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

Responsable

François Rechenmann [directeur de recherche, INRIA]

Chercheurs, enseignants-chercheurs et ingénieurs permanents

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

Christophe Bruley [ingénieur expert, INRIA]
Stéphane Bruley [ingénieur associé, INRIA]
Stéphane Declere [ingénieur expert, INRIA]
Véronique Dupierris [ingénieur expert, INRIA]
Gaël Faroux [ingénieur expert, INRIA]
Céline Hernandez [ingénieur associé, INRIA]
Anne Morgat [chercheur associé, INRIA]

Post-doctorants

Sandrine Hughes [ATER UCBL]
Violaine Pillet [INRIA]
Nadia Pisanti [bourse ERCIM]
Sébastien Provencher [bourse du Gouvernement du Québec]

Doctorants

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]
Vincent Daubin [allocataire du Ministère de la Recherche, directeurs de thèse : Guy Perrière, Manolo Gouy]
Jean-François Dufayard [allocataire du Ministère de la Recherche, directeurs de thèse : Manolo Gouy, 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]
Thibault Parmentier [allocataire du Ministère de la Recherche, directrice de thèse : Danielle Ziébelin]
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]
Fabienne Thomarat [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 [ISB-SIB, Genève]
Charles Metivier [société Aureus Pharma, Paris (à temps partiel)]
Erwan Reguer [CEA, Grenoble]

Assistante de projet

Françoise de Coninck

2 Présentation et objectifs généraux

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

2.1 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) et analyse (algorithmes). 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 et de transmission de l'information génétique, ainsi que les relations que ces entités entretiennent entre elles, en particulier au sein des réseaux géniques et métaboliques.

2.2 Axes de recherche

Les travaux de l'équipe se structurent ainsi en onze axes principaux : annotation des génomes et modélisation des gènes ; organisation des génomes et cartographie comparée ; cartographie comparée, phylogénie et évolution ; modélisation de processus évolutifs ; modélisation dynamique des réseaux d'interaction géniques ; modélisation du métabolisme intermédiaire ; modélisation des données de protéome et protéomique expérimentale ; extraction d'informations à partir de textes ; environnement didactique en bioinformatique.

2.2.1 Modélisation des gènes et inférence de motifs

Participants : Laurent Guégen, Guy Perrière, Nadia Pisanti, Marie-France Sagot [Correspondant], Alain Viari [Correspondant].

La prédiction de gènes (en particulier ceux codants pour les protéines) à partir d'un génome brut est une problématique située très en amont au sein d'HELIX puisqu'elle va alimenter de nombreuses autres problématiques du projet. Cette question se pose en des termes différents suivant la nature de l'organisme étudié. Pour ce qui concerne les eucaryotes, un gène peut être vu comme une suite de facteurs particuliers, appelés exons, séparés par les introns. Les contraintes lexicales strictes portant sur ces facteurs ne sont généralement pas suffisantes pour déterminer sans ambiguïté ou erreur la structure d'un gène. Une de nos hypothèses de travail est que les suites d'exons d'exemplaires différents d'un même gène (sur un même génome ou sur des organismes différents) sont mieux conservées au cours de l'évolution que les autres parties du génome.

En termes algorithmiques, identifier un gène eucaryote peut donc être vu comme un problème d'optimisation combinatoire d'une analyse (*parsing*) simultanée sur deux chaînes couplée à une comparaison de ces chaînes. Concernant les procaryotes, la question ne se pose pas en terme de recherche de la structure du gène (dépourvue d'exons dans ce cas) mais en terme de point de référence. En effet, les meilleures méthodes actuellement utilisées (basées sur l'emploi des chaînes de Markov) nécessitent une étape d'apprentissage à partir de gènes connus, étape problématique dans le cas de gènes atypiques (par leur taille ou leur composition). Nos travaux dans ce domaine visent à éviter cette étape d'apprentissage, en particulier en couplant l'utilisation de chaînes de Markov avec des algorithmes de classification non supervisée.

Un autre champ entrant dans le cadre de l'annotation de génomes, concerne la recherche de structures exceptionnelles (au sens d'un modèle ; d'une certaine façon, les gènes sont un cas particulier de telles structures exceptionnelles). Ce champ touche, plus généralement, à l'analyse exploratoire de données et met donc en œuvre des méthodes statistiques. Nous nous intéressons, dans ce cadre, plus particulièrement à la mise en œuvre de l'analyse discriminante des correspondances ainsi qu'aux méthodes de segmentation sous contraintes.

2.2.2 Inférence de motifs et de structures

Participant : Marie-France Sagot [Correspondant].

L'inférence de motifs est importante pour l'étude de la régulation de l'expression et de la fonction des gènes et pour la détermination de la structure des génomes. Ces motifs cor-

respondent soit à des sites d'interaction de macromolécules impliquées dans divers processus biologiques, soit à des segments apparaissant répétés, de manière dispersée ou contiguë, le long d'un génome.

Ces éléments ont été conservés au cours de l'évolution à cause de leur rôle fonctionnel, cependant cette conservation peut n'être que partielle, voire impliquer des caractéristiques plus physico-chimiques que la séquence elle-même.

La diversité des conditions et environnements de vie des espèces exige une flexibilité du vivant face aux changements rapides qui peuvent survenir, et une richesse dans le répertoire des moyens pouvant conduire à la survie de l'espèce. Cette flexibilité et richesse se traduisent par une combinatoire des éléments, agissant souvent de manière synergique, impliqués dans certains processus biologiques. Les motifs à inférer sont ainsi souvent multiples, c'est-à-dire composés de plusieurs motifs simples que l'on appelle *boîtes*. L'ordre des boîtes d'un motif multiple est important, de même que la distance entre deux motifs. Nous avons appelé ces motifs *structurés*.

Enfin, un nouvel axe de travail concernant l'inférence de structures secondaires d'ARN est également en cours de développement.

2.2.3 Organisation des génomes et cartographie comparée

Participants : Gisèle Bronner, Christophe Bruley, Christian Gautier [Correspondant], Adel Khelifi, Anne Morgat [Correspondant], Dominique Mouchiroud, Bruno Spataro, Alain Viari.

La cartographie génomique comparée vise à associer les séquences génomiques et ce que l'on sait de l'information qu'elles contiennent à leur organisation spatiale dans différentes espèces. En effet, au cours de l'évolution, l'organisation générale des génomes est remaniée par des réarrangements des chromosomes : un fragment d'un chromosome peut venir s'insérer dans un autre chromosome. Il n'y a donc pas de correspondance simple entre les chromosomes de différentes espèces, y compris au sein d'un groupe taxonomique comme les mammifères. Cependant, ces modifications laissent intacts des fragments relativement petits, les segments conservés, au sein desquels on retrouve les gènes orthologues dans des arrangements voisins. La prise en compte simultanée des relations d'homologie entre gènes et entre segments conservés est essentielle dans le processus d'enrichissement des informations sur un génome à partir des informations sur un autre génome.

Au niveau informatique, la cartographie comparée soulève de nombreux problèmes, les plus importants étant des problèmes algorithmiques liés à la comparaison de permutations d'objets (en l'occurrence les gènes au sein d'un segment conservé ou d'une structure plus large comme un chromosome) et des problèmes de représentation et de gestion de connaissances très complexes, à la fois par la structure des objets en cause et par le nombre des relations qui les relient. Par ailleurs, l'existence de multiples types de repérage le long des génomes (cartes cytogénétiques, génétiques, d'hybrides d'irradiation, etc.), la prise en compte des différents types d'homologie, des pseudogènes et des différentes séquences répétées conduit à un système complexe d'entités et de relations qu'il est nécessaire de modéliser précisément. Nos travaux, dans ce domaine, portent donc sur l'élaboration de modèles permettant de représenter ces connaissances dans

le cas des organismes eucaryotes et procaryotes. Dans ce dernier cas, nous nous sommes plus 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).

2.2.4 Cartographie comparée, phylogénie et évolution

Participants : Jean-François Dufayard, Laurent Duret, Manolo Gouy [Correspondant], Guy Perrière, Nadia Pisanti, Marie-France Sagot [Correspondant], Marina Zelwer.

Ce thème prolonge le précédent par la prise en compte explicite de l'aspect temporel des phénomènes biologiques. La reconstitution des relations temporelles à partir des données actuelles constitue le question centrale de la phylogénie et donc une clef pour comprendre les phénomènes évolutifs. Cette reconstitution peut impliquer un calcul de distance entre organismes ou la reconstruction des évènements qui ont eu lieu. Les évènements considérés peuvent être ponctuels (mutations) ou impliquer des segments entiers d'un génome (remaniements ou réarrangements). Parallèlement aux travaux coutumiers de reconstitution phylogénétique, nos plus récents travaux dans ce domaine ont porté sur : (1) la mise au point d'un algorithme de réconciliation d'arbres phylogénétique, par l'introduction d'un nombre minimal de noeuds de duplications dans un arbre de spéciation ; (2) la localisation des points de remaniements ; (3) le calcul d'une distance de remaniements entre deux arbres phylogénétiques et (4) le calcul d'une distance de remaniements entre génomes multi-chromosomiques.

2.2.5 Modélisation de processus évolutifs

Participants : Manolo Guy, Jean Lobry [Correspondant], Guy Perrière, Gwenael Piganeau.

Ce thème porte essentiellement sur la modélisation de processus évolutifs (au niveau mutationnel et/ou sélectif) permettant d'expliquer les variations observées de la composition en bases le long des génomes bactériens. Outre l'aspect fondamental du problème (comprendre l'origine de ces processus), des applications pratiques ont d'ores et déjà émergé, concernant, entre autres, la prédiction de la localisation de l'origine de réplication des chromosomes bactériens et la mise en évidence de l'influence de l'hétérogénéité compositionnelle sur la détection des transferts horizontaux de gènes. Enfin, la modélisation de l'évolution de fragments génomiques soumis à la fois à des processus de mutation et de sélection a permis une quantification de l'influence des recombinaisons dans l'efficacité de la sélection.

2.2.6 Modélisation de familles de séquences homologues

Participants : Laurent Duret [Correspondant], Manolo Guy, Guy Perrière [Correspondant].

L'étude des familles de gènes homologues présents dans les génomes de divers organismes est l'un des outils les plus puissants de l'analyse comparative des génomes. C'est pourquoi nous développons des statégies et des outils informatiques permettant de classifier systématiquement

tous les gènes connus dans les banques de données de séquences en familles de gènes homologues. Actuellement, ce travail est réalisé dans deux contextes biologiques différents : d'une part pour les gènes de vertébrés, avec la banque HOVERGEN, et d'autre part pour les gènes de procaryotes (bactéries et archées) avec la banque HOBACGEN. Ces systèmes informatiques sont constitués d'une partie génération des familles homologues (comparaison de toutes les séquences avec elles-mêmes, délimitation des familles, alignement multiple de leurs membres, calcul de leur phylogénie) et d'une partie consultation de la banque. Les développements récents ont concerné la migration de la banque HOVERGEN vers l'architecture développée pour HOBACGEN, permettant ainsi un accès en mode client-serveur à la banque HOVERGEN. Ces banques ont d'autre part été plusieurs fois mises à jour au cours de l'année 2001, opérations demandant chacune de l'ordre de deux semaines de temps de calcul et beaucoup d'expertise manuelle des données générées. Le logiciel d'interrogation de ces banques fait l'objet d'une diffusion très demandée par la communauté (par exemple le serveur HOBACGEN a été installé dans 65 laboratoires au plan international).

2.2.7 Modélisation dynamique des réseaux de régulation génique

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

La plupart des propriétés importantes d'un organisme vivant émerge 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 positives et négatives, 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 leur application en collaboration avec des biologistes.

2.2.8 Modélisation des données de protéome et protéomique expérimentale

Participants : Anne Morgat, 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 (c'est-à-dire les catalyseurs biologiques des réactions métaboliques évoquées au paragraphe précédent). 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 HAMAP (cf. section collaborations internationales) visant la ré-annotation de l'ensemble des protéomes bactériens (une soixantaine à ce jour).

Parallèlement, nous nous intéressons, en collaboration avec le CEA Grenoble et dans le cadre de la plateforme de protéomique rhône-alpine, 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.

2.2.9 Modélisation de données génomiques et post-génomiques : projet Panoramix

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

La communauté des biologistes dispose à l'heure actuelle de plus de 50 génomes procaryotes entièrement séquencés et plusieurs centaines de projets de séquençage sont en cours d'achèvement (`http://wit.integratedgenomics.com/GOLD`).

