INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Action HELIX

Informatique et génomique

Rhône-Alpes

THÈME 3A

Table des matières

1 Composition de l'équipe

Chercheurs INRIA permanents

François Rechenmann [directeur de recherche, responsable du projet]

Alain Viari [directeur de recherche]

Hidde de Jong [chargé de recherche]

Chercheurs et ingénieurs INRIA non permanents

Stéphane Declere [ingénieur expert (à l'Atelier de Bio-Informatique, Paris 6)]

Céline Hernandez [ingénieur associé]

Anne Morgat [chercheur associé]

Ingénieurs détachés

Charles Metivier [(à temps partiel), société Aureus Pharma, Paris]

Post-doctorants

Hélène Rivière-Rolland [action de développement GénoStar]

Doctorants

Frédéric Boyer [allocataire du Ministère de la Recherche, directeurs de thèse : Laurent Trilling, université Joseph Fourier, et Alain Viari, INRIA]

Gisèle Bronner [allocataire du Ministère de la Recherche, UMR 5558, Lyon]

Denys Proux [CIFRE, XRCE, Meylan]

Assistante de projet

Françoise de Coninck

2 Présentation et objectifs généraux

Mots clés : 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 de l'action 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 cinq axes majeurs : la modélisation des réseaux géniques et métaboliques ; la cartographie comparée ; la protéomique ; l'extraction d'informations à partir de textes.

2.2.1 Modélisation des réseaux géniques et métaboliques

Participants : Frédéric Boyer, Hidde de Jong, Céline Hernandez, Hélène Rivière-Rolland, Anne Morgat, Alain Viari.

La modélisation et la simulation des systèmes régulateurs géniques sont soumises à deux contraintes importantes. D'une part, les mécanismes sous-jacents aux interactions du réseau ne sont pas bien connus, ce qui implique que des modèles cinétiques complets ne peuvent pas être construits. D'autre part, des informations quantitatives sur les paramètres caractérisant les interactions sont rarement disponibles. Par conséquent, les méthodes de simulation numérique sont difficiles à appliquer ici et il faut recourir à des méthodes capables de traiter des modèles qualitatifs.

Le métabolisme peut être défini comme l'ensemble des réactions biochimiques à l'intérieur d'une cellule vivante, qui contrôlent la croissance, le fonctionnement (synthèses et dégradations de molécules) et la division de cette cellule.

Une quantité importante de données existe d'ores et déjà sur le métabolisme, mais ces données ne sont généralement pas formalisées. Il est donc nécessaire, dans un premier temps de modéliser ces connaissances, afin de les intégrer dans des programmes de traitement automatisé, puis dans un deuxième temps, d'une part de relier ces données métaboliques aux séquences génomiques, aux structures protéiques, à la régulation de l'expression des gènes, à l'organisation des génomes et à leur évolution, et d'autre part d'inférer, à partir de ces liens, des connaissances métaboliques sur l'ensemble des organismes, ce qui contribuera à améliorer le processus d'annotation de nouveaux génomes.

2.2.2 Protéomique

Participant : Alain Viari.

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.

Réciproquement, les données génomiques brutes (séquences des chromosomes) fournissent une information topologique (localisation des gènes codant pour les protéines étudiées, par exemple) souvent fondamentale (par exemple, dans l'identification de maladies génétiques), mais en partie déconnectée de cette dimension « expression ».

L'objectif de ces travaux de recherche est de réconcilier ces deux aspects fondamentaux du fonctionnement cellulaire en croisant des données d'expression issues d'expériences de protéomique et les données de séquences chromosomiques complètes. Il s'agit de localiser des étiquettes protéinique sur de l'ADN génomique. Une étiquette est constituée d'une portion de séquence peptidique flanquée de deux parties de séquences inconnues, mais de masse totale connue.

2.2.3 Cartographie comparée et synténies

Participants : Gisèle Bronner, Anne Morgat, François Rechenmann, Alain Viari.

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. Un second axe de travail concerne la recherche et l'analyse de groupes de gènes dont la localisation chromosomique est conservée entre deux ou plusieurs espèces bactériennes (synténies bactériennes). Comme précédemment, la définition de cette conservation doit prendre en compte des opérations de permutation et d'insertion/délétion des gènes. Cet axe est directement connecté au thème 2.2.1 dans la mesure où ces groupes de synténie sont en mesure de fournir d'importantes informations sur le rôle fonctionnel, sur la co-expression ou la co-régulation des gènes qui les constituent

2.2.4 Extraction d'informations à partir de textes

Participants : Denys Proux, François Rechenmann.

