AMBASSADE DE FRANCE AUX ETATS-UNIS
MISSION POUR LA SCIENCE ET LA TECHNOLOGIE
CONSULAT GENERAL DE SAN-FRANCISCO

LA BIOINFORMATIQUE EN CALIFORNIE

Recherches et Applications, un apercu

FEVRIER 2004

Christophe LEROUGE
Attaché pour la Science et la Technologie

Armand RENUCCI
Attaché pour la Science et la Technologie

Gilbert DELEAGE, Professeur, Institut de Biologie et Chimie des protéines, Lyon.
Alexander BOCKMAYR, Professeur, Université de Nancy.
Jean-Loup RISLER, Directeur de Recherche au CNRS, UMR 8116, Evry.

RESUME

En septembre 2003, un groupe d'experts français s'est rendu en Californie pour évaluer le développement des recherches et les nouvelles initiatives dans le domaine de la bio-informatique et de ses applications. Des centres de recherches universitaires ainsi que des entreprises ont été visités.
La mission a pu constater le dynamisme de ce secteur aux Etats-Unis qui bénéficie de la politique générale des agences de recherche fédérales en faveur des nouvelles technologies. Ces recherches initialement menées dans un cadre académique ont désormais trouvé un large champ d'applications dans le secteur industriel auquel concourre le développement d'entreprises spécialisées.

LA BIOINFORMATIQUE EN CALIFORNIE
Recherches et Applications, un aperçu

A l'initiative du service pour la science et la technologie du Consulat Général de France à San Francisco, un groupe d'experts français s'est rendu en Californie en septembre 2003. La mission avait pour objectif d'évaluer le développement de la Bio-Informatique ('BioIT') dans de multiples programmes et les nouvelles initiatives plus spécifiques à ce domaine dans les centres d'excellence universitaires mais également dans les entreprises de cet état qui concentre 40% de l'activité biotechnologie aux USA.

La mission s'est déroulée sur une période de temps réduite, focalisée sur la Californie. Cette évaluation n'est donc que partielle mais permet néanmoins d'illustrer de manière significative les efforts très importants consacrés à ce domaine sur l'ensemble des Etats-Unis.

I - Présentation :

La bio-informatique 'BioIT' est par définition une activité à caractère inter-disciplinaire dont le développement récent est lié aux bouleversements de l'ère dite génomique. En effet les capacités de séquençage massif de l'ADN ont donné accès à ce jour à plus de 170 génomes allant des microorganismes à l'homme (http://wit.integratedgenomics.com/GOLD/).

Afin de décrypter l'information présente dans les gènes et les protéines, le traitement et l'analyse des séquences a été naturellement pris en charge par l'outil informatique, Afin d'identifier, comparer et classer cette énorme masse d'informations, la mise en place de bases de données avec de nouveaux outils de gestion et d'intégration a été nécessaire. Parallèlement l'outil informatique s'est avéré indispensable d'une part pour approcher la fonction biochimique des protéines en relation avec leur structure, dont le nombre élucidé est en progression incrémentale, et d'autre part pour déterminer leur fonction biologique liée à leur position dans des réseaux d'interactions génétiques et biochimiques complexes.

La bioinformatique a donc un caractère essentiellement moléculaire et peut être structurée en trois grands thèmes interactifs dont l'intégration est un des enjeu de la BioIT : Séquence, Structure et Fonction**.** De nombreuses applications de ces outils sont attendues dans le secteur des biotechnologies, et en particulier dans le domaine médical.

D'un point de vue plus technologique, la BioIT propose des outils pour stocker et traiter des données biologiques et biochimiques :

- Acquisition des données, création et diffusion de bases de données spécialisées, associées à des problèmes de fouille des données ;
- Développement de logiciels pour l'analyse, la comparaison et la modélisation de ces données ;
- Développement de logiciels permettant l'intégration de données variées et la simulation de processus biologiques et/ou biochimiques.

II - Etat des lieux :

1 - La situation en Californie/USA :

a/ Les enjeux

La BioIT est devenu un outil indispensable en recherche biologique avec des applications majeures dans tous les domaines de la biotechnologie et en premier lieu le domaine santé où l'industrie pharmaceutique est contrainte de développer de nouveaux médicaments avec une réduction substantielle des coûts et des délais ; la BioIT apporte de nouvelles solutions à toutes les phases : recherche de cibles pharmacologiques, criblage de composants et tests précliniques et cliniques. D'autres secteurs bénéficieront certainement des avancées dans ce domaine, l'industrie agrobiologique (amélioration des espèces et des variétés), la gestion globale des ressources vivantes (biodiversité…) et la parapharmacie (industrie cosmétique)... Un impact similaire à ce qui s'est passé dans les autres secteurs industriels quand les outils informatiques y ont été introduits (modélisation de pièces automobiles, aéronautiques…) est attendu.

b/ Les grands axes

Sur le socle des séquences acquises, les laboratoires et entreprises visités en Californie s'intéressent principalement aux problèmes de structure et de fonction des protéines au niveau biochimique et dans le cadre de réseaux macromoléculaires et cela dans de multiples systèmes biologiques. Les secteurs d'activité évalués au cours de cette mission révèlent sur un plan technologique d'importantes préoccupations liées à l'intégration des données et l'offre de capacité de calcul.

