INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Projet helix

Informatique et génomique

Rhône-Alpes

THÈME 3A

Table des matières

1. Composition de l'équipe

L'équipe du projet Helix est localisée à Montbonnot (Grenoble) et sur le campus de la Doua à Villeurbanne (Lyon). La composante lyonnaise est constituée par l'équipe « Biométrie moléculaire, évolution et structure des génomes » dirigée par Manolo Gouy, DR CNRS, au sein de l'UMR 5558 CNRS - université Claude Bernard (UCB) « Biométrie et biologie évolutive », dirigée par Christian Gautier, professeur à l'UCB. Le groupe SwissProt, dirigé par Amos Bairoch au sein du SIB (« Swiss Institute of Bioinformatics ») à Genève, est « équipe associée » au projet Helix.

Responsable

François Rechenmann [directeur de recherche, INRIA]

Chercheurs, enseignants-chercheurs et ingénieurs permanents

Stéphane Delmotte [assistant ingénieur, CNRS]
Laurent Duret [chargé de recherche, CNRS]
Christian Gautier [professeur, université Claude Bernard]
Philippe Genoud [maître de conférences, université Joseph Fourier]
Manolo Gouy [directeur de recherche, CNRS]
Laurent Gueguen [maître de conférences, université Claude Bernard]
Hidde de Jong [chargé de recherche, INRIA]
Jean Lobry [maître de conférences, université Claude Bernard]
Dominique Mouchiroud [professeur, université Claude Bernard]
Michel Page [maître de conférences, université Pierre Mendès-France]
Guy Perrière [chargé de recherche, CNRS]
François Rechenmann [directeur de recherche, INRIA]
Marie-France Sagot [chargé de recherche, INRIA]
Bruno Spataro [ingénieur de recherche, CNRS]
Alain Viari [directeur de recherche, INRIA]
Danielle Ziébelin [maître de conférences, université Joseph Fourier]

Chercheurs et ingénieurs non permanents

Eric Coissac [maître de conférences, université Paris 6, en délégation à l'INRIA]
Christophe Bruley [ingénieur expert, INRIA]
Stéphane Bruley [ingénieur associé, INRIA]
Antoine Brun [ingénieur expert, INRIA]
Pierre-Emmanuel Ciron [ingénieur expert, INRIA]
Véronique Dupierris [ingénieur expert, INRIA]
Céline Hernandez [ingénieur associé, INRIA]
Agnès Iltis [ingénieur expert, INRIA]
Anne Morgat [chercheur associé, INRIA, puis ingénieur expert, SIB]
Thimotée Silvestre [ingénieur sur contrat, CNRS]
Erik Wessel [ingénieur associé, INRIA]

Post-doctorants

Sandrine Hughes [ATER UCBL]
Stéphanie Merriene [INRIA]
Simon Penel [contrat européen]
Violaine Pillet [INRIA]
Nadia Pisanti [bourse ERCIM]
Sébastien Provencher [bourse du Gouvernement du Québec]
Raquel Tavarès [bourse du gouvernement portugais]

Doctorants

Grégory Batt [allocation couplée ENS Lyon, directeurs de thèse : Hidde de Jong, François Rechenmann]
Frédéric Boyer [allocataire du Ministère de la Recherche, directeurs de thèse : Laurent Trilling, Alain Viari]
Gisèle Bronner [allocataire du Ministère de la Recherche, directeurs de thèse : Christian Gautier, François Rechenmann]
Alexandra Calteau [allocataire du Ministère de la Recherche, directeurs de thèse : Guy Perrière, Manolo Gouy]
Vincent Daubin [allocataire du Ministère de la Recherche, directeurs de thèse : Guy Perrière, Manolo Gouy]
Stéphane Declere [convention CIFRE, société GENOME express, directeurs de thèse : Alain Viari et Pierre Netter, université Paris 6]
Jean-François Dufayard [allocataire du Ministère de la Recherche, directeurs de thèse : Manolo Gouy, François Rechenmann]
Christelle Gonindar [allocataire du Ministère de la Recherche, directeurs de thèse : Christian Gautier, Didier Piaud, François Rechenmann]
Adel Khelifi [allocataire du Ministère de la Recherche, directeur de thèse : Dominique Mouchiroud]
Gabriel Marais [Allocation Couplée ENS Lyon, directeurs de thèse : Dominique Mouchiroud, Laurent Duret]
Julien Meunier [allocation couplée ENS Lyon, directeur de thèse : Laurent Duret]
Vincent Navratil [bourse INRA, directeur de thèse : Christian Gautier]
Gwenaele Piganeau [allocation couplée ENS Lyon, directeurs de thèse : Christian Gautier, Laurent Duret]
Loïc Ponger [allocataire du Ministère de la Recherche, directeur de thèse : Dominique Mouchiroud]
Marie Semon [allocataire du Ministère de la Recherche, directeur de thèse : Laurent Duret]
Fabienne Thornarat [allocataire du Ministère de la Recherche, directeur de thèse : Manolo Gouy]
Marina Zelwer [allocataire du Ministère de la Recherche, directeurs de thèse : Maxime Crochemore et Marie-France Sagot]

Membres extérieurs

Jean Dina [XRCE, Meylan]
Eric Fanchon [chargé de recherche, CNRS, IBS, Grenoble]
Gilles Faucherand [société GENOME Express, Meylan]
Corinne Lachaize [SIB, Genève]
Erwan Reguer [CEA, Grenoble, puis ingénieur expert INRIA]

Assistante de projet

Françoise de Coninck
La partie lyonnaise d'Helix utilise les ressources humaines de l'UMR 5558, en particulier en terme de secrétariat.

2. Présentation et objectifs généraux

2.1. Introduction

Mots clés : *bioinformatique, biologie, génomique, génome, protéomique, protéome, génomique comparative, cartographie comparée, synténie, métabolisme, régulation génique, annotation des génomes, représentation des connaissances, modèles dynamiques, simulation, extraction d'informations, évolution.*

2.2. Contexte et objectifs du projet

La dualité diversité/unité qui caractérise le Vivant fait jouer à l'informatique, et aux moyens de modélisation spécifiques qu'elle apporte, un rôle privilégié en biologie, certainement comparable au rôle qu'ont joué les mathématiques en physique. Ainsi, la bioinformatique ne se limite plus à l'analyse des séquences, mais cherche à exploiter et à recouper des données hétérogènes dont les origines expérimentales se diversifient. Pour ce faire, elle associe étroitement modélisation (bases de données et de connaissances, systèmes dynamiques) et analyse (algorithmique, statistiques). Les méthodes qu'elle propose se doivent d'être efficaces, mais surtout fiables et pertinentes.

Au sein du projet Helix, la bioinformatique est vue comme l'ensemble des méthodes et des outils informatiques destinés à modéliser, analyser et visualiser les diverses entités impliquées dans les processus d'expression, de transmission et d'évolution de l'information génétique. Les relations que ces entités entretiennent entre elles jouent un rôle majeur dans cette modélisation. Elles résultent, par exemple, de l'appartenance d'entités biologiques à un même réseau génique ou métabolique, de leur voisinage au sein d'un génome donné ou de l'existence d'une évolution commune. Cette diversité est difficile à prendre en compte dans une perspective purement mathématique et contribue à l'importance de la collaboration informatique/biologie.

2.3. Axes de recherche

Le projet Helix réunit sur deux sites (Lyon et Grenoble) deux cultures bioinformatiques différentes, issues historiquement de l'informatique pour la partie grenobloise et de la biologie pour la partie lyonnaise. Cependant, une longue tradition de collaboration entre les deux groupes assure une cohérence à l'ensemble, tant sur le plan des méthodes informatiques que sur celui des thèmes biologiques. La représentation des connaissances est sans doute le meilleur exemple de la dualité méthodologique entre les deux groupes, alors que la génomique comparative apparaît au centre des thématiques biologiques. Cette « intrication » des deux groupes ne se limite pas, bien entendu, à ces deux thèmes historiques. Six axes de recherche, la plupart transversaux entre les groupes, organisent notre activité en termes d'objectifs biologiques :

1. Evolution des espèces et des familles de gènes
2. Organisation spatiale (le long du génome) de l'information génomique
3. Annotation syntaxique et fonctionnelle des génomes
4. Modélisation dynamique des réseaux d'interactions géniques
5. Protéomique : aide à l'acquisition et modélisation des données
6. Extraction d'informations à partir de textes
7. Environnement didactique informatisé en bioinformatique et en biométrie.

La recherche méthodologique concerne principalement la représentation des connaissances, l'algorithmique, les systèmes dynamiques et les probabilités et statistiques.

Enfin, le développement de deux plates-formes joue un rôle intégrateur majeur :

- Genostar est une plate-forme bioinformatique de génomique exploratoire qui permet de valoriser de nombreux résultats de recherches antérieurs, tant au niveau des méthodes d'analyse que des outils de modélisation des données et des connaissances. Ce projet mobilise d'importantes forces de conception et développement logiciels.
- PBIL (Pôle BioInformatique Lyonnais) en collaboration avec l'IBCP (UMR CNRS 5086) fédère les développements méthodologiques du groupe lyonnais. Cette plate-forme devrait rapidement devenir rhône-alpine et donc intégrer l'ensemble d'Helix. Il convient également de souligner, dans ce cadre, les besoins importants en ressources de calcul et le rôle majeur, dans ce contexte, de la collaboration grandissante avec le centre de calcul de l'IN2P3 (par exemple pour la comparaison régulière et systématique de toutes les séquences protéiques les unes avec les autres).

Les deux plates-formes Genostar et PBIL offrent des services complémentaires et ne visent pas le même type de relations avec les biologistes. La première offre une modélisation sophistiquée permettant une appropriation efficace des méthodes d'analyse par le biologiste mais relève, en revanche, d'un effort important d'implémentation logicielle. PBIL est une plate-forme Web directement utilisable par n'importe quel chercheur, mais n'offrant pas de possibilité de personnalisation des stratégies d'analyse.

2.3.1. Évolution des espèces et des familles de gènes

Participants : Vincent Daubin, Jean-François Dufayard, Laurent Duret, Christian Gautier, Manolo Gouy [Correspondant], Guy Perrière, Gwenael Piganeau, Nadia Pisanti, Marie-France Sagot [Correspondant], Fabienne Thornarat, Marina Zelwer.

2.3.1.1. Préambule biologique

L'évolution est la caractéristique première des systèmes vivants. La diversité biologique résulte de la succession de deux processus indépendants : un processus d'« erreur » permet que l'information génétique transmise à un descendant soit légèrement différente de l'information présente chez l'organisme parental ; un processus de fixation au cours duquel une très petite fraction de ces erreurs vont voir leur fréquence augmenter dans les populations jusqu'à devenir la « norme ».

L'analyse de ces deux processus et de leurs conséquences dans les génomes soustendent une partie importante de la bioinformatique moléculaire ; elle sera de ce fait présente dans presque tous les thèmes d'Helix. On peut en particulier citer :

- *La reconstitution de l'histoire du Vivant.* La distance entre génomes s'accroît avec le temps de divergence, ce qui permet des estimations de la topologie et de la métrique de l'arbre du Vivant. La complexité des deux processus (« erreur » et « fixation ») conduit à des problèmes mathématiques difficiles, principalement dans le champ des probabilités et statistiques. Cependant, la prise en compte de l'ordre des gènes (voir plus loin la partie « cartographie ») conduit à des distances entre permutations ancrant la méthodologie associée dans l'algorithmique.
- *L'annotation des génomes.* Les génomes portent des informations très diverses, tant dans la nature des fonctions associées que dans le mode d'expression de cette information. Le processus de fixation, voire même le processus d'erreur, dépend de ces deux caractéristiques et conduit à l'existence de traces « diagnostiques », au sein des génomes, de la nature de l'information. Ces traces ne sont, bien entendu, interprétables que dans une perspective évolutive.

2.3.1.2. Approches méthodologiques

De manière parfois un peu arbitraire, la biologie sépare l'étude de l'évolution en l'étude du « *pattern* » et celle du « *process* », ce qui revient principalement à séparer la détermination de l'arbre du Vivant (le « *pattern* ») des mécanismes évolutifs eux-mêmes (le « *process* »). Nous avons adopté cette séparation par commodité, même si la reconstitution de l'arbre implique une compréhension de l'ensemble des mécanismes de l'évolution.

- *Reconstitution de l'arbre du Vivant, phylogénie.* Les séquences génomiques ou protéiques ont le même « format » quel que soit l'organisme. Leur comparaison permet, *a priori*, de reconstituer l'ensemble de l'arbre du Vivant. Cependant, la complexité mathématique des processus (au mieux markoviens non homogènes sur un arbre, mais en réalité plus complexes) nécessite des procédures d'estimations approchées. Helix s'intéresse tout particulièrement à l'introduction des connaissances sur les mécanismes évolutifs dans la modélisation des processus et donc dans l'estimation de la topologie et de la métrique de ces arbres. La partie 6.1 donnera des exemples de positionnement de ramifications très anciennes de l'arbre du Vivant. Les séquences ne sont pas les seuls objets biologiques utilisables dans la reconstruction de l'histoire de la Vie. L'ordre même des gènes se modifie progressivement et la comparaison des permutations permet la construction de distances phylogénétiquement pertinentes. Les problèmes méthodologiques soulevés sont principalement l'estimation de ces distances en termes de nombre d'opérations élémentaires (biologiquement possibles) nécessaires pour passer d'une permutation à une autre. L'algorithmique y est fortement mise à contribution.
- *Gestion de familles d'arbres.* La reconstruction phylogénétique utilisant la séquence d'un gène donné fournit un arbre. Actuellement, on connaît 6000 familles de gènes ayant plus de 4 séquences et donc 6000 arbres différents. La gestion et la comparaison de ces arbres est un problème informatique tout à fait caractéristique des collaborations entre les deux sites d'Helix (*cf.* partie 6.1).
- *Processus évolutifs.* Le changement d'une base présente dans le génome parental en une autre base présente dans le descendant résulte d'une interaction entre un processus purement physique (action mutagène d'une radiation par exemple) et de multiples processus biologiques. Le plus évident de ces processus est la réparation : la plupart des modifications du génome sont réparées,

mais pas toutes avec la même efficacité. Des interaction plus complexes peuvent exister au travers de situations qui favorisent ou défavorisent les erreurs (méthylation de certaines bases, temps plus ou moins long de séparation des deux brins de l'ADN, ...). La fixation d'une mutation est elle-même soumise à de multiples contraintes. Les plus connues sont celles dites de la sélection naturelle qui traduisent le fait que, si un individu porteur d'une mutation a plus de descendants que les autres, la mutation a une plus forte probabilité de se fixer. D'autres, plus subtiles, modulent l'effet de la sélection naturelle (taille des populations, recombinaison). Helix s'intéresse à l'ensemble de ces phénomènes, aussi bien dans le domaine bactérien qu'eucaryote. Il convient de souligner que ce domaine est assez caractéristique des relations entre recherche fondamentale et appliquée en biologie en citant (voir « Résultats nouveaux ») la détermination de l'origine de réplication ou le phénomène des transferts horizontaux, actuellement considéré comme majeur dans les domaines de la santé ou de l'agronomie.

2.3.2. Organisation spatiale (le long du génome) de l'information génomique

Participants : Gisèle Bronner, Eric Coissac, Christian Gautier, Laurent Gueguen, Adel Khelifi, Jean Lobry [Correspondant], Anne Morgat, Dominique Mouchiroud, Guy Perrière [Correspondant], Nadia Pisanti, Marie-France Sagot [Correspondante], Bruno Spataro, Alain Viari.

2.3.2.1. Préambule biologique

Les équipes fondatrices d'Helix ont participé très tôt (il y a plus de 20 ans) à la mise en évidence de fortes hétérogénéités au sein d'un même génome, aussi bien dans ses caractéristiques biologiques que statistiques. Il apparaît, en particulier, que des gènes voisins sur le génome partagent le plus souvent de multiples propriétés tant structurales (taille et nombre d'introns par exemple) que statistiques (fréquences des bases et des codons). Dans certains cas, cette structure de voisinage a pu trouver des interprétations en terme de processus biologiques. Par exemple, chez les bactéries, une telle structure résulte partiellement du mécanisme même de réplication du génome. Ce mécanisme, déjà cité au paragraphe précédent, ne génère pas les mêmes types d'erreurs tout au long du génome. Par contre, d'autres structures « régionales » résistent encore à la découverte des mécanismes qui les engendrent et les maintiennent. Le cas le plus caractéristique est celui des isochores des vertébrés. La prise en compte de cette structure est essentielle dans l'annotation des séquences ; elle correspond à la modification de multiples caractères (fréquences des bases, structure des gènes, nature des éléments transposables, ...).