Si les bases de données biologiques se sont considérablement développées ces dernières années, tant en taille qu'en domaines couverts, 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. Ainsi, les travaux de l'action Helix visent à extraire automatiquement des données sur les interactions moléculaires à partir de résumés d'articles.

L'approche suivie repose sur une analyse syntaxique (« shallow parsing ») des textes et la construction de graphes conceptuels qui représentent le contenu des phrases relativement aux requêtes posées. Ce travail se fait en collaboration étroite avec le Centre de Recherche Européen de Xerox (XRCE) à Meylan.

2.3 Relations internationales et industrielles

L'action Helix entretient des liens forts avec l'équipe « Biométrie et biologie évolutive » de l'UMR CNRS 5558 à Lyon, en particulier sur le thème de la cartographie comparée.

Elle est en contact avec les différentes équipes de bioinformatique française, en particulier au sein des différentes génopoles.

Elle participe 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.

L'action Helix bénéficie également 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.

Enfin, l'équipe Helix assiste Aureus Pharma dans la conception de son système d'aide à l'identification de cibles thérapeutiques. Elle accueille, à temps partiel, un ingénieur de cette jeune société.

3 Fondements scientifiques

Le thème fédérateur des travaux de l'action Helix est la modélisation des connaissances biologiques. Le principe est de représenter explicitement et formellement toute entité qui est manipulée par les programmes d'analyse et de visualisation. Cette explicitation est un préalable indispensable dans un domaine où un terme, même aussi fondamental que « gène », est interprété différemment selon les contextes. La formalisation permet par ailleurs d'appliquer des mécanismes de vérification, de requêtes et d'inférence particulièrement utiles dans une démarche qui reste résolument exploratoire.

Les modèles à objets sont bien adaptés à la représentation des diverses entités biologiques et les membres de l'équipe bénéficient d'une longue expérience de conception de bases de

connaissances à objets dans plusieurs domaines biologiques. Cette activité de conception a toujours été menée en interaction forte avec des équipes de biologie, autour d'une problématique bien définie.

Mais il est vite apparu indispensable de représenter finement les relations entre les entités, que ce soit par exemple la position relative de « marqueurs » sur un chromosome, l'activation ou la répression d'un gène par le produit d'un autre, ou encore les interactions entre enzymes et substrats. L'équipe Helix met donc en œuvre un modèle de connaissances dans lequel les relations sont explicites, hiérarchiquement organisées, possèdent des attributs et sont susceptibles d'être l'objet de requêtes. Ainsi, la plupart des bases de connaissances développées au sein d'Helix ou en partenariat avec d'autres équipes, utilisent, le modèle AROM conçu et développé au sein de l'action Romans.

Cette représentation explicite des relations ente entités conduit à la modélisation des réseaux, géniques et métaboliques, et à leur simulation. Compte tenu de l'insuffisance des données sur les interactions moléculaires, il est nécessaire de recourir à des simulations qualitatives par lesquelles le comportement dynamique d'un réseau peut être étudié sans avoir à disposer des valeurs numériques des paramètres, mais seulement de leurs valeurs relatives.

Par ailleurs, de nouveaux algorithmes sont conçus et développés pour analyser de nouvelles catégories de données, telles que celles qui sont obtenues en protéomique à l'aide de spectromètres de masse.

4 Domaines d'applications

Par essence même de l'action 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. cerevisiae, 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) précède de quelques mois celui de l'Homme (plus de trois milliards de « lettres ») dont il n'existe actuellement qu'une version de travail (« 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 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. De plus, d'autres classes de données peuvent être analysées 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 Metabolic Viewer

Participant : Alain Viari [Correspondant].

Metabolic Viewer est un programme d'affichage à la volée des voies métaboliques. Il permet de générer et visualiser de manière non statique des graphes représentant les enchaînements de réactions métaboliques.

Les nœuds du graphe correspondent aux composés biochimiques et aux enzymes intervenant dans ces réactions et les arcs aux transformations biochimiques. Des motifs principaux présents dans les voies métaboliques (actuellement les cercles qui symbolisent des cycles) sont détectés et disposés automatiquement grâce à un algorithme récursif de placement par force. Les autres nœuds se disposent ensuite autour de ces motifs grâce à ce même algorithme, jusqu'à l'obtention d'une conformation stable. Metabolism Viewer autorise ensuite l'utilisateur à modifier manuellement le graphe obtenu.

5.2 GNA

Participant : Hidde de Jong [Correspondant].