1 - L'activité la plus ancienne et la plus traditionnelle est liée à la comparaison et l'analyse des génomes toujours fort utile avec de nombreuses équipes travaillant sur cette base (Pattern Recognition, Data Mining) partout aux Etats-Unis. On peut citer en particulier la recherche de gènes humains homologues à des gènes à valeur thérapeutique (eg cibles pharmacologiques) dans des organismes modèles, et les travaux d'évolution et de phylogénie au Super Computer Center de l'Université de San Diego (SDSC).

2 - L'analyse structurale des protéines : un des objectifs majeurs en BioIT est de prédire à partir de la séquence primaire des protéines (déduite de la séquence ADN codante) les structures tridimensionnelles correspondantes ; cette activité s'appuie sur les 23.000 structures connues. De nombreuses approches sont développées :

- Identification et détermination de la structure putative à partir des séquences codantes du génome (BioX - Stanford) ;
- Classification des protéines sur la base de motifs structuraux : développement d'algorithmes originaux dans le domaine du traitement de l'électrostatique pour la modélisation structurale des protéines (BioX - Stanford), application à l'ingénierie des protéines (eg modifier la spécificité et/ou les caractéristiques cinétiques d'enzymes) ;
- Prédiction des domaines fonctionnels et détermination de sites actifs : (BioX - Stanford), application à la recherche de protéines cibles (Triad, Cengent) ;
- Analyse du repliement des protéines et des modifications conformationelles avec utilisation du calcul distribué, dédié au repliement des protéines (BioX - Stanford), application à de nombreuses pathologies (prions, maladies neurodégénératives), analyse

des interactions ligand-protéine (crible virtuel de composants pharmacologiques, Cengent).

3 - La modélisation de la dynamique moléculaire intra et inter-cellulaire. L'objectif à terme est de mieux appréhender les réseaux d'interactions moléculaires qui sous-tendent les fonctions physiologiques au niveau cellulaire et à des niveaux d'organisation plus complexes.

- Au niveau fondamental l'université de Stanford montre des développements novateurs avec la modélisation mathématique couplée et confrontée à l'expérimentation biologique des interactions entre un petit nombre de gènes permettant d'expliquer des phénomènes biologiques complexes (Ecole de Médecine, Institut Beckman).
- Des modélisations plus globales de l'activité cellulaire de microorganismes, ou de cellules eucaryotes impliquant un grand nombre de gènes liés à la signalisation, la division cellulaire et l'activité métabolique ont également été réalisées à Stanford (Ecole de Medecine, Institut Beckman), au SRI (Palo Alto), au SDSC (San Diego Supercomputer Center), chez Genomatica (San Diego). La modélisation des voies de signalisation hépatiques et cardiaques est soutenue par le SDSC. Les applications dans le domaine des biotechnologies sont nombreuses : détermination et validation de cibles thérapeutiques, production de médicaments par des microorganismes, une plateforme de simulation qui intègre les différentes échelles fonctionelles de la molécule à l'organisme complet est réalisée par la société Entelos (San Francisco), permettant des tests cliniques in silico sur des groupes de patients virtuels.

4 - Parallèlement à ces axes de recherches sur des thématiques purement biologiques, de nombreux travaux portent sur le développement de l'outil informatique lui-même. En particulier les méthodes de calcul distribué (Grid computing) connaissent un développement très fort afin de permettre aux biologistes de disposer de la puissance de calcul nécessaire au traitement des masses considérables de données. Ce sujet mobilise les équipes académiques pour le calcul haute performance scientifique au service de la simulation numérique mais aussi les entreprises d'informatiques. A ce titre les grands constructeurs ont développé une gamme de produits complète répondant aux besoins des laboratoires de biologie, dans les universités, les entreprises de biotechnologie ou les grands groupes pharmaceutiques : bases de données, logiciels de traitement et de fouille des données, matériels informatiques dédiés (serveurs, grappes), outils de stockage massif de données… Dans ce domaine trois grandes tendances apparaissent :

- des architectures plus décentralisées et plus orientées vers les données : développement du peer-to peer et des bases de données ;
- des outils et des services spécifiques au domaine scientifique concerné (biologie, physique des particules, géophysique…) ;
- et des composants standardisés permettant la communication entre les sites (fusion des concepts du grid computing et des web services commerciaux).

c/ Une stratégie et des moyens affichés

Les moyens humains et matériels aux USA liés à l'application de la BioIT dans des programmes très variés apparaissent considérables, tous les laboratoires académiques ou industriels bénéficient de moyens très importants en terme de financement et d'outils de calcul.