Enfin, il faut souligner qu'au cours de l'évolution l'organisation spatiale des gènes est modifiée par des processus biologiques, encore incomplètement connus, mais qui génèrent de nombreuses modifications, parmi lesquelles : des permutations entre gènes relativement voisins, des inversions (retournements) de segments, des duplications, des mouvements à grande distance. Il est alors nécessaire de définir une distance biologiquement pertinente entre les arrangements correspondant à des espèces différentes afin d'en déduire des distances évolutives entre espèces (voir la section 2.2) ou afin de comparer le taux de réarrangements pour des régions différentes du même génome. L'informatique joue également pleinement son rôle d'outil de modélisation en proposant des définitions formelles de régions « homologues ». Ainsi nous nous sommes particulièrement intéressés à l'élaboration d'une définition opérationnelle pour les synténies bactériennes (groupes de gènes orthologues dont l'organisation spatiale est conservée entre deux espèces bactériennes).

2.3.2.2. Approches méthodologiques

Le double problème de la modélisation et de l'analyse des données de cartographies génomiques interpelle la représentation des connaissances, les statistiques et l'algorithmique. Le système de représentation des connaissances AROM est utilisé aussi bien à Grenoble sur les bactéries (Genostar, GEB) qu'à Lyon sur les eucaryotes (GemCore). L'analyse de la structuration spatiale des génomes fait appel à la fois à des méthodes de corrélation (corrélation non paramétriques sur graphe de voisinage et processus de Markov) et de partitionnement ou segmentation (y compris dans le champ de l'analyse de données vectorielle). Les problèmes difficiles d'algorithmique émergent dans la construction et le calcul de distances entre cartes.

2.3.3. Annotation syntaxique et fonctionnelle des génomes

Participants : Frédéric Boyer, Stéphane Bruley, Laurent Gueguen, Anne Morgat [Correspondante], Guy Perrière, Marie-France Sagot [Correspondante], Alain Viari [Correspondant].

2.3.3.1. Préambule biologique

Si des expériences permettent d'apporter des résultats importants dans l'organisation de l'information le long des séquences génomiques, leur coût en temps et en moyens n'en fait pas un moyen d'analyse systématique de l'ensemble des génomes séquencés. La prédiction de cette information à partir de la séquence et de données expérimentales, même parcellaires, est donc absolument nécessaire à la valorisation des efforts de séquençage. Cette prédiction repose sur l'ensemble des connaissances biologiques>

L'annotation d'un génome fait référence à trois objectifs biologiques différents :

1. L'annotation syntaxique concerne l'identification de zones d'intérêt sur la séquence. Il s'agit typiquement de la recherche des zones codant potentiellement pour des protéines ou des ARNt, de la recherche de signaux de régulation de l'expression génétique et, d'une manière générale, de la localisation de motifs lexicaux ou structuraux caractérisés. A l'heure actuelle, l'annotation syntaxique des gènes codant pour des protéines des génomes procaryotes ne pose plus de réelles difficultés. En revanche, l'identification de signaux de régulation ou de « petites » structures régulatrices d'ARN reste un problème pertinent.
2. L'annotation fonctionnelle concerne l'attribution d'une (ou plusieurs) fonction(s) biologique(s) aux signaux détectés au niveau précèdent. L'exemple typique en est l'attribution d'un rôle fonctionnel aux produits protéiques des gènes ou la caractérisation fonctionnelle d'une séquence opératrice. Lorsqu'il n'existe pas de données expérimentales associées à une séquence polypeptidique (le produit d'un gène), la stratégie classique consiste à effectuer un criblage des bases de séquences afin d'identifier des séquences fortement similaires et à attribuer, par analogie, leur(s) fonction(s) à la séquence requête. Les résultats d'une telle stratégie sont des hypothèses de travail qu'il convient de valider expérimentalement. Réalisée automatiquement, cette stratégie d'assignation de fonctions présente de nombreuses limites. Par exemple, il est nécessaire d'évaluer au cas par cas la pertinence de la similarité entre les séquences comparées. D'autre part, cette stratégie est totalement dépendante de la qualité des données présentes dans les bases de séquences publiques utilisées lors du criblage (problème de propagation des erreurs). Enfin, les relations entre les entités manipulées ne sont pas exploitées. Ainsi, on n'exploite encore que trop peu ou pas systématiquement le fait que des enzymes (protéines ayant la fonction de catalyser des transformations chimiques) intervenant dans une même voie métabolique (ensemble de réactions chimiques couplées) tendent à être groupés en opérons (groupe de gènes co-transcrits et donc co-localisés sur le chromosome).
3. L'annotation relationnelle concerne l'identification des relations existant entre les objets caractérisés (individuellement) aux deux niveaux précédents. Ces relations sont de natures diverses. Il peut s'agir par exemple de leur implication dans un processus cellulaire commun (participation à une même voie métabolique, à une même voie de transport), ou d'une interaction physique (interaction protéine-protéine). Les informations qui doivent être manipulées à ce niveau d'annotation - opérons, régulons, graphes représentant des chemins réactionnels ou des assemblages moléculaires - sont plus complexes que les seules données de séquences et réclament donc un traitement particulier. Les objets manipulés et les relations qu'ils entretiennent présentent généralement un plus haut degré d'abstraction et de structuration (par exemple, un graphe décrivant un réseau métabolique). Il se pose alors deux problèmes majeurs : d'une part, le problème de leur représentation formelle, c'est-à-dire leur modélisation, et d'autre part le problème de leur instanciation.

Concernant l'aspect modélisation, force est de constater que si plusieurs initiatives ont déjà vu le jour avec l'objectif de représenter ces informations nouvelles - EcoCyc ou KEGG (http://www.genome.ad.jp/kegg/) pour les données métaboliques, RegulonDB pour les données d'opérons - ces efforts ne sont pour l'instant que

peu ou pas concertés, au point qu'il est pratiquement impossible de dépasser le stade du simple « pointeur » lorsqu'on désire lier entre elles les différentes sources d'information. Par-delà les aspects purement techniques (liés aux choix technologiques opérés par les différents groupes de recherche), un problème de fond est que les modèles employés (lorsqu'ils existent) ne sont pas toujours explicites ou compatibles entre eux ; il ne suffit pas d'appeler un objet « gène » ou « enzyme » ou « opéron » pour qu'il représente la même chose dans plusieurs bases de données.

2.3.3.2. Approches méthodologiques

La démarche qui reste la plus utilisée utilise le fait que les processus évolutifs sont différents en fonction de la nature de la séquence. La caractérisation quantitative et qualitative de la similarité entre deux séquences homologues donne ainsi de bons arguments quant à la nature de l'information génétique que portent ces séquences. Les méthodes permettant de tirer parti de cette situation sont très variées. Dans la partie 6.1 pourront être trouvés des exemples très fortement algorithmiques ou, au contraire, utilisant une modélisation complexe mais moins facilement mathématisable de la réalité biologique. Des modélisations probabilistes, très en vue dans ce domaine (processus de type markovien par exemple), commencent à se développer au sein du projet, en particulier au travers de collaborations avec le projet IS2 (Rhône-Alpes).

Les approches fonctionnelles et relationnelles donnent encore plus de place à la modélisation, y compris de connaissances, dont l'utilisation demande une expertise « manuelle ». Helix s'attache particulièrement dans ce cadre au monde bactérien avec le développement du projet GEB (GenoExpertBacteria, *ex* projet Panoramix) qui se concentre actuellement autour des trois problématiques suivantes :

- Modélisation des éléments génétiques (gènes, signaux, opérons, ...) ;
- Modélisation des protéines (modélisation des modifications post-traductionnelles et des assemblages moléculaires) ;
- Modélisation du métabolisme intermédiaire (modélisation statique des acteurs moléculaires intervenant dans ce processus cellulaire).

Le niveau relationnel met également en oeuvre des aspects plus algorithmiques (en particulier sur les graphes représentant les réseaux métaboliques) dont on trouvera des exemples dans la partie 6.1.

Enfin, Helix s'intéresse aussi, dans le cadre du projet européen ORIEL, à la question de la normalisation de l'échange de ces informations, par la mise en place de spécifications et d'implémentation de schémas XML.

2.3.4. Modélisation dynamique des réseaux de régulation génique

Participants : Grégory Batt, Céline Hernandez, Hidde de Jong [Correspondant], Michel Page.

La plupart des propriétés importantes d'un organisme vivant émergent des interactions entre ses gènes, ses protéines, ses molécules messagères et d'autres constituants. Il s'ensuit que la compréhension du fonctionnement d'un organisme passe par l'élucidation des réseaux d'interactions impliqués dans la régulation génique, le métabolisme, la transduction des signaux et d'autres processus cellulaires et inter-cellulaires.

L'étude des réseaux de régulation génique a été fortement stimulée par l'introduction récente des technologies génomiques permettant, entre autres, de mesurer simultanément le niveau d'expression de tous les gènes d'un organisme. Outre ces nouveaux outils expérimentaux, des méthodes formelles pour la modélisation et la simulation des systèmes de régulation génique sont indispensables. La plupart des réseaux intéressants implique un grand nombre de gènes connectés par des boucles de rétroaction positive et négative, si bien qu'une compréhension intuitive de la dynamique de ces systèmes est difficile à obtenir. Des méthodes formelles de modélisation et de simulation, assistées par des outils informatiques, peuvent contribuer à l'élucidation d'un réseau d'interactions.

Afin de répondre aux besoins des biologistes énoncés ci-dessus, nous développons des méthodes pour la modélisation et la simulation de réseaux de régulation génique, des outils informatiques basés sur ces méthodes, et des applications en collaboration avec des biologistes.

2.3.5. Protéomique : aide à l'acquisition et à la modélisation des données

Participants : Erwan Reguer, Alain Viari [Correspondant].

On désigne usuellement par protéome l'ensemble des protéines potentiellement exprimées dans un organisme ou exprimées dans des conditions physiologiques données. L'ambition des projets de protéomique repose à la fois sur le très grand nombre de protéines à analyser et sur la capacité à identifier des protéines peu abondantes dans la cellule. L'objectif ultime de ces études est de fournir des informations sur la réponse du protéome à une molécule, un stress, voire à la destruction d'un ou plusieurs gènes, en tentant d'appréhender cette réponse dans sa globalité (par l'analyse de toutes les protéines exprimées) et non plus de façon fragmentaire.

Le premier volet de nos travaux dans ce thème concerne la modélisation du protéome avec un accent particulier sur les assemblages moléculaires des enzymes. La difficulté est ici de donner une représentation informatique explicite de situations biologiques parfois très complexes et souvent décrites de manière ambiguë dans la littérature elle-même. Ce travail de modélisation est effectué conjointement avec le projet GEB et avec le projet HAMAP coordonné par le SIB à Genève (*cf.* la partie 8.3) visant la réannotation de l'ensemble des protéomes bactériens (une centaine à ce jour).

Parallèlement, nous nous intéressons, en collaboration avec le CEA Grenoble et dans le cadre de la plate-forme nationale de protéomique, aux aspects liés à l'obtention des données expérimentales et, plus particulièrement, à la mise au point d'algorithmes permettant la localisation rapide de fragments peptidiques, obtenus par spectrométrie de masse en tandem, sur des chromosomes complets. L'objectif de l'ensemble de ces travaux est de tenter de réconcilier deux aspects complémentaires du fonctionnement cellulaire : génome et protéome, en croisant des données d'expression issues d'expériences de protéomique et les données de séquences chromosomiques complètes.

2.3.6. Extraction d'informations à partir de textes

Participants : Jean Dina, Violaine Pillet, François Rechenmann [Correspondant].

Si les bases de données se sont considérablement développées et diversifiées ces dernières années, un volume considérable d'informations n'est encore disponible que sous la forme de textes en langage naturel, en particulier d'articles de revues spécialisées. Malgré les progrès appréciables en matière d'analyse et de compréhension d'énoncés en langage naturel, l'extraction de données et de connaissances à partir de textes reste une tâche difficile à automatiser. Dans le domaine de la biologie, plusieurs équipes se sont ainsi concentrées sur le problème d'extraire des données à partir de textes biologiques. À titre d'exemple, plusieurs projets ont pour objectif d'identifier, dans les textes, des données sur les interactions protéine-protéine ou les interactions gène-protéine. D'autres tentent d'extraire des relations spécifiques entre les entités moléculaires telles que la localisation cellulaire des protéines ou encore les interactions entre gènes ou protéines et médicaments. Quelle que soit la problématique abordée, la première tâche de tous les systèmes construits est l'identification des noms de ces gènes ou de ces protéines dans les textes des articles analysés.

Helix participe ainsi au projet BioMiRe, qui vise le développement d'un outil de reconnaissance des noms d'entités biologiques et son expérimentation en vraie grandeur sur des corpus de textes concernant plusieurs espèces. Ces entités concernant les noms de gènes, de protéines, d'ARNs et d'espèces.

L'identification des noms d'entités biologiques n'est pas une tâche facile et ceci pour plusieurs raisons. Tout d'abord parce que la manière dont sont décrits les gènes et les protéines dans la littérature n'est pas homogène. En effet, il existe plusieurs dénominations possibles pour désigner une seule et même entité. De plus, même si de nombreux efforts sont faits pour mettre en place une nomenclature des noms de gènes et de protéines, et même si la plupart des dénominations sont répertoriées dans différentes bases de données, les auteurs des articles ont une fâcheuse tendance à utiliser, pour nommer les entités biologiques, des variantes lexicales ou des abréviations qui leur sont souvent propres. Une autre difficulté est celle de l'ambiguïté. En effet, de nombreux noms de gènes ou de protéines sont identiques à des termes du langage courant ou bien identiques à la terminologie biologique employée dans les textes. L'étude du contexte est donc indispensable pour déterminer à quelle catégorie (nom d'entité biologique ou mot courant) appartient le terme. Une dernière difficulté est celle de l'ambiguïté due à l'homonymie des noms de gènes et de protéines. Un nom détecté dans un texte peut donc désigner plusieurs gènes ou protéines différents, que ce soit au sein de la même espèce ou entre espèces différentes. Il est donc important de pouvoir détecter dans un texte, non seulement les noms de gènes et de protéines, mais aussi les noms d'espèces auxquelles appartiennent ces entités.

Le projet BioMiRe est mené en collaboration avec le Centre de Recherche Européen de Xerox (XRCE) à Meylan et deux équipes de l'INRA, à Versailles et à Gand (Belgique). Il a bénéficié du soutien du Ministère de la Recherche, Direction de la Technologie, dans le cadre du programme GenHomme.

2.3.7. Environnement didactique en bioinformatique

Participants : Philippe Genoud [Correspondant], Stéphanie Merriene, Anne Morgat, Alain Viari, François Rechenmann, Danielle Ziébelin.

Le nombre de méthodes et outils d'analyse de données en biologie moléculaire, déjà important, ne cesse de croître. Malheureusement, il est bien souvent difficile d'appréhender ces méthodes et de maîtriser les paramètres qui les accompagnent. Une bonne compréhension des algorithmes que ces méthodes mettent en oeuvre faciliterait leur emploi ainsi que l'obtention de résultats plus pertinents.

Nous avons donc entrepris le développement d'un environnement didactique pour expliquer les algorithmes de bioinformatique, indépendamment les uns des autres ou au travers de stratégies y faisant appel. Cet environnement est un des éléments du projet « École de l'ADN » proposé par le Centre de Culture Scientifique Technique et Industrielle (CCSTI) de Grenoble. Il est destiné à un public assez large. S'il s'adresse principalement aux étudiants de terminale et premières années d'études supérieures en biologie, il pourra aussi être utilisé comme complément dans des filières d'informatique proposant des options en bioinformatique, voire comme outil d'auto-formation.

2.4. Relations internationales et industrielles

Le groupe SwissProt dirigé par A. Bairoch à l'Institut Suisse de BioInformatique (SIB) à Genève est « équipe associée » au projet Helix. Les deux groupes entretiennent des liens forts sur le thème de l'annotation/réannotation des protéomes bactériens (projet HAMAP et projet HERBS) ainsi que sur l'extraction d'information à partir de textes (en vue d'assister les annotateurs de la banque de protéines SwissProt dans leur travail).

Le projet Helix participe au projet européen ORIEL (IST), qui constitue le volet « recherche » du projet eBioSci, ainsi qu'au réseau ESF (*European Science Foundation*) intitulé « Experimental and in silico Analysis of Biomolecular Interactions ».

Le projet Helix est en contact avec les différentes équipes de bioinformatique françaises, en particulier au sein des différentes génopoles.

Au premier rang des relations industrielles figurent les sociétés GENOME express et Hybrigenics, partenaires du consortium Genostar.

La société GENOME express est également partenaire, avec le CEA, du projet de protéomique.

Le Centre Européen de Recherche de Xerox (XRCE) à Meylan intervient de façon déterminante sur le thème de l'extraction d'informations à partir de textes.