Les entrées de GNA (Genetic Network Analyzer) se composent du modèle mathématique d'un réseau d'interactions géniques et d'un état qualitatif initial. Les résultats de simulation sont produits sous forme d'un graphe des états 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.

5.3 Metabolic KB

Participant : Alain Viari [Correspondant].

Metabolic KB est une base de connaissances dédiée à la représentation et la gestion des données du métabolisme intermédiaire. Son implémentation repose sur le modèle AROM développé dans l'action ROMANS. A l'heure actuelle, elle contient près de 5000 composés chimiques regroupés en 3000 réactions, pour la plupart issues de données publiques (bases KEGG et SwissProt).

5.4 PSKB

Participante : Anne Morgat [Correspondant].

PSKB (Prokaryotic Synteny Knowledge Base) est une base de connaissances dédiée à la représentation et la gestion des données de synténie bactérienne. Son implémentation repose sur

le modèle AROM développé dans l'action ROMANS. Elle regroupe les informations d'orthologie et de synténie concernant tous les couples d'organismes procaryotes entièrement séquencées à ce jour (environ 40 génomes).

6 Résultats nouveaux

6.1 Modélisation et simulation des réseaux géniques et métaboliques

Participants : Frédéric Boyer, Hidde de Jong, Céline Hernandez, Hélène Rivière-Rolland, Anne Morgat, Alain Viari.

Dans la méthode de simulation qualitative de réseaux d'interactions géniques développée par Hidde de Jong au sein de l'action Helix et Michel Page de l'action Romans, les réseaux sont modélisés par une classe d'équations différentielles linéaires par morceaux dans lesquelles le taux d'expression d'un gène est défini en fonction de la concentration des protéines régulatrices codées par d'autres gènes. Les interactions sont représentées, non pas par une description détaillée des réactions biochimiques qui interviennent dans la régulation, mais par des fonctions en escalier permettant de faire abstraction des mécanismes de réaction. Les fonctions en escalier sont de bonnes approximations des courbes sigmoïdes dont la validité a été montrée expérimentalement.

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. À cette fin, l'espace de phase est divisé en états qualitatifs correspondant aux états fonctionnels du système. La forme mathématique simple des équations différentielles linéaires par morceaux permet de déterminer des transitions possibles entre les états qualitatifs en raisonnant sur les égalités et les inégalités entre paramètres. La simulation qualitative consiste en la génération de tous les états qualitatifs atteignables par une ou plusieurs transitions à partir d'un état initial. Puisque le nombre d'états qualitatifs est fini, chaque succession de transitions d'état mène, soit à un état stable, soit à un cycle entre un ensemble d'états. À cause de la nature qualitative du modèle, un état qualitatif peut avoir plusieurs états successeurs, ce qui peut donner lieu à une structure arborescente de la simulation et donc à plusieurs états stables ou cycles. Des expériences de simulation ont montré que des modèles de réseaux d'interactions d'environ 20 composants, liés par des boucles de rétroaction, peuvent être traités.

L'implantation de la méthode a été réalisée en Java dans l'environnement GNA (Genetic Network Analyzer), doté d'interfaces de visualisation des réseaux et des graphes d'états développées par C. Hernandez et par S. Maza (dans le cadre de son stage de magistère). Le tout a été testé sur des modèles simples de réseaux bien connus, comme la décision entre lyse et lysogénie chez le bactériophage? Actuellement, GNA est utilisé pour la modélisation et la simulation d'autres processus de régulation bactériens, moins connus et plus complexes, comme l'induction du cycle lytique chez le bactériophage Mu et le début de la sporulation chez B. subtilis. Ces travaux sont effectués en collaboration avec l'équipe de Hans Geiselmann du laboratoire « Plasticité et Expression des Génomes Microbiens» (CNRS EP2029) de l'université Joseph Fourier (Grenoble).

Par ailleurs, une base de connaissances sur les données métaboliques a été développée. Elle met en oeuvre le système AROM développé par l'action Romans, qui utilise des associations n-aires pour représenter les relations entre les objets et les références à d'autres objets. L'instanciation de la base a été réalisée à partir de données publiques, extraites des bases de données KEGG et Swiss-Prot. La cohérence des données dans la base a ensuite été testée grâce à des programmes d'alignement de séquences peptidiques, issus du domaine public ou bien développés par des chercheurs de l'équipe. La visualisation graphique des réseaux est réalisée « à la volée» à l'aide du programme Metabolic Viewer (cf paragraphe 5.1).