Les grandes agences fédérales qui traditionnellement orientent la recherche aux Etats-Unis ont clairement affiché la BioIT comme prioritaire. C'est le cas bien entendu des NIH (National Institute of Health) et de la NSF (National Science Foundation) mais aussi du DOE (Department of Energy). En effet le DOE gère les laboratoires fédéraux et les grands équipements scientifiques qui sont librement accessibles à la communauté scientifique. Il a lancé un vaste programme baptisé « Genomes to Life » dans le cadre duquel seront développés des outils de calcul et d'imagerie à haute résolution. L'objectif est de mettre à disposition des infrastructures complètes de capture, de traitement et d'archivage automatique des données. Des centres de ressources offrant la capacité de calcul nécessaire sont ainsi accessibles aux scientifiques américains.

1 - Moyens Matériels

Les moyens financiers apportés par les Universités et les agences gouvernementales sont à la hauteur des besoins et des enjeux des sciences du vivant soutenus au plus haut niveau politique (doublement du budget des NIH en 5 ans).

Les visites effectuées ont, en particulier, permis d'apprécier l'ampleur des moyens informatiques mis à disposition des chercheurs américains. A l'initiative de la NSF, les laboratoires sont regroupés dans le **NPACI** (National Partnership for Advanced Computational Infrastructure : il concerne 48 institutions américaines), financé à la hauteur de 30 millions de $ pour développer des ressources informatiques utilisables par l'ensemble de la communauté scientifique américaine. Les travaux portent sur la construction d'un « Teragrid » qui consiste à relier les différents grands centres de calcul américains, de manière à constituer une super machine virtuelle unique. Le réseau comprend à l'heure actuelle 4 nœuds, centres de ressources (SDSC à San Diego, Caltech, Urbana Champaign, Argonne National Lab) pour une capacité de calcul de l'ordre de 20 Teraflops, une capacité de stockage de 500 TeraOctets reliés par un backbone avec un débit visé de 40Gbit/s sur des distances de 4000km. D'autres centres de calculs sont progressivement intégrés au dispositif, comme par exemple le PSC (Pittsburgh Supercomputing Center). La NSF réfléchit à l'ouverture de son réseau à une coopération internationale.

Le SDSC, en particulier, est l'archétype du centre de ressource à l'échelle américaine. Il héberge l'ordinateur IBM Blue Horizon (1.152 processeurs capables de 1,7 trillion d'opérations par seconde). La grande originalité du SDSC est d'abriter des équipes de recherches dans différents secteurs scientifiques *: High computing, Integration of quantitative science clusters group, Data knowledge, Imaging, Geoscience, Geological time, Database coupling, Ontologies,* avec un programme de biosciences très important http://biology.sdsc.edu.

Les attentes sont grandes parmi les chercheurs américains. Reste cependant la mise en œuvre de ces solutions de travail coopératif qui n'ont pas encore atteint un niveau optimal.

2 - Moyens Humains

Liés aux moyens matériels, les moyens humains sont très importants ; de manière classique on retrouve un nombre impressionnant de post-doctorants, pour beaucoup bien sûr étrangers. Cette force de frappe conduit évidemment à une production scientifique plus importante.

L'interdisciplinarité est réelle et effective, on retrouve aussi bien en académie que dans l'industrie des équipes qui regroupent des biologistes, des ingénieurs, des informaticiens et des mathématiciens avec une très grande flexibilité et réactivité dans leurs échanges.

3 - Une dynamique établie

La Bio-Informatique américaine montre incontestablement une grande visibilité, elle s'intègre dans le contexte plus large du développement de la biologie aux Etats-Unis, avec une acquisition croissante de données que les nouvelles technologies (eg microsytèmes) pourraient encore accélérer. Les moyens mis en œuvre et les investissements réalisés offrent à la communauté scientifique des outils de calculs de grande capacité, et contribuent à l'établissement de bons groupes de recherche capables d'attirer massivement les meilleurs étudiants et postdocs, dont beaucoup viennent d'Europe. Ce caractère profondément interdisciplinaire de la BioIT et cette volonté de rapprocher les disciplines se retrouvent au niveau de la formation, comme à l'université de San Diego, qui abrite le Super Computer Center et qui est l'une des premières dans le monde à avoir mis en place un programme de formation doctorale en biologie des systèmes ; ou à Stanford dans le nouveau centre BioX nouvellement créé où la formation multidisciplinaire des graduates inclut technologie, biologie et informatique.