Le projet Helix bénéficie d'un *grant* du Welcome Trust, impliquant des équipes du King's College de Londres, l'université de Marne La Vallée et l'INRIA Rhône-Alpes.

3. Fondements scientifiques

Plus encore que dans d'autres domaines scientifiques, l'informatique est appelée à jouer en biologie moléculaire deux rôles complémentaires et indissociables : d'une part offrir des modèles pour représenter les nombreuses classes d'entités impliquées ainsi que les relations qu'elles entretiennent, d'autre part proposer des méthodes pour identifier et caractériser ces entités et leurs relations à partir des données expérimentales.

Expliciter et formaliser est une nécessité dans un domaine qui se distingue par une grande diversité, tant des problématiques scientifiques que des entités impliquées. Un terme aussi central que « gène » possède ainsi des acceptions et des interprétations très différentes selon la problématique adoptée et donc le point de vue retenu. L'interopérabilité des bases de données biologiques et celle des programmes d'analyse, requise pour la confrontation et l'intégration de toutes les données et connaissances impliquées, passe par la représentation

explicite et formelle de ces entités et de leurs relations. La complexité du domaine conduit à la conception et au développement de modèles adaptés, en particulier en ce qui concerne la représentation des relations, tant statiques que dynamiques, entre entités. La modélisation des données et des connaissances est ainsi au coeur de la problématique du projet Helix.

Mais il convient également de concevoir et de développer les méthodes d'analyse adaptées aux diverses classes de données produites, au premier rang desquelles figurent bien entendu les séquences génomiques et protéiques. La disponibilité de plusieurs dizaines de génomes complets modifie quelque peu les démarches d'analyse, qui peuvent d'une part viser l'exhaustivité (identifier par exemple tous les gènes d'un organisme), d'autre part confronter et recouper les connaissances portant sur plusieurs organismes simultanément afin de les compléter, en tenant compte de la distance qui sépare ces organismes dans les arbres phylogénétiques.

Mais les séquenceurs ne constituent plus la seule source de données biologiques au niveau moléculaire. L'émergence des différents dispositifs d'étude des transcriptomes, tels que les « puces à ADN », et des protéomes, tels que l'électrophorèse bidimensionnelles et les diverses techniques de spectrométrie, est à l'origine d'un flux de données nouvelles, pour lesquelles il est nécessaire de concevoir des méthodes et des démarches d'analyse à la fois pertinentes et efficaces.

Enfin, l'étude des relations entre les entités biologiques que sont les gènes et leurs produits conduit à la reconstruction des réseaux de régulation de l'expression des gènes et des réseaux métaboliques.

Les outils d'analyse des données biologiques que cherche à mettre en oeuvre le projet Helix font ainsi appel à l'algorithmique des chaînes de caractères, des arbres et des graphes, dans un contexte dominé par les approches probabilistes et statistiques.

4. Domaines d'application

Par essence même du projet Helix, ses travaux de recherche sont tous motivés par des problématiques issues des sciences du Vivant, et plus particulièrement de la génomique.

L'information nécessaire au développement et au maintien de tout organisme vivant est contenue dans son génome, matérialisé au sein de chacune des cellules par une ou plusieurs macromolécules d'ADN, enchaînements d'acides nucléiques de quatre types différents symbolisés par les lettres, A, C, G et T. Le contenu informationnel d'un génome peut ainsi être représenté comme un texte, écrit dans l'alphabet de ces quatre lettres.

Plus d'une centaine de génomes bactériens ont fait l'objet d'un séquençage exhaustif ; leur « texte », composé de plusieurs millions de « lettres », est donc connu. D'autres génomes plus longs sont également disponibles, tels que celui de la levure (*S. cerevisae*, 14 millions de lettres, premier organisme eucaryote complètement séquencé) ou celui du nématode (*C. elegans*, 100 millions, premier organisme pluricellulaire complètement séquencé) ; celui de la drosophile *D. melanogaster* (160 millions) a précédé de quelques mois celui de l'Homme (plus de trois milliards de lettres) et, plus récemment, de la souris.

Mais disposer de ces séquences ne suffit pas, encore faut-il les interpréter, les annoter. Il s'agit d'abord d'identifier les gènes, c'est-à-dire les zones qui codent pour les protéines, puis de comprendre la fonction de ces protéines, mais aussi les réseaux d'interactions qui contrôlent l'expression des gènes suivant les besoins de l'organisme. Au-delà encore, il est fondamental de comprendre comment ces différentes structures se sont mises en place ou ont été modifiées au cours de l'évolution. Dans le domaine de la biologie, il est en effet impossible d'ignorer cette composante historique, car c'est elle qui façonne les objets que l'on manipule. L'étude des processus évolutifs, à un niveau global (phylogénie) ou mécanistique (modélisation de processus mutationnels ou sélectifs) est donc un passage obligé.

Dans l'ensemble de ces travaux, il est fondamental de ne pas limiter l'information aux seules données de la génomique, c'est-à-dire les séquences. D'autres classes de données doivent être également mises en oeuvre et recoupées avec les résultats d'analyse de ces séquences. C'est en particulier le cas des données expérimentales obtenues à l'aide de « bio-puces » (*DNA chips*), de gels 2D ou de la spectrométrie de masse (*proteomics*), ainsi que des données de la littérature concernant notamment les réseaux de régulation ou les voies métaboliques.

5. Logiciels

5.1. Genetic Network Analyzer (GNA)

Participants : Grégory Batt, Céline Hernandez, Hidde de Jong [Correspondant], Michel Page.

Une méthode de simulation qualitative des réseaux de régulation génique a été implémentée en Java, dans un outil baptisé Genetic Network Analyzer (GNA). Les entrées de GNA se composent du modèle mathématique d'un réseau de régulation génique et d'un état qualitatif initial. Les résultats de simulation sont produits sous forme d'un graphe des états qualitatifs atteignables à partir de l'état initial et des transitions possibles entre ces états. Afin de faciliter l'utilisation du simulateur, une interface graphique permet de visualiser des réseaux d'interactions et d'analyser les résultats de simulation. La version 5.0 de GNA a été déposée auprès de l'APP. Elle est disponible à travers le Web (http://www-helix.inrialpes.fr/gna).

5.2. Smile

Participants : Laurent Marsan, Marie-France Sagot [Correspondante], Sébastien Wirth.

Smile implémente un algorithme d'inférence de motifs à partir d'un ensemble de séquences biologiques (nucléiques ou de protéines). Il permet d'identifier des motifs, écrits sur l'alphabet des séquences ou sur un alphabet physico-chimique (c'est-à-dire constitué de sous-ensembles de l'alphabet des séquences, y compris le symbole *don't care*), satisfaisant un certain nombre de contraintes dont le nombre minimum de séquences où ce motif doit être présent (« quorum ») et le taux de substitution autorisé entre un motif et chacune de ses occurrences. Les motifs sont dit structurés ; cela veut dire qu'ils sont composés d'un nombre quelconque de parties (« boîtes ») séparées par des intervalles de distance dont, soit les bornes, soit l'étendue est spécifiée par l'utilisateur. Le nombre de « boîtes » composant un motif structuré est également spécifié par l'utilisateur, de même que le quorum et le taux maximum de substitutions. Le contenu des boîtes est, bien sûr, inconnu au départ. Le but de Smile est de déterminer tous ceux pour lesquels les motifs continuent de vérifier les contraintes introduites. En 2002, l'algorithme a été étendu afin de pouvoir traiter des « méta-différences », c'est-à-dire des différences (délétions essentiellement) portant, non sur le contenu des boîtes cette fois, mais sur les boîtes elles-mêmes. Il est ainsi possible désormais de traiter un nombre variable de boîtes.

5.3. Utopia

Participants : Philippe Blayo, Pierre Peterlongo, Marie-France Sagot [Correspondante].

Utopia est un logiciel de détection de gènes. Il a été initialement élaboré afin de déterminer la structure de gènes orphelins, homologues entre eux, en utilisant une méthode par *pure homology*. L'algorithme sous-jacent réalise un alignement doublement épissé de deux séquences (l'épissage a lieu sur les deux séquences) en se basant sur un modèle de gène générique : un gène débute par un codon *start* (par exemple « ATG »), se termine par un codon stop et chacun de ses exons est bordé à gauche et à droite respectivement par « AG » et « GT ». Courant 2002, l'algorithme a été étendu afin de pouvoir inférer plus d'un gène à la fois. Les gènes à découvrir doivent apparaître dans le même ordre sur les deux séquences pour que l'algorithme puisse les identifier. Une version est en cours permettant de traiter des gènes incomplets. La version courante d'Utopia, plus des scripts de post-traitement, est libre d'accès aux académiques.

5.4. Druid

Participants : Marie-France Sagot [Correspondante], Marina Zelwer.

Druid prend en entrée un alignement de séquences nucléiques d'espèces proches ou de souches différentes d'une même espèce et détecte la présence éventuelle de points de recombinaison le long de l'alignement ainsi que leur localisation. Il a été considérablement modifié dans le courant 2002 afin de mieux traiter le problème des tests multiples. Par ailleurs, il n'utilise plus maintenant aucune routine externe et fonctionne de façon autonome.

5.5. FactorTree

Participants : Julien Allali, Marie-France Sagot [Correspondante].

FactorTree construit, à partir d'un texte quelconque, un index appelé k-*depth factor tree*. Il s'agit d'un arbre de tous les facteurs de longueur au plus k d'un texte dont la construction s'inspire de celle de l'arbre des suffixes de Esko Ukkonen. Il s'agit d'une construction *on-line*. L'arbre des facteurs de longueur au plus k permet d'économiser en espace lorsque son utilisation ultérieure ne porte que sur des motifs (connus ou non selon qu'il s'agit d'une recherche ou d'une inférence), dont la longueur maximale est fixée. L'économie varie selon le type de texte considéré. Pour des textes en langage naturel, il peut atteindre 70-80% pour k entre 10 et 20. Pour des séquences biologiques, le gain est en général plus petit, mais peut atteindre 30-40%. Il est désormais possible d'indexer des chromosomes entiers de génomes eucaryotes (les tests ont porté sur des séquences ayant jusqu'à entre 30 et 40 millions de paires de bases).

5.6. Satellites

Participant : Marie-France Sagot [Correspondante].

Satellites est un programme de détection de répétitions en tandem (c'est-à-dire apparaissant de manière contiguë) dans une séquence biologique (ADN ou, en version prototypale, de protéine). Les répétitions sont approximatives : un nombre maximum d'erreurs (substitutions, insertions et délétions) est ainsi autorisé. Ce nombre est spécifié par l'utilisateur. Satellites est le fruit d'une collaboration avec Gene Myers.

5.7. EMKov

Participant : Alain Viari [Correspondant].

EMKov est un logiciel de recherche de gènes bactériens. Il combine l'utilisation de chaînes de Markov et l'algorithme EM (*Expectation Maximisation*) ; il permet ainsi de s'affranchir de l'étape d'apprentissage préalable sur un jeu de gènes connus. Les tests d'EMKov sur les données de génomes annotés ont montré qu'il présente une excellente précision (de 80 à 95%). Par ailleurs, par construction, EMKov propose un partitionnement des gènes ainsi trouvés suivant leurs propriétés markoviennes, permettant ainsi la détection de gènes « atypiques ».

5.8. MTDP

Participants : Matthieu Vignes, Alain Viari [Correspondant].

MTDP est un logiciel développé en collaboration avec le projet IS2 et qui implémente un modèle de Markov « parcimonieux » (*Mixture of Transition Distributions*), périodique et généralisé (issu des travaux de Berchtold et Raftery). Au contraire des chaînes de Markov « complètes », le nombre de paramètres du modèle croît linéairement avec l'ordre. Il est donc destiné à la caractérisation de dépendances à longue distance sur des chaînes nucléiques ou protéiques dans les conditions où les chaînes de Markov traditionnelles ne sont pas correctement paramétrables (grands ordres markoviens ou grands alphabets). Son emploi est actuellement testé dans le cadre de la recherche de gènes eucaryotes.

5.9. DomainProteix

Participants : Erwan Reguer, Alain Viari [Correspondant].

DomainProteix est un logiciel de segmentation de protéines en domaines structuraux, implémentant un algorithme proposé initialement par Zu et Gabow et basé sur une représentation sous la forme d'un graphe de flot des structures protéiques.

5.10. TagMap

Participants : Erwan Reguer, Alain Viari [Correspondant].

TagMap est un logiciel développé dans le cadre du projet PepMap, mené en collaboration avec le CEA Grenoble (J. Garin, M. Ferro) et la société GENOME express (T. Vermat) et soutenu par le Ministère de la Recherche, Direction de la Technologie, dans le cadre du programme GenHomme. Il permet de localiser rapidement, sur des génomes eucaryotes complets (*A. thaliana*, *H. sapiens*), des étiquettes peptidiques produites par spectrométrie de masse (analyse MS/MS) à partir de la digestion trypsique de protéines issues de ces espèces. La version actuelle permet de localiser, en quelques minutes, plusieurs centaines d'étiquettes simultanément sur des chromosomes de plusieurs millions de bases. Elle est destiné à s'insérer à terme dans une chaîne de traitement « à haut-débit » des données de protéomique.

5.11. Herbs

Participants : Corinne Lachaize [Correspondante], Anne Morgat, Alain Viari.

Herbs (HAMAP *Expert Rule Based System*) est un système d'aide à la réannotation de protéomes bactériens complets, développé en collaboration avec le SIB (Genève) dans le cadre du projet HAMAP. Herbs est un système à base de connaissances implémenté sur le système Jess (*Java Expert System Shell*) du Sandia National Laboratories. Il dispose d'une base de règles portant sur les propriétés physiologiques et métaboliques de l'organisme étudié (« si l'organisme est une cyanobactérie alors le complexe du photosystème I doit être présent »), ainsi que d'une base de faits constituée des annotations fonctionnelles associées aux gènes de l'organisme étudié (« ce gène est annoté comme étant la sous-unité *psaM* du photosystème I »). Ces annotations sont produites par la chaîne de réannotation automatique d'HAMAP. Herbs permet donc de valider l'ensemble des annotations fonctionnelles en détectant les gènes « manquants » (le gène devrait être présent, mais il n'a pas été détecté) ou les gènes « inattendus » (le gène ne devrait pas être présent et il a pourtant été détecté). Herbs est actuellement testé sur les voies de biosynthèse des acides aminés et des acides nucléiques.

5.12. BOX

Participants : Antoine Brun, Anne Morgat, Alain Viari [Correspondant].

BOX (pour Bio Oriel XML-schema) n'est pas à proprement parler un logiciel, mais une suite de spécifications et de schémas XML permettant l'échange normalisé de données de génomique et post-génomique au format XML. La version actuelle de BOX (1.1) permet la représentation de données de génomique (annotations syntaxiques) et de métabolisme (annotations fonctionnelles). Elle est en cours d'évaluation par les partenaires du projet ORIEL.

5.13. GenoExpertBacteria (GEB)

Participants : Frédéric Boyer, Christophe Bruley, Stéphane Bruley, Anne Morgat [Correspondant], Alain Viari, Erik Wessel.

GenoExpertBacteria est une plateforme d'expertise des données génomiques et métaboliques d'organismes bactériens. Il intègre une base de connaissances (issue de nos travaux antérieurs sur Panoramix) et une interface graphique d'analyse et d'exploration permettant le travail d'expertise. Bien que pouvant fonctionner de manière autonome, GEB est néanmoins prévu pour être intégré dans la plateforme Genostar. Dans ce cadre, il dispose alors des fonctionnalités d'interopérabilité fournies par Genostar et, en particulier, des services proposés par les autres modules (GenoAnnot , GenoLink et GenoBool).

5.14. Environnement Didactique en Bioinformatique (EDB)

Participants : Philippe Genoud [Correspondant], Stéphanie Merriene, Anne Morgat, Alain Viari, François Rechenmann, Danielle Ziébelin.

L'Environnement Didactique en Bioinformatique (EDB) est développé dans le langage Java. Il a été conçu de manière générique, sous la forme d'un *framework* permettant le pilotage d'algorithmes et des interfaces explicatives associées à partir d'un contenu pédagogique au format HTML. La conception de la partie

pédagogique peut être ainsi complètement dissociée de la conception de la partie algorithmique et peut être réalisée au moyen d'outils d'édition standards, sans recours à un langage de programmation. EDB est donc une instanciation de ce *framework*, dédiée à la biologie moléculaire. Plusieurs algorithmes (avec leurs interfaces explicatives) pour la recherche de zones codantes dans des séquences génomiques ont d'ores et déjà été développés et intégrés : recherche de motifs dans une séquence, y compris exprimés à l'aide d'expressions régulières, analyse du biais de codage par le test du Chi2, *dotplots* et alignement de deux séquences par programmation dynamique. Une stratégie combinant ces différents algorithmes, ainsi qu'une interface cartographique pour la visualisation des résultats, ont été développées.

