Proposition d'une option " bio-informatique "
de la maîtrise d'informatique

Responsable (nom, adresses électronique et laboratoire):
Danielle Ziebelin
Universite Joseph Fourier
Projet Helix - INRIA Rhone-Alpes
655 Avenue de L'Europe -Montbonnot
38334 Saint Ismier Cedex
Danielle.Ziebelin@inrialpes.fr or Danielle.Ziebelin@imag.fr
Tel 33 (0) 4 76 61 53 68 Fax 33 (0) 4 76 61 54 08

Membres de l'équipe pédagogique :
UFR IMA Laurent Trilling, Anestis Antoniadis, Stéphane Girard, Frédérique Leblanc
UFR de biologie : Francoise Giroud, Hans Geiselmann, Stéphanie Blanc-Manel, Véronique Stoppin.
TIMC-IMAG : Jacques Demongeot, Olivier Cohen.
INRIA RA : Hidde de Jong, Eric Coissac, François Rechenmann.

Présentation
La demande de méthodes et d'outils d'analyse des données sur les génomes n'a cessé de croître ces derniers mois, conséquence de l'augmentation du flux de données en provenance des projets de séquençage exhaustif. Depuis plus de 25 ans, des efforts ont porté sur le développement de méthodes d'analyse de ces données génomiques. Mais l'efficacité et la pertinence de ces méthodes sont encore très souvent limitées. À titre d'exemple, la recherche des régions qui codent pour les protéines (les gènes) est encore un problème ouvert dans le cas des génomes eucaryotes, dont l'Homme. Les meilleures méthodes actuelles, mêmes combinées entre elles, produisent plus d'un tiers de faux positifs. Par ailleurs, les sources de données expérimentales, en particulier au niveau du transcriptome (" puces à ADN ", séquençage d'EST, etc.) et du protéome (gels 2D, spectrométrie de masse en tandem, etc) se diversifient. Plus généralement, l'éventail des problématiques biologiques qui font appel à l'informatique s'élargit. Des algorithmes et des modèles adaptés à ces nouvelles données et à ces nouvelles problématiques doivent ainsi être conçus et mis en œuvre.

Il est donc indispensable de former des informaticiens aux méthodes et outils de la bio-informatique et de les ouvrir aux problématiques biologiques, afin qu'ils puissent concevoir et développer de nouvelles méthodes d'analyse, mais aussi adapter et intégrer des méthodes existantes ou encore concevoir des systèmes d'informations pertinents.

Simultanément, les biologistes doivent être évidemment formés à l'emploi raisonné de ces méthodes. Il faut cependant souligner que le nombre d'étudiants concernés dans les deux cas est fort différent. En effet, si on peut imaginer à terme que tout biologiste doit, lors de sa formation initiale, avoir été formé à l'emploi des outils bio-informatiques, le nombre d'informaticiens à former à la bio-informatique restera plus limité. Il sera cependant amené à augmenter afin de répondre aux problèmes de plus en plus complexes que les biologistes, toujours plus compétents pour utiliser de façon pertinente les outils mis à leur disposition, ne manqueront pas d'exprimer.

L'option " bio-informatique " proposée vise à répondre à ce besoin de formation des informaticiens. Le marché de l'emploi correspondant est constitué des sociétés pharmaceutiques, agro-alimentaires et bio-technologiques. Ces sociétés, soit disposent d'un service de bio-informatique, soit centrent leur activité sur ce domaine, souvent en tant que sous-traitantes des premières.

Les étudiants ayant choisi l'option " bio-informatique " et poursuivant en 3e cycle pourront orienter leur recherche dans ce domaine, au sein des deux écoles doctorales concernées : EDIV et EDMI. Il existe en effet une communauté de recherche extrêmement active en bio-informatique. Cette communauté possède ses revues propres, telles que *Bioinformatics* ou *J. Computationnal Biology*, et ses conférences internationales, telles que ISMB (*Int. Conf. on Intelligent Systems for Molecular Biology*) ou RECOMB (*Int. Conf. on Computationnal Molecular Biology*).

Structure et contenu

L'option est structurée en 4 unités d'enseignement de 30 heures cours et 30 heures de TD chacun. Son contenu est détaillé dans le tableau joint. Les chapitres s'enchaînent naturellement et séquentiellement, mais l'ordre n'est pas total. Pour chaque chapitre, on indique de gauche à droite :

- la problématique biologique. Les notions de biologie spécifiquement nécessaires au chapitre correspondant sont présentées et viennent compléter celles qui auront été introduites au chapitre 1.
- les techniques et modèles informatiques ou statistiques. Dans plusieurs cas, les algorithmes ou les modèles sont spécifiques de l'approche bio-informatique ; dans d'autres, ils sont replacés dans ce contexte ou adaptés.
- les travaux pratiques. Il s'agit essentiellement de programmer les algorithmes fondamentaux de chacune des grandes classes d'algorithmes en bio-informatique.
- les travaux dirigés sur machine. Les étudiants utilisent des outils existants et les appliquent sur des données biologiques en vraie grandeur.

L'unité d'enseignement 1 correspond aux chapitres 1 et 2. Elle introduit les notions de biologie moléculaire indispensables et présente la problématique de l'acquisition et de la gestion des données génomiques.