Les liens traditionnellement forts avec l'industrie, illustrés par la création récente de nouvelles entreprises dans ce secteur, contribuent à la très grande réactivité des équipes BioIT face aux enjeux des nouvelles thématiques fondamentales et appliquées.

4 - Un marché pour les constructeurs informatiques.

Les grands constructeurs informatiques (HP, Sun, IBM, EMC2…) ont eux aussi clairement identifié les besoins des équipes de biologistes qui requièrent des capacités de calcul et d'archivage des données accrues. Les entreprises ont ainsi développé une gamme de solutions complète matérielles et logicielles. Des services se mettent également en place pour proposer des bases de données avec des interfaces adaptées.

2 - Comparaison avec la France

Les recherches en Bio IT effectuées dans les laboratoires américains sont du même type que celles qu'on peut trouver en France, où les problèmes de séquences occupent de manière naturelle une place très importante dans la recherche et dans l'enseignement supérieur. En effet, des actions concertées (ACI) ont été lancées par le ministère de la recherche dans le domaine BioIT, soit pilotées par les STIC (ACI GRID, ACI Masses de données), soit pilotées par les sciences de la vie (ACI IMPBio). Le réseau genHomme soutient un projet de grille (RUGBI) unissant des industriels, des laboratoires de physique et de biologie et le centre de calcul (CC-IN2P3) de l'IN2P3 du CNRS. Les organismes comme l'INRIA affichent aussi une priorité dans ce domaine. La bioinformatique a été aussi présente dans le cadre du réseau des génopôles.

Les moyens disponibles, en proportion, ne sont pas négligeables comparés à ceux dont disposent les Américains. Des infrastructures informatiques existent en France (CC-IN2P3, IDRIS, CINES) mais ces centres ne sont essentiellement que des centres de ressources. Ils sont le plus souvent découplés des projets de recherche ou alors adossés à une seule discipline (physique des particules par exemple).

Au-delà des questions de moyens, une grande différence réside dans la question de l'interdisciplinarité : les informaticiens français, par exemple, sont tout à fait compétents dans l'élaboration d'algorithmes, mais ne développent pas d'interface dont les biologistes ont besoin. Les structures de recherche ne favorisent pas cette interdisciplinarité, les structures administratives et/ou scientifiques contribuent plutôt à maintenir les cloisons. Elles se retrouvent bien entendu au niveau de la formation dans les universités et les écoles. Tous ces éléments contribuent à rendre la bioinformatique française relativement peu visible, même si la France participe à de nombreux projets internationaux au niveau européen. A titre d'exemples, la France est impliquée dans des programmes européens pilotés par le CERN à Genève portant sur des projets de middleware ou d'infrastructure de Grid computing avec comme applications biologiques : l'imagerie médicale, la physique des particules, la bioinformatique (DataGRID, EGEE). Dans le domaine de la BioIT sciences de la vie, la France est en retard même si quelques laboratoires français (INRA, CNRS) ont été associés à un projet déposé par l'EBI portant sur le Grid computing en bioinformatique. Dans le domaine des bases de données, la France participe à de nombreux projets européens relevant de la bioinformatique comme, par exemple, les bases PRODOM, CAZY, euHCVDB, IMGT. Enfin dans certaines niches comme la bioinformatique appliquée à la virologie, la France a été à l'origine d'initiatives (exemple le réseau d'excellence Virgil piloté par l'INSERM à Lyon, réunissant plus de 50 partenaires en Europe).

III – Quelques propositions :

1 - Interactions académie/industrie.

Comme dans bien d'autres domaines il serait souhaitable de contribuer à développer les partenariats entre l'académie et l'industrie (« transfert d'outils ») ainsi que le transfert de technologies. Un encouragement réel pour le personnel académique à développer une activité économique est nécessaire, à la manière de ce qui se passe aux USA.

2- Moyens humains et matériels.

Au niveau de la formation : favoriser l'éducation aux interfaces entre l'informatique, la technologie, et la biologie dans les cursus universitaires afin de former des ingénieurs et des scientifiques avec des compétences larges ; dans ce contexte les cursus actuels en biologie comportent trop peu de mathématiques et d'informatique.

Au niveau recherche : favoriser l'émergence de quelques groupes (même peu nombreux) avec une masse critique leur assurant une réelle compétitivité internationale, et favoriser l'approche pluri-disciplinaire sur le modèle du centre BioX à Stanford, avec l'idée de mettre en synergie dans un même lieu des équipes d'origines et d'horizons très divers.