5.15. PBIL

Participants : Laurent Duret, Manolo Gouy, Guy Perrière [corréspondant].

PBIL est un site web (http://pbil.univ-lyon1.fr/) pour l'analyse bioinformatique des séquences biologiques particulièrement selon la perspective de l'approche comparative développé en collaboration entre le projet HELIX à Lyon et l'équipe de conformation des protéines, de l'UMR CNRS 5086.

5.16. HOBACGEN, HOVERGEN et ORTHOGEN

Participants : Laurent Duret, Manolo Gouy, Guy Perrière [corréspondant], Jean-François Dufayard, Dominique Mouchiroud.

HOBACGEN et HOVERGEN sont des bases de données de familles de séquences homologues dédiées aux bactéries et aux vertébrés. Les données sont constituées de familles homologues (obtenues par comparaison de toutes les séquences avec elles-mêmes, délimitation des familles, alignement multiple de leurs membres, calcul de leur phylogénie). L'évolution de ces logiciels par l'intégration d'interfaces écrites en Java et le développement de procédures automatiques de mise à jour ont permis une utilisation plus large du logiciel (actuellement quatre bases de données extérieures à Helix utilisent ce logiciel). HOBACGEN et HOVERGEN sont par ailleurs installées sur une vingtaine de sites dans le monde.

Cette comparaison massive (une des trois réalisées de manière systématique en France) met en jeu une collaboration avec le centre de calcul de l'IN2P3. Elle conduit actuellement à plus de 6000 familles de plus de 4 gènes (et donc à autant d'arbres non-triviaux). J.-F. Dufayard a développé des outils de gestion et de requête de cet ensemble d'arbres. Ce travail est essentiel au maintien de la seule base d'orthologues homme/souris utilisant la phylogénie pour analyser les relations d'homologie entre gènes. Les bases sont consultables en mode client-serveur (http://pbil.univ-lyon1.fr/databases/hobacgen.html).

5.17. ORILOC

Participant : Jean Lobry [corréspondant].

ORILOC est un programme d'identification de l'origine et du site de terminaison de la réplication dans les génomes bactériens séquencés (http://pbil.univ-lyon1.fr/software/oriloc.html).

5.18. JADIS

Participant : Dominique Mouchiroud [corréspondant].

JADIS est une application Java de calcul de distances entre séquences, développé en collaboration avec Isabelle Gonçalves à l'Atelier de BioInformatique (ABI, université Paris 6) (http://pbil.univ-lyon1.fr/software/jadis.html).

5.19. ACNUC

Participant : Manolo Gouy [corréspondant].

ACNUC est un SGBD dédié à la gestion des séquences génomiques. Son développement a débuté en 1980 et il sert depuis à la fois d'outil d'interrogation et de couche basse pour l'ensemble des logiciels développés

à Lyon. Il reste le seul logiciel permettant l'interrogation, transparente pour l'utilisateur, des sous-séquences des séquences présentent dans les banques. Mise à jour quotidiennement, cette base est intensivement utilisée à travers l'interface du PBIL.

5.20. SeaView

Participant : Manolo Gouy [corréspondant].

SeaView est un éditeur d'alignements multiples de séquences nucléotidiques et protéiques. Il associe de manière cohérente matrices de points et édition « manuelle » des alignements.

5.21. GeM

Participants : Gisèle Bronner, Bruno Spataro [corréspondant].

GeM est une base de connaissances (développée sous AROM) contenant l'ensemble des données disponibles sur l'Homme et la souris. La base GeM résulte du travail réalisé dans le cadre des thèses de G. Bronner (maintenant en séjour postdoctoral à Upsala) et B. Spataro (actuellement IR au CNRS). GeM permet de modéliser et d'analyser les données de cartographies génomiques sur les vertébrés.

6. Résultats nouveaux

6.1. Évolution des espèces et des familles de gènes

Participants : Vincent Daubin, Jean-François Dufayard, Laurent Duret, Christian Gautier, Manolo Gouy [Correspondant], Guy Perrière, Gwenael Piganeau, Nadia Pisanti, Marie-France Sagot [Correspondant], Fabienne Thornarat, Marina Zelwer.

6.1.1. Deux résultats en phylogénie

La reconstitution de l'arbre du Vivant à partir des séquences soulève de nombreux problèmes méthodologiques principalement liés à la non-homogénéité des processus. Une analyse prenant en compte les variations importantes de vitesses le long des branches a permis à F. Thomarat et M. Gouy de préciser la position taxonomique des microsporidies. Ces organismes sont des eucaryotes parasites intracellulaires de nombreux hôtes eucaryotes, dont ils peuvent être des agents pathogènes. Les microsporidies font partie des eucaryotes sans mitochondries ce qui a suggéré leur divergence très précoce (avant l'endosymbiose mitochondriale). En fait, les auteurs ont montré, en utilisant la séquence complète d'une microsporidie, une divergence beaucoup plus récente et un positionnement de ces organismes parmi les champignons. Une démarche symétrique, impliquant un très grand nombre de familles de gènes relatives à un grand nombre d'espèces, a conduit V. Daubin et G. Perrière à proposer une nouvelle phylogénie des bactéries contestant une divergence ancienne des thermophiles.

6.1.2. Quelques résultats informatiques sur les arbres phylogénétiques

Dans le cadre des logiciels HOVERGEN et HOBACGEN, J.-F. Dufayard a développé un algorithme générique de recherche de motifs dans une banque de données d'arbres phylogénétiques appelé TQuest. Bien que la recherche de motifs non-ordonnés dans un arbre soit un problème non polynomial, cette recherche peut être réalisée dans des temps raisonnables (de l'ordre de quelques dizaines de secondes pour parcourir 6000 arbres de la base de données HOVERGEN). Un éditeur graphique permet de décrire le motif (sous-arbre) recherché et de spécifier les contraintes sur les noeuds et branches de ce motif (présence ou non de duplications, présence ou non d'un groupe taxonomique donné, *etc.*). Il est possible directement depuis cette interface de sélectionner automatiquement des gènes, non seulement sur des critères d'orthologie, mais également sur n'importe quel autre critère de topologie d'arbre. Cet outil pourra notamment être utilisé pour repérer des topologies aberrantes, signes de possibles transferts horizontaux.

Cette approche comporte cependant une approximation, à savoir qu'elle postule que toutes les régions d'un même gène ont la même histoire. En fait, des remaniements peuvent avoir lieu au sein d'une famille, ce qui

implique que certains segments d'une même séquence peuvent avoir des histoires évolutives différentes. Ainsi, segmenter un gène en fragments suivant l'histoire évolutive est un problème biologiquement important mais particulièrement difficile au niveau algorithmique. Un compromis doit en effet être trouvé entre des méthodes de reconstruction d'arbres plus ou moins sophistiquées, mais aussi plus ou moins rapides. Ce travail est mené dans le cadre de sa thèse par M. Zelwer sous la direction de M.-F. Sagot et de M. Crochemore.

6.1.3. Un exemple de résultat sur les processus évolutifs

L'évolution repose sur deux processus, le premier crée des mutations (des erreurs) et le second permet la fixation de certaines d'entre elles. La modélisation mathématique de ces processus est complexe, mais repose sur des travaux déjà anciens (tels que ceux de Kimura et Malécot). La fixation correspond au fait que le processus atteint un état absorbant. Cependant, le temps d'absorption et le choix entre les états absorbants dépendent de la sélection. La situation est cependant encore un peu plus intéressante car l'efficacité de cette sélection est elle-même sous la dépendante à la fois de la taille de la population et de l'intensité d'un mécanisme génétique essentiel : la recombinaison (la recombinaison est le mécanisme qui mélange les informations venant du père et de la mère). Plus la recombinaison est forte, plus la sélection est efficace.

Le résultat nouveau auquel Helix a participé est la mise en évidence d'un mécanisme de fixation ne faisant pas intervenir la sélection, mais qui mathématiquement se comporte « presque » comme de la sélection. Ce mécanisme est le biais de conversion génique qui conduit à convertir un allèle par celui qui lui fait face dans un individu hétérozygote et ceci à l'occasion d'une recombinaison. Ce mécanisme va également dépendre de la taille de la population et de l'intensité de la recombinaison, posant un nouveau défi dans le cadre de la démonstration de l'existence d'une sélection (et donc d'un rôle fonctionnel).

6.1.4. Segmentation de séquences et phylogénie

Le problème de la détection et localisation de points de remaniements peut être vu comme celui de la recherche d'un partitionnement en colonnes optimal d'une matrice, représentant un alignement de séquences, de façon à pouvoir expliquer l'évolution de ces séquences de la manière la plus économique possible lorsque les opérations prises en compte sont les mutations ponctuelles.

Dans le cadre du travail de thèse de M. Zelwer ; nous avons élaboré une première approche statistique du problème, débouchant sur le programme Druid. En 2002 cette approche a été largement améliorée par l'introduction, en particulier, d'une correction de Bonferoni à la valeur estimée de la signification statistique. La spécificité de Druid est ainsi améliorée. Une analyse comparative entre Druid et d'autres méthodes de détection et de localisation de points de remaniements a produit des résultats intéressants. À la question plus simple : « Y a-t-il eu remaniement ? » (sans tentative de localisation des remaniements qui ont pu survenir), les résultats de Druid sont équivalents à ceux des meilleures méthodes. En absolu, la spécificité de Druid est d'environ 95% et sa sensibilité d'environ 15% à 100% selon les conditions (en gros, selon la distance d'évolution entre les espèces considérées : plus cette distance est grande, moins les méthodes sont sensibles). À la question maintenant plus complexe : « Où sont les points de remaniements ? », Druid est la seule parmi les meilleures méthodes qui essaie d'y répondre par des moyens automatiques, c'est-à-dire non graphiques. Dans ce cas, Druid localise entre 50% et 80% des points vrais. Par ailleurs, entre 0% et 30% des points inférés ne sont pas localisés près de points vrais. Le travail d'amélioration de Druid doit se poursuivre dans les années à venir, à la fois sur le plan théorique et pratique, notamment au travers de la collaboration avec E.M. Rodrigues (doctorante du département de Mathématiques et Statistiques de l'université de São Paulo au Brésil, co-encadrée par Y. Wakabayashi, professeur dans le département, et M.-F. Sagot).

6.1.5. Distances d'arbres

Les calculs de distance basés sur une différence de topologie qui peuvent être trouvés dans la littérature considèrent un parmi trois types d'opérations : échange de sous-arbres voisins, coupure de sous-arbre et réinsertion à un autre endroit de l'arbre de départ, ou bissection d'arbre et reconnection. Le problème est alors de trouver le plus petit nombre de l'un de ces trois types d'opérations permettant de passer d'un arbre à un autre. Le second type, appelé SPR (pour *Subtree Prune and Regraft*) est particulièrement intéressant pour modéliser des remaniements et c'est celui que nous avons commencé à explorer. Un article datant de

1997 indiquait que la distance SPR est égale à la taille de la forêt de concordance maximale moins un (cette taille est appelée MAF en anglais pour *Maximum Agreement Forest*) entre deux arbres. Une forêt de concordance entre deux arbres est la forêt obtenue de l'un ou l'autre arbre par une suite de coupures et de contractions d'arêtes. Le MAF est la forêt de concordance ayant le plus petit nombre de composantes. L'article donnait en outre un algorithme d'approximation de ratio 3 permettant de calculer le MAF de deux arbres phylogénétiques (enracinés, binaires, étiquetés aux feuilles et non ordonnés). Nous avions montré que l'algorithme fourni par les auteurs avait en réalité un ratio de 4 et nous avions donné un nouvel algorithme dont le ratio d'approximation est effectivement de 3. L'algorithme, appelé MAFALDA, a été implanté en C. En 2002, nous avons montré qu'il existait un autre algorithme d'approximation dont le ratio, de 3, était beaucoup plus simple à prouver. Nos analyses récentes ont montré toutefois que ce deuxième algorithme, bien qu'ayant le même ratio théorique que notre premier, est en pratique beaucoup moins performant. Il atteint, en effet, le ratio de 3 pour des nombres de transferts de sous-arbres (c'est-à-dire de remaniements entre les espèces) plus faibles que le précédent. Ces analyses ont également montré que notre premier algorithme avait un ratio en pratique plus proche de 1.5 pour un nombre raisonnable de transferts (99) et proche de 2 pour un nombre élevé de remaniements (399).

Sur le plan théorique, un chercheur néo-zélandais, M. Steel, avait montré que la distance SPR n'est pas toujours égale à la taille du MAF moins un. Nous avons déjà réussi à montrer que le SPR est, soit strictement égal à la taille du MAF, soit égal à la taille du MAF moins un, et travaillons actuellement à caractériser dans quels cas nous avons l'un ou l'autre.

6.2. Organisation spatiale (le long du génome) de l'information génomique

Participants : Gisèle Bronner, Eric Coissac, Christian Gautier, Laurent Gueguen, Adel Khelifi, Jean Lobry [Correspondant], Anne Morgat, Dominique Mouchiroud, Guy Perrière [Correspondant], Nadia Pisanti, Marie-France Sagot [Correspondante], Bruno Spataro, Alain Viari.

J. Lobry avait déjà montré une dissymétrie des brins chez les procaryotes induisant une structuration spatiale suffisamment forte pour permettre une bonne estimation de l'origine de réplication (*cf.* le logiciel ORILOC). V. Daubin et G.Perrière ont récemment montré une variation du contenu en C+G le long de ces mêmes génomes. Cette variation entraîne une surestimation des transferts horizontaux par toutes les méthodes basées sur une hétérogénéité du code génétique.

6.3. Annotation syntaxique et fonctionnelle des génomes

Participants : Frédéric Boyer, Stéphane Bruley, Laurent Gueguen, Anne Morgat [Correspondante], Guy Perrière, Marie-France Sagot [Correspondante], Alain Viari [Correspondant].

6.3.1. Modélisation des gènes et inférence de motifs

Récemment, le programme Utopia de recherche de gènes eucaryotes a été étendu afin de pouvoir traiter des séquences génomiques contenant plus d'un gène. Si les gènes communs à deux séquences sont co-linéaires, Utopia peut ainsi désormais *a priori* les identifier tous. Nos temps de calcul sont souvent plus importants que ceux des approches traditionnelles, mais notre méthode demeure la seule exacte existant actuellement, et surtout la seule à n'utiliser aucune autre information que la séquence afin d'inférer les gènes. Nos analyses comparatives indiquent que la prédiction peut demeurer bonne même dans ce cas. Cela devrait permettre ainsi à Utopia d'identifier des gènes ayant des caractéristiques inhabituelles par rapport à la moyenne.

6.3.2. Inférence de motifs structurés

Dans le courant 2002, Smile (algorithme d'inférence de motifs structurés, c'est-à-dire composés de diverses parties séparées par des distances variables) a été étendu afin de permettre les « méta-différences ». Les méta-différences sont des différences portant non plus sur les symboles individuels des séquences traitées (et qui pourraient correspondre à des mutations ponctuelles), mais sur les boîtes elles-mêmes. Si l'on cherche à inférer des motifs avec p boîtes, et si le nombre maximum de méta-différences autorisées est d, alors les occurrences

des motifs identifiés par l'algorithme peuvent avoir entre $p - d$ et p boîtes. Les méta-différences modélisent explicitement les suppressions, mais pas les insertions. Une version contrainte de Smile (pour deux boîtes au maximum) est également utilisable à travers le Web sur le site de l'Institut Pasteur grâce à une interface très conviviale développée par C. Letondal.

Smile utilise une structure de données, appelée « arbre des facteurs de profondeur au plus k » (en anglais, *k-depth factor tree*). L'arbre des facteurs a été mis au point en collaboration avec J. Allali (doctorant de l'université de Marne La Vallée, co-encadré par M. Crochemore et M.-F. Sagot). L'idée algorithmique sous-jacente peut être appliquée à pratiquement toutes les méthodes de construction d'arbre des suffixes existantes, y compris des versions très compactes utilisant des systèmes de codage astucieux permettant de diminuer parfois assez considérablement la taille de l'arbre. La technique que nous avons développée avec J. Allali permet d'économiser encore plus en espace. Le gain dépend du degré de structuration du texte. Pour des facteurs de longueur entre 10 et 20, ce gain va de 10% environ sur des textes quasi-aléatoires à 70-80% sur des textes en langue naturelle. Pour ce qui concerne les séquences biologiques, la nouvelle structure permet à l'heure actuelle d'indexer des chromosomes entiers contenant jusqu'à 130 millions de bases.