Le prolongement naturel de ce travail est la conception et le développement de mécanismes d'inférence de réseaux : il s'agit de construire ou de compléter un réseau métabolique d'une espèce en utilisant soit le réseau connu d'une autre et les relations d'homologie entre les constituants, soit par une reconstruction *ab-nihilo* à partir de l'ensemble des réactions possibles et de considérations de bilan énergétique. Ce travail est actuellement en cours dans le cadre de la thèse de Frédéric Boyer.

6.2 Protéomique

Participant : Alain Viari.

L'objectif est ici de localiser directement des informations d'expression (étiquettes protéiques) sur les séquences génomiques sans passer par une reconstruction de l'organisation génique du chromosome, reconstruction qui pose actuellement de nombreux problèmes théoriques et pratiques. Une étiquette protéique est constituée d'une portion de séquence peptidique flanquée de deux parties de séquences inconnues, mais de masse totale connue.

Un premier algorithme de localisation rapide, sur de l'ADN génomique, des zones potentiellement codantes pour chaque étiquette a été conçu. La difficulté essentielle provient de l'organisation en mosaïque des gènes eucaryotes (introns/exons) qui interdit de faire l'hypothèse que les étiquettes couvrent une région contiguë du chromosome. Une seule étiquette peut s'avérer insuffisante pour localiser de manière unique un peptide sur un chromosome eucaryote complet (en raison du nombre important de « hits» possibles). En revanche, lorsque plusieurs étiquettes sont disponibles, il devient possible, à partir de l'analyse statistique de leur distribution sur le chromosome (tests d'aggrégation), de proposer une (ou quelques) localisation(s) possible(s). Néanmoins, la stratégie de localisation précédente échoue lorsque la répartition des exons du gène cible est très dispersée. Il devient alors nécessaire de faire intervenir d'autres informations telles que les ESTs (Expressed Sequence Tags). Les premiers résultats, obtenus sur le chromosome 22 humain sont très encourageants et permettent d'ores et déjà de valider l'approche.

D'un point de vue expérimental, le projet repose sur la plate-forme instrumentale développée au Laboratoire de Chimie des Protéines (Jérôme Garin, CEA, Grenoble) constituée d'un nano- chromatographe liquide (nano-LC) couplé à un spectromètre de masse (Q-Tof). Cette technique, extrêmement novatrice, permet en effet de produire rapidement une grande quantité d'informations à partir d'échantillons biologiques ciblés (par opposition aux techniques de type gel 2D, plus lentes).

6.3 Cartographie comparée et synthénies

Participants : Gisèle Bronner, Anne Morgat, François Rechenmann, Alain Viari.

Les travaux sur la cartographie comparée sont menés en collaboration étroite avec l'équipe de l'UMR 5558, en particulier à travers le co-encadrement de la thèse de Gisèle Bronner par François Rechenmann et Christian Gautier. Ils portent sur le développement du système GemCore.

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 été menés dans le cadre du stage de magistère d'informatique de Jean-François Dufayard, en collaboration avec Laurent Duret de l'UMR 5558 à Lyon. Ils 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.

Une nouvelle activité de génomique comparative a débuté avec l'arrivée en octobre d'Anne Morgat au sein de l'équipe. L'objectif de cette activité est la description des segments synténiques conservés entre plusieurs espèces bactériennes et l'utilisation de ces connaissances à des fins d'inférence de fonction. Une chaîne de traitement a été développée pour la recherche systématique des relations d'orthologie et de synténie entre deux génomes bactériens complets. La définition de l'orthologie repose sur la comparaison deux à deux de tous les produits des gènes de chaque bactérie avec l'autre. La définition de la synténie repose sur la recherche d'un chemin (liant les gènes appartenant à un même groupe appelé «synton») dans un graphe. Ces deux opérations, effectuées pour toutes les paires de génomes bactériens complets (environ 40 génomes actuellement), sont relativement coûteuses en temps de calcul. Nous envisageons, à terme, d'utiliser dans ce but la grappe de PC disponible dans l'UR. Enfin, l'ensemble des informations (gènes, protéines, relations d'orthologies, syntons) a été modélisé et le schéma conceptuel a été instancié dans le modèle AROM. Une interface spécialisée permettant d'émettre des requêtes et de visualiser graphiquement les syntons a également été développée.

6.4 Extraction d'informations à partir de textes

Participants : Denys Proux, François Rechenmann.