3 - Au niveau stratégique : une meilleure définition des objectifs.

Développer des logiciels complets, avec une interface utilisable, en particulier, par les biologistes permettant leur distribution. L'intérêt est de rendre visible et de mettre en commun les résultats .

Réduire les cloisonnements disciplinaires, afin de favoriser les échanges biologistes /informaticiens essentiels dans ce champ interdisciplinaire en développement rapide et particulièrement réactif aux innovations. L'organisation encore trop verticale du CNRS, par exemple, est certainement un frein à cet objectif.

Inviter des scientifiques américains, afin d'expertiser les efforts récemment mis en œuvre en France dans le domaine. Cela permettrait de réorienter par la suite les financements sur les projets les plus stratégiques et des groupes les plus performants. Bien que la coopération avec l'Europe et plus spécifiquement la France apparaisse encore très limitée, nos interlocuteurs américains souhaitent développer des interactions, en particulier avec les capacités de calcul des grids européens.

ANNEXES

Centres de recherche et laboratoires académiques visités

1 - Université de Stanford[1]

a/ Pr. Harley Mac Adams, Département de Biologie du Développement, Ecole de Mèdecine Institut Beckman

Dans son laboratoire, où il pratique depuis longtemps l'association entre étudiants de formation très diverse, le Pr Mac Adams étudie et modélise l'architecture des réseaux génétiques de régulation bactériens par une représentation de type circuits. La « logique » est codée par des circuits en couplant les interactions génétiques et biochimiques, la « mémoire » est représentée par l'ADN et les « routines » sont activées par la fabrication de protéines. L'objectif est d'identifier les molécules régulatrices, leur mécanisme d'action et les interactions qui caractérisent le réseau ; à plus long terme obtenir des modèles capables de simuler le comportement d'un système biologique et d'appréhender ainsi les mécanismes physico-chimiques sous-jacents aux différents processus physiologiques de la cellule.

Il utilise comme modèle biologique la régulation du cycle cellulaire d'un unicellulaire procaryotique *Caulobacter Crescentus*. Ce processus est étroitement lié à la vie bactérienne car cette bactérie se divise de manière permanente dans un milieu de croissance approprié avec une fréquence rapide (pour la bactérie *E. Coli* toutes les 20 minutes en conditions optimales) ; *C. Crescentus* possède 2500 (sur 4500) gènes exprimés liés à sa division. Cette bactérie présente plusieurs avantages pour cette étude, bien sûr facile à manipuler ; sa division est asymétrique et génere une cellule fille mobile, pourvue d'un flagelle, tandis que l'autre reste fixée. De plus les divisions dans une population peuvent être synchronisés permettant un tri des cellules mère et filles homogène et ainsi caractériser l'état transcriptionnel global de chaque stade (réalisé sur puces Affymetrix) ; la fonction des gènes ainsi identifiés peut être facilement testée grâce à l'obtention de mutants nuls.

Ces études ont permis de mettre en évidence un circuit maître sous la dépendance du gène CtrA, qui après activation contrôle 126 gènes (cellule fille mobile), un second gène maître GcrA est défini après la protéolyse de Ctra (il contribue à la mise en place de la cellule fille fixée) et contrôle 112 gènes dont seulement 11 sont partagés avec CTra.
L'architecture de ces circuits de régulation montre plusieurs choses :
- Tous les gènes liés au cycle cellulaire sont sous la dépendance de très peu de gènes régulateurs (2 ou 3).
- L'expression des gènes est régulée de manière très précise dans le temps et aussi en fonction de la géométrie cellulaire.
- Les gènes sont organisés du point de vue régulation en modules, leur expression est donc coordonnée (activés ou réprimés ensemble, la mise place du flagelle de la cellule mobile correspond à l'activation de 41 gènes). C'est un des éléments qui assure une réponse rapide au stimulus qui induit la division.

[1] Department of Developmental Biology, School of Medicine, Room 1125-CCSR, Stanford, CA 94305

b/ Pr. Claire Tomlin, Département d'aéronautique et d'astronautique

C. Tomlin qui a une formation en ingénierie astronautique et aéronautique dirige une équipe interdisciplinaire où des informaticiens et des physiciens s'adressent à des questions biologiques.