Enfin, Smile continue d'être extensivement utilisé sur des organismes procaryotes. Actuellement, notre attention se concentre sur *Mycobacterium tuberculosis* et *Mycobacterium leprae* d'une part, et sur *Synechocystis* PCC 6803 et *Escherichia coli* d'autre part. *Mycobacterium tuberculosis* et *Mycobacterium leprae* sont deux génomes très proches en termes évolutifs, mais dont l'un, *Mycobacterium leprae*, semble être en train de « rétrécir » (environ 1500 gènes présents dans *Mycobacterium tuberculosis* ont été « perdus » dans *Mycobacterium leprae* qui, par contre, en a acquis 165 « nouveaux »). Notre approche est dans ce cas comparative. Ce travail a été réalisé en collaboration avec A. Cariou (stagiaire en 5e année de l'École Vétérinaire de Maison-Alfort). Le travail sur *Synechocystis* PCC 6803 et *Escherichia coli* procède par analogie. Le système de régulation des gènes étant bien mieux connu chez *Escherichia coli*, cet organisme est utilisé afin d'explorer les potentialités et limites de nos algorithmes d'inférence de signaux régulateurs avant de les appliquer à *Synechocystis* PCC 6803, dans un travail réalisé par H. de Jong, A. Morgat et M.-F. Sagot d'Helix avec J. Geiselmann de l'université Joseph Fourier à Grenoble et J. Houmard de l'ENS Ulm. Des résultats assez surprenants ont été récemment obtenus lors d'une première analyse de certains de ces signaux chez *E. coli*. Ces résultats pourraient conduire à des développements assez considérables, en termes à la fois des algorithmes d'inférence et de la biologie.

6.3.3. Base de motifs

Les motifs considérés jusqu'à présent dans le contexte d'une base sont des motifs avec des jokers (en anglais, des *don't care*). Étant donné une chaîne ou un ensemble de chaînes, le nombre de motifs avec des jokers peut être exponentiel. Plusieurs heuristiques, basées sur certaines propriétés pertinentes à la biologie ou sur la notion de quorum, essayent de réduire le nombre de motifs intéressants afin de permettre un post-traitement de ceux identifiés. Parmi les diverses méthodes permettant de sélectionner les motifs, nous nous intéressons à celles qui se basent sur les notions de maximalité et de spécificité. De façon informelle, un motif est maximal si son extension par des symboles à gauche ou à droite, ou le remplacement d'un de ses jokers par un symbole, lui fait perdre des occurrences. Malheureusement, la notion de maximalité ne suffit pas à limiter le nombre de motifs intéressants ainsi qu'il l'a été montré par Parida *et al.* L'introduction de la notion de base, à nouveau par Parida, a alors représenté une avancée importante dans ce domaine. A la suite d'une analyse poussée de l'approche proposée par Parida *et al.* et des difficultés pratiques et théoriques qu'elle entraînait, nous avons proposé une notion alternative de base pour un quorum de 2 qui permet de générer tous les motifs maximaux. Notre base présente des caractéristiques intéressantes telles que, (a) la base est un sous-ensemble de la base définie par Parida ; (b) elle est vraiment linéaire, le nombre de motifs étant inférieur à n et le nombre d'occurrences de ces motifs étant inférieur à $2n$; (c) elle est symétrique puisque la base de l'inverse de la chaîne s est composée de l'inverse des motifs dans la base de s ; (d) la base est calculable en temps polynomial, en fait en temps $O(nlogn)$.

Ce travail, apparemment théorique et avec des résultats surprenants, a des conséquences importantes pour l'inférence de motifs même dans un contexte plus large que celui de motifs avec des jokers, bien que certains

ne deviendront visibles que sur le long terme. Le travail a été réalisé dans le cadre d'un projet Da Vinci avec N. Pisanti et R. Grossi de l'université de Pise, M. Crochemore de l'université de Marne La Vallée et M.-F. Sagot.

6.3.4. Reconstruction ab initio de voies métaboliques

Dans le cadre du travail de thèse de F. Boyer, nous nous intéressons au problème de la reconstitution de voies métaboliques à partir des seules données de réactions biochimiques (reconstruction *ab initio*). Ce problème présente de nombreuses formulations dans la littérature qui souffrent toutes (y compris celle dans laquelle nous nous étions initialement engagés) d'une forte sensibilité aux données initiales (en d'autres termes d'un manque de robustesse). Cette année, nous avons entièrement reformulé ce problème sous un angle très différent en faisant intervenir, non plus les composés chimiques, mais les flux d'atomes que traduisent les réactions.

Ainsi, dans cette approche, reconstruire un chemin métabolique d'un substrat S à un produit P revient à rechercher les chemins du graphe métabolique permettant de transférer le maximum d'atomes dans le minimum d'étapes. Le nouvel algorithme, issu de cette formulation, fonctionne ainsi en deux étapes. Dans la première étape, on recherche les correspondances atomiques au sein des réactions biochimiques élémentaires (cette correspondance est connue du chimiste mais ne figure malheureusement pas explicitement dans les bases de données telles que KEGG, il convient donc de l'expliciter). Dans une seconde étape, on cherche toutes les compositions d'injections partielles (des composés vers les composés) permettant de transférer au moins n atomes en au plus q étapes. Deux approches algorithmiques ont été testées : l'une basée sur un *branch and bound* énumérant toutes les chemins et l'autre basée sur la construction d'un automate (acceptant toutes et seulement toutes les solutions précédentes). La seconde approche s'avère extrêmement satisfaisante en pratique puisqu'il devient possible de traiter des chemins de taille 10 en quelques secondes dans une base telle que KEGG (quelques milliers de composés et de réactions). Par ailleurs, et c'est là le point le plus important pour nous, cette approche s'avère beaucoup moins sensible aux données initiales. Elle ne nécessite, en particulier, pas de distinction entre des composés « normaux » et des composés « réservoirs » (c'est-à-dire disponibles en quantité illimité) et résiste mieux au bruit (erreur de stochiométrie ou réactions parasites). Notre objectif pour l'avenir est double : sur le plan pratique, nous souhaitons maintenant intégrer ce noyau algorithmique au sein du module GEB (5.13) et la sous-section suivante) afin de le rendre accessible aux biologistes ; sur le plan théorique, nous souhaitons aller plus loin et envisager maintenant la question (difficile) de la reconstitution en absence de certaines réactions, c'est-à-dire de l'inférence de nouvelles réactions.

6.3.5. Modélisation des données de génomique et de métabolisme

En 2002, le projet Panoramix (*cf.* rapport d'activité 2001) a évolué vers le projet GenoExpertBacteria (sous l'influence, en particulier, du projet inter-EPST (2001-2003) de réannotation de deux génomes de référence (*B. subtilis* et *Synechocystis*). Ce projet est réalisé en collaboration avec trois laboratoires experts de biologie, à l'Institut Pasteur (Hong-Kong et Paris), à l'Ecole Normale Supérieure (Ulm) et à l'université Joseph Fourier. Il se propose d'atteindre les deux objectifs suivants :

- développer une base de connaissances dédiée à l'annotation de génomes procaryotes. Cette base doit permettre de représenter des données génomiques, protéiques et métaboliques.
- expertiser les données pour deux génomes de référence (B. subtilis et Synechocystis).

Dans ce but, nous avons choisi de rassembler les trois bases de connaissances initialement développées dans le cadre de Panoramix (Genomix, Proteix et Metabolix) en un seul et unique schéma et de développer les interfaces d'accès sous la forme d'un module compatible avec l'environnement Genostar. Utilisé seul, ce module permet de naviguer dans la base de connaissances et de visualiser les données à l'aide d'interfaces graphiques spécialisées. Son utilisation au sein de la plate-forme Genostar permet d'étendre ces fonctionnalités, par exemple d'accéder aux outils d'annotation syntaxique de GenoAnnot (prédiction de séquences codantes, recherche de terminateurs, ...) ou aux outils d'analyse statistique de GenoBool. Ceci afin de mettre à disposition des utilisateurs biologistes le maximum d'outils en un seul et même environnement.

Les visualiseurs spécialisés ont été développés sous la forme de composants graphiques Java (Java Beans) :

- MolBean est un visualiseur de structure 2D (formule plane) de composés chimiques ;
- ReacBean est un visualiseur de réactions chimiques ;
- PathwayBean permet de représenter des voies métaboliques (pathways) sous la forme d'un graphe métabolique ;
- CartoBean permet une représentation cartographique d'éléments génétiques.

Une première version du module GEB est opérationnelle et sera mise à disposition des partenaires du projet début 2003.

6.3.6. Modélisation des éléments transposables dans les génomes séquencés

La compréhension de la dynamique des éléments transposables (ET), de leurs interactions avec les gènes de l'hôte et entre eux, de leur distribution au sein des génomes, de leur structure et de leur évolution, soulève de nombreuses questions qui ne peuvent être résolues par une approche ponctuelle de quelques éléments chez une espèce, mais nécessitent une approche globale à partir d'un grand nombre de séquences et d'informations sur ces séquences dans plusieurs organismes. D'où la nécessité de mettre en place une base de connaissances des ETs qui résumerait les informations obtenues à partir des génomes séquencés.

Le projet FlyGATE répond à cette nécessité en incluant :

- le développement avec AROM d'un schéma des connaissances impliquées, appelé GATE (*Genome Analysis for Transposable Elements*), développé par C. Rizzon. Il décrit la biologie des ETs (struture, protéines intrinsèques, classification par familles) et prend en compte leur organisation spatiale par le lien avec GemCore, développé en collaboration avec G. Bronner.
- la construction de la base de connaissances FlyGATE basée sur GATE et dédiée aux ETs du génome de *Drosophila melanogaster*. FlyGATE sera bientôt interrogeable via internet.

6.4. Modélisation dynamique des réseaux de régulation génique

Participants : Grégory Batt, Céline Hernandez, Hidde de Jong [Correspondant], Michel Page.

Puisque des informations quantitatives sur les paramètres cinétiques et les concentrations sont rarement disponibles, les méthodes de simulation numérique ne sont pas applicables à l'analyse de réseaux de régulation génique. Afin de faire face à ce problème, une méthode pour la simulation qualitative a été développée au sein du projet Helix. De manière similaire à certains travaux réalisés en biomathématique, les systèmes de régulation génique sont modélisés par une classe d'équations différentielles linéaires par morceaux ayant des propriétés mathématiques favorables. Au lieu de donner une valeur numérique exacte aux paramètres du modèle, ces derniers sont contraints par des relations d'égalité et d'inégalité qui sont exploitées afin de prédire les comportements qualitatifs possibles du réseau.

Les équations différentielles traitées par la méthode ont des discontinuités dans leur second membre. Avec J.-L. Gouzé (INRIA Sophia-Antipolis) et T. Sari (université de Haute Alsace, Mulhouse) les problèmes mathématiques liés aux discontinuités ont été étudiés, en se basant sur le concept de solutions de Filippov. Cette approche, largement utilisée pour des problèmes similaires en automatique, permet de traiter les discontinuités d'une façon rigoureuse et pratique. La généralisation de la méthode de simulation qualitative qui en résulte a été décrite dans une communication pour la *European Conference on Artificial Intelligence (ECAI)* et dans un rapport de recherche INRIA.

La nouvelle version de la méthode de simulation qualitative a été implémentée dans la version 5.0 de l'outil *Genetic Network Analyzer (GNA)*. Le logiciel a été déposé auprès de l'APP et est disponible à travers le Web (http://www-helix.inrialpes.fr/gna). Une article décrivant cette version de GNA paraîtra dans la revue *Bioinformatics* début 2003. Afin de faciliter l'utilisation du simulateur, nous avons poursuivi le développement d'un éditeur de modèles de simulation. Céline Hernandez a réalisé, avant la fin de son contrat d'ingénieur associé, un prototype d'un tel éditeur qui est actuellement repris par M. Page et H. de Jong pour l'intégrer dans GNA.

Suite à son stage de DEA en Informatique, G. Batt a commencé une thèse au sein du projet Helix. Le sujet de cette thèse est le développement d'une méthode de validation, complémentaire à la méthode de simulation. Afin de valider un modèle d'un réseau de régulation génique, les prédictions des comportements qualitatifs possibles du réseau sont comparées avec le profil d'expression mesuré expérimentalement. La validation de modèles de réseaux de régulation génique est également au coeur d'un projet retenu cette année dans le cadre de l'appel d'offre du programme Bioinformatique inter-EPST. Ce projet s'inscrit dans une collaboration entre bioinformaticiens et biologistes de l'ENS Paris, de l'INRIA Rhône-Alpes et de l'université Joseph Fourier et fait suite à un premier projet financé au cours de la période 2000-2002.

Dans le cadre de ces deux projets inter-EPST, nous travaillons avec les laboratoires de J. Geiselmann (université Joseph Fourier) et de J. Houmard (ENS Paris) sur la modélisation de la régulation globale de la transcription chez les bactéries *Escherichia coli* et *Synechocystis* PCC 6803. Nous essayons de comprendre comment la réponse de la bactérie à des signaux provenant de l'extérieur émerge des réseaux d'interactions entre gènes, protéines et molécules messagères. Afin d'élucider ces réseaux, nous utilisons GNA en combinaison avec des méthodes expérimentales pour tester les prédictions du simulateur, ainsi que des méthodes bioinformatiques classiques, comme l'analyse des séquences. Cette année nous avons également continué la modélisation d'un autre système bactérien, l'initiation de la sporulation chez la bactérie *Bacillus subtilis*.

6.5. Protéomique : aide à l'acquisition et modélisation des données

Participants : Anne Morgat, Erwan Reguer, Alain Viari [Correspondant].

Le projet PepMap, financé par le Ministère de la Recherche, a été engagé en collaboration avec le CEA Grenoble (Laboratoire de Chimie des Protéines, responsable J. Garin) et la société GENOME express (responsable du projet, T. Vermat). Du point de vue expérimental, le projet repose sur la plate-forme instrumentale développée au CEA à Grenoble, constituée d'un nano-chromatographe liquide (nano-LC) couplé à un spectromètre de masse (nano-electrospray / Q-TOF). Ce type de technique, extrêmement novatrice, permet en effet de générer rapidement une grande quantité d'informations à partir d'échantillons biologiques ciblés. Cette plate-forme a déjà reçu un soutien financier de la région Rhône-Alpes (appel d'offre « Thématiques Prioritaires 2000-2002 ») pour l'achat d'un second Q-TOF et de matériel de robotisation. Elle constitue par ailleurs le fer de lance de la plate-forme protéomique grenobloise dans le cadre de la génopole Rhône-Alpes. PepMap constitue le volet bioinformatique complémentaire de cette plate-forme technologique. L'objectif principal est de fournir un ensemble de modules logiciels destinés à l'exploitation des données de type « étiquettes protéiques » fournies par la spectrométrie de masse. Nous nous intéressons plus particulièrement à la localisation directe de ces étiquettes sur l'ADN génomique (chromosomes eucaryotes complets) sans passer par une reconstruction de la structure génique du chromosome, reconstruction qui pose encore de nombreux problèmes théoriques et pratiques.

En pratique, les étiquettes sont produites à partir des fragments trypsiques de la (ou des) protéines à analyser, séparés par chromatographie liquide couplée un spectromètre de masse (Q-TOF). Chaque fragment trypsique fournit ainsi une étiquette constituée d'une portion de séquence peptidique (obtenue grâce à l'analyse du spectre de fragmentation (N et C terminales) du peptide) flanquée de deux parties de séquence inconnue, mais de masse totale connue.

Le module central du système, baptisé TagMap, a pour but de localiser rapidement une étiquette sur de l'ADN chromosomique. La difficulté provient ici de l'organisation en mosaïque des gènes eucaryotes qui interdit de faire l'hypothèse que l'étiquette couvre une portion contiguë du chromosome.

Compte tenu de la taille des chromosomes humains (quelques dizaines à centaines de Mb) et du flux de production des étiquettes par le Q-TOF (environ 100 étiquettes/heure/Q-TOF), il convient de soigner particulièrement les performances en temps de l'algorithme de localisation. La complexité temporelle de l'algorithme actuel est linéaire avec la taille des chromosomes, en log du nombre d'étiquettes et indépendant de leur taille. Ceci permet de localiser, en quelques minutes, plusieurs centaines d'étiquettes simultanément sur des chromosomes de plusieurs millions de bases et permet donc d'envisager son inclusion dans une chaîne de traitement « à haut débit » des données de protéomique.