L'approche suivie par Helix et XRCE repose sur une analyse des textes qui met en œuvre les outils linguistiques développés par XRCE et exploite des connaissances sur le domaine (« ontologies »). Il existe maintenant un prototype d'un système d'extraction d'informations

à partir de textes qui a été validé sur un corpus de plusieurs centaines de phrases décrivant ou non des interactions moléculaires. Ce corpus a été préparé par l'équipe de Bernard Jacq à Marseille à partir de commentaires extraits de la base de données FlyBase sur la drosophile. Les tests se sont poursuivis sur des résumés extraits de la base documentaire Medline. Le bénéfice de la redondance de la littérature scientifique apparaît alors : si une interaction n'a pas été détectée dans un résumé, elle peut l'être dans un autre où elle est décrite différemment.

En parallèle, un projet, moins ambitieux et destiné à court terme à un logiciel opérationnel, a été monté. L'analyse de plusieurs projets, en particulier d'annotation de génomes, montre en effet que le souci de nombreuses équipes est de relier leurs données expérimentales aux informations contenues dans les articles de référence. Dans toutes ces situations, le lien privilégié est constitué par les noms des entités pertinentes. Ainsi, la reconnaissance des noms de gènes, de protéines ou de voies métaboliques est la phase incontournable de tout travail visant à exploiter les connaissances de la littérature spécialisée.

Dans ce contexte, un nouveau projet rassemble XRCE, deux équipes de l'INRA (à Versailles et à Gand) et l'équipe HELIX pour développer un module logiciel capable de reconnaître, dans un texte scientifique tel qu'un article ou un commentaire dans une base de données, les noms de ces différentes entités biologiques et de l'expérimenter en vraie grandeur sur des corpus de textes concernant plusieurs organismes.

6.5 Le projet GénoStar

Participants : Anne Morgat, François Rechenmann, Hélène Rivière-Rolland, Alain Viari.

L'objectif du projet GénoStar est de concevoir, développer et expérimenter un environnement modulaire de génomique exploratoire. Il est mené par un consortium qui associe l'Institut Pasteur de Paris, les sociétés Hybrigenics (Paris) et Genome Express (Grenoble) et l'INRIA. GénoStar est une action de développement de l'INRIA.

La modularité de l'environnement se concrétise par l'existence de plusieurs applications distinctes, mais qui bénéficient de services communs offerts par le noyau de l'environnement, GénoCore.

Dans la première phase, de deux ans, du projet, trois applications seront conçues et réalisées :

- GénoAnnot est dédiée à l'annotation de génomes procaryotes et eucaryotes. Les données d'entrée de cette application sont des séquences chromosomiques complètes ou partielles. Les méthodes qu'elle propose, enchaînées au sein de stratégies différentes suivant les objectifs visés et les organismes concernés, permettent d'identifier et de visualiser les régions d'intérêt biologique, telles que les régions codantes et les signaux de régulation.

- GénoLink est destinée à prolonger le processus d'annotation amorcé par GénoAnnot vers la caractérisation de la (ou des) fonction(s) des gènes identifiés, en exploitant des relations de voisinage entre gènes et en rapprochant ainsi des informations de sources diverses sur les gènes et leurs produits. Par exemple, deux gènes peuvent être considérés comme voisins parce qu'ils sont proches sur le chromosome, mais aussi parce que leurs

produits interviennent dans la même voie métabolique, ou encore parce qu'ils sont cités dans un même article.

- GénoBool permet d'explorer des ensembles de données hétérogènes à travers l'application de techniques d'analyse multifactorielle. Les données sont rendues homogènes par l'utilisation de codages booléens spécifiques à chaque type de données génomiques rencontré. GénoBool est ainsi susceptible de faire apparaître de nouvelles relations de voisinage intégrables dans GénoLink.

La conception des applications GénoAnnot et GénoLink a débuté, en collaboration respectivement avec Genome Express et Hybrigenics.

L'action Helix a ainsi développé avec Genome Express une structure de classes et d'associations décrivant les diverses entités qui interviennent dans le processus d'annotation d'un génome bactérien complet et les relations que ces entités entretiennent.

Simultanément, la société Hybrigenics a conçu le schéma conceptuel de l'application GénoLink et a commencé à l'instancier, ce qui conduit à des bases de très grande taille.

7 Contrats industriels (nationaux, européens et internationaux)

L'action 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 cet automne 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 est attendu en 2001. Enfin, GénoStar est une action de développement de l'INRIA, assurée d'un soutien pendant 3 ans (2000-2002).

7.2 GénoPlante

L'action 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, d'une durée de deux ans 2000-2001. Son rôle dans ce projet est de développer, à partir d'ImaGene, un prototype d'un environnement d'aide à l'annotation de génomes végétaux, en intégrant des méthodes proposées par les autres partenaires (INRA : Christine Gaspin

et Pierre Rouzé ; Institut Pasteur : Marie-France Sagot) et de l'expérimenter sur le génome de l'arabette.