Elle cherche à modéliser des systèmes de signalisation complexes pour des réseaux multi-cellulaires, elle développe des modèles mathématiques pouvant rendre compte de la dynamique complexe qui intervient dans la transduction et la propagation des signaux entre cellules. La valeur prédictive des modèles mathématiques établis est testée *in vivo* au niveau moléculaire en utilisant un couplage à la protéine GFP qui permet de visualiser la présence et la distribution des protéines impliquées en relation avec le phénotype observé**.**

Le modèle biologique utilisé est la différenciation de la cellule alaire chez la mouche Drosophile qui est caractérisée par un poil unique orienté. La disponibilité de mutants a permis d'identifier plusieurs gènes impliqués dans la différenciation de la cellule alaire et la mise en place du poil.
Ce processus dépend en particulier de la propagation de signaux entre cellules, la composante extracellulaire est transduite au misn en partie par des récepteurs membranaires de type frizzled qui présente une distribution asymétrique. Un modèle d'influence intégrant ces gènes a été développé à l'aide d'équations différentielles partielles ou ordinaires, les simulations réalisées montrent qu'il est possible de décrire ainsi un processus cellulaire apparement complexe, ce type d'approche présente une utilité certaine pour disséquer des mécanismes moléculaires. Néanmoins s'il permet d'éliminer certaines hypothèses il ne permet pas de prouver que le modèle testé reflète une réalité.

c/ Pr. Jim Ferrel, Département de pharmacologie moléculaire

J. Ferell est un physicien qui développe depuis 5 ans des modèles mathématiques à base d'équations différentielles pour décrire la transduction de signaux cellulaires et mieux comprendre les systèmes de signalisation complexes au niveau cellulaire.

Il s'intéresse aux signaux qui induisent la différenciation appelée maturation de l'ovocyte utilisé comme modèle. L'ovocyte se différencie irréversiblement à partir d'une cellule germinale en passant par plusieurs stades bien définis ; J.Ferell étudie plus particulièrement la transition qui détermine la première phase de méiose. Cette transition est induite par la progestérone qui entraîne une signalisation en cascade correspondant à l'activation successive de protéines–kinases. Deux cascades sont impliquées : la cascade MAP kinase qui induit à son tour la cascade Cdc2–CycB ; une fois activées ces deux cascades interagissent et présentent des "feedbacks" positifs et négatifs, dans ce contexte l'élimination de la progestérone n'entraîne pas l'arrêt de la cascade MAPk qui demeure activée en permanence permettant la progression de la maturation.

Au niveau moléculaire, sur la base des propriétés des feedback observés, la différenciation de l'ovocyte peut se décrire aisément (les 31 equations d'états considérées au total dans la cascade MAPk peuvent être simplifiées par un paramètre global selon 2 états) comme un système bistable « auto-alimenté » avec hytérésis (mémoire) et irréversibilité où une réponse permanente établie à partir d'un stimulus transitoire (l'hormone) et indépendante de tout signal extérieur supplémentaire génère une progression irréversible de l'ovocyte.

d/ BioX

Stanford vient, fin 2003, d'inaugurer un nouveau centre de recherche biomédical. Baptisé BIO-X, il regroupera dans un même lieu des équipes pluridisciplinaires. Cette réalisation est la traduction concrète de la volonté d'associer la biologie aux autres disciplines : physique, chimie, informatique. De façon très pragmatique, les universitaires américains veulent créer les conditions nécessaires à une recherche située à l'interface des disciplines traditionnelles. Au-delà des objectifs de recherche, la démarche interdisciplinaire est inscrite dans la formation des étudiants. Il s'agit pour les doctorants d'acquérir une culture transversale où la réalisation de leur projet demande un travail (ou des interactions fortes) avec plusieurs laboratoires disciplinairement différents. Dans la même optique, un premier objectif est de donner une formation biologique sur le campus à des ingénieurs travaillant dans le BIO-X.

Le fonctionnement du BIO-X prévoit, en outre, d'associer les industriels aux travaux menés dans le centre grâce à un programme de partenariat : participation à des séminaires, rencontres régulières avec les chercheurs, accès aux informations sur les recherches et les résultats, possibilité d'orienter les travaux.

Dans ce nouveau centre pluridisciplinaire, nous avons rencontré un groupe de chercheurs qui travaillent sur l'étude et la prédiction de la structure tridimensionnelle des protéines.

- **Dr. P. Koehl**, Français du CNRS s'intéresse à la prédiction de la structure des protéines à partir de la séquence primaire (déduite des séquences DNA). Il cherche à définir des paramètres structuraux qui permettent la comparaison et la classification automatique des structures ; il développe en parallèle de nouveaux outils pour améliorer la description et la modélisation, par exemple, les paramètres qui permettent le calcul de l'énergie minimum d'une structure en solution aqueuse. (Minimisation de la surface de la protéine, calcul des forces électrostatiques s'exerçant sur la structure par des méthodes de tesselation de Voronoï).