Paradoxalement, si l'obtention de la séquence chromosomique ne pose plus d'importantes difficultés techniques, l'annotation de ces génomes reste encore problématique. Du point de vue des objectifs biologiques, on peut distinguer trois niveaux de complexité croissante :

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, 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.
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.
3. l'annotation relationnelle concerne l'identification des relations existantes entre les objets caractérisés (individuellement) aux niveaux précédents (1 et 2). Ces relations sont de nature diverse. 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 objets manipulés à ce

niveau présentent généralement un plus haut degré d'abstraction et de structuration (par exemple un graphe décrivant un réseau métabolique).

L'assignation de fonctions aux produits des gènes (annotation fonctionnelle) s'effectue encore essentiellement par recherche de similarité avec des séquences existantes et n'exploite que peu ou pas les relations qui viennent d'être évoquées, par exemple, on n'exploite encore que trop peu systématiquement le fait que les enzymes intervenant dans les mêmes voies métaboliques tendent à être regroupés en opérons. Les informations qui doivent être manipulées à ce niveau d'annotation, que nous qualifions de niveau « relationnel » – opérons, régulons, graphes représentant des chemins réactionnels, des assemblages moléculaires – sont plus complexes que les seules données de séquences et réclament donc un traitement particulier.

Actuellement, le traitement de ces informations pose deux problèmes majeurs : le premier concerne leur représentation formelle, c'est-à-dire leur modélisation, et le second 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 (`http://ecocyc.panbio.com`) ou KEGG (`http://www.genome.ad.jp/kegg/`) pour les données métaboliques, RegulonDB (`http://www.cifn.unam.mx/Computational_Genomics/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). Après la question de modélisation des données biologiques qui impose une importante activité de formalisation du domaine biologique que l'on cherche à représenter, vient le problème de l'instanciation des données. Pour être de bonne qualité, cette étape nécessite absolument une expertise humaine approfondie, un travail qui n'a, jusqu'à présent, été réalisé que sur peu d'organismes et dans des domaines biologiques délimités (EcoCyc et RegulonDB pour la bactérie modèle *E. coli* par exemple).

Dans le cadre du projet Panoramix, nous nous concentrons plus particulièrement autour des trois problématiques suivantes :

1. Génomique comparative et étude des synténies bactériennes
2. Modélisation du métabolisme intermédiaire
3. Modélisation des données concernant les protéines et les complexes protéiques

Il s'agit, pour chacune de ces problématiques, de formaliser le domaine biologique correspondant, de le traduire en un schéma de classes/relations, d'instancier ce schéma à partir de données publiques (essentiellement à partir de bases de données existantes) ou de résultats de calculs et de développer un ensemble d'outils d'exploitation des données (outils de requête et de visualisation).

2.2.10 Extraction d'informations à partir de textes

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

Si les bases de données biologiques 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 des 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. Plusieurs équipes se sont ainsi concentrées sur le problème d'extraire des données sur les interactions entre gènes ou entre protéines. 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. La difficulté du problème résulte de la fâcheuse tendance des chercheurs à faire usage, pour nommer les entités biologiques dans leurs publications, de variantes lexicales ou d'abréviations qui leur sont souvent propres ; de plus, l'importance du flux de nouvelles données expérimentales entraîne un décalage dans la mise à jour des « dictionnaires » de noms. L'approche repose par conséquent sur l'utilisation conjointe de techniques d'analyse de textes (lexiques de la langue, shallow parsing), de lexiques spécialisés et d'algorithmes de comparaison et de recherche de noms voisins (typiquement par des méthodes de programmation dynamique). Ces techniques devraient ainsi permettre de détecter des noms d'entités biologiques non répertoriés dans les lexiques spécialisés.

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).

2.2.11 Environnement didactique en bioinformatique

Participants : Gaël Faroux, Philippe Genoud [Correspondant], 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 œuvre faciliterait l'emploi de celles-ci ainsi que l'obtention de résultats plus pertinents. L'objectif de ce projet est donc le développement d'un environnement didactique pour expliquer les algorithmes de bio-informatique, indépendamment les uns des autres ou au travers de stratégies y faisant appel. Ce logiciel est un des éléments du projet *Ecole de l'ADN* proposé par le Centre de Culture Scientifique Technique et Industrielle (CCSTI) de Grenoble. Il est destiné à un public assez large et s'adresse principalement aux étudiants de terminale et premières années d'études supérieures en biologie. Il pourrait aussi être utilisé comme complément dans des filières d'informatique proposant des options en bio-informatique, voire comme outil d'auto-formation.

2.3 Relations internationales et industrielles

Le groupe d'Amos Bairoch à l'Institut Suisse de BioInformatique (SIB) à Genève est équipe associée avec le projet HELIX. Les deux groupes entretiennent des liens forts sur les thèmes des protéomes bactériens (projet HAMAP) et 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 est en contact avec les différentes équipes de bioinformatique françaises, en particulier au sein des différentes génopoles.

Il participe également au réseau ESF (European Science Foundation) intitulé *Experimental and in silico Analysis of Biomolecular Interactions*.

Au premier rang des relations industrielles figurent les partenaires Genome Express et Hybrigenics du consortium GénoStar.

Le projet HELIX bénéficie d'un contrat dans le cadre de GénoPlante, en partenariat avec des équipes de l'INRA (Toulouse et Gand) et de l'Institut Pasteur (Paris).

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 à 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 cœur 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 œuvre 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'applications

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.

Plusieurs dizaines 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. cerevisiæ*, 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 précèdé de quelques mois celui de l'Homme (plus de trois milliards de lettres) dont il n'existe actuellement qu'une version de travail préliminaire (*draft*).

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. Encore au delà, 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 (les séquences). D'autres classes de données doivent être également

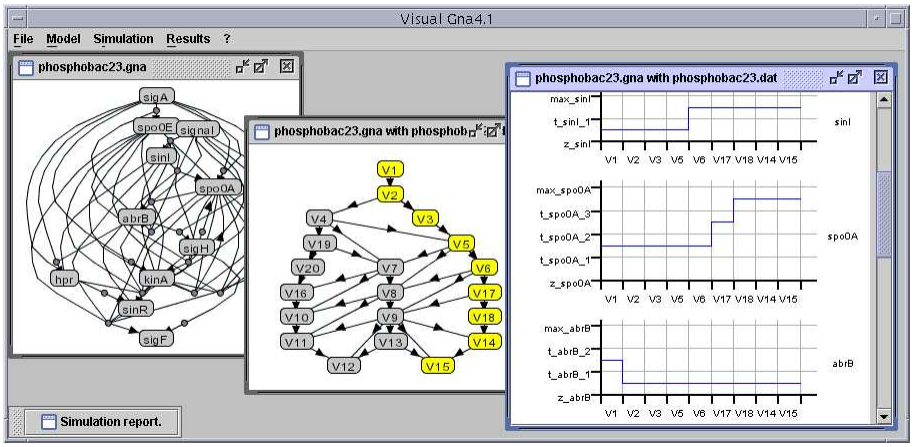

FIG. 1 – Simulation d'un réseau de régulation génique en utilisant GNA. La fenêtre de gauche montre les gènes et les interactions du réseau contrôlant l'initiation de la sporulation chez *B. subtilis*. La fenêtre du milieu contient le graphe des transitions entre états résultant de la simulation du réseau pour des conditions initiales induisant la sporulation, tandis que la fenêtre de droite montre une séquence d'états qualitatifs sélectionnée dans le graphe des transitions.

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 (*proteomics*) ou de la spectrométrie de masse, 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 : 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 (figure 1). La version 4.1 de GNA a été déposée auprès de l'APP. Elle est disponible à travers le Web (`http://www-helix.inrialpes.fr/gna`).

5.2 Druid

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

Druid est un algorithme qui prend en entrée un alignement de séquences nucléiques d'es-

pèces proches ou de souches différentes d'une même espèce et détecte la présence éventuelle de points de recombinaison le long de l'alignement de même que leur localisation. Druid utilise deux algorithmes disponibles dans PAUP (un logiciel regroupant plusieurs algorithmes d'analyse phylogénétique) : DnaPars et le test de Farris. L'algorithme est utilisable à travers une interface web (`http://bioweb.pasteur.fr/seqanal/interfaces/druid.html`). Le développement du logiciel Druid a été mené à l'Institut Pasteur et à l'université de Marne-la-Vallée ; il se poursuit maintenant au sein d'HELIX à Lyon.

5.3 Environnement Didactique en Bioinformatique (EDB)

Participants : Gaël Faroux, Philippe Genoud [Correspondant], 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 effectuée au moyen d'outils d'édition standards sans recours à un langage de programmation. L'EDB est donc une instanciation de ce framework, dédiée à la biologie moléculaire. Actuellement plusieurs algorithmes (avec leurs interfaces explicatives) pour la recherche de zones codantes dans des séquences génomiques ont d'orés et déjà été développés et intégrés : recherche de motifs dans une séquence, analyse du biais de codage par le test du Chi 2, 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 (figure 2).

5.4 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. Le développement de Satellites a été dirigé par Marie-France Sagot, dans le contexte de sa collaboration avec l'université de Marne-la-Vallée.

5.5 Panoramix

Participants : Frédéric Boyer, Stéphane Bruley, Anne Morgat [Correspondante], Alain Viari.

Le projet *Panoramix* vise à fédérer les activités de modélisation autour des trois problématiques suivantes :

1. Génomique comparative et étude des synténies bactériennes

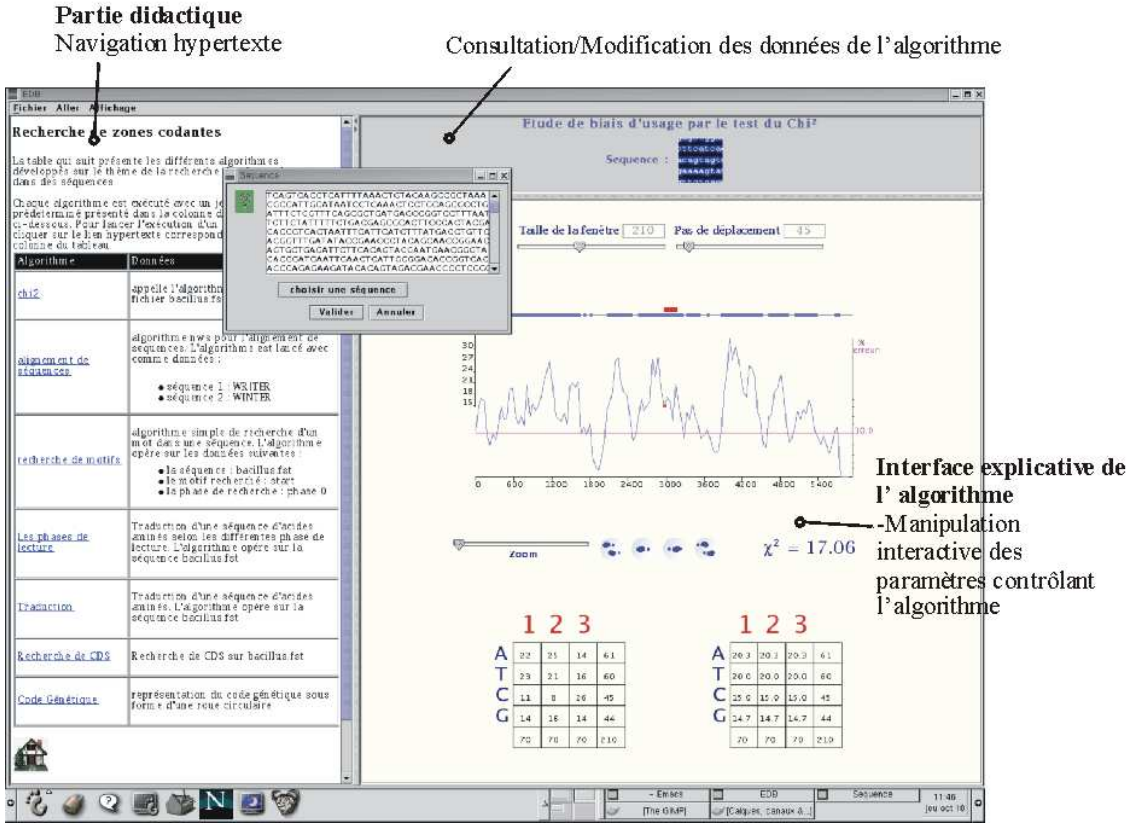

FIG. 2 – Les différents composants de l'interface graphique de l'Environnement Didactique en Bioinformatique.