L'unité d'enseignement 2 correspond au chapitre 3, qui regroupe les méthodes d'analyse de séquences, y compris la grande classe des méthodes statistiques, en particulier les modèles de Markov.

L'unité d'enseignement 3 est relatif à l'approche comparative : comment utiliser les données et les connaissances sur des espèces pour compléter celles d'une espèce voisine ? Comment déterminer le degré de voisinage d'une espèce et de ses gènes avec d'autres espèces ? Cette unité d'enseignement regroupe les chapitres 4 et 5.

L'unité d'enseignement 4 est une unité d'enseignement " avancé ". Elle regroupe les chapitres 6, 7, et 8. Le chapitre 6, sur la prédiction de structures, touche la biologie structurale. Les chapitres 7 et 8 présentent des approches qui sont encore, pour la plupart, au stade de la mise au point, voire de la recherche.

Proposition d'une option " bio-informatique " de la maîtrise d'informatique (144 heures)

Cours/TD		**Sujets de TP / travail personnel**	**Sujets de TD sur machine** (en général, mise en œuvre sur des données en vraie grandeur)
Problématique biologique	**Techniques informatiques ou statistiques**		
1 Rappels de biologie, unité d'enseignement 1 – La cellule procaryote et eucaryote : architecture et composants – L'information génétique : du gène aux protéines – La régulation de l'expression des gènes – Le métabolisme		– Étude de cours interactifs sur le Web (en anglais) – Programmation de la traduction ADN -> chaîne d'acides aminés dans les 6 phases de lecture	

2 Les données génomiques, unité d'enseignement 1 – Les techniques de séquençage – Les problèmes informatiques liés au séquençage – Banques et bases de séquences – Modélisation de la fonction des gènes – Recherche dans les bases de séquences : similarité de séquences – Ressources documentaires	– Algorithme de reconstruction de *contigs* – Les bases de données génomiques et protéiniques – Ontologies – Comparaison et alignement de séquences 2 à 2 : algorithmes et heuristiques – Extraction d'informations à partir de textes	– Lecture et formatage de données (langages de *script*) – Programmation de l'algorithme simple d'alignement de séquences avec *gaps* (programmation dynamique)	– Consultation de banques et de bases de séquences génomiques et protéiniques – Utilisation des programmes de type BLAST et FASTA – Interrogation de bases documentaires (Medline)
3 L'analyse de séquences, unité d'enseignement 2 – Détection des régions codantes, des régions régulatrices de l'expression, et d'autres objets d'intérêt biologique	– Recherche de motifs (expressions régulières), avec et sans " erreurs " – Inférence de motifs – Analyse statistique des séquences : profiles, modèles de Markov	– Programmation d'un algo. de recherche de motifs – Programmation d'une méthode de détection de gènes par analyse du biais de codage	– Utilisation d'outils de détection de gènes sur des séquences de génomes bactériens
4 Etude des familles de gènes et de protéines, unité d'enseignement 3	– Algorithmes et heuristiques d'alignement multiple (N séquences simultanément)	– Programmation d'un algorithme d'alignement multiple (sur des séquences courtes)	– Utilisation de programmes d'alignement multiple et étude d'une famille de gènes

5 Génomique comparative et phylogénie, unité d'enseignement 3 – Évolution et classification des espèces – Reconstruction d'arbres phylogénétiques – Réconciliation d'arbres des espèces et phylogénétiques – Cartographie comparée	– Alignement multiple – Calcul d'une matrice de distance, détermination d'un arbre – *Mapping* d'arbres – Ordres et permutations sur les cartes	– Programmation du calcul de la matrice de distances et de l'arbre, dessin de l'arbre obtenu – Connexion à l'unité d'enseignement d'alignement précédent	– Consultation de bases de données taxonomiques – Modélisation UML de cartes – Utilisation de programmes de reconstruction d'arbres phylogénétiques
6 Prédiction de structures 2D et 3D, unité d'enseignement 4 – ARN – protéines	– Propagation-satisfaction de contraintes – Réseaux de neurones formels – *Threading* – Modélisation des structures, bases de données	– Programmation d'un algorithme de prédiction de structure 2D d'ARN par propagation et satisfaction de contraintes (Prolog ?) – Visualisation de la structure 2D	– Consultation de bases de structures – Utilisation d'outils de visualisation de structures
7 Réseaux de régulation et réseaux métaboliques, unité d'enseignement 4	– Modèles booléens, modèles qualitatifs – Simulation – Inférence de réseaux (isomorphisme de graphes, recherche de cliques maximales)	– Programmation d'un simulateur de réseaux booléens	– Etude et utilisation de bases de données métaboliques (Kegg, EcoCyc)
8 Introduction au transcriptome et au protéome, unité d'enseignement 4 – ESTs – Techniques expérimentales : " puces à ADN ", gels 2D d'électrophorèse, spectrométrie de masse	– Analyse statistique des données d'expression – Analyse de données : *clustering* – Mapping d'EST et d'étiquettes peptidiques	– Programmation d'un algorithme de *mapping* d'EST	– Visite des plates-formes de transcriptome et de protéome de la génopole Rhône-Alpes à Grenoble