7.3 Aureus Pharma

L'action Helix bénéficie également d'un contrat d'étude avec la jeune société Aureus Pharma. Elle accueille à ce titre un ingénieur de cette société, Charles Métivier. Ce contrat consiste à mener l'étude de faisabilité technique du système informatique de recherche de cibles thérapeutiques que projette Aureus Pharma.

7.4 XRCE

Le centre européen de recherche de Xerox (XRCE) est le partenaire privilégié de l'équipe Helix sur le thème de l'extraction d'informations à partir de textes. Ces travaux sont menés dans le cadre d'une convention CIFRE qui s'achève en décembre 2000, avec la soutenance de thèse prochaine de Denys Proux.

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 et de l'UMR 5558 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.

Accueil au sein de Helix de Frédéric Boyer, doctorant de Laurent Trilling, professeur à l'université Joseph Fourier sur le sujet de l'inférence de réseaux métaboliques.

Accueil à temps partiel au sein de Helix d'Eric Fanchon, chercheur CNRS à l'IBS (Institut de Biologie Structurale) sur la modélisation et la classification de structures tertiaires («folds») de protéines.

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 béta- lactamines » a été sélectionné et financé.

Sur la protéomique, collaboration avec Jérôme Garin (LCP : Laboratoire de Chimie des Protéines, CEA), 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.

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 Bio-Informatique) à Paris 6 (Joël Pothier), à l'Institut Pasteur (Marie-France Sagot), 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 ».

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 ».

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).

L'action Helix, ainsi que l'équipe de l'UMR 5558, participent au projet HAMAP d'annotation automatique de protéomes bactériens, à l'initiative de l'Institut Suisse de Bioinformatique (Amos Bairoch) à Genève.

Dans le cadre du programme d'actions intégrées franco-néerlandais Van Gogh, l'action Helix poursuit 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. L'action est planifiée sur deux ans, du printemps 1999 au printemps 2001, et prévoit des visites entre les deux projets des chercheurs participants.

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

9 Diffusion de résultats

9.1 Animation de la communauté scientifique

François Rechenmann a organisé la session « sciences du vivant » lors des journées « Doc-Forum – La biennale du savoir » à Lyon, le 27 janvier. Christian Gautier (université Claude Bernard, Lyon), Régine Vignes-Lebbe (université Paris 6) et Simon Tillier (Museum Natio-

nal d'Histoire Naturelle, Paris) ont fait des exposés sur le thème « Acquisition et gestion des connaissances dans les sciences du Vivant »

Première réunion du groupe de travail IMPG « Bases de données, interfaces et ontologies», les 5, 6 et 7 juin 2000 à l'université Paris 6, Ce groupe est animé par Philippe Bessières (INRA), Emmanuel Barillot (GénoPlante) et François Rechenmann

La première réunion du projet HAMAP d'annotation automatique de protéomes bactériens s'est tenue dans les locaux de l'unité de recherche Rhône-Alpes, les 20 et 21 novembre 2000. Plus de 25 personnes, en provenance d'équipes de bioinformatique françaises et suisses ont participé.

9.2 Enseignements universitaires

Maîtrise de biologie, filière « mathématiques-informatique », université Claude Bernard, Lyon cours de François Rechenmann : « La modélisation informatique des connaissances», 14h

DEA de biologie cellulaire et moléculaire de l'université Joseph Fourier (Grenoble), interventions de François Rechenmann et Alain Viari sur la bioinformatique, de 1h 30 chacune.

Hidde de Jong, François Rechenmann et Alain Viari sont intervenus dans l'option « bioinformatique» de la maîtrise d'informatique à l'Université Joseph Fourier (Grenoble).

Alain Viari intervient également au DEA de Biophysique de l'université Paris 6 (cours de 8h), ainsi qu'au DEA de Génétique de l'université Paris 6 (3h), aux DEAs de Biologie et Informatique (option bioinformatique) de Marseille (3h). Il est, par ailleurs, responsable de l'option «analyse de séquences» du module de maîtrise (Biophysique Moléculaire) de l'université Paris 6 (8 h). Enfin, il intervient pour 3h de cours à la formation permanente « bioinformatique» de l'université Paris 6.