- Dr. Vijay Pande (http://folding.stanford.edu) s'intéresse à la prédiction de la conformation des protéines. Il utilise des méthodes de calcul distribué à large échelle (large scale distributed computing) sur Internet : projet « Folding@home » (500.000 cpu utilisés)

- Dr. Doug Brutlag http://brutlag.stanford.edu. Le groupe de Doug Brutlag développe des algorithmes de recherche et de comparaison de signatures dans les motifs structuraux conservés, et ce à différents niveaux d'organisation des protéines. Il s'agit d'une activité «classique» dans le domaine de la bioinformatique structurale. Ses programmes sont vendus sous licence à Genset, Sanofi et Novartis.

Certains de ses programmes et bases de données sont accessibles sur le web :
- http://brutlag.stanford.edu/emotif/ (Alignment de motifs) ;
- http://brutlag.stanford.edu/3motif (Alignement de motifs) ;
- http://brutlag.stanford.edu/eblocks et http://brutlag.stanford.edu/eproteome (Bases de données de microenvironnements) ;
- http://brutlag.stanford.edu/chimera (Facteur de transcription) ;
- http://brutlag.stanford.edu/biosprector (Superposition des structures des protéines) ;

- Pr. J-C. Latombe développe des méthodes de calcul afin d'accélérer et d'optimiser la prédiction de structures. Un des objectifs est de pouvoir prédire et suivre l'évolution

structurale d'une protéine au cours du temps, principalement pendant le repliement juste après synthèse, et dans le cas d'une interaction avec un ligand qui induit un changement de conformation local (permet le positionnement d'un acide aminé du site actif ou le dégagement d'un site dans un autre domaine de la protéine). Ses méthodes reposent sur les parcours aléatoires (molecular motion probabilistic roadmaps/ Stochastic Roadmap Sampling) : http://robotics.stanford.edu/~apaydin/software.html

2- Le ' Stanford Research Institute' (SRI) (Menlo Park)[2]

Le SRI est un institut de recherche indépendant et sans but lucratif de 1200 personnes qui comprend 5 divisions : sciences physiques, systèmes d'ingénierie, politiques éducatives, biopharmacie et informatique.

- **P. Karp** dirige un groupe de bioinformatique avec 4 thésards et 2 développeurs. Il a établi différentes bases de données pour les réseaux macromoléculaires, en particulier pour les voies métaboliques d'environ 70 microorganismes. Il a développé de nombreux outils et bases de données : les suites xxxCYC : BioCyc Knowledge Library http://BioCyc.org, EcoCYC (sert 115,000 pages par mois), HumanCYC, Encyclopedia of E. coli genes and metabolism, Metacyc (Metabolic pathway database) et des outils d'analyse (« Pathway tools ») qui représentent actuellement 15 bases de données de voies métaboliques PGDB (pathway genome database).
Au niveau de la recherche en bioinformatique, l'équipe développe des stratégies pour identifier et combler les « trous » présents dans les voies métaboliques. Pour cela, elle utilise une stratégie d'analyse génomique comparative basée sur les banques SWISS-PROT et PIR. Par cette approche, environ 20-25% des « trous » peuvent être comblés. En collaboration avec l'équipe de P. Lincoln, P. Karp travaille aussi sur une approche symbolique pour la biologie des systèmes (« Symbolic Systems Biology »), qui cherche à mettre en œuvre des méthodes de logique et de vérification formelle au profit de la modélisation et de l'analyse des systèmes biologiques.

3 – UC San Diego[3]

Nous avons visité sur le campus de l'université de San Diego le Super Computer Center SDSC. Il s'agit d'une très grosse infrastructure informatique, créée en 1985, qui abrite 400 personnes avec un budget annuel de 60 millions de $.
Une partie importante de l'activité correspond à celui d'un centre de ressources offrant des capacités de calcul et de stockage de données à l'ensemble de la communauté scientifique américaine. A ce titre le SDSC est tête de pont du **NPACI** (National Partnership for Advanced Computational Infrastructure, http://www.npaci.edu) et est un nœud du réseau Teragrid de la NSF (qui consiste à relier les différents grands centres de calcul américains de manière à constituer une super machine virtuelle unique). Le SDSC abrite notamment l'ordinateur IBM « Blue Horizon » (15 Tbits en parallèle sur 1152 processeurs IBM Power 3), des clusters IBM P4, et un centre d'archivage de données HPSS de 500 Toctets.

Sur le plan scientifique le SCCSD abrite plusieurs groupes de recherche dans des secteurs variés, une partie importante est dédiée à la biologie avec le programme 'Integrative

[2] SRI International, Room EK207, 3333 Ravenswood Avenue, Menlo Park, CA 94025-3493
[3] Sequoyah Hall, 9500 Gilman Drive, La Jolla, CA 92093-0527

Biosciences' (directeur P.Bourne) qui regroupe plusieurs équipes dans les domaines de la neurobiologie, les systèmes complexes, les interactions moléculaires et la crystallographie, la génomique structurale, la protéomique et la chimioinformatique.