2. Modélisation du métabolisme intermédiaire
3. Modélisation des données concernant les protéines et les complexes protéiques

D'un point de vue pratique, le développement de ces thèmes s'est concrétisé par la conception et l'implémentation de quatre bases de connaissances spécialisées. Ces bases ont été implémentées à l'aide du système de représentation par objets AROM développé dans l'action Romans (INRIA-RA). Dans ce système, inspiré d'UML, les associations entre les objets sont représentées explicitement, elles peuvent, comme les objets, posséder des attributs (ce qui permet de caractériser une relation, indépendamment des entités qu'elle connecte), enfin les associations peuvent être n-aires. Le système AROM présente d'autres caractéristiques qui s'avèrent très utiles dans la pratique, comme la possibilité d'attacher une expression algébrique à un attribut.

Genomix est une base rassemblant les informations concernant les gènes de 49 génomes bactériens complets ainsi que les informations d'orthologie, de paralogie et de synténie bactériennes portant sur ces gènes. Cette base est instanciée à partir des données publiques produites pour chacun des projets de séquençage. Les données d'homologie et de synténies bactériennes résultents de calculs effectués au sein du projet.

Metabolix est une base rassemblant les données du métabolisme intermédiaire, soit l'ensemble des réactions biochimiques, des enzymes catalysant ces réactions et des principales voies métaboliques. Cette base est instanciée à partir des bases de données KEGG (`http://www.genome.ad.jp/kegg/`) et ENZYME (`http://www.expasy.org/enzyme/`).

Organix est une base rassemblant les informations sur plus de 5000 composés chimiques impliqués dans le métabolisme cellulaire.

Protéix est une base rassemblant les informations sur les protéines. Les assemblages moléculaires y sont représentés explicitement. Cette base est en cours d'instanciation.

La figure 3 présente un résumé synoptique de la situation relative et de la complémentarité des quatre bases.

5.6 Smile

Participant : Marie-France Sagot [Correspondante].

Smile est un algorithme d'inférence de motifs à partir d'un ensemble de séquences biologiques (nucléiques ou de protéines), développé avec Gene Myers dans le contexte de la collaboration entre l'Institut Pasteur et l'université de Marne-la-Vallée. Il permet d'identifier des motifs, écrits sur l'alphabet des séquences ou sur un alphabet physico-chimique (*i.e.* 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 à 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 l'algorithme est de déterminer toutes celles pour lesquelles les motifs continuent de vérifier les contraintes introduites.

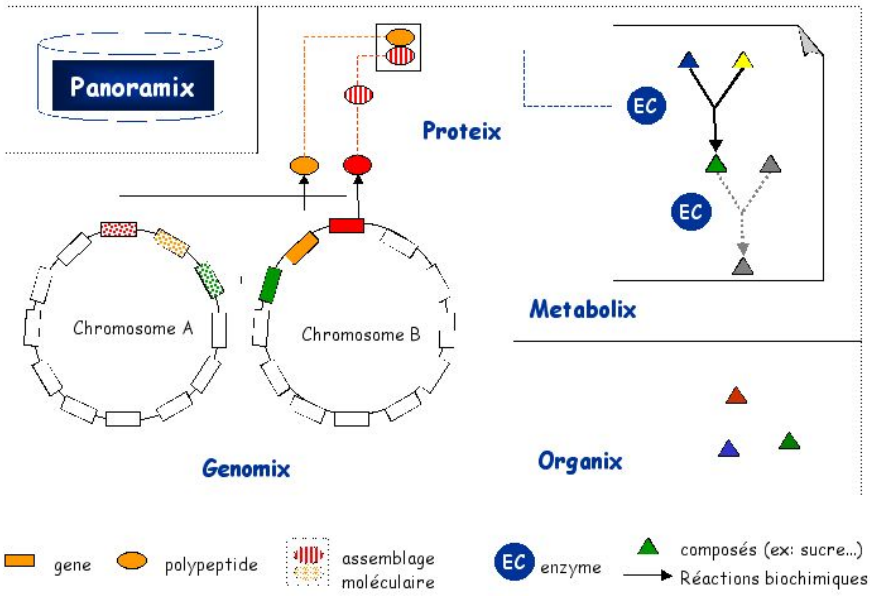

FIG. 3 – Projet Panoramix : Intégration des bases de connaissances Genomix, Proteix, Metabolix et Organix.

5.7 EMKov

Participant : Alain Viari [Correspondant].

EMKov est un logiciel de recherche de gènes bactériens développé par Jean Romanet dans le cadre de son stage de DEA. EMKov combine l'utilisation de chaînes de Markov et l'algorithme EM (Expectation Maximisation) et 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, montrent qu'il présente une sensibilité supérieure à 95

5.8 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.

5.9 Utopia

Participant : Marie-France Sagot [Correspondante].

Utopia est un algorithme de détection de gènes, développé avec Philippe Blayo dans le contexte de la collaboration entre l'Institut Pasteur et l'université de Marne-la-Vallée. 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 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 (e.g. ATG), se termine par un codon STOP et chacun de ses exons est bordé à gauche et à droite respectivement par AG et GT.

5.10 PBIL

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

PBIL est un site web (`http://pbil.univ-lyon1.fr/`) pour l'analyse bio-informatique 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.

5.11 HOBACGEN et HOVERGEN

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

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). Les bases sont consultables en mode client-serveur (`http://pbil.univ-lyon1.fr/databases/hobacgen.html`).

5.12 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`).

5.13 JADIS

Participant : Dominique Mouchiroud [corréspondant].

JADIS est une application Java de calcul de distances entre séquences, développé en collaboration avec Isabelle Gonçalves (`http://pbil.univ-lyon1.fr/software/jadis.html`).

6 Résultats nouveaux

6.1 Modélisation des gènes et inférence de motifs

Participants : Laurent Guégen, Guy Perrière, Nadia Pisanti, Marie-France Sagot [Correspondant], Alain Viari [Correspondant].

6.1.1 Prédiction des gènes

Utopia est un algorithme exact de recherche de gènes eucaryotes par *pure homology* réalisé en collaboration avec Philippe Blayo (doctorant de l'université de Marne-la-Vallée, co-encadré par Maxime Crochemore et Marie-France Sagot). L'algorithme effectue une optimisation combinatoire d'une analyse (*parsing*) simultanée sur deux chaînes couplée à une comparaison de ces chaînes. L'analyse se base sur un modèle de gène générique et complètement indépendant de l'organisme eucaryote considéré. L'algorithme a une complexité en temps proportionnelle au produit des longueurs de deux séquences et, ce qui est souvent plus important en pratique, une complexité en espace linéaire en la longueur de la plus courte des deux chaînes.

Utopia a été testé sur plusieurs familles intra ou inter-espèces de gènes de plantes. La structure (*i.e.* le nombre d'exons) des gènes appartenant à une même famille intra-espèce (on parle de gènes paralogues, c'est-à-dire ayant un ancêtre commun dont ils dérivent par duplication) ou inter-espèces (on parle dans ce cas de gènes orthologues, c'est-à-dire ayant un ancêtre commun dont ils dérivent par spéciation) peut être différente. Le degré de similarité entre les gènes d'une même famille varie de 90% à 30%. Ces gènes peuvent en outre présenter des erreurs de séquençage. Les résultats obtenus sont très prometteurs. Malgré le fait que les débuts et fins des gènes sont souvent mal localisés, nos résultats semblent dans tous les cas aussi bons que ceux des meilleurs programmes disponibles de prédiction de gènes par comparaison de séquences génomiques, y compris les plus récents. Ces résultats sont dans certains cas bien meilleurs, soit parce que la plupart des autres méthodes ne permettent pas d'identifier des gènes par une approche comparative lorsque ces gènes ont une structure différente, soit parce que notre algorithme est à la fois plus sensible et plus général. Nos temps de calcul sont souvent

plus lents mais notre méthode est la seule exacte existant actuellement, et, surtout, la seule à n'utiliser aucune autre information afin d'inférer les gènes. Cela devrait permettre à Utopia d'identifier des gènes ayant des caractéristiques inhabituelles par rapport à la moyenne (les autres algorithmes se basent en effet toujours sur un apprentissage au préalable de cette « moyenne » afin de localiser de nouveaux gènes).

De très nombreuses perspectives concernent ce sujet. La première vise à améliorer l'algorithme de comparaison deux-à-deux pour le rendre plus efficace et également plus souple en permettant, par exemple, l'identification de plus d'un gène à la fois, ainsi que de gènes incomplets ou que la nature peut épisser (c'est-à-dire, « parser ») de diverses façons différentes. Nous commençons également à explorer une méthode simultanée d'inférence de motifs et inférence de gènes afin de permettre une approche multiple pour une telle prédiction. Les résultats préliminaires obtenus avec Utopia ont servi à montrer qu'une telle démarche permettrait d'améliorer la détermination des exons constituant un gène, en particulier d'affiner le positionnement des frontières exactes de ces exons, tout en demeurant dans le cadre d'une approche complètement générique de cette prédiction (c'est-à-dire, à la fois indépendante de tout organisme et ne faisant pas appel à des connaissances « apprises » au préalable sur des exemples).

6.1.2 Inférence de motifs structurés

L'identification de motifs structurés peut être abordée à plusieurs niveaux de complexité. Le plus simple part du texte génomique vu comme une chaîne de caractères sur un alphabet Σ à quatre lettres (les nucléotides des molécules d'ADN ou d'ARN) ou à vingt lettres (les acides aminés composant les protéines) et représente les motifs correspondant aux sites ou répétitions à identifier comme des mots définis soit sur Σ, soit sur $\mathcal{P}(\Sigma)$, c'est-à-dire, sur l'ensemble de tous les sous-ensembles de Σ. Ces mots ont des occurrences dans une ou plusieurs chaînes. Chaque occurrence peut présenter un nombre limité de différences (substitutions, insertions, suppressions) par rapport au motif m (c'est-à-dire, être à une certaine distance d'édition de m) . Dans le cas où l'alphabet des motifs est $\mathcal{P}(\Sigma)$, la distance entre un motif m et un facteur u d'une chaîne devient le minimum des distances entre u et les mots dans m.

Formellement, un motif structuré est donc une paire (m, d) où :

- m est une p-tuple de motifs simples $(m_1, \ldots, m_p)$ (les p boîtes) ;
- d est une $(p-1)$-tuple de paires $((d_1, \delta_1), \ldots, (d_{p-1}, \delta_{p-1}))$ (les $p-1$ intervalles de distance) ;

avec p un entier positif, $m_i \in \Sigma^+$ et d_i, δ_i des entiers non négatifs.

Le terme d_i représente une distance entre les boîtes et $\pm\delta_i$ un intervalle autorisé autour de cette distance. Le nombre maximum d'erreurs (la distance d'édition maximale) autorisé par boîte peut être fixé à une valeur différente selon la boîte. Enfin, un taux maximum d'erreur sur l'ensemble des boîtes peut être considéré. Cela permet de prendre partiellement en compte une possible corrélation entre les boîtes.

L'algorithme Smile que nous avons développé permet, étant donné un texte (un ensemble de séquences biologiques) et un ensemble de contraintes (nombre de boîtes, distances entre boîtes, taux maximum d'erreurs autorisés par boîte et globalement), d'infèrer tous les motifs présents dans le texte qui satisfaisont ces contraintes. L'algorithme a été réalisé en collaboration avec Laurent Marsan (doctorant de l'université de Marne-la-Vallée, co-encadré par Maxime

Crochemore et Marie-France Sagot, soutenance prévue au début 2002).

Smile a été extensivement testé par le passé sur des organismes procaryotes (*Escherichia coli*, *Bacillus subtilis* et *Helicobacter pylori*) afin de détecter des séquences promotrices et régulatrices de la transcription de l'ADN en ARN. Deux publications dans des revues biologiques en sont issues. Ce travail se poursuit à l'heure actuelle sur deux autres familles de bactéries. La première concerne les cyanobactéries, en particulier *Synechocystis* PCC 6803 dont le comportement au niveau génomique peut sembler par certains aspects intermédiaire entre celui d'un procaryote et celui d'un eucaryote. Cette analyse est menée en collaboration, entre autres, avec plusieurs membres du projet Helix. Un premier motif fortement significatif a été récemment détecté dont la fonction potentielle est en cours d'investigation en laboratoire de biologie. La seconde famille de bactéries à laquelle nous nous intéressons est celle des mycobactéries, en particulier *Mycobacterium tuberculosis* et *Mycobacterium leprae*. Il s'agit là de 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 également comparative. Ce travail est en train d'être réalisé en collaboration avec Anne Cariou (stagiaire de l'École Vétérinaire de Maison-Alfort).