9.3 Participation à des colloques, séminaires, invitations

Hidde de Jong a présenté le poster «Simulation of genetic regulatory systems : A qualitative approach» par H. de Jong et M. Page lors de la «8th International Conference on Intelligent Systems for Molecular Biology», ISMB 2000, à San Diego (États-Unis) du 19 au 23 août 2000

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 3 au 17 janvier, du 21 au 28 mai et du 25 septembre au 4 octobre 2000). Lors de la première visite, il a rencontré le 12 janvier les bioinformaticiens et biomathématiciens au «Centre for Biometrics» du CPRO- DLO (« Centre for Plant Breeding and Reproduction Research, Wageningen University & Research Centre »).

Hidde de Jong a effectué deux visites à l'INRIA Sophia-Antipolis (projet Comore) dans le cadre du programme d'actions intégrées franco-néerlandais Van Gogh (du 23 au 25 juillet et du 23 au 25 novembre 2000). Lors de la deuxième visite, il a également présenté l'exposé «Simulation of genetic regulatory systems: A qualitative approach».

Hélène Rivière-Rolland et Gisèle Bronner ont présenté leurs travaux, respectivement sur « Modélisation par objets et relations des connaissances sur les réseaux métaboliques » et « Modélisation des connaissances et cartographie comparée chez les mammifères » lors des journées du groupe de travail « Bases de données, interfaces et ontologies » de l'action IMPG

(Informatique Mathématique Physique pour la Génomique), à l'université Paris 6, le 5 juin 2000.

École d'été « Analyse in silico de séquences génomiques », Marseille, juillet 2000, intervention de François Rechenmann sur le thème « La modélisation informatique des connaissances biologiques »

Séminaire de François Rechenmann à l'université de Munich (Institut für Informatik), le 15 février : « Object-based knowledge modeling : entities, relations and tasks »

Dans le cadre du séminaire « Algorithmique et biologie » de l'Institut Pasteur, 13-15 mars 2000, exposés de Hidde de Jong « Qualitative simulation of genetic regulatory systems » et de François Rechenmann « Acquisition and representation of knowledge on interaction networks »

Séminaire de François Rechenmann à l'IBS (Institut de Biologie Structurale, Grenoble), le 11 février 2000 : « Génomique et recherche en informatique »

Dans le cadre du colloque « Mathématiques et biologie », organisé à l'ENS Ulm, les 26 et 27 mai, François Rechenmann a fait un exposé sur « L'analyse informatique des données génomiques »

Journées « 100 ans de l'INPG », ENS-IMAG, Grenoble, exposé de François Rechenmann : « Enjeux et besoins de la bioinformatique », 13 octobre 2000

Journée ARTEB « Expression des gènes et bioinformatique », exposés d'Alain Viari « De la séquence d'ADN à la modélisation des interactions géniques » et François Rechenmann « Stratégie bioinformatique nationale et régionale », Lyon, 5 décembre 2000

Séminaire d'Alain Viari à l'INRA de Bordeaux (laboratoire des mycoplasmes, invitation A. Blanchard) le 25 avril 2000 : « Des sacs à gènes aux génomes : a propos de l'organisation des génomes bactériens »

Intervention d'Alain Viari au séminaire « Genoplante » du 25 au 26 juillet 2000 à la Grande Motte : « Stratégies et Outils pour l'annotation de génomes complets ».

10 Bibliographie

Ouvrages et articles de référence de l'équipe

[1] H. de Jong, M. Page, « Qualitative simulation of large and complex genetic regulatory systems », *in : Proc. 14th Europ. Conf. Artif. Intell. (ECAI 2000)*, W. Horn (éditeur), IOS Press, p. 141–145, 2000.

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

[3] 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 Res. 9*, 11, 1999, p. 1116–1127.

[4] D. Proux, F. Rechenmann, L. Julliard, « A pragmatic information extraction strategy for gathering data on genetic interactions », *in : Proc. 8th Int. Conf. Intell. Syst. Mol. Biol. (ISMB 2000)*, AAAI Press, p. 279–285, 2000.

[5] E. Rocha, A. Danchin, A. Viari, « Translation in Bacillus subtilis: roles and trends of initiation and termination, insights from a genome analysis », *Nucleic Acids Res. 27*, 17, 1999, p. 3567–3576.

[6] E. Rocha, A. Danchin, A. Viari, «Universal replication biases in bacteria», *Mol. Microbiol. 32*, 1, 1999, p. 11–16.

Articles et chapitres de livre

[7] G. Achaz, E. Coissac, A. Viari, P. Netter, «Analysis of intrachromosomal duplications in yeast Saccharomyces cerevisiae: a possible model for their origin», *Mol. Biol. Evol. 7*, 8, 2000, p. 1268–1275.