De nombreux travaux portent sur la constitution de bases de données :

- PDB : 'Protein data bank', http://alpha.rcsb.org/pdb, dont le premier objectif est d'assigner une structure 3D aux protéines putatives déduites des séquences génomiques disponibles dans les bases de données. Des outils de visualisation et de prédiction de structures sont accessibles. Le projet occupe 27 personnes dont l0 à San Diego.
- Encyclopedia of life : http://eol.sdsc.edu a pour ambition d'élargir l'annotation des protéines au-delà de la séquence et de la structure pour intégrer les données fonctionnelles et biochimiques, les interactions et le fonctionnement global de la cellule.
- Joint Center for structural genomics.
- Protein Kinase resource.
- Tree of life (Phylogénie).

Au sein du SDSC le **Pr B. Palsson,** du Département de Bioengineering, est un des protagonistes de la nouvelle biologie des systèmes, http://systemsbiology.ucsd.edu. Il a en particulier participé à l'élaboration du programme post-génomique « From genomes to life » (http://doegenomestolife.org) du Department of Energy (103 millions de dollars attribués à l'appel d'offres 2002). Il travaille sur la modélisation des voies du métabolisme cellulaire, notamment chez *E. coli, H. influenza, H. pylori* (pathogènes) et *S. cerevisae,* avec pour objectif de définir de réseaux de régulations avec une forte valeur prédictive. En réunissant diverses bases de données, il opère une reconstruction de la physiologie bactérienne en intégrant les paramètres physico-chimiques des réactions et des enzymes. En particulier, il a développé de nouvelles méthodes informatiques (modélisation sous contraintes). Les modèles établis ont une valeur prédictive vérifiée expérimentalement in vivo dans différentes situations : adaptation évolutive au niveau métabolique en fonction des conditions du milieu, sélection de phénotypes liés à l'inactivation de gènes définis. Les applications en biotechnologie ont guidé la création de la société Génomatica (cf infra).

Entreprises visitées.

Région de San Diego

1 - Genomatica.[4]

Cette société est une émanation du SCCSD : en effet le professeur B. Palsson de UCSD en est un des co-fondateurs. Genomatica a obtenu en 2003 le prix 'Technical Insights Award for Technology Leadership' délivré par Frost and Sullivan (une société majeure dans le domaine du consulting et du marketing aux USA).

Cette société a développé une plate-forme informatique, 'SimPheny', où la dynamique du métabolisme cellulaire est modélisée, permettant des simulations et des prédictions. Appliquée à plusieurs modèles cellulaires, elle peut permettre en particulier d'aider à définir des souches productrices d'un métabolite particulier (eg la production de médicaments).

[4] Genomatica, Inc., 5405 Morehouse Drive, Suite 210, San Diego, CA 92121. www.genomatica.com-

2 - Triad Therapeutics.[5]

La société Triad est pilotée par le Dr Hugo Villar avec, dans le comité scientifique, K. Wutrich et Barry Hoenin. La société est spécialisée dans la phase initiale du processus de Drug Discovery depuis la définition de la cible jusqu'au criblage de composants pharmacologiques.

L'effectif est de 50 personnes : 7 administratifs, 7 biologistes, 3 bioinformaticiens (modélisateurs et bioanalystes), 4 informaticiens (maintenance + développement ciblé), 16 chimistes. L'équipement en RMN est d'un 500 et 700 MHz Brucker et d'un cluster 64 processeurs. La surface est de 9000m², dont une partie est prévue pour le développement de la société. Triad travaille sur des cibles en partenariat avec des grandes sociétés pharmaceutiques ; elle développe aussi ses propres projets en interne, et fait largement appel à la BioIT.

Un de leurs objectifs est de développer des composants (eg inhibiteurs) spécifiquement dirigés contre des microorganismes pathogènes. Ils ont focalisé leur activité sur deux types d'enzymes essentielles pour de multiples fonctions cellulaires : les oxydo-réductases et les protéines-kinases. La stratégie développée est 'Parallel or Object-Oriented Drug Discovery'. Elle tire avantage d'un site actif commun à tous les membres de ces familles où se fixe, pour les kinases, l'ATP utilisé pour la phosphorylation du substrat qui est différent pour chaque protéine ; le substrat se fixe sur un site adjacent à celui de l'ATP. La connaissance détaillée de ce site permet d'en obtenir une signature RMN fine et de déterminer ainsi comment est orienté un analogue de l'ATP. Une librairie combinatoire de bi-ligands géométriquement orientée peut être synthétisée (un analogue de l'ATP associé avec un des 10000 composants disponibles à