ont mis l'accent sur le développement de technologies pour générer des données (programmes de séquençage, développement des puces à ADN, etc...) et pour les stocker (création de réseaux de bases de données biologiques). L'ère qui s'ouvre à présent, souvent qualifiée de "post−génomique", vise à analyser ces données et à en extraire de nouvelles connaissances biologiques. Elle est tout à la fois prometteuse d'importants progrès en biologie et médecine, et redoutable par la complexité des problèmes qu'elle soulève.
La post−génomique a rapidement progressé ces dernières années, fournissant par exemple des algorithmes pour localiser les gènes au sein des génomes ou prédire leur fonction au sein de la cellule. Cependant, ces méthodes sont encore loin d'être parfaites, et de nouveaux champs d'investigation prometteurs, comme l'analyse des données d'expression fournies par les puces à ADN, semblent à peine développés. Des outils mathématiques de plus en plus avancés sont utilisés en post−génomique, et par l'originalité des problèmes qu'elle pose, la post−génomique promeut le développement de nouvelles méthodes mathématiques.
Afin d'illustrer concrètement cette synergie entre post−génomique et mathématiques je présenterai brièvement deux projets de recherche en cours au Centre de Bioinformatique de l'université de Kyoto, dans le laboratoire du Professeur Kanehisa. Ce laboratoire gère la base de données KEGG (Kyoto Encyclopedia of Genes and Genomes) et est un des principaux centres de recherche Japonais en bioinformatique. Des informations plus détaillées sur ces projets peuvent être obtenues sur la toile: http://web.kuicr.kyoto−u.ac.jp/~vert/ .

De la modélisation à la classification de séquences biologiques
La production de données biologiques est particulièrement féconde en séquences de lettres. Ces séquences représentent la structure primaire de macromolécules comme l'ADN, l'ARN ou les protéines. Elles s'expriment sous forme de suites finies dans un alphabet de 4 lettres pour l'ADN et l'ARN, et de 20 lettres (les acides aminées) pour les protéines. Afin de manipuler et d'analyser ces séquences biologiques, des modèles probabilistes ont très tôt été utilisés (1). Un modèle probabiliste est une distribution de probabilité sur l'ensemble des séquences. Pour modéliser un certain "type" de séquences (comme par exemple une certaine famille structurelle ou fonctionnelle de protéines, ou bien les régions codantes au sein d'un génome), il est courant de construire un modèle probabiliste qui affecte de grandes probabilités aux séquences qui appartiennent au "type" considéré, et de faibles probabilités aux autres. Les modèles les plus classiques pour modéliser les séquences biologiques sont les chaînes de Markov et les chaînes de Markov cachées, qui ont atteint un niveau de complexité remarquable.

Souvent, le véritable but de ces modèles est de déterminer si une nouvelle séquence (par exemple une nouvelle protéine) appartient ou non au type considéré. Une approche naturelle (dite Bayésienne) consiste à calculer la probabilité de cette nouvelle séquence sous le modèle probabiliste qui a été construit, et à vérifier si cette probabilité est élevée ou non. Malgré des succès certains cette méthode est limitée par le fait que les modèles probabilistes ne sont souvent pas assez précis pour représenter toute la subtilité du "type" considéré, car le nombre d'exemples utilisés lors de l'estimation du modèle (par voie statistique) est souvent trop faible par rapport à la complexité du "type". En d'autres termes les modèles probabilistes fournissent une bonne image "grossière" mais manquent parfois de précision.
Afin d'obtenir de meilleures performances en classification, il est possible de combiner ces modèles avec une famille d'algorithmes appelée "Support Vector Machines" (SVM). Les SVM ont été développées dans les années 1990 (2) et sont actuellement considérées comme les algorithmes de classification les plus efficaces dans de nombreux domaines d'application, surpassant notamment les réseaux de neurone. Les objets à classer sont d'abord projetés dans un espace Euclidien de grande dimension, puis séparés linéairement. Pour fonctionner, l'algorithme a juste besoin de connaître un "noyau", c'est–à–dire la fonction "produit scalaire" entre les projections de deux objets quelconques dans l'espace de grande dimension. En d'autre termes le noyau contient les informations sur la géométrie de l'espace dans lequel les objets sont projetés. De nombreux noyaux ont été développés quand les objets à classer sont des vecteurs de nombres réels (permettant de classer des images par exemple), mais très peu de noyaux ont été conçus quand les objets sont des séquences discrètes.
Une question naturelle se pose donc: est–il possible de définir un noyau pour des séquences biologiques qui utilise "au mieux" l'information capturée par les modèles probabilistes? Nous avons récemment proposé une méthode générale pour construire un tel noyau, qui permet d'engendrer une géométrie Euclidienne sur l'espace des séquences à partir d'une distribution de probabilité (3). Cette construction créée de nouveaux liens entre probabilités, géométrie et informatique, et fournit des classificateurs performants pour les séquences biologiques.

Analyser des réseaux complexes

Un second thème de recherche nécessitant des développements mathématiques est l'analyse des réseaux complexes qui apparaissant de plus en plus dans les bases de données biologiques. Un exemple de réseau complexe est le graphe dont les noeuds sont tous les gènes actuellement connus dans toutes les espèces biologiques (plus de 200.000 gènes début 2001), et dont les arètes entre gènes représentent différentes relations : similarité des séquences, proximité au sein d'un génome, appartenance à un même cycle métabolique, co–expression dans certaines conditions etc... L'étude des propriétés qualitatives et quantitatives d'un tel graphe est susceptible de fournir des connaissances nouvelles concernant le fonctionnement de systèmes biologiques complexes, mais les outils pour y parvenir sont encore à développer.
La base de données KEGG contient par exemple le graphe dont les arètes entre gènes représentent les similarités entre séquences (4). Nous développons actuellement une méthode pour prédire la fonction d'un gène à partir de la topologie du graphe autour de ce gène. En effet, le voisinage d'un gène renseigne sur les espèces qui contiennent un autre gène "ressemblant " au gène initiale, et fournit donc une indication sur la fonction de ce gène à travers la conservation de cette fonction au sein de certaines espèces lors de l'évolution. La taille du graphe rend difficile toute approche qui chercherait à caractériser systématiquement sa topologie locale par un nombre fini de paramètre (car le nombre de paramètre augmenterait exponentiellement avec la taille du graphe local), mais des méthodes stochastiques sous–optimales semblent en mesure de fournir des algorithmes performants. Ici encore, le développement de nouvelles méthodes est tirée par un nouveau type de besoin apparaissant en post–génomique.

Références

[1] R. Durbin et al, "Biological sequence analysis: Probabilistic models of proteins and nucleic acids", Cambridge University Press, 1998.
[2] V. Vapnik, "Statistical Learning Theory", Wiley, 1998.
[3] J.–P. Vert, "Support Vector Machince prediction of signal peptide cleavage site", to appear.
[4] SSDB : the Sequence Similarity Database, http://ssdb.genome.ad.jp .