En termes informatiques, Smile est actuellement un algorithme précurseur dans la mesure où il constitue la seule méthode exacte permettant de détecter des motifs composés de deux boîtes. Par ailleurs, il est la seule méthode existante qui peut identifier des motifs présentant plus de deux parties. La partie extraction utilise une structure de données nouvelle, l'arbre des facteurs d'un texte, et est efficace en temps et très économe en espace. L'arbre des facteurs a été mis au point en collaboration avec Julien Allali (doctorant de l'université de Marne-la-Vallée, co-encadré par Maxime Crochemore et Marie-France Sagot).

Les perspectives concernant l'inférence de motifs sont très nombreuses. Elles portent sur : 1. la notion d'une « base de motifs » (sous-ensemble des motifs solutions d'une instance particulière d'inférence qui permet de retrouver tous les autres motifs qui sont solutions de l'instance à partir d'une certaine opération algébrique portant sur l'espace des motifs), 2. celle de « méta-différences » (différences portant non plus sur chaque boîte d'un motif – substitutions, insertions ou délétions de lettres – mais sur le motif lui-même – substitutions, insertions ou délétions de boîtes), 3. celle de motifs mixtes lexicaux et structuraux, 4. celle d'un ensemble minimal de motifs recouvrant par rapport à un ensemble de séquences (c'est-à-dire, tel que toute séquence possède une occurrence d'un motif de l'ensemble recouvrant), et, enfin, 5. sur la prise en compte de l'évolution dans la détection de motifs.

6.1.3 Comparaison et inférence de structures secondaires d'ARNs

La structure d'une molécule d'ARN est déterminante pour la fonction que celle-ci exerce au sein d'une cellule. Les divers éléments de base pouvant composer une telle structure sont les suivants :

- une *hélice* correspond à une succession de bases appariées par la liaison Watson Crick A$\leftrightarrow$ T, C$\leftrightarrow$ G (c'est ce que l'on appelle un palindrome biologique).
- une *boucle* est une suite de bases non-appariées joignant les deux parties d'une même hélice.

- une *boucle multiple* est le point de jonction de différentes hélices.
- un *renflement* est une suite de bases non appariées comprise entre deux hélices.
- une *boucle interne* correspond à deux renflements compris entre deux hélices.
- une *tige* est une structure comprennant une hélice et zéro, une ou plusieurs boucles internes et renflements. Une tige est ainsi un ensemble d'hélices, boucles internes et renflements se trouvant entre deux boucles multiples ou entre une boucle multiple et une boucle.
- un *pseudo-noeud* est une suite de bases appariées mettant en jeu des bases appartenant à deux boucles du type simple, multiple, interne ou renflement.

Tous les éléments sauf le dernier composent ce que l'on appelle la structure secondaire d'un ARN.

Le problème qui nous concerne à terme est celui de l'inférence de sous-structures maximales communes à un ensemble d'ARNs en utilisant comme seule information au départ les chaînes de nucléotides représentant ces molécules. Ce problème repose toutefois, pour partie, sur un autre, également difficile et relativement mal résolu actuellement, qui concerne la comparaison de structures d'ARNs. Ces structures ont pu être obtenues de manière expérimentale ou prédites de façon plus ou moins précise et complète. C'est ce second problème que nous avons abordé initialement, en considérant dans un premier temps des structures secondaires d'ARN dans lesquelles, cependant, les pseudo-noeuds sont également pris en compte.

Développée en collaboration avec Julien Allali (université de Marne-la-Vallée), notre approche se base sur l'idée d'effectuer cette comparaison à des niveaux de plus en plus fins. Afin de modéliser le repliement de l'ARN , nous avons introduit une nouvelle structure que nous avons appelée MIGAL pour *MultIple GrAphe Layer*. Cette structure est une composition de plusieurs graphes, quatre en tout, qui sont ordonnés entre eux. Chacun de ces graphes est indépendant, mais il existe une relation d'inclusion entre un noeud d'un graphe et certains noeuds des graphes qui l'entourent dans l'ordre. Chaque graphe représente un niveau d'abstraction par rapport à la séquence de l'ARN. L'information présente aux différents niveaux est résumée dans le tableau ci-dessous.

Graphe	Description	Noeud	Arc
0	Graphe représentant le réseau de boucles multiples	Boucles multiples	Tiges entre 2 boucles multiples
1	Graphe représentant le réseau des boucles multiples	Tiges, boucles, boucles multiples	Lien entre 2 éléments de structure
2	Graphe représentant la structure secondaire au niveau des tiges	Tiges, boucles, boucles multiples	Lien entre 2 éléments de structure
3	Graphe représentant la structure secondaire au niveau des hélices	Hélices, boucles, boucles internes, renflements, boucles multiples, pseudo-noeuds	Lien entre 2 éléments de structure

La comparaison de deux structures d'ARNs commence au niveau 0. Ce sera uniquement si des éléments communs auront été détectés à ce niveau que la comparaison pourra se poursuivre aux niveaux plus fins.

La structure elle-même a été implantée (en C++, l'algorithme est appelé MIGAL, comme la structure). Elle est efficace en temps (8 secondes pour représenter 10000 ARNs de longueur moyenne 361 nucléotides) et en espace (complexité en $O(4n)$ si n est la longueur totale des chaînes d'ARNs). Nous travaillons à l'heure actuelle sur la formalisation d'une distance locale à chacun de ces niveaux, distance entre deux, puis entre plus de deux, structures d'ARNs.

6.2 Organisation des génomes et cartographie comparée

Participants : Gisèle Bronner, Christophe Bruley, Christian Gautier [Correspondant], Adel Khelifi, Anne Morgat [Correspondant], Dominique Mouchiroud, Bruno Spataro, Alain Viari.

GemCore est un système à base de connaissances dédié à l'analyse de l'organisation de l'information génétique dans les génomes. Deux points de vue sont privilégiés : celui de la modélisation de l'organisation spatiale (monodimensionelle), qui est complexe, car étudiée à plusieurs niveaux de granularité, et celui de la comparaison interspécifique.

Ces deux points de vue ont conduit à la spécification d'un premier schéma conceptuel exprimé en classes et associations et qui a fait l'objet d'une instanciation, essentiellement à l'aide de données sur les génomes murin et humain. Plusieurs requêtes complexes ont pu être exprimées et traitées.

Dans le même contexte, des travaux sur la réconciliation d'arbres phylogénétiques ont conduit à l'obtention d'un environnement d'aide à la réconciliation qui a été évalué en vraie grandeur. Les algorithmes originaux qu'il incorpore permettent d'aider le biologiste à distinguer les gènes orthologues (issus d'un événement de spéciation) des gènes paralogues (issus d'un événement de duplication), étape fondamentale dans toute analyse comparative.

L'objectif de l'activité de génomique comparative dans le domaine des procaryotes est la description des segments synténiques conservés entre plusieurs espèces bactériennes et l'utilisation de ces connaissances à des fins d'inférence de fonction. De manière informelle, on définit un synton comme l'ensemble maximal de couples de gènes dont l'organisation chromosomique est conservée entre deux chromosomes (deux espèces). On entend par organisation chromosomique conservée, une localisation similaire des gènes dans chacun des deux chromosomes bactériens étudiés. Au niveau informatique, la recherche des syntons doit prendre en compte des opérations de permutation et d'insertion/délétion des gènes. En pratique, nous avons reformulé ce problème comme celui de la recherche d'un double chemin dans un graphe. Une illustration de la recherche de syntons est donnée par la figure suivante. Une chaîne de traitement a été développée pour la recherche systématique des relations d'orthologie et de synténie entre deux couples de génomes bactériens.

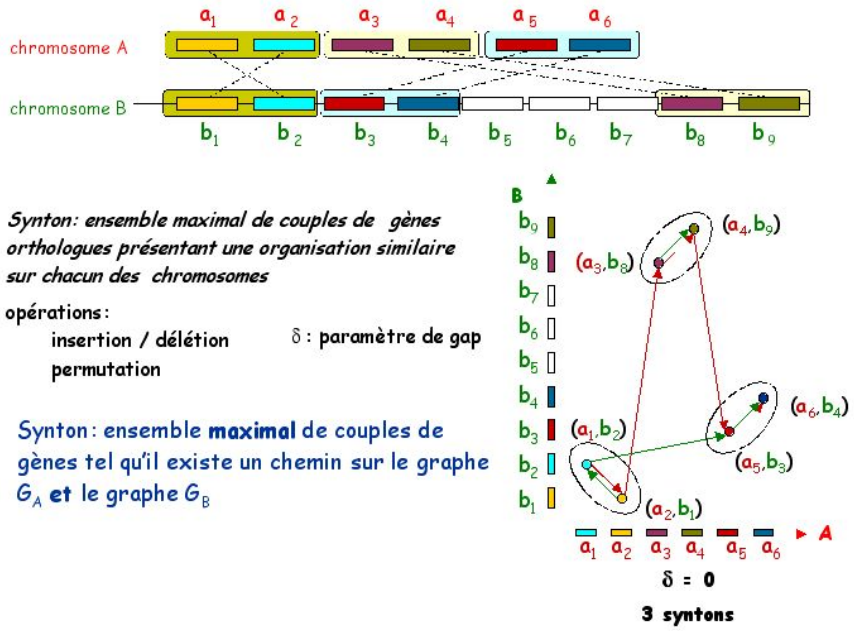

FIG. 4 – Exemple de recherche de syntons

6.3 Cartographie comparée, phylogénie et évolution

Participants : Jean-François Dufayard, Laurent Duret, Manolo Gouy [Correspondant], Guy Perrière, Nadia Pisanti, Marie-France Sagot [Correspondant], Marina Zelwer.

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 telle sorte 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. Nous avons élaboré une première approche statistique du problème. Ce travail a été réalisé en collaboration avec Pierre Darlu (directeur CNRS dans le laboratoire U535 de l'INSERM). Une implantation (en C) de l'algorithme, appelé DRUID, est directement utilisable à travers une interface web (`http://bioweb.pasteur.fr/seqanal/interfaces/druid.html`). L'algorithme a été appliqué à la fois sur des jeux biologiques divers, trouvés dans la litérature ou fournis par des biologistes, et sur des séquences sur lesquelles ont été simulés des évènements de remaniement et de mutations ponctuelles. Les résultats obtenus indiquent que DRUID est très sensible (il « rate » très peu de points de remaniement). Il semble permettre de détecter des points de remaniement dans des cas difficiles que la plupart des autres approches ne sont pas capables de traiter. Cela concerne en particulier les situations, très fréquentes dans la nature, où plusieurs remaniements ont eu lieu au niveau, par exemple, d'un gène chez des espèces ou sous-espèces proches évolutivement. L'algorithme présente en outre l'avantage de se baser sur une évaluation statistique de la prédiction des points de remaniement. Cela est rarement le

cas dans d'autres méthodes mais n'est pas sans présenter des difficultés. Notre approche est également limitée par le fait que l'évolution de chaque partition est actuellement mesurée par la longueur des arêtes d'un arbre phylogénétique, et donc la distance entre deux arbres par une simple différence entre leurs longueurs (c'est-à-dire, le nombre de mutations le long des arêtes des arbres). On ne prend ainsi pas du tout en compte une éventuelle différence topologique. Une des perspectives concernant ce travail vise à explorer des distances de remaniement entre arbres phylogénétique qui prennent en compte la topologie des arbres.

Cette exploration a déjà été initiée sur le plan théorique dans un travail en collaboration avec Estela Maris Rodrigues (doctorante du département de Mathématiques et Statistiques de l'université de São Paulo au Brésil, co-encadrée par Yoshiko Wakabayashi, professeur dans le département, et Marie-France Sagot). 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 bisection 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 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 avons montré dans un article accepté et présenté à une conférence d'informatique (APPROX) que l'algorithme fourni par les auteurs avait en réalité un ratio de 4 et nous avons donné un nouvel algorithme dont le ratio d'approximation est effectivement de 3. L'algorithme a été implanté en C. Par ailleurs, un autre chercheur a montré que la distance SPR n'est pas toujours égale à la taille du MAF moins un. Nous conjecturons qu'elle est, soit strictement égale à la taille du MAF, soit égale à la taille du MAF moins un et travaillons actuellement sur une preuve de cette conjecture. Nous essayons en même temps de trouver un algorithme d'approximation de meilleur ratio pour résoudre le problème du MAF.