Plus récemment, nous nous sommes intéressés à une étape « amont » de cette chaîne : celle qui concerne la production des étiquettes à partir des spectres de masse MS/MS. Il s'agit là d'un problème difficile (habituellement connu sous le nom de « *de novo sequencing* »), mais qui constitue un réel « goulot d'étranglement » puisque, jusqu'à présent, cette étape était conduite manuellement (ou semi-manuellement) dans la chaîne des logiciels propriétaires fournis par les constructeurs de spectromètres. Dans l'optique de la production d'étiquettes courtes (quelques acides aminés), nous avons pu proposer un algorithme (de type *branch and bound*) très satisfaisant en termes de performances et permettant de s'affranchir de toute intervention de l'opérateur. L'étape délicate dans ce type d'algorithme est moins la production de toutes les étiquettes possibles que l'évaluation de leur plausibilité compte tenu du spectre expérimental (en d'autres termes ce type de programme trouve fréquemment les « bonnes » étiquettes, mais se montre souvent incapable de dire quelles sont les meilleures). Nous avons contourné ce problème délicat en couplant ce programme (Taggor) au programme TagMap, c'est-à-dire en utilisant les chromosomes pour « séparer le bon grain de l'ivraie » (les « bonnes étiquettes » étant alors celles qui s'aggrègent dans une même région d'un chromosome). Les premières évaluations de cette approche s'avèrent extrêmement prometteuses. Dans une expérience portant sur une fraction hydrophobe du plasmalemme d'*Arabidopsis thaliana*, nous avons pu ainsi retrouver l'essentiel des protéines identifiées à partir des étiquettes produites manuellement par les expérimentateurs. Notre objectif est maintenant de coupler plus intimement Taggor et TagMap au sein d'une chaîne de traitement totalement automatisée.

6.6. Extraction d'informations à partir de textes

Participants : Jean Dina, Violaine Pillet, François Rechenmann [Correspondant].

Le système développé dans le cadre du projet BioMiRe a pour fonction de détecter, dans les textes scientifiques, les noms d'entités biologiques (gènes, protéines, ARNs et espèces) et les liens entre ces entités. L'architecture générale de ce système est composée d'un contrôleur central qui gère quatre modules :

1. Un module de sélection des documents ;
2. Un module de traitement des documents ;
3. Un module de gestion des informations ;
4. Un module interface d'interrogation.

6.6.1. Module de sélection des documents

Le module de sélection des documents est une interface permettant de sélectionner les documents que l'on veut analyser à partir de la source documentaire PubMed. L'utilisateur saisit une requête qui est transmise à la base de données PubMed et la liste des documents satisfaisant la requête est alors proposée. Les documents sont tout d'abord transformés en format XML, puis transmis au module de traitement des documents.

6.6.2. Module de traitement des documents

Le module de traitement des documents effectue la reconnaissance des noms d'entités biologiques. Il découpe chaque phrase des textes en unités élémentaires appelés *tokens* et mène une analyse lexicale complétée de règles contextuelles sur chacun d'eux. L'analyse lexicale est obtenue en utilisant un analyseur morphologique et un segmenteur NTM (Normalisation, Tokénisation, Morphologie), tous deux développés par XRCE. Cette analyse permet, grâce à l'utilisation de plusieurs lexiques, d'attribuer à chacun des *tokens* une ou plusieurs étiquettes grammaticales. Les lexiques créés pour cette analyse sont au nombre de quatre : le lexique « anglais » contenant les termes de base de l'Anglais (400000 formes) ; le lexique « terminologie biologique » qui complète le lexique anglais et contient des termes spécifiques au domaine biologique (9600 formes) ; le lexique « noms d'entités » qui rassemble les noms de gènes (86950 gènes, pour les quatre espèces : Homme, souris, drosophile et *Arabidopsis thaliana*), les noms de protéines (83969 protéines pour la majorité des espèces) et les noms d'espèces (170000 espèces, lignées comprises) ; le lexique « ambigu » qui rassemble tous les termes communs aux trois lexiques (anglais, biologique, noms d'entités).

Une phase de désambiguïsation de certaines parties du discours permet ensuite d'attribuer une seule étiquette grammaticale pour chacun des *tokens* à l'aide d'un outil basé sur les modèles de Markov cachés (HMM) et développé par XRCE. Enfin, une série de règles contextuelles est appliquée pour déterminer parmi les différents *tokens* ceux qui sont des noms d'entités biologiques. Ces règles sont des combinaisons de mots-clés (490 mots-clés déterminés) qui forment des *patterns* (97 *patterns* construits). Deux classes de *patterns* ont été mises en place : les *patterns* positifs déterminant un *token* en tant que nom d'entité biologique (gène, protéine ou ARN) et les *patterns* restrictifs identifiant un *token* comme n'étant pas un nom d'entité biologique.

Pour mettre en place le système de reconnaissance des noms d'entités biologiques, définir les règles contextuelles et évaluer les performances du système, un corpus d'apprentissage de 130 résumés et un corpus de test de 56 résumés ont été créés. Ces corpus contiennent des résumés d'articles scientifiques provenant de la base de données Medline. Chaque résumé a été annoté manuellement pour localiser les noms d'entités biologiques présents. Chaque nom d'entité est alors identifié par un tag spécifique.

Ce module de reconnaissance est capable de détecter des noms d'entités biologiques avec un taux de rappel de 84,3% et un taux de précision de 78,9%.

6.6.3. Module de gestion des informations

Ce module gère une base de données relationnelle (Postgresql) où sont stockées les informations relatives aux documents XML : les méta-informations du document, telles que titre, journal, date et auteurs ; les noms d'entités détectés, ainsi que les positions de ces noms dans le texte (le numéro de la phrase et la place dans la phrase) ; les termes qui composent les noms d'entités.

6.6.4. Module interface d'interrogation

Cette interface permet d'interroger et de visualiser les documents analysés. L'interrogation des documents se fait à l'aide d'un langage prédicatif. Il permet d'effectuer des requêtes plus ou moins complexes pour rechercher des noms d'entités biologiques, mais aussi des relations existant entre ces noms d'entités. L'utilisateur peut tout d'abord spécifier la classe et/ou la valeur de l'entité recherchée (gène, protéine, RNA et espèce). Il peut aussi spécifier la position absolue (titre, résumé, introduction ou conclusion) du nom d'entité recherché ou alors le lien de structure (phrase ou paragraphe) entre deux noms d'entités. L'utilisateur peut ajouter une contrainte de distance (en nombre de mots ou de phrases) entre deux noms d'entités ou d'ordre d'apparition dans le texte de deux noms d'entités. Il lui est possible aussi d'exclure la présence d'un nom entre deux autres noms d'entités, ou de spécifier que le nom d'entité recherché appartient ou n'appartient pas à une espèce. Toutes ces possibilités peuvent être combinées entre elles.

Pour chaque requête formulée, la liste des instances qui la satisfont est affichée. En sous-arborescence de chacune des instances sont listés les documents contenant l'instance. Il suffit de sélectionner l'un des documents pour obtenir son affichage complet. Pour chaque affichage, les noms d'entités recherchés sont mis en surbrillance.

6.7. Environnement didactique en bioinformatique

Participants : Philippe Genoud [Correspondant], Stéphanie Merriene, Anne Morgat, Alain Viari, François Rechenmann, Danielle Ziébelin.

Une première maquette de l'EDB (Environnement Didactique en Bioinformatique) est opérationnelle. Elle se présente sous la forme d'une application Java autonome, qui permet, d'une part d'accéder à un contenu pédagogique hypertexte, d'autre part d'illustrer les points de ce cours par des « manips » qui mettent en oeuvre des algorithmes bioinformatiques. Le contenu pédagogique permet de guider l'apprenant dans son exploration de ces algorithmes bioinformatiques (sélection de l'algorithme, sélection des données, activation ou non de la possibilité de modifier les données, choix de l'interface graphique de visualisation). Une version entièrement accessible à travers un navigateur Web est simultanément développée.

La maquette actuelle intègre six modules sur le thème de l'annotation de séquences géno-miques : présentation du code génétique, explication des différentes phases de lecture, recherche de motifs (simples, avec erreurs, exprimés sous la forme d'expressions régulières), *dotplot* avec fenêtre glissante et lissage,

alignement de deux séquences, recherche de zones codantes par analyse du biais de codage (test du chi2), et stratégie de recherche de zones codantes combinant, au travers d'une interface cartographique, les algorithmes de recherche de motifs, d'alignement de séquences et d'analyse du biais de codage.

L'Environnement Didactique en Bioinformatique sera progressivement utilisé dans plusieurs filières d'enseignement de l'UJF : option « bioinformatique » de la maîtrise d'informatique et DESS CCI (Compétence Complémentaire en Informatique). Par ailleurs, une première expérience d'atelier « bioinformatique » au sein de l'« École de l'ADN » (CCSTI, Grenoble) est prévue au printemps 2003. Le public est ici constitué d'élèves de classes scientifiques de lycée.

Le développement d'EDB a bénéficié d'un soutien du GRECO (Grenoble Campus Ouvert). Des démonstrations ont ainsi été faites lors des journées du GRECO, à Grenoble, ainsi que sur le stand de l'INRIA Rhône-Alpes lors de la Fête de la Science les 18, 19 et 20 octobre.

6.8. Projet GénoStar

Participants : Christophe Bruley, Pierre-Emmanuel Ciron, Véronique Dupierris, Gilles Faucherand, Agnès Iltis, François Rechenmann, Alain Viari [Correspondant].

L'objectif du projet Genostar est de concevoir, développer et expérimenter un environnement modulaire de génomique exploratoire. La modularité du système doit lui permettre d'évoluer facilement et rapidement à la suite de l'apparition de nouvelles catégories de données ou de nouvelles méthodes d'analyse.

Dans la première phase, de deux ans, du projet, trois applications ont été développées : GenoAnnot est dédiée à l'annotation de génomes, procaryotes dans un premier temps, puis eucaryotes ; GenoLink est destinée à prolonger le processus d'annotation amorcé par GenoAnnot vers la caractérisation des fonctions des gènes identifiés ; GenoBool permet d'explorer des ensembles de données hétérogènes à travers l'application de techniques d'analyse multifactorielle.

Ces trois applications, ainsi que d'autres qui seraient ultérieurement conçues et développées (le module GEB en est le premier exemple concret), communiquent entre elles ; elles échangent données et résultats grâce au noyau GenoCore, qui permet de décrire les objets étudiés et leurs relations, ainsi que les stratégies d'analyse. GenoCore gère également la persistance des données et des connaissances et assure leur édition et leur visualisation grâce à des interfaces graphiques.

Les travaux de conception et de développement menés en 2001 et 2002 ont permis l'obtention d'une première version de l'environnement qui a fait l'objet d'une présentation et d'une démonstration publiques dans l'auditorium de l'Institut Pasteur le 21 mai. Depuis le 15 novembre, cette version est distribuée aux laboratoires publics de recherche français qui en font la demande.

Les contributions du projet Helix concernent essentiellement GenoCore et les applications GenoAnnot et GenoBool, en interaction forte avec la société GENOME express, partenaire grenoblois du consortium Genostar

L'application GenoCore repose sur le système AROM de représentation et de gestion d'objets et de relations. Elle étend et complémente ses fonctionnalités à travers l'adjonction de modules spécialisés. C'est ainsi qu'a été développé un module de définition et de gestion de types construits. Le premier exemple d'un tel type est le type « séquence », qui autorise la manipulation de longues séquences, génomiques et protéiques, à travers des opérateurs appropriés. De même, un module dédié à la gestion de la mémoire et de la persistance a été intégré. Il permet une gestion efficace de grandes bases, contenant plusieurs millions d'objets. Enfin, un module graphique de requêtes est en cours de développement. Il permet de sélectionner un ensemble d'instances de classes et de relations qui satisfont des contraintes exprimées sur les valeurs de leurs attributs et de leurs rôles.

Le développement de GenoAnnot, application dédiée à l'annotation de génomes entiers, passe par l'élaboration de son ontologie, c'est-à-dire par l'explicitation des entités concernées, qu'elles soient informatiques, par exemple des motifs détectés sur la séquence, ou biologiques, telles que les gènes et leurs constituants, et de leurs relations. Cette ontologie a été construite, tant pour les génomes procaryotes qu'eucaryotes. Les méthodes d'annotation ont été organisées en stratégies. C'est le module de tâches de GenoCore qui accepte

la description de ces stratégies et les exécute à la demande de l'utilisateur. Les résultats de ces méthodes, c'est-à-dire les entités détectées sur la séquence d'un génome donné, sont affichées sur l'interface cartographique. Là encore, compte tenu du nombre d'objets impliqués et de la réactivité attendue du système, les critères d'efficacité sont primordiaux.

Enfin, l'application GenoBool comprend trois modules principaux : le « tableur », qui permet la sélection, la visualisation et la manipulation des valeurs, extraites d'objets d'une base GenoCore, sur lesquelles porte l'analyse ; les « codeurs » qui rendent ces valeurs homogènes, par exemple en les transcrivant sous forme booléenne ; et enfin, l'interface de visualisation des résultats de l'application de méthodes classiques d'analyse de données. Ces méthodes sont elles aussi organisées en stratégies, qui sont incorporées à l'application, aidant ainsi l'utilisateur à exploiter au mieux ses fonctionnalités.

7. Contrats industriels

7.1. GénoStar

Le projet Genostar est conduit par un consortium de quatre membres :

- la société Hybrigenics, Paris ;
- la société GENOME express, Grenoble ;
- l'Institut Pasteur, Paris ;
- l'INRIA.

Ce consortium a signé à l'automne 1999 un accord sur le développement et la valorisation de l'environnement. Le projet a obtenu le soutien du programme « Génomique » du Ministère de la Recherche, à travers une aide à la génopole Institut Pasteur de Paris, puis un soutien complémentaire de la Direction de la Technologie (programme GenHomme). Enfin, Genostar est une action de développement de l'INRIA et a bénéficié à ce titre d'un soutien pendant trois ans (2000-2002).

7.2. Contrats express

Le projet PepMap rassemble la société GENOME express, le LCP (Laboratoire de Chimie des Protéines /CEA, J. Garin) et l'INRIA Rhône-Alpes. Il a abouti au développement d'un module de *mapping* d'étiquettes peptidiques sur le génome qui porte les régions codantes aux protéines analysées. Ce projet a bénéficié d'un soutien du Ministère de la Recherche, Direction de la Technologie.

7.3. XRCE

Le Centre Européen de Recherche Xerox (XRCE) est le partenaire privilégié du projet Helix sur le thème de l'extraction d'informations à partir de textes, dans cadre en particulier du projet BioMiRe, qui a abouti cette année à un module de reconnaissance de noms d'entités biologiques dans les articles scientifiques. Ce projet est soutenu par le Ministère de la Recherche (Direction de la Technologie) et implique l'INRIA, le Centre de Recherche Européen de Xerox (XRCE) à Meylan, et deux équipes de l'INRA, à Versailles et à Gand (Belgique).

8. Actions régionales, nationales et internationales

8.1. Actions régionales

Les activités d'Helix s'inscrivent dans le cadre de la génopole Rhône-Alpes. Le projet bioinformatique mis en avant par la génopole est l'analyse comparative des génomes, cadre dans lequel s'inscrivent les travaux d'Helix sur la cartographie comparée.

Une collaboration scientifique majeure implique J. Geiselmann du CERMO (université Joseph Fourier, Grenoble). Elle porte sur la modélisation et la simulation des interactions géniques. Ce projet bénéficie d'un soutien du programme « Bioinformatique » inter-EPST (CNRS-INRA-INRIA-INSERM).

Le projet Helix accueille à temps partiel E. Fanchon, chercheur CNRS à l'IBS (Institut de Biologie Structurale) sur la modélisation et la classification de structures tertiaires (*folds*) de protéines.

Sur la protéomique, une collaboration avec J. Garin (LCP : Laboratoire de Chimie des Protéines, CEA) est poursuivie, avec un soutien du Ministère de la Recherche, Direction de la Technologie, dans le cadre du programme GenHomme. Le projet soutenu rassemble la société GENOME express, le LCP/CEA et l'INRIA Rhône-Alpes.

L. Duret participe au projet « *C. elegans* : Organisme modèle », dans le cadre de l'appel d'offres « Projets Thématiques Prioritaires » de la région Rhône-Alpes (coordonnateur L. Ségalat, CGMC, Lyon).

8.2. Actions nationales

Les membres de l'équipe Helix sont en relation avec les différents groupes français de bioinformatique, dans les universités ou les organismes de recherche, en particulier à l'ABI (Atelier de BioInformatique) à Paris 6 (J. Pothier), à l'INRA à Jouy-en-Josas (P. Bessières), Gif-sur-Yvette (C. Thermes), Evry (C. Médigue), Toulouse (C. Gaspin), Marseille (G. Fichant et Y. Quentin) et Gand en Belgique (P. Rouzé). Bien entendu, l'équipe souhaite en tout premier lieu renforcer les interactions avec les projets INRIA déjà engagés en bioinformatique.