[8] H. de Jong, M. Page, C. Hernandez, H. Geiselmann, S. Maza, «Modeling and simulation of genetic regulatory networks», *ERCIM News 43*, 2000, p. 18–19.

[9] P. Hoyne, L. Edwards, A. Viari, L. Maher 3rd, «Searching genomes for sequences with the potential to form intrastrand triple helices», *J. Mol. Biol. 302*, 4, 2000, p. 797–809.

[10] F. Rechenmann, C. Gautier, «Donner un sens au génome», *La Recherche*, juin 2000, p. 39–45.

[11] F. Rechenmann, «From data to knowledge», *Bioinformatics 16*, 5, 2000, p. Bioinformatics, editorial.

[12] H. Rivière-Rolland, L. Taloc, D. Ziébelin, F. Rechenmann, A. Viari, «Modelling metabolism knowledge using objects and associations», *ERCIM News 43*, 2000, p. 21.

[13] E. Rocha, A. Danchin, A. Viari, «The DB case: pattern matching evidence is not significant», *Mol. Microbiol. 37*, 1, 2000, p. 216–218.

[14] E. Rocha, P. Guerdoux-Jamet, I. Moszer, A. Viari, A. Danchin, «Implication of gene distribution in the bacterial chromosome for the bacterial cell factory», *J. Biotechnol. 78*, 3, 2000, p. 209–219.

Communications à des congrès, colloques, etc.

[15] G. Bronner, C. Gautier, F. Rechenmann, «An entity association model for comparative mapping», *in : Proc. German Conf. Bioinformatics*, Logos Verlag, p. 133–138, 2000.

[16] G. Bronner, C. Gautier, F. Rechenmann, «GeMCore, une base de connaissances dédié à la cartographie des génomes de mammifères», *in : Actes 1ères Journées Ouvertes : Biologie, Mathématiques, Informatique (JOBIM)*, p. 55–64, Montpellier, 2000.

[17] H. de Jong, M. Page, «Qualitative simulation of large and complex genetic regulatory systems», *in : Proc. 14th Europ. Conf. Artif. Intell. (ECAI 2000)*, W. Horn (éditeur), IOS Press, p. 141–145, 2000.

[18] H. de Jong, M. Page, «Qualitative simulation of large and complex genetic regulatory systems», *in : Working Notes 14th Int. Workshop Qualitative Reasoning (QR 2000)*, J. Flores (éditeur), p. 32–39, Morélia, Mexico, 2000.

[19] C. Gautier, S. Tillier, R. Vignes-Lebbe, F. Rechenmann, «Acquisition et gestion des connaissances dans les sciences du Vivant», *in : Les Savoirs Déroutés : Experts, Documents, Supports, Règles, Valeurs et Réseaux Numériques*, Les Presses de l'Enssib – Association Doc-Forum – La Biennale du savoir, p. 17–23, 2000.

[20] D. Proux, F. Rechenmann, L. Julliard, « Muninn: a pragmatic information extraction system », *in : Proc. String Processing and Information Retrieval Conf. (SPIRE 2000)*, IEEE Computer Society, p. 236–241, 2000.

[21] D. Proux, F. Rechenmann, L. Julliard, « A pragmatic information extraction strategy for gathering data on genetic interactions », *in : Proc. 8th Int. Conf. Intell. Syst. Mol. Biol. (ISMB 2000)*, AAAI Press, p. 279–285, 2000.

[22] C. Roux, D. Proux, F. Rechenmann, L. Julliard, « An ontology enrichment method for a pragmatic information extraction system gathering data on genetic interactions », *in : Working Notes 1st Workshop on Ontology Learning (OL 2000), held in conjunction with 14th Europ. Conf. Artif. Intell. (ECAI 2000)*, Berlin, 2000.

[23] I. Vatcheva, H. de Jong, N. Mars, « Selection of perturbation experiments for model discrimination », *in : Proc. 14th Europ. Conf. Artif. Intell. (ECAI 2000)*, W. Horn (éditeur), IOS Press, p. 191–195, 2000.

[24] I. Vatcheva, H. de Jong, N. Mars, «Selection of perturbation experiments for model discrimination », *in : Working Notes 14th Int. Workshop Qualitative Reasoning (QR 2000)*, J. Flores (éditeur), p. 24–31, Morélia, Mexico, 2000.

Rapports de recherche et publications internes

[25] H. de Jong, « Modeling and simulation of genetic regulatory systems: A literature review », *rapport de recherche n° RR-4032*, INRIA Rhône-Alpes, Montbonnot Saint-Martin, 2000.