Le problème du calcul d'une distance de remaniements entre génomes multi-chromosomiques est équivalent à celui du calcul d'une distance entre deux partitions différentes d'un même ensemble d'objets lorsque les opérations permises sont la fusion de deux classes en une, la fission d'une classe en deux et l'échange réciproque d'éléments entre deux classes. Un article a été soumis qui étudie cette distance. Nous y montrons à l'aide d'une preuve directe que le diamètre de cette distance est de $2n - 4$ où n est le nombre de gènes dans l'un des deux génomes. Nous fournissons également un algorithme exact permettant de résoudre une version contrainte du problème. Cet algorithme a une meilleure complexité qu'un autre précédemment établi par deux chercheurs de l'université de Stanford. Nous travaillons actuellement au calcul d'une distance entre génomes qui considère non seulement la composition en gènes des chromosomes, mais surtout l'ordre et l'orientation des gènes le long des chromosomes. Les algorithmes existants qui prennent ces informations en compte ne considèrent en général qu'un sous-ensemble

de tous les types de réarrangements observés (ceux-ci peuvent inclure l'inversion d'un segment, la duplication, l'insertion, la suppression, la transposition ou déplacement vers un autre endroit dans le génome et la translocation de segments, c'est-à-dire l'échange de matériel génétique intra ou inter-espèces). La justification pour un tel choix est le plus souvent algorithmique plutôt que biologique. Nous souhaitons aller vers des modèles plus réalistes en biologie. Ces modèles sont encore à définir et pourront varier selon le problème considéré : distance entre génomes ou entre opérons (un opéron est un ensemble de gènes qui sont exprimés, c'est-à-dire traduits, toujours en même temps et dans les mêmes proportions. Les gènes d'un opéron occupent des places voisines dans un génome mais des réarrangements peuvent avoir lieu entre gènes d'un même opéron). Par ailleurs, nous commençons à réfléchir sur le calcul de distance entre génomes basés sur les segments conservés. Ce travail se fera en collaboration avec divers membres d'Helix, en particulier à Lyon. Il implique dans un premier temps obtenir une définition adaptée aux organismes eucaryotes de la notion de synton établie pour les organismes procaryotes.

6.4 Modélisation de processus évolutifs

Participants : Manolo Guy, Jean Lobry [Correspondant], Guy Perrière, Gwenael Piganeau.

Le modèle générique des taux de substitution des quatre bases nucléiques comporte 12 paramètres. Le nombre total de modèles dérivés par simplification est égal, à une reparamétrisation près, au nombre de partitions d'un ensemble de 12 éléments, soit 4,213,597 modèles possibles. Un travail d'étudiant de DEA a consisté à caractériser tant du point de vue des hypothèses biologiques que des propriétés mathématiques le petit sous-ensemble des modèles publiés dans la littérature. Ces connaissances ont été organisées sous la forme d'un graphe exprimant les dépendances entre les modèles.

6.5 Modélisation de familles de séquences homologues

Participants : Laurent Duret [Correspondant], Manolo Guy, Guy Perrière [Correspondant].

Le séquençage systématique de plusieurs dizaines de génomes bactériens permet désormais d'envisager l'analyse comparative de génomes complets. Aussi, dans la lignée du travail réalisé sur HOVERGEN, nous avons entrepris le développement d'une banque de données de gènes homologues bactériens : HOBACGEN (Perrière et al., 2000). Cette banque contient toutes les séquences d'eubactéries, d'archées ainsi que le génome complet de la levure, avec une classification par familles des gènes homologues (au total 183000 séquences classées en 18700 familles dans la version 9, avril 2001). Les alignements protéiques et les arbres phylogénétiques sont calculés pour chaque famille. Cette banque est basée sur le modèle d'HOVERGEN auquel nous apportons plusieurs améliorations. Ainsi, HOBACGEN inclut à la fois les séquences nucléiques provenant d'EMBL et les séquences protéiques provenant de SWISS-PROT. Par ailleurs, pour assurer une meilleure diffusion de la banque, nous avons développé une nouvelle interface en langage Java qui permet d'utiliser HOBACGEN sur la

majorité des plates-formes existantes (stations de travail UNIX, micro-ordinateurs Macintosh ou Windows). Cette interface permet d'utiliser HOBACGEN à travers le réseau Internet, ce qui évite à l'utilisateur de devoir installer la banque de données sur son propre ordinateur. Enfin, la procédure de mise à jour a été entièrement automatisée afin d'éviter les expertises manuelles coûteuses en temps. HOBACGEN est disponible publiquement depuis février 1999 (`http://pbil.univ-lyon1.fr/databases/hobacgen.html`). Elle a été installée sur une vingtaine de sites dans le monde. HOBACGEN est actuellement utilisée dans notre groupe pour analyser les phénomènes de transfert de gènes chez les procaryotes et pour étudier la phylogénie des bactéries.

Les outils que nous avons développés pour HOBACGEN (procédure de classification en famille de gènes homologues, interface graphique, système client-serveur) ne sont pas spécifiques des bactéries. Ils sont également utilisés pour des projets au sein de notre groupe ou en collaboration avec d'autres groupes : analyse du génome d'une microsporidie, analyse de génomes bactériens (F. Chétaouni, F. Kunst, Institut Pasteur), NuReBase (base de données de récepteurs nucléaires, M. Robinson ENS-Lyon), RTK-db (base de données de récepteurs tyrosine kinase, G. Mouchiroud CGMC Lyon), RPSdb (base de données sur les rétropseudogènes de mammifères). Par ailleurs nous sommes en train de transférer la base de données HOVERGEN sous la structure développée pour HOBACGEN.

A l'aide d'ACNUC, il est possible de retrouver de façon automatique dans HOVERGEN ou HOBACGEN l'ensemble des gènes homologues communs à différents taxons. Par contre, l'interprétation des relations d'homologie (orthologie ou paralogie) nécessite une expertise manuelle. Bien que les interfaces graphiques d'HOVERGEN et HOBACGEN simplifient grandement l'analyse des arbres phylogénétiques, le nombre énorme de familles de gènes à expertiser rend ce travail extrêmement fastidieux. Par exemple, pour retrouver dans HOVERGEN les 4500 gènes orthologues entre homme et souris, il a été nécessaire d'analyser manuellement plus de 4000 familles de gènes. Nous avons donc entrepris de développer des outils permettant l'analyse automatique des arbres phylogénétiques.

Jean-François Dufayard a développé un nouvel algorithme d'analyse automatique d'arbre, appelé RAP, capable de prendre en compte des incertitudes à la fois sur l'arbre des gènes et sur l'arbre des espèces (Dufayard et al., 2000). Par ailleurs, RAP prend en compte non-seulement la topologie, mais également les longueurs de branches pour repérer d'éventuelles paralogies cachées.

A la suite de ce travail, Jean-François 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 complexe (au sens algorithmique, i.e. 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). A l'aide d'un éditeur graphique, il est possible de décrire le motif (sous-arbre) recherché, et de spécifier les contraintes sur les nœuds et branches de ce motif (présence ou non de duplications, présence ou non d'un groupe taxonomique donné, etc.). Cet outil est en cours d'intégration dans la nouvelle interface graphique d'HOVERGEN et HOBACGEN. Ainsi, il sera possible directement depuis cette interface de sélectionner automatiquement des gènes dans ces bases de données, 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 sur

HOBACGEN pour repérer des topologies aberrantes, signes de possibles transferts horizontaux

6.6 Modélisation dynamique des réseaux de régulation génique

Participants : 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 Jean-Luc Gouzé (INRIA Sophia-Antipolis) et Tewfik Sari (université de Mulhouse) les problèmes mathématiques liés aux discontinuites 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 rigoreuse et pratique. En outre, nous avons commencé 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 méthode de simulation a été implémentée en Java, dans un outil baptisé *Genetic Network Analyzer (GNA)*. Afin de faciliter l'utilisation du simulateur, nous avons continué le développement d'une interface graphique permettant de visualiser des réseaux d'interactions et d'analyser les résultats de simulation. Un éditeur pour aider à la spécification d'un modèle de simulation est en cours de développement. La version 4.1 de GNA a été déposée auprès de l'APP. Elle est disponible à travers le Web (`http://www-helix.inrialpes.fr/gna`).

En collaboration avec des biologistes, GNA est utilisé pour étudier des processus de régulation bactériens, comme l'initiation de la sporulation chez *B. subtilis*. Dans le cadre de son stage de licence à l'ENS de Lyon, Grégory Batt a étendu un modèle du réseau des gènes et des interactions contrôlant l'initiation de la sporulation chez la bactérie *Bacillus subtilis*, développé en 2000. Cette extension permet de prendre en compte des interactions protéine-protéine impliquées dans la transduction et l'intégration des signaux provenant de l'environnement par un phosphorelai.

Un nouveau projet, impliquant plusieurs membres de l'équipe HELIX et les laboratoires de J. Geiselmann à l'université Joseph Fourier (Grenoble) et de J. Houmard à l'ENS (Paris), a récemment débuté dans le cadre de l'action pluri-organisme *Bioinformatique*. Planifié sur deux ans, il concerne la modélisation et la simulation de la transduction des signaux par les nucléotides cycliques chez la cyanobactérie *Synechocystis* PCC 6803. Nous essayons de comprendre comment la réponse de la bactérie à des signaux provenant de l'extérieur émerge d'un réseau d'interactions entre gènes, protéines, et molécules messagères. Afin d'élucider ce réseau, nous

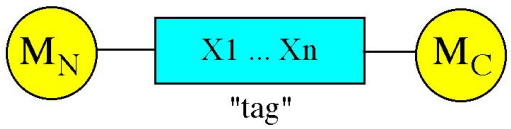

FIG. 5 – Représentation schématique d'une « étiquette peptidique ».

utilisons GNA en combinaison avec des méthodes expérimentales, comme la construction des mutants et des mesures du taux d'expression des gènes, ainsi que des méthodes bioinformatiques classiques, comme l'analyse des séquences.

6.7 Modélisation des données de protéome et protéomique expérimentale

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, reponsable Jérome Garin) et la société Génome-Express (responsable du projet, Thierry Vermat). Du point de vue expérimental, le projet repose sur la plateforme instrumentale développée au CEA-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 plateforme 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 plateforme protéomique grenobloise dans le cadre de la génopole Lyon-Grenoble. PepMap constitue le volet bioinformatique complémentaire de cette plateforme 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 (figure 5).

L'analyse informatique se décompose alors en deux étapes, associées à deux modules logiciels complémentaires.

Le but du premier module, baptisé PEP-MAP, est 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

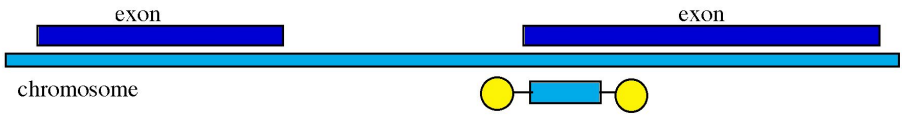

FIG. 6 – Positionnement d'une « étiquette peptidique » sur le chromosome. Problème posé par la structure exonique des gènes.

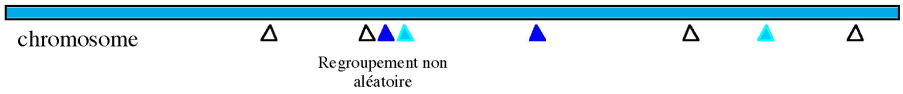

FIG. 7 – Regroupement des « étiquettes peptidiques » d'une protéine

chromosome.

Compte tenu de la taille des chromosomes humains (35 Mb pour le 22 qui constitue un « petit » chromosome) 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.

Une seule étiquette est, bien entendu, insuffisante pour localiser de manière unique un peptide sur un chromosome eucaryote complet, en raison du nombre important de « touches » (*hits*) possibles. En revanche, lorsque plusieurs étiquettes sont disponibles il devient possible, à partir de l'analyse statistique de la répartition des « touches » sur le chromosome, de proposer une (ou quelques) localisation possible. Le deuxième module MAP-EVAL est chargé d'évaluer la signification statistique de l'aggrégation des « touches » associées à plusieurs étiquettes par l'étude de leur répartition sur le chromosome (figure 7). Deux méthodologies seront testées : la première repose sur des tests statistiques classiques d'aggrégation, la seconde sur l'approche *r-scan* developpée par Samuel Karlin à l'université de Stanford (Karlin et Brendel, *Science*, 257, 39-49, 1992).

Le projet, d'une durée de 18 mois a été découpé en quatre phases : (i) évaluation des besoins et établissement d'un cahier des charges, (ii) développement d'un prototype , (iii) évaluation expérimentale du prototype et (iv) implémentation du produit final. Nous en some actuellement à la phase iv. Les premiers résultats obtenus avec le prototype sont extrêmement encourageants, puisque nous avons pu montrer la faisabilité de l'approche sur le génome d'*Arabidopsis thaliana* et d'*Homo sapiens*. Dans le cas d'*Arabidopsis thaliana* des tests en « aveugle » sur des données expérimentales réelles nous ont permis d'assigner sans ambiguité 10 protéines dans un mélange de 12. Par ailleurs d'autres expérimentations nous ont permis d'identifier, dans une fraction hydrophobe du plasmalemme d'*Arabidopsis*, des protéines qui n'étaient pas jusqu'à présent considérées comme appartenant à ce compartiment cellulaire.

6.8 Modélisation de données génomiques et post-génomiques : projet Panoramix

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

Génomique comparative et étude des synténies bactériennes (base Genomix)

Une chaîne de traitement a été développée pour la recherche systématique des relations d'orthologie et de synténie (§6.2 : description de l'algorithme) entre deux couples de génomes bactériens. Ces deux opérations, effectuées pour toutes les paires de chromosomes bactériens disponibles (52 chromosomes actuellement), sont relativement coûteuses en temps de calcul. Ces résultats ont ensuite été sauvegardés dans la base Genomix. Il est important de noter que ceci a constitué l'une des premières applications de « grande taille » sous AROM (environ deux millions d'objets), ce qui a permis, en collaboration avec les concepteurs d'AROM d'augmenter significativement les performances du système. Enfin une interface de consultation est en cours de développement. Cette interface est destinée à valoriser la base de connaissances vers un public de biologistes.

Modélisation du métabolisme intermédiaire (bases Metabolix et Organix)

Le travail effectué en 2001 sur les bases Metabolix et Organix a porté plus particulièrement sur :

- la modification et l'amélioration du schéma initial, en particulier pour établir la liaison avec les autres bases de connaissances développées (figure 3), et pour représenter plus précisément les données à notre disposition.
- la vérification de la cohérence des données. En effet, les premiers résultats ont montré que les données publiques ayant servi à la première instanciation de la base s'avèrent peu fiables et souvent incohérentes. Le travail a consisté à croiser les informations en provenance de plusieurs sources : KEGG (`http://www.genome.ad.jp/kegg/`), ENZYME (`http://www.expasy.ch/enzyme`) et CCD (`http://www.chemper.com`) afin d'améliorer la fiabilité des données des bases Metabolix et Organix.

Modélisation des données concernant les protéines et les complexes protéiques (base Proteix)

Un modèle permettant de représenter explicitement les protéines matures (modification post-traductionnelles et assemblages moléculaires a été définie). Il n'existe, à ce jour aucune base de données modélisant ces informations, notre travail a donc consisté à définir une stratégie d'analyse combinant différentes approches :

- extraction d'information à partir des enregistrements de la base Swissprot
- analyse par similarité de séquences
- analyse du contexte chromosomique

Enfin, au cours du développement de ces trois bases, deux questions se sont naturellement rapidement posées : (1) l'inter-opérabilité des bases et (2) la fiabilité des données biologiques.

Afin de répondre à ces deux questions, nous avons soumis en septembre 2001 le projet Panoramix dans le cadre de l'appel d'offres multi-organisme *Bioinformatique* (CNRS-INRA-INSERM-INRIA). Ce projet présente un double aspect. Sur le plan méthodologique, il propose de fédérer les trois bases Genomix, Metabolix et Proteix au sein d'un même système (Panoramix) permettant aux biologistes de croiser les différents types d'informations sur l'ensemble des différents génomes bactériens actuellement disponibles. Sur le plan biologique, il propose d'expertiser les données de Panoramix sur deux génomes de référence : *Bacillus subtilis* et *Synechocystis sp.* Le projet est constitué autour d'un partenariat entre Helix et trois laboratoires-experts de biologie, à l'Institut Pasteur (Hong-Kong et Paris) ; à l' Ecole Normale Supérieure (Ulm) et à l'université J. Fourier. Il a été financé pour deux ans, à partir de novembre 2001.

6.9 Extraction d'informations à partir de textes

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

Détecter dans un texte des noms d'entités biologiques, par exemple de gènes ou de protéines, est une première étape indispensable à l'extraction d'informations, mais non suffisante. À titre d'exemple, reconnaître un nom de gène sans le relier à l'espèce correspondante est peu utile. Le principe retenu dans le projet BioMiRe consiste par conséquent à rechercher dans les textes, non seulement des occurrences, mais des co-occurences de ces noms, en précisant de plus leurs positions absolues ou relatives. Des essais sur des corpus expertisés devraient permettre à terme de déterminer les requêtes pertinentes, par exemple pour associer un gène et l'espèce correspondante, mais aussi un gène et ses lieux d'expression, voire sa ou ses fonctions.

La première étape du projet a ainsi consisté à spécifier les noms d'entités et les relations à détecter. Les noms d'entités biologiques sont les noms de gènes, de protéines, d'espèces et de lieu d'expression des gènes (par exemple, des noms de tissus ou d'organes). Les relations portent sur la position absolue (présence d'un nom dans le titre, le résumé, l'introduction ou le corps du texte) ou relative (dans la même phrase ou le même paragraphe), sur la proximité (distance en mots ou en paragraphes entre noms), et l'ordre (en amont ou en aval d'un autre nom).

L'architecture du logiciel a été spécifiée, ainsi que l'interface qui permet d'exprimer, sous forme graphique ou textuelle, une requête sur un corpus et d'étudier les résultats. Le résultat d'une requête est ainsi un ensemble de fragments de textes dans lesquels les noms spécifiés dans la requête figurent dans des positions précisées par les relations.

Par ailleurs, des lexiques de noms d'entités biologiques ont été créés pour les quatre espèces retenues : l'Homme, la souris, la mouche drosophile et la plante *Arabidopsis thaliana*. Pour établir un lexique de 114 000 noms de gènes, différentes sources d'informations accessibles par Internet ont été utilisées. Les dictionnaires de noms de protéines (93 000 protéines répertoriées), d'espèces et de termes spécifiques au domaine ont été constitués de façon similaire. Enfin, un corpus de 300 phrases a été constitué. Pour chaque phrase, les noms de gènes, de protéines, d'espèces et de localisation ont été détectés manuellement. Ce corpus servira à tester les dictionnaires et les outils linguistiques.

6.10 Environnement didactique en bioinformatique

Participants : Gaël Faroux, Philippe Genoud [Correspondant], 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. L'EDB permet d'accéder à un contenu pédagogique hypertexte (HTML) pouvant être situé sur n'importe quel serveur http. Ce contenu pédagogique permet de guider l'apprenant dans son exploration des 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). Cette maquette intègre six modules sur le thème de l'annotation de séquences génomiques : présentation du code génétique, explication des différentes phases de lecture, recherche de motifs (START, STOP, RBS, motif choisi par l'utilisateur), alignement de deux séquences, recherche de zones codantes par analyse du biais de codage (test du chi 2), 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'utilisation de l'Environnement Didactique en Bioinformatique est envisagée dès l'année 2002 dans différentes filières d'enseignement de l'UJF : option bioinformatique de la maîtrise d'informatique, en DESS CCI (Compétence Complémentaire en Informatique). Un travail de finalisation (amélioration des interfaces graphiques des modules, rédaction et réalisation d'un contenu pédagogique) est néanmoins nécessaire afin de rendre cet environnement véritablement exploitable dans un contexte didactique. Dans cet optique une demande de soutien a été déposée auprès du GRECO (Grenoble Campus Ouvert). Sont également envisagées la réalisation de nouveaux modules, en particulier sur le thème de la reconstruction d'arbres phylogénétiques, ainsi que l'évolution de l'environnement pour le rendre directement accessible au travers du Web.

6.11 Projet GénoStar

Participants : Christophe Bruley, Véronique Dupierris, Gilles Faucherand, François Rechenmann [Correspondant], Alain Viari.

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 sont 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, 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 visent l'obtention d'une première version de l'environnement pour les tout premiers mois de l'année 2002. Ceux réalisés au sein 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 de requêtes est en cours de développement. Il permettra 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. Il s'agit à présent d'introduire dans l'application les méthodes d'annotation, 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. Simultanément, une première version de l'interface de visualisation des entités détectées sur la séquence d'un génome donné, dite « interface cartographique », a été développée. Là encore, compte tenu du nombre d'objets impliqués et de la réactivité attendue du système, les critères d'efficacité sont primordiaux.

Enfin, le développement de l'application GenoBool a commencé avec la spécification des trois modules principaux qui le composent : 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. À terme, ces méthodes pourront être organisées en stratégies, qui seront incorporées à l'application, aidant ainsi l'utilisateur à exploiter au mieux ses fonctionnalités.

7 Contrats industriels (nationaux, européens et internationaux)

Le projet HELIX est engagée dans deux partenariats industriels majeurs, à travers le projet GénoStar et un projet soutenu par GénoPlante.

7.1 GénoStar

Le projet GénoStar est conduit par un consortium de quatre membres :
- la société Hybrigenics, Paris ;
- la société Génome Express, Grenoble ;
- l'Institut Pasteur, Paris ;
- l'INRIA.

Ce consortium a signé à l'automne 2000 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. Un soutien complémentaire de la Direction de la Technologie a été obtenu en 2001. Enfin, GénoStar est une action de développement de l'INRIA, assurée d'un soutien pendant 3 ans (2000-2002).

7.2 GénoPlante

Le projet HELIX participe au projet intitulé « Outils informatiques pour la prédiction de gènes et l'annotation de génomes – Application au génome d'*Arabidopsis thaliana*», financé par GénoPlante sur 2000-2001. Le rôle de HELIX dans ce projet est de développer, à partir d'ImaGene, un prototype d'environnement d'aide à l'annotation de génomes végétaux, en intégrant des méthodes proposées par les autres partenaires et de l'expérimenter sur le génome de l'arabette.

7.3 XRCE

Le Centre Européen de Recherche Xerox (XRCE) est le partenaire privilégié su projet HELIX sur le thème de l'extraction d'informations à partir de textes. Le partenariat a débuté dans le cadre d'une convention CIFRE, qui s'est achevée en décembre 2000, suivie par la soutenance de thèse de Denys Proux le 9 avril 2001. Il continue avec le projet BioMiRe, qui est soutenu par le Ministère de la Recherche (Direction de la Technologie) et qui 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 Hans Geiselmann du CERMO (université Joseph Fourier, Grenoble). Elle porte sur la modélisation et la simulation des interactions géniques. Ce projet vient de recevoir un soutien dans le cadre de l'appel d'offres *Bioinformatique* CNRS-INRA-INRIA-INSERM.

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

de protéines.

Le projet poursuit une collaboration avec l'IBS (Institut de Biologie Structurale, CEA/CNRS/UJF, UMR 5075), concrétisée par une réponse à l'appel d'offres *Programmes thématiques prioritaires* de la Région Rhône-Alpes, avec le CHU et Genome Express. Le projet « Résistance aux bétalactamines » a été sélectionné et financé.

Sur la protéomique, une collaboration avec Jérôme Garin (LCP : Laboratoire de Chimie des Protéines, CEA) est poursuivi, avec un soutien du Ministère de la Recherche, Direction de la Technologie, dans le cadre de l'appel d'offres *Bioinformatique*. Le projet soutenu rassemble la société Genome Express, le LCP/CEA et l'INRIA Rhône-Alpes.

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

Plusieurs collaborations scientifiques sont en cours avec la société Genome Express sur l'annotation de génomes bactériens. Elles portent notamment sur l'amélioration d'algorithmes de recherche de zones codantes et de régions structurées (terminateurs rho-indépendants).

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 (Joël Pothier), à l'INRA à Jouy-en-Josas (Philippe Bessières), Gif-sur-Yvette (Claude Thermes), Évry (Claudine Médigue), Toulouse (Christine Gaspin), Marseille (Gwenaelle Fichant et Yves Quentin) et Gand en Belgique (Pierre Rouzé). Bien entendu, l'équipe souhaite en tout premier lieu renforcer les interactions avec les projets INRIA déjà engagés en bioinformatique, en particulier au sein de l'ARC REMAG (recherche et extraction de motifs pour l'analyse génomique).

Marie-France Sagot coordonne le projet « Régulation, Synténie et Pathogénicité – Algorithmes et Expérimentations » dans le cadre du programme multi-organisme *Bioinformatique* CNRS-INRA-INRIA-INSERM. Les partenaires d'HELIX sont l'Institut de Biologie Physico-Chimique de Paris (Anne 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, Marie-France Sagot et Alain Viari participent au projet « Détection des exons/introns dans le génome humain ». Le projet implique des équipes du CGM de Gif-sur-Yvette (Claude Thermes, co-coordonnateur), Atelier de BioInformatique de l'université de Paris VI, Laboratoire Génome et Informatique de l'université de Versailles et 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 même programme multi-organismes (*Bioinformatique* CNRS-INRA-INRIA-INSERM), Hidde de Jong coordonne le projet « Modélisation et simulation de réseaux de régulation génique : La transduction des signaux par les nucléotides cycliques chez la cyanobactérie *Synechocystis* PCC6803 ». Plusieurs membres d'HELIX participent 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).

Enfin, plus récemment, le projet « Panoramix : Fédération de bases de connaissances pour la génomique et expertise sur deux génomes bactériens de référence » a également été accepté dans le cadre du même programme (*Bioinformatique* CNRS-INRA-INRIA-INSERM) (coordonnatrice Anne Morgat) et financé sur deux ans à partir de fin 2001.

8.3 Actions européennes et internationales

Au niveau européen, l'équipe participe au réseau ESF (European Science Foundation) intitulée *Experimental and in silico Analysis of Biomolecular Interactions*, en particulier avec l'Institut Pasteur (Antoine Danchin, actuellement à Hong-Kong), la société LION (Heidelberg, Allemagne), le laboratoire CNB-CSIC à Madrid (Alfonso Valencia) et l'université Tor Vergata à Rome (Manuela Helmer Citterich).

Le projet HELIX participe au projet HAMAP d'annotation automatique de protéomes bactériens, à l'initiative de l'Institut Suisse de Bioinformatique (Amos Bairoch) à Genève.

Marie-France Sagot participe à un Projet CNPq, « *Problemas de Otimizacao Combinatória : algorítmos e aplicaçöes* » (« Problèmes en Optimisation Combinatoire : Algorithmes et Applications »), avec le Département d'Informatique, Institut de Mathématiques et Statistiques, université de Sao Paulo, Brésil (coordonnatrice : Yoshiko Wakabayashi, professeur à l'université de Sao 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 France et 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.

Dans le cadre du programme d'actions intégrées franco-néerlandais Van Gogh, du printemps 1999 au printemps 2001, le projet HELIX a poursuivi une coopération scientifique avec le projet Plinius du Département d'Informatique à l'université de Twente (Pays-Bas). Le coopération entre HELIX et Plinius a pour but, d'une part, le développement de techniques de modélisation applicables à l'analyse de données scientifiques et, d'autre part, l'échange d'expériences obtenues en appliquant ces techniques dans les domaines de la biologie moléculaire et des sciences des matériaux.

Nadia Pisanti et Marie-France Sagot collaborent avec Roberto Grossi de l'université de Pise sur un problème d'inférence de motifs. Maxime Crochemore de l'Institut Gaspard Monge de l'université de Marne-la-Vallée fait également partie de cette collaboration.

Alain Viari entretient une collaboration avec James Maher du *Department of Biochemistry and Molecular Biology* (Mayo Foundation, Rochester, États-Unis) sur la recherche de zones structurées (triple-hélices) dans les génomes procaryotes complets.

Laurent 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).

Laurent Duret est correspondant du projet européen de grille de calcul *DataGrid* : WP10 applications à la biologie moléculaire et à l'imagerie médicale.

9 Diffusion de résultats

9.1 Animation de la communauté scientifique

Laurent Duret a soutenu une thèse de Habilitation à Diriger les Recherches, intitulée « Organisation, fonctionnement et évolution des génomes de métazoaires : mais où est donc passée la sélection naturelle ? », le 21 novembre 2001 à l'université Lyon 1.

Marie-France Sagot anime le séminaire *Algorithmique et Biologie* (`http://www-igm.univ-mlv.fr/~sagot/AlgoBio/index.html`).

Christian Gautier est directeur du programme de recherche *Bioinformatique* des établissements publics, scientifiques et techniques CNRS, INRA, INRIA et INSERM.

L'équipe participe à l'action IMPG (*Informatique, Mathématiques et Physique pour la Génomique*), soutenue par le Ministère chargé de la recherche. François Rechenmann y anime, avec Philippe Bessières (INRA) et Emmanuel Barillot (Infobiogène) le groupe de travail *Bases de données, interfaces et ontologies*.

Manolo Gouy est représentant de la France au conseil de direction du *Global Biodiversity Information Facility (GBIF)*.

Manolo Gouy est membre du conseil scientifique de l'Institut Français de la Biodiversité, membre du jury d'attribution de l'ATIPE *Biodiversité* du CNRS, et membre du Conseil National des Universités, section 67.

François Rechenmann a organisé, avec Christian Gautier et Marie-France Sagot, l'atelier INSERM 122 *Bioinformatique : méthodes et pratiques pour l'analyse de l'information génomique*, Lyon, 11-12 janvier 2001. Cet atelier a rassemblé une dizaine d'intervenants, devant une centaine de participants. Un support de cours, puis un CD-ROM reprenant les exposés et les transparents, ont été édités par l'INSERM et font l'objet d'une diffusion.

François Rechenmann a organisé une journée INRIA - XRCE, 26-27 avril 2001, dans les locaux de l'unité de recherche Rhône-Alpes. Elle constituait simultanément un séminaire du groupe de travail *Bases de données, interfaces et ontologies* de l'action IMPG (*Informatique, Mathématiques et Physique pour la Génomique*).

Laurent Duret a été co-président du comité scientifique de la conférence JOBIM 2001 à Toulouse.

Hidde de Jong est membre du comité de programme du *Sixteenth International Workshop on Qualitative Reasoning* qui se tiendra à Barcelone en 2002.

9.2 Enseignements universitaires

François Rechenmann a donné des cours en maîtrise de biologie, filière *Mathématiques-informatique*, université Claude Bernard, Lyon, (14h).

Marie-France Sagot a donné des cours dans le module *Informatique du génome*, DEA d'Informatique Fondamentale et Applications, université de Marne-la-Vallée (10h). Elle intervient également dans la maîtrise d'informatique (2h).

Marie-France Sagot a donné des cours à l'INSA-Lyon, 4ème année, spécialité *Bioinformatique et modélisation* (8h).

Hidde de Jong, Anne Morgat, François Rechenmann, Marie-France Sagot, Alain Viari et Danielle Ziébelin sont intervenus dans l'option *Bioinformatique* de la maîtrise d'informatique

à l'université Joseph Fourier (Grenoble) (4-8h).

Alain Viari intervient également dans plusieurs DEA (DEA de Génétique université Paris 6 (3h) ; DEA de Biologie et Informatique (option bioinformatique), Marseille (3h) ; Ecole Doctorale de Grenoble (10h).

9.3 Participation à des colloques, séminaires, invitations

En tant que professeur invité, Marie-France Sagot est intervenue dans le colloque de Mathématiques, Institut de Mathémathiques Pures et Appliquées (IMPA), Rio de Janeiro, Brésil, 27-31 Juillet 2001.

Invitée par Ricardo Baeza-Yates, Marie-France Sagot a fait un exposé lors du séminaire au Laboratoire d'Informatique de l'université de Santiago, Chili, dans la semaine du 16 au 19 octobre.

Marie-France Sagot a présenté l'exposé « Un lac, deux villes et trois rivières : Algorithmique combinatoire et biologie moléculaire » lors d'un séminaire au Laboratoire de Biométrie et Biologie Évolutive (8 novembre 2001). Elle est également intervenue lors du séminaire d'Informatique de l'Institut Gaspard Monge, université de Marne-la-Vallée (13 novembre 2001). Le titre de l'exposé : « Quelques résultats autour du problème de la Forêt d'Accord Maximum ».

Laurent Duret a organisé l'atelier de Formation INSERM en Bioinformatique, Santiago, Chili (16-19 Octobre 2001). Interventions de Laurent Duret et Marie-France Sagot.

Laurent Duret, Christian Gauthier, François Rechenmann et Marie-France Sagot ont organisé l'atelier de Formation INSERM « Bio-informatique : méthodes et pratiques pour l'analyse de l'information génomique » (Lyon, 11-12 janvier). Interventions de Laurent Duret, Manolo Gouy et Marie-France Sagot.

Laurent Duret a organisé l'atelier de Formation INSERM « Bio-informatique : phase pratique » (Lyon, 9-11 mai). Interventions de Laurent Duret, Manolo Gouy et Guy Perrière.

Laurent Duret a organisé l'atelier de Formation INSERM « Bio-informatique : alignement de séquences et prédiction de gènes » (Lyon, 25-27 septembre). Interventions de Laurent Duretet Guy Perrière.

Laurent Duret a été conférencier invité au *5th Anton Dohrn Workshop : Natural selection and the neutral theory*, Ischia (Italie), 24-27 octobre 2001. Le titre de son exposé était « Isochore organization of mammalian genomes : selection or neutral evolution ? ».

Manolo Gouy a été conférencier invité au *Workshop on Population genetics at the molecular level*, Montréal (Canada), 8-10 Mars 2001.

Manolo Gouy et Laurent Duret ont été conférenciers invités lors de la *Conférence Jacques Monod*, Aussois, 26-28 Avril 2001. Le titre de leur exposé était « Gene and genome duplications and the evolution of novel gene functions ».

Marie-France Sagot et Alain Viari ont donné deux séries de cours de 3h à la *First International Conference on BioComputing* qui s'est tenue du 4 au 9 Juin 2001 à Poznan (Pologne).

Frédéric Boyer, Gisèle Bronner, Céline Hernandez, Hidde de Jong, Anne Morgat et Violaine Pillet ont participié à la *9th International Conference on Intelligent Systems for Molecular Biology (ISMB 2001)*, Copenhague (Danemark), 19-26 juillet 2001. Anne Morgat et Frédéric Boyer ont présenté le poster « Representation and integration of metabolic and genomic data : the

Panoramix project » (réalisé avec Hélène Rivière-Rolland, Danielle Ziébelin, François Rechenmann et Alain Viari). Ils ont également participé à la conférence satellite *Bio-ontologies*. Hidde de Jong et Céline Hernandez ont présenté le travail sur la simulation qualitative de l'initiation de la sporulation chez *Bacillus subtilis*, réalisé avec Johannes Geiselmann et Michel Page, lors du *Satellite Meeting on Computer Modeling of Cellular Processes (SIGSIM-01)*. Violaine Pillet a présenté « A generic statistical method for information extraction in genomics », réalisé avec Denys Proux et François Rechenmann, lors du *Satellite Meeting on Biological Research with Information Extraction & Open-Access Publications (BRIE & OAP)*. Plusieurs membres d'équipe ont contribué au poster « Modelling genomic annotation data using objects and associations : the GenoAnnot project » (Hélène Riviere-Rolland, Gilles Faucherand, Christophe Bruley, Anne Morgat, Magalie Roux-Rouquie, Claudine Medigue, François Rechenmann, Alain Viari, Yves Vandenbrouck).

Hidde de Jong a effectué trois visites à l'université de Twente (Pays-Bas) dans le cadre du programme d'actions intégrées franco-néerlandais Van Gogh (du 1 au 8 janvier, du 8 au 16 mai et du 29 au 31 août 2001). Dans le cadre du même programme, il a également effectué une visite à l'INRIA Sophia-Antipolis (projet Comore), 18-22 avril 2001.

François Rechenmann a donné l'exposé « Introduction à la bio-informatique » lors du séminaire du DEA ECD, en collaboration avec la filière *Bioinformatique et Modélisation* de l'INSA de Lyon et l'université Claude Bernard.

François Rechenmann a donné une conférence intitulée « La bio-informatique : l'analyse informatique de l'information génomique », CARA, Lyon, 1 mars 2001.

François Rechenmann a fait un exposé sur « Acquisition, modélisation, gestion et analyse des données génomiques » lors de la journée du pôle numérique, Château de la Baume, Grenoble, 9 mars 2001.

François Rechenmann a fait l'exposé « Modélisation et simulation des réseaux de régulation de l'expression des gènes », ENST Paris, ParisTech, 29 mars 2001 (avec la contribution de Hidde de Jong).

François Rechenmann a donné une conférence intitulée « Informatique et génomique » lors du séminaire au Centre de Recherche Européen de Xerox (XRCE, Meylan), 12 juin 2001.

François Rechenmann a animé un atelier lors de la journée *Biotechno*, Palais des Congrès, Lyon, 22 juin, 2001.

François Rechenmann a donné une conférence invitée intitulée « Informatique et génomique » lors de la plateforme AFIA, Grenoble, 27 juin 2001.

François Rechenmann a participé à la table ronde « Comment gérer la masse de données obtenues » lors de la journée de conférence *Bioinformatique et bio-industrie – Quels enjeux pour l'Europe ?*, organisée par *L'Usine Nouvelle*, Evry, 18 janvier 2001.

Frédéric Boyer, Gisèle Bronner, Jean-François Dufayard, Céline Hernandez, Hidde de Jong, Anne Morgat, Marie-France Sagot et Danielle Ziébelin ont participé aux *Journées Ouvertes Biologie Informatique Mathématiques (JOBIM 2001)*, Toulouse, 29 mai - 1 juin 2001. Lors de cette réunion, Hidde de Jong a présenté le papier « Qualitative simulation of the initiation of sporulation in *B. subtilis* » (réalisé avec Johannes Geiselmann, Céline Hernandez et Michel Page).

Hidde de Jong a présenté le papier « Qualitative simulation of genetic regulatory networks : Method and application » par H. de Jong, M. Page, C. Hernandez et J. Geiselmann lors du

Fifteenth International Workshop on Qualitative Reasoning (QR-01), San Antonio (États-Unis), 17-19 mai 2001.

Michel Page a présenté le papier « Qualitative simulation of genetic regulatory networks : Method and application » par Hidde de Jong, Michel Page, Céline Hernandez et Johannes Geiselmann lors du *Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01)*, Seattle (États-Unis), 4-8 août 2001.

Hidde de Jong a donné l'exposé « Simulation qualitative de réseaux de régulation génique » au Laboratoire Génétique et Cancer (CNRS UMR5641), Lyon, 9 mars 2001.

Hidde de Jong a donné une conférence invitée lors de l'atelier CONSENSUS au Génopole d'Evry, 11 avril 2001.

Hidde de Jong a donné un exposé sur la « Qualitative simulation of the initiation of sporulation in *B. subtilis* » à l'École Supérieure d'Ingénieurs de Luminy (ESIL), universite de la Mediterranée, Marseille, 12 juin 2001.

Frédéric Boyer et Anne Morgat ont présenté le poster « Representation and integration of metabolic and genomic data for ab initio reconstruction of metabolic pathways » lors du *2nd workshop on Computation of Biochemical Pathways and Genetic Networks*, Villa Bosch, Heideleber, 21-22 Juin 2001. Anne Morgat a également animé la discussion de la session « Genetic networks ».

Anne Morgat a fait un exposé sur « Representation and integration of genomic and metabolic data : the Panoramix project » lors de la réunion « Data integration in functional genomics and proteomics » de l'*European Science Foundation (ESF)*, Genève, 15-17 Octobre 2001.

Frédéric Boyer, Céline Hernandez, Hidde de Jong, Anne Morgat, Sébastien Provencher, Marie-France Sagot et Alain Viari ont participé au séminaire *Algorithmique et Biologie* à l'université Claude Bernard (Lyon), 20-21 septembre 2001. Dans le cadre de ce séminaire, Hans Geiselmann et Hidde de Jong ont fait un exposé « Qualitative simulation of the initiation of sporulation in *B. subtilis* ». Anne Morgat a présenté son travail sur les bases de données fédérées dans Panoramix.

Frédéric Boyer, Céline Hernandez, Hidde de Jong, Anne Morgat, Sébastien Provencher, Marie-France Sagot et Alain Viari ont participé à la première réunion du groupe de travail *Bioinformatique fonctionnelle des systèmes de régulations génétiques* de l'action *Informatique, Mathématique, Physique pour la Génomique (IMPG)*, 30-31 mars 2001, Marseille. Exposés de Hidde de Jong et d'Anne Morgat.

Plusieurs membres du projet ont participé à une réunion du groupe de travail *Bases de données, interfaces et ontologies* de l'action *Informatique, Mathématique, Physique pour la Génomique (IMPG)*, Grenoble, 26-27 avril 2001. Exposé d'Anne Morgat (« Knowledge bases for genomics data »).

Alain Viari a fait un exposé sur le thème « Outils et stratégies d'annotation de génomes » à l'occasion des *Journées de Biologie Végétale* organisées dans le cadre de la Formation Permanente par le CNRS et l'INRA, Carry-Le-Rouet, 18-23 mars 2001.

Alain Viari a donné une conférence intitulée « Modèles : de la représentation des connaissances à la simulation », à l'occassion des *Journées Post-Génomique de la Doua*, Lyon, 5-6 avril, 2001.

Manolo Gouy a participé au jury de thèse de Catherine Letondal, Institut Pasteur, Paris, 27 Septembre 2001.

Laurent Duret a participé au jury de thèse de Gwenael Piganeau, université Claude Bernard, Lyon, Septembre 2001.

Hidde de Jong a participé au jury de thèse d'Ivayla Vatcheva, université de Twente, Enschede (Pays-Bas), 31 août 2001.

Hidde de Jong a participé au jury de thèse de Didier Morel, université Joseph Fourier, Grenoble, 19 décembre 2001.

François Rechenmann a participé au jury de thèse de Denys Proux, université de Bourgogne, Dijon, 9 avril 2001.

10 Bibliographie

Ouvrages et articles de référence de l'équipe

[1] H. de Jong, M. Page, C. Hernandez, J. Geiselmann, « Qualitative simulation of genetic regulatory networks : Method and application », *in : Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI-01*, B. Nebel (éditeur), Morgan Kaufmann, p. 67–73, San Mateo, CA, 2001.

[2] L. Duret, G. Perrière, M. Gouy, « HOVERGEN : Database and software for comparative analysis of homologous vertebrate genes », *in : Bioinformatics Databases and Systems*, S. Letovsky (éditeur), Kluwer Academic Publishers, Boston, 1999, p. 13–29.

[3] N. Galtier, N. Tourasse, M. Gouy, « A nonhyperthermophilic common ancestor to extant life forms », *Science 282*, 1999, p. 220–221.

[4] J. Lobry, « Asymmetric substitution patterns in the two DNA strands of bacteria », *Molecular Biology and Evolution 13*, 5, 1996, p. 660–665.

[5] L. Marsan, M.-F. Sagot, « Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification », *Journal of Computational Biology 7*, 2000, p. 345–362.

[6] G. Matassi, P. Sharp, C. Gautier, « Chromosomal location effects on gene sequence evolution in mammals », *Current Biology 9*, 15, 1999, p. 786–791.

[7] C. Médigue, F. Rechenmann, A. Danchin, A. Viari, « Imagene : an integrated computer environment for sequence annotation and analysis », *Bioinformatics*, 15, 1999, p. 2–15.

[8] C. Médigue, M. Rose, A. Viari, A. Danchin, « Detecting and analyzing DNA sequencing errors : toward a higher quality of the Bacillus subtilis genome sequence », *Genome Research 9*, 11, 1999, p. 1116–1127.

[9] G. Perrière, L. Duret, M. Gouy, « HOBACGEN : Database system for comparative genomics in bacteria », *Genome Research 10*, 2000, p. 379–385.

[10] D. Proux, F. Rechenmann, L. Julliard, « A pragmatic information extraction strategy for gathering data on genetic interactions », *in : Proceedings of the 8th International Conference on Intelligent Systems in Molecular Biology (ISMB 2000)*, AAAI Press, p. 279–285, 2000.

Articles et chapitres de livre

[11] P. Blayo, P. Rouzé, M.-F. Sagot, « Orphan gene finding – An exon assembly approach », *Theoretical Computer Science*, 2001.

[12] G. Bronner, B. Spataro, C. Gautier, F. Rechenmann, « GeMCore, a knowledge base dedicated to mapping mammalian genomes », *in : Computational Biology*, O. Gascuel et M. Sagot (éditeurs), *X-164*, Springer, 2001, p. 12–23.

[13] G. Bronner, B. Spataro, M. Page, C. Gautier, F. Rechenmann, « Modeling comparative mapping using objects and associations : Modeling comparative mapping using objects and associations », *Computers and Chemistry*, 2001.

[14] I. Chambaud, R. Heilig, S. Ferris, V. Barbe, D. Samson, F. Galisson, I. Moszer, K. Dybvig, H. Wroblewski, A. Viari, E. Rocha, A. Blanchard, « The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis* », *Nucleic Acids Research 29*, 10, 2001, p. 2145–2153.

[15] L. Gueguen, « Segmentation by maximal predictive partitioning according to composition biases », *in : Computational Biology*, O. Gascuel et M. Sagot (éditeurs), *X-164*, Springer, 2001, p. 32–44.

[16] S. Guidon, G. Perrière, « Intragenomic base content variation is a potential source of biases when searching for horizontally transferred genes », *Molecular Biology and Evolution 18*, 9, 2001, p. 1838–1840.

[17] G. Perrière, J. Thioulouse, « Use of Correspondence Discriminant Analysis to predict the subcellular location of bacterial proteins », *Computational Methods and Programs in Biomedecine*, 2001.

[18] G. Piganeau, D. Mouchiroud, L. Duret, C. Gautier, « Expected relationship between the silent substitution rate and GC content : Implication for the evolution of isochores », *Journal of Molecular Evolution*, 2001.

[19] E. Rocha, A. Danchin, A. Viari, « Evolutionary role of restriction/modification systems as revealed by comparative genome analysis », *Genome Research 11*, 6, 2001, p. 946–958.

[20] E. M. Rodrigues, M.-F. Sagot, Y. Wakabayashi, « Some approximation results for the maximum agreement forest problem », *in : Approximation, Randomization and Combinatorial Optimization : Algorithms and Techniques (APPROX & RANDOM 2001)*, M. Goemans, K. Jansen, J. D. P. Rolim, et L. Trevisan (éditeurs), *Lecture Notes in Computer Science, 2129*, Springer-Verlag, 2001, p. 159–169.

Communications à des congrès, colloques, etc.

[21] V. Daubin, M. Gouy, G. Perrière, « A phylogenetic approach using supertrees to reconstruct prokaryotic history », *in : Recueil des Actes des Journées Ouvertes Biologie Informatique Mathématiques, JOBIM 2001*, L. Duret, C. Gaspin, T. Schiex (éditeurs), p. 3–9, Toulouse, 2001.

[22] H. de Jong, J. Geiselmann, C. Hernandez, M. Page, « Qualitative simulation of the initiation of sporulation in *B. subtilis* », *in : Recueil des Actes des Journées Ouvertes Biologie Informatique Mathématiques, JOBIM 2001*, L. Duret, C. Gaspin, T. Schiex (éditeurs), p. 187–194, Toulouse, 2001.

[23] H. de Jong, M. Page, C. Hernandez, J. Geiselmann, « Qualitative simulation of genetic regulatory networks : Method and application », *in : Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI-01*, B. Nebel (éditeur), Morgan Kaufmann, p. 67–73, San Mateo, CA, 2001.

[24] H. de Jong, M. Page, C. Hernandez, J. Geiselmann, « Qualitative simulation of genetic regulatory networks : Method and application », *in : Proceedings of the Fifteenth International Workshop on Qualitative Reasoning, QR-01*, G. Biswas (éditeur), p. 134–141, San Antonio, TX, 2001.

[25] I. Vatcheva, O. Bernard, H. de Jong, J.-L. Gouzé, N. Mars, « Discrimination of semi-quantitative models by experiment selection : Method and application in population biology », *in : Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI-01*, B. Nebel (éditeur), Morgan Kaufmann, p. 121–127, San Mateo, CA, 2001.

[26] I. Vatcheva, O. Bernard, H. de Jong, J.-L. Gouzé, N. Mars, « Discrimination of semi-quantitative models by experiment selection : Method and application in population biology », *in : Proceedings of the Fifteenth International Workshop on Qualitative Reasoning, QR-01*, G. Biswas (éditeur), p. 121–127, San Antonio, TX, 2001.

Rapports de recherche et publications internes

[27] H. de Jong, J. Geiselmann, C. Hernandez, M. Page, « Genetic Network Analyzer : A tool for the qualitative simulation of genetic regulatory networks », *rapport de recherche n°RR-4262*, INRIA Rhône-Alpes, Montbonnot Saint-Martin, 2001, `http://www.inria.fr/rrrt/rr-4262.html`.

[28] N. Pisanti, M.-F. Sagot, « Further thoughts on the synteny distance between genomes », *rapport de recherche n°IGM-00-14*, Institut Gaspard Monge, Université de Marne-la-Vallée, 2001, accepté pour *Algorithmica*.