M.-F. Sagot coordonne le projet « Régulation, Synténie et Pathogénicité - Algorithmes et Expérimentations » dans le cadre du programme « Bioinformatique » inter-EPST (CNRS-INRA-INRIA-INSERM). Les partenaires d'Helix sont l'Institut de Biologie Physico-Chimique de Paris (A. Vanet, co-coordonnatrice) et Institut Gaspard Monge de l'université de Marne La Vallée. L'objectif du projet est d'aborder les aspects à la fois algorithmiques et expérimentaux liés à l'expression des gènes et aux réarrangements génomiques dans le but de mieux comprendre leur fondement ainsi que leur relation avec la pathogénicité.

Dans le même programme, M.-F. Sagot et A. Viari participent au projet « Détection des exons/introns dans le génome humain ». Le projet implique des équipes du CGM de Gif-sur-Yvette (C. Thermes, co-coordonnateur), de l'Atelier de BioInformatique de l'université de Paris 6, du Laboratoire « Génome et Informatique » de l'université de Versailles et l'Institut de Mathématiques de Luminy à Marseille. L'objectif du projet est la création d'un algorithme reproduisant au plus près le fonctionnement de la machinerie d'épissage.

Toujours dans le programme « Bioinformatique » inter-EPST (CNRS-INRA-INRIA-INSERM), H. de Jong a coordonné le projet « Modélisation et simulation de réseaux de régulation génique : La transduction des signaux par les nucléotides cycliques chez la cyanobactérie *Synechocystis* PCC 6803 ». Plusieurs membres d'Helix ont participé à ce projet (C. Hernandez, M. Page, A. Morgat, M.-F Sagot, A. Viari) ; les partenaires sont des équipes de l'université Joseph Fourier (Grenoble) et de l'ENS (Paris). Les résultats obtenus ont fait l'objet d'une présentation lors du premier colloque bilan du programme « Bioinformatique » inter-EPST, à Paris, les 14 et 15 octobre. Suite à ce projet, un nouveau projet intitulé « Validation de réseaux de régulation génique : La régulation globale de la trnacription chez *Escherichia coli* et *Synechocystis* » sera financé dans la cadre du programme « Bioinformatique » inter-EPST (2002-2004).

Le projet « Panoramix : fédération de bases de connaissances pour la génomique et expertise sur deux génomes bactériens de référence » bénéficie également du soutien de ce programme. Coordonné par A. Morgat, il est financé sur deux ans depuis fin 2001.

M.-F. Sagot coordonne le projet « Algorithms for Modelling, Search and Inference Problems in Molecular Biology » dans le cadre du programme « Bioinformatique » inter-EPST (CNRS-INRA-INRIA-INSERM). Ce projet commence fin 2002 et durera deux ans. Les partenaires d'Helix sont membres de neuf laboratoires français et six laboratoires européens (en Belgique, Finlande, Grande Bretagne, Italie et Suède). L'objectif du projet est d'utiliser des approches combinatoires ou statistiques/probabilistes afin d'identifier certains des principaux éléments ou processus supports de la vie, et d'analyser et modéliser les relations temporelles ou spatiales pouvant exister entre eux. La première partie recouvre essentiellement quatre problèmes d'inférence

dans les séquences biologiques en relation avec l'expression des gènes et l'étude de la structure d'un génome, c'est-à-dire sa division en des unités bien identifiées (gènes de protéines ou d'ARNs, répétitions, *etc.*). La seconde partie concerne l'analyse phylogénétique de grandes familles de séquences homologues, l'étude et l'inférence d'évolutions réticulaires, la détection de segments conservés et de distances de réarrangements entre génomes, l'inférence de structures secondaires et tertiaires d'ARN, l'analyse de réseaux génétiques et, enfin, la reconstruction *ab initio* et l'évolution des réseaux métaboliques.

M.-F. Sagot participe au projet « Découverte de motifs dans les séquences biologiques » dans le cadre du programme inter-EPST « Bioinformatique » CNRS-INRA-INRIA-INSERM. Ce projet est coordonné par J. Nicolas de l'IRISA. Le projet a un double objectif : (1) proposer des algorithmes performants de découverte de motifs, dont la mise au point sera effectuée à la fois sur des modèles aléatoires et sur quelques problèmes particuliers de biologie ; (2) disposer d'une plate-forme d'expérimentation permettant l'utilisation et la comparaison des algorithmes en liaison avec les banques de données biologiques.

L'équipe Helix a bénéficié d'un contrat ANVAR dans le cadre d'une collaboration entre A. Vanet et M.-F. Sagot. L'objectif de ce contrat est le développement d'algorithmes pour la détection de positions corrélées dans un alignement de séquences avec application à la découverte de vaccins, traitements ou tests de diagnostics contre des pathogénicités d'origine virale ou bactérienne.

8.3. Actions européennes et internationales

Au niveau européen, l'équipe participe au projet ORIEL (IST) constituant le volet « recherche » du projet e-BioSci.

Le projet Helix participe au projet HAMAP d'annotation automatique de protéomes bacté-riens, à l'initiative de l'Institut Suisse de Bioinformatique (A. Bairoch) à Genève.

M.-F. Sagot participe à un Projet CNPq, « Problèmes en Optimisation Combinatoire : Algorithmes et Applications » avec le Département d'Informatique, Institut de Mathématiques et Statistiques, université de São Paulo, Brésil (coordonnatrice : Y. Wakabayashi, professeur à l'université de São Paulo).

L'équipe Helix bénéficie d'un grant du Welcome Trust, impliquant des équipes du King's College à Londres, l'université de Marne La Vallée et l'INRIA Rhône-Alpes. L'objectif du projet est l'échange de chercheurs entre la France et l'Angleterre en vue de collaborations, en particulier pour l'étude de la combinatoire des mots et l'élaboration d'algorithmes permettant de traiter certains problèmes en biologie.

N. Pisanti et M.-F. Sagot collaborent avec R. Grossi de l'université de Pise sur un problème d'inférence de motifs. M. Crochemore de l'Institut Gaspard Monge de l'université de Marne La Vallée fait également partie de cette collaboration.

L. Duret participe au projet « The European Molecular Biology Linked Original Resources : TEMBLOR » financé par l'Union Européenne dans le cadre du programme *Quality of Life and Management of Living Resources* (QLRT-2001-00015). Le projet est coordonné par R. Apweiler (EBI, Hinxton, UK).

L. Duret est correspondant du projet européen de grille de calcul DataGrid : WP10 « Applications à la biologie moléculaire et à l'imagerie médicale ».

M.-F. Sagot est, depuis le 1er novembre 2002, *Visiting Research Fellow* du King's College à Londres.

L'équipe Helix à travers M.-F. Sagot participe à un Projet CNPq, « Problèmes en Optimisation Combinatoire : Algorithmes et Applications », avec le Département d'Informatique, Institut de Mathématiques et Statistiques, université de São Paulo, Brésil. La coordonnatrice du projet est Y. Wakabayashi, professeur à l'université de São Paulo.

L'équipe Helix participe à un projet « Leonardo da Vinci » avec R. Grossi et N. Pisanti de l'université de Pise en Italie, M. Crochemore de l'université de Marne La Vallée en France et M.-F. Sagot de l'INRIA Rhône-Alpes.

9. Diffusion des résultats

9.1. Animation de la communauté scientifique

C. Gautier est chargé de mission pour la bioinformatique au CNRS et responsable du programme inter EPST Bioinformatique.

M.-F. Sagot et L. Duret sont membres du comité d'organisation de la conférence française annuelle de bioinformatique JOBIM.

M. Gouy est membre du CNU, section 67. Il est également membre du conseil scientifique de l'Institut Français de la Biodiversité.

G. Perrière est membre titulaire de la section 67 de la Commission de Spécialistes de l'Enseignement Supérieur de l'UCBL. Depuis septembre 2001, il est second vice-président de cette même commission. Il est également membre du conseil de gestion de l'IFR 41 (Institut des Sciences et Méthodes de l'Écologie et de l'Évolution) de l'UCBL. Il est aussi membre du comité d'organisation du groupe IMPG « Classification et Phylogénie » avec A. Guénoche et R. Christen et du groupe de travail du CNRS dirigé par C. Michau sur les logiciels libres (2000-2001).

H. de Jong est membre du comité de programme du *Seventeenth International Workshop on Qualitative Reasoning* qui se tiendra au Brésil en 2003. Il est également membre titulaire de la Commission de Spécialistes (section 27) du « Centre de Mathématiques et d'Informatique » (université de Provence, Marseille).

M.-F. Sagot anime le séminaire « Algorithmique et Biologie » (http://www.inrialpes.fr/helix/people/sagot/AlgoBio). Trois séries ont été organisées déjà à Lyon, « Régulation, métabolisme et structure génomique » en Septembre 2001, « Dynamique des Génomes et Évolution » en avril 2002, « Statistiques et Probabilités en Génomique » en octobre 2002. La prochaine série est prévue pour avril 2003 et aura pour thème : « Algorithmique et Combinatoire en Biologie ». Chaque série comporte une quinzaine d'exposés, dont la moitié environ est de chercheurs français et l'autre moitié d'invités étrangers.

M.-F. Sagot est co-présidente du comité de programme et membre du comité d'organisation de la *European Conference on Computational Biology* (ECCB) qui aura lieu du 27 au 30 septembre 2003 à Paris.

M.-F. Sagot est membre du comité scientifique de l'ACI Cryptologie (Ministère de la Recherche), du comité IMPG (Informatique, Mathématique et Physique pour la Génomique), du Ministère de la Recherche, du comité scientifique du cours « Informatique en Biologie » de l'Institut Pasteur et du cours « *Computational Biology* » de l'université du Chili à Santiago, Chili.

M.-F. Sagot est membre nommé (suppléant) de la Commission d'Evaluation de l'INRIA. Elle a participé à la section d'audition Rhône-Alpes lors du concours CR2 en 2002.

M.-F. Sagot a été en 2002 membre du comité de programme des conférences internationales *Combinatorial Pattern Matching* (CPM'02), RECOMB'02, *Workshop on Algorithms in BioInformatics* (WABI'02), *String Processing and Information REtrieval* (SPIRE'02), *European Conference on Computational Biology* (ECCB'02), et de la conférence national en bioinformatique *JOBIM'02*.

F. Rechenmann est membre de l'*Editorial board* du journal *Bioinformatics*. Il est membre du Comité de Coordination des Sciences du Vivant (CCSV), du comité scientifique du programme « Bioinformatique » inter-EPST, du comité de pilotage de l'action IMPG (Informatique, Mathématique et Physique pour la Génomique), du comité technique « Bioinformatique » de GenoPlante. Il est coordinateur de la bioinformatique au sein du Réseau National des Génopoles et responsable de la plate-forme bioinformatique de Rhône-Alpes Genopole. Il est également le responsable scientifique du consortium Genostar. Il est membre de la Commission de Spécialistes (27e section) de l'UFR « Informatique et Mathématiques Appliquées » (université Joseph Fourier).

A. Viari est membre du Comité de Coordination des STIC (CCSTIC) et membre nommé (collège A) à la Section 65 (biologie cellulaire) du Conseil National des Universités.

D. Ziébelin a organisé le *First International Workshop on Bioinformatics ISMIS'02 (XIIIe International Symposium on Methodologies for Intelligent Systems)*, Lyon, 26 juin.

9.2. Enseignements universitaires

C. Gautier, D. Mouchiroud, J. Lobry, L. Gueguen, P. Genoud, D. Ziébelin et M. Page sont enseignants-chercheurs universitaires et assurent un plein service d'enseignant.

L. Duret et G. Perrière enseignent au DEA « Différentiation, génétique et immunologie », Lyon.

M. Gouy enseigne au DEA « Analyse et modélisation des systèmes biologiques », Lyon.

L. Duret, G. Perrière et M. Gouy enseignent en 4e année de la filière « Biochimie » de l'INSA, Lyon.

G. Perrière enseigne en maîtrise en sciences biologiques et médicales, Lyon.

M. Gouy enseigne en maîtrise de biologie des organismes et des populations et en magistère de biologie, Lyon.

H. de Jong a enseigné dans l'option « Bioinformatique » de l'ESIL (4e année) à l'université de la Méditerranée (Marseille) (4h de cours et de TP). Avec C. Hernandez, il est également intervenu dans la filière « Bioinformatique Modélisation » à l'INSA (Lyon), 4e année (12h de cours et de TP).

M.-F. Sagot a enseigné 12h en DEA d'« Informatique Fondamentale et Applications » de l'université de Marne La Vallée ; 20h en 4e année de la filière « Biologie, Informatique et Modélisation » de l'INSA de Lyon ; 2h en 4e année du cours de Magistère en Informatique de l'École Normale Supérieure de Lyon ; 2h en 4e année de l'Institut des Sciences de l'Ingénieur de l'université Blaise Pascal à Clermont Ferrand ; 3h en 4e année de l'Institut National d'Agronomie de Paris-Grimont (INA-PG) et 2h en DEA d'« Informatique » du LIRMM à l'université de Montpellier.

F. Rechenmann a enseigné « La modélisation informatique des connaissances » en maîtrise de biologie, filière « Mathématiques-Informatique », université Claude Bernard, Lyon, (14h).

H. de Jong, A. Morgat, F. Rechenmann, M.-F. Sagot et D. Ziébelin sont intervenus dans l'option « Bioinformatique » de la maîtrise d'« Informatique » à l'université Joseph Fourier (Grenoble).

A. Viari a présenté deux séminaires (janvier 2002 et décembre 2002) sur l'annotation de génomes au DEA « Biologie-Santé » de Montpellier, ainsi qu'un séminaire sur la modélisation au magistère « Biotechnologie-Santé » de Grenoble.

9.3. Participation à des colloques, séminaires, invitations

Séminaire de H. de Jong à l'Institut de Biologie de Lille, 27 février, « Simulation qualitative de l'initiation de la sporulation chez *B. subtilis* ».

Conférence invitée de H. de Jong et de J. Geiselmann,à l'école inter-disciplinaire *Imaging, modelling and manipulating transcriptional regulatory networks (ImmaGene)*, 17-22 octobre, « Qualitative simulation of genetic regulatory networks ».

Interventions de H. de Jong dans la phase théorique et la phase pratique de l'atelier de formation INSERM 138, *From the bioinformatic analysis of functional genomic data to the dynamical analysis of molecular networks*, 11-13 et 16-18 octobre, La-Londe-les-Maures et Marseille, « Simulation qualitative de réseaux de régulation génique ».

Présentation de H. de Jong à l'*European Conference on Mathematical and Theoretical Biology (ECMTB)*, 2-6 juillet, Milan, « Qualitative simulation of the initiation of sporulation in Bacillus subtilis » (avec J. Geiselmann, C. Hernandez et M. Page).

Présentation de H. de Jong au *15th European Conference on Artificial Intelligence (ECAI-02)*, 21-26 juillet, Lyon, « Dealing with discontinuities in the qualitative simulation of genetic regulatory networks » (avec J.-L. Gouzé, C. Hernandez, M. Page, T. Sari et J. Geiselmann).

Présentation de H. de Jong au *16th International Workshop on Qualitative Reasoning (QR 2002)*, 10-12 juin, Sitges, Barcelone, « Dealing with discontinuities in the qualitative simulation of genetic regulatory networks » (avec J.-L. Gouzé, C. Hernandez, M. Page, T. Sari et J. Geiselmann).

Présentation de C. Hernandez aux *Journées Ouvertes Biologie Informatique Mathématiques, JOBIM 2002*, Saint-Malo, 10-12 juin, « VisualGNA : A graphical interface for the qualitative simulation of genetic regulatory networks » (poster réalisé avec H. de Jong, M. Page et J. Geiselmann).

Interventions de J. Geiselmann, M. Page et G. Batt dans le séminaire Interface Biologie, Informatique et Modélisation (IBIM), Montbonnot, 14 juin. « Simulation qualitative des réseaux de régulation génique : le cas de l'initiation de la sporulation chez B. subtilis », « Validation de modèles qualitatifs : Application aux réseaux de gènes ».

Séminaire de G. Batt et H. de Jong au LORIA, Nancy, 20 novembre, « Simulation qualitative de réseaux de régulation génique ».

Séminaires de H. de Jong au Deutsches Krebsforschungszenter (DKFZ) et chez Lion Bioscience, Heidelberg, 19-20 février, « Qualitative simulation of genetic regulatory networks ».

Présentation de H. de Jong au colloque de fin des projets du programme « Bioinformatique » inter-EPST, Ministère de la Recherche, Paris, 14 octobre, « Modélisation et simulation de réseaux de régulation génique : la transduction des signaux par les nucléotides cycliques chez la cyanobactérie Synechocystis PCC 6803 ».

Participation de H. de Jong au jury de thèse de S. Jaeger, université de la Méditerranée, Marseille, 25 octobre.

Présentation de D. Éveillard à l'International Workshop on Regulatory Proteins Interplay and Traffic on DNA, Evry, 12-13 juin, « Modeling the effects of SR proteins on alternative splicing » (poster réalisé avec D. Ropers, H. de Jong et A. Bockmayr).

Intervention d'A. Morgat dans la phase théorique de l'atelier de formation INSERM 138, *From the bioinformatic analysis of functional genomic data to the dynamical analysis of molecular networks*, 11-13 octobre, La-Londe-les-Maures, « Représentation explicite de connaissances biologiques. Application aux données de génomiques et post-génomiques ».

Séminaire d'A. Morgat à l'INSA, Lyon, 15 novembre, « Représentation et intégration de données génomiques ».

Le 21 mai, le consortium Genostar a fait une présentation et une démonstration publiques de la plate-forme dans l'auditorium de l'Institut Pasteur à Paris. Interventions de F. Rechenmann, A. Viari et A. Morgat. Le 14 juin, cette présentation a été faite à nouveau dans l'amphithéatre de l'unité de recherche INRIA Rhône-Alpes.

Séminaire de M.-F. Sagot au Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, « Détection de signaux et de gènes », Montpellier, 24 janvier.

Séminaire de M.-F. Sagot à l'INRIA Rhône-Alpes, « Un lac, deux villes et quatre rivières : Algorithmes combinatoires et biologie moléculaire », 31 janvier.

Séminaire de M.-F. Sagot au Laboratoire de Probabilité, Combinatoire et Statistiques (LACS), « Some approximation results for the Maximum Agreement Forest (MAF) problem », université de Lyon 1, 7 février.

Séminaire de M.-F. Sagot à l'Instituto Gulbenkian de Ciência de Lisbonne, « Algorithmes combinatoires et biologie moléculaire », Lisbonne, 26 novembre.

M.-F. Sagot était orateur invité au Dagstuhl Workshop on Computational Biology, Dagstuhl, Allemagne, 17-22 novembre.

Exposé de F. Rechenmann devant le *Visiting Commitee* de l'INRIA : « Bioinformatics - the Helix research team : modeling and analyzing genomic data », 14 janvier.

Exposés sur la bioinformatique de F. Rechenmann à Autrans, le 13 mars, à l'invitation du CEA/DSV, puis le 20 mars à l'invitation du LETI (CEA).

Présentation de F. Rechenmann des activités bioinformatiques du réseau national des génopoles lors de la rencontre France-Brésil de bioinformatique, organisée à Lyon le 2 avril.

Exposé de F. Rechenmann devant le Conseil de la Recherche de l'Ecole des Mines de Paris, 23 mai, « La bioinformatique : modélisation et analyse des données génomiques ».

Présentation de F. Rechenmann du projet Genostar lors des journées « Génomique sans frontières », à Genève, 24 juin.

Conférence de F. Rechenmann, « Bioinformatics : Modeling and analyzing biological data » au *First International Workshop on Bioinformatics ISMIS'02 (XIIIe International Symposium on Methodologies for Intelligent Systems)*, Lyon, France, 26 juin.

Les 19 et 20 septembre, XRCE et l'INRIA Rhône-Alpes ont organisé conjointement à Grenoble un séminaire international intitulé « Information search and extraction for text documents in life sciences ».

F. Rechenmann a fait une conférence intitulée « From data to knowledge : data mining, text mining and knowledge modeling in biology ».

Le 15 octobre, exposé de F. Rechenmann sur la bioinformatique devant la commission TIC de l'Académie des Technologies, à la Maison de la Chimie à Paris.

Le 12 novembre, présentation de F. Rechenmann de la plate-forme Genostar lors de la journée « Informatique et sciences de la vie » organisée par « L'Usine Nouvelle » et HP, à Paris.

Le 28 novembre, dans le cadre des « 50 ans de l'informatique à Grenoble », exposé de F. Rechenmann sur « La bioinformatique : Modélisation et analyse des données génomiques ».

Le 19 décembre, exposé sur « Bioinformatique : l'INRIA Rhône-Alpes et ses partenaires » lors de la journée de célébration des 10 ans de l'unité de recherche Rhône-Alpes.

Le 15 janvier, 5es Rencontres Plantes-Bactéries, INRA, Aussois, exposé de A. Viari sur l'organisation des génomes bactériens.

Le 8 novembre, workshop ESF *Ontologies for Biology*, Heidelberg, exposés de A. Viari sur l'intégration des données de génomique et post-génomique dans Genostar.

Organisation par D. Ziébelin du séminaire « Travaux de recherche et développement autour de la plate-forme AROM » avec les interventions de C. Capponi, « La classification » ; J. Chabalier, « Isymod : une base de connaissances pour l'analyse des systèmes intégrés » ; J. Gensel, M. Villanova et H. Martin, « Accès progressif dans les représentations de connaissances » ; P. Genoud, « Web-Arom2 » ; D. Ziébelin « LMA2 » ; A. Morgat, « Panoramix : le point de vue d'un utilisateur », Grenoble, 25 et 26 novembre.

Contribution d'A. Culhane, G. Perrière et D. Higgins à l'atelier *Toward the Functional Analysis of Microarrays*, « Between group eigen analysis : a simple and flexible class prediction method for gene expression data », Manchester, 27-28 mars.

Contribution de V. Daubin et G. Perrière à l'*Annual Meeting of the Society for Molecular Biology and Evolution*, « G+C3 structuring along the genome : a universal feature in bacteria », Sorrento, 13-16 juin.

Contribution de L. Duret, C. Gautier et D. Mouchiroud au séminaire *Algorithmique et biologie : Dynamique des Génomes et Evolution*, « Isochore organization of mammalian genomes : selection or neutral evolution ? », Lyon , 3-5 avril.

Présentation de L. Duret, « Detecting genomic features under weak selective pressure : the example of codon usage in animals and plants », *Bioinformatics and Computational Biology*, Madrid (Espagne), 25-26 avril.

Présentation de L. Duret, « La disparition des isochores riches en GC dans les génomes de mammifères », *Petit Pois Déridé : XXIVème réunion du Groupe de Génétique et Biologie des Populations*, Montpellier, 27-30 août.

Présentation de L. Duret, « Detecting genomic features under weak selective pressure : the example of codon usage in animals and plants », *European Conference on Computational Biology*, Saarbrücken (Allemagne), 6-9 octobre.

Présentation de L. Duret, « Vanishing GC-rich isochores in mammalian genomes », *Meeting of the Society for Molecular Biology and Evolution*, Sorrento (Italie), 13-16 juin

Présentation de M. Gouy, « Genome-scale phylogenetic analysis of bacteria », Bioinformatics, Bergen, Norvège, 4-7 avril.

Contribution de J. Lobry et D. Chessel à la *38th Meeting of Polish Society for Biochemistry*, « Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria », Wroclaw, Poland, 18-22 septembre.

Présentation de G. Perrière, « G+C heterogeneity along bacterial genomes : evolutionary implications and consequences for lateral gene transfer detection », *38th Meeting of the Polish Biochemical Society*, Wroclaw, 18-22 septembre.

Séminaire de L. Duret, « Evolution de la composition en base des génomes de mammifères : processus neutre ou sélectif ? », Laboratoire d'Ecologie, université Pierre et Marie Curie, Paris, 19 avril.

Séminaire de L. Duret, « Multiple Sequence Alignment », International School on Computational Biology, Le Havre, 28-31 octobre.

Séminaire de M. Gouy, « Introduction to methods for molecular phylogeny », International School on Computational Biology, Le Havre, 28-31 octobre.

Séminaire de M. Gouy, université de Barcelone (Espagne), « Phylogenetic analysis of the genome of the microsporidian *Encephalitozoon cuniculi* », 15 février.

Séminaire de M. Gouy, université de Genève, département de zoologie, « Analyse phylogénéntique de la séquence complète du génome de la microsporidie *Encephalitozzon cuniculi* », 27 septembre.

Séminaire de G. Perrière, « Banques de données de séquences biologiques », Laboratoire de Spectrométrie de Masse Bio-Organique, Strasbourg, 22 janvier.

Séminaire de G. Perrière, « Organisation of G+C content along the genome : a common feature in bacteria », Department of Biochemistry, université de Cork, 7 octobre.

10. Bibliographie

Articles et chapitres de livre

[1] J. BALTER, A. LABARRE-VILA, D. ZIEBELIN, C. GARBAY. *A knowledge-driven agent-centred framework for data mining in EMG.* in « C.R. Biologies », volume 325, 2002, pages 375-382.

[2] G. BRONNER, B. SPATARO, M. PAGE, C. GAUTIER, F. RECHENMANN. *Modeling comparative mapping using objects and associations.* in « Comput. Chem. », numéro 5, volume 26, 2002, pages 413-420.

[3] C. CHUREAU, M. PRISSETTE, A. BOURDET, V. BARBE, L. CATTOLICO, L. JONES, A. EGGEN, P. AVNER, L. DURET. *Comparative sequence analysis of the X-inactivation centre region in mouse, human and bovine.* in « Genome Res. », volume 12, 2002, pages 894-908.

[4] V. DAUBIN, M. GOUY, G. PERRIÈRE. *A phylogenomic approach to bacterial phylogeny : Evidence for a core of genes sharing common history.* in « Genome Res. », volume 12, 2002, pages 1080-1090.

[5] H. DE JONG. *Modeling and simulation of genetic regulatory systems : A literature review.* in « J. Comp. Biol. », numéro 1, volume 9, 2002, pages 69-105.

[6] J. DUARTE, G. PERRIÈRE, V. LAUDET, M. ROBINSON-RECHAVI. *NuReBase : Database of nuclear hormone receptors.* in « Nucleic Acids Res. », volume 30, 2002, pages 364-368.

[7] L. DURET. *Evolution of synonymous codon usage in metazoans.* in « Curr. Opin. Genet. Dev. », volume 12, 2002, pages 640-649.

[8] J.-R. LOBRY, N. SUEOKA. *Asymmetric directional mutation pressures in bacteria.* in « Genome Biol. », numéro 10, volume 3, 2002.

[9] C. MATHE, M.-F. SAGOT, T. SCHIEX, P. ROUZE. *Current methods of gene prediction, their strengths and weaknesses.* in « Nucleic Acids Res. », numéro 19, volume 30, 2002, pages 4103-4117.

[10] C. MATHÉ, T. SCHIEX, P. ROUZÉ, P. BLAYO, M.-F. SAGOT. *Gene finding in eukaryotes.* éditeurs Q. LU, M. WEINER., in « Cloning and Expression Technologies », Eaton Publishing, 2002.

[11] A. MORGAT, F. RECHENMANN. *Modélisation des données biologiques.* in « Médecine/Sciences », numéro 3, volume 18, 2002, pages 366-374.

[12] C. MOUGEL, J. THIOULOUSE, G. PERRIÈRE, X. NESME. *A mathematical method for determining genome divergence and species delineation using AFLP.* in « Int. J. Syst. Evol. Microbiol. », volume 52, 2002, pages 573-586.

[13] G. PERRIÈRE, J. THIOULOUSE. *Use and misuse of correspondence analysis in codon usage studies.* in « Nucleic Acids Res. », volume 30, 2002, pages 4548-4555.

[14] G. PIGANEAU, D. MOUCHIROUD, L. DURET, C. GAUTIER. *Expected relationship between the silent substitution rate and the GC content : Implications for the evolution of isochores.* in « J. Mol. Evol. », volume 54, 2002, pages 129-133.

[15] N. PISANTI, M.-F. SAGOT. *Further thoughts on the syntenic distance between genomes.* in « Algorithmica », volume 34, 2002, pages 157-180.

[16] L. PONGER, D. MOUCHIROUD. *CpGProD : Identifying CpG islands associated with transcription start sites in large genomic mammalian sequences.* in « Bioinformatics », numéro 4, volume 18, 2002, pages 631-633.

[17] F. RECHENMANN. *GenoStar : A bioinformatics platform for exploratory genomics.* in « ERCIM news », october, 2002.

[18] C. RIZZON, G. MARAIS, M. GOUY, C. BIÉMONT. *Recombination rate and the distribution of transposable elements in the Drosophila melanogaster genome.* in « Genome Res. », numéro 3, volume 12, 2002, pages 400-407.

[19] M.-F. SAGOT, Y. WAKABAYASHI. *Pattern inference under many guises.* éditeurs B. REED, C. SALES., in « Recent Advances in Algorithms and Combinatorics », Springer-Verlag, Berlin, 2002.

[20] F. TAHI, M. GOUY, M. REGNIER. *Automatic RNA secondary structure prediction with a comparative approach.* in « Comput. Chem. », numéro 5, volume 26, 2002, pages 521-530.

[21] D. THIEFFRY, H. DE JONG. *Modélisation Analyse et simulation des réseaux génétiques.* in « Médecine/Sciences », numéro 4, volume 18, 2002, pages 492-502.

[22] C. VIVARÈS, M. GOUY, F. THOMARAT, G. MÉTÉNIER. *Functional and evolutionary analysis of a eukaryotic parasitic genome.* in « Curr. Opin. Microbiol. », volume 5, 2002, pages 499-505.

Communications à des congrès, colloques, etc.

[23] V. DAUBIN, G. PERRIÈRE. *G+C3 structuring along the genome : A universal feature in Bacteria.* in « Actes des 3es Journées Ouvertes : Biologie, Informatique et Mathématiques (JOBIM 2002) », IMPG, éditeurs J. NICOLAS, C. THERMES., pages 65-73, Rennes, 2002.

[24] H. DE JONG, J.-L. GOUZÉ, C. HERNANDEZ, M. PAGE, T. SARI, H. GEISELMANN. *Dealing with disconti-*

nuities in the qualitative simulation of genetic regulatory networks. in « Working Notes of 16th International Workshop on Qualitative Reasoning, QR 2002 », éditeurs N. AGELL, J. ORTEGA., pages 67-74, Sitges, Barcelona, Spain, 2002.

[25] H. DE JONG, J.-L. GOUZÉ, C. HERNANDEZ, M. PAGE, T. SARI, J. GEISELMANN. *Dealing with discontinuities in the qualitative simulation of genetic regulatory networks.* in « Proceedings of 15th European Conference on Artifical Intelligence, ECAI-02 », IOS Press, éditeurs F. VAN HARMELEN., pages 412-416, Amsterdam, 2002.

[26] J.-F. DUFAYARD, L. DURET, G. PERRIÈRE, F. RECHENMANN. *Pattern matching in phylogenetic trees.* in « Actes des 3es Journées Ouvertes : Biologie, Informatique et Mathématiques (JOBIM 2002) », IMPG, éditeurs J. NICOLAS, C. THERMES., pages 239-245, Rennes, 2002.

[27] J. GRASSOT, G. PERRIÈRE, D. MOUCHIROUD. *RTKdb : Database of receptor tyrosine kinase.* in « Actes des 3es Journées Ouvertes : Biologie, Informatique et Mathématiques (JOBIM 2002) », IMPG, éditeurs J. NICOLAS, C. THERMES., pages 199-213, Rennes, 2002.

[28] C. HERNANDEZ, H. DE JONG, J. GEISELMANN, M. PAGE. *VisualGNA : A graphical interface for the qualitative simulation of genetic regulatory networks.* in « Actes des 3es Journées Ouvertes : Biologie, Informatique et Mathématiques (JOBIM 2002) », IMPG, éditeurs J. NICOLAS, C. THERMES., pages 335-336, Rennes, 2002.

[29] C. RIZZON, G. BRONNER, M. GOUY, C. BIÈMONT. *GATE : An object-oriented model dedicated to the analysis of transposable elements in eukaryote genomes.* in « Actes des 3es Journées Ouvertes : Biologie, Informatique et Mathématiques (JOBIM 2002) », IMPG, éditeurs J. NICOLAS, C. THERMES., Rennes, 2002.

[30] C. RIZZON, M. MARAIS, M. GOUY, C. BIÈMONT. *Recombination rate and the distribution of transposable elements in the Drosophila melanogaster genome.* in « Molecular Evolution, Evolution, Genomics, Bioinformatics (ISME/SMBE) », Sorrento, 2002.

Rapports de recherche et publications internes

[31] H. DE JONG, J. GEISELMANN, G. BATT, C. HERNANDEZ, M. PAGE. *Qualitative simulation of the initiation of sporulation in B. subtilis.* rapport technique, numéro RR-4527, INRIA, 2002, http://www.inria.fr/rrrt/rr-4527.html.

[32] H. DE JONG, J.-L. GOUZÉ, C. HERNANDEZ, M. PAGE, T. SARI, H. GEISELMANN. *Qualitative simulation of genetic regulatory networks using piecewise-linear models.* rapport technique, numéro RR-4407, INRIA, 2002, http://www.inria.fr/rrrt/rr-4407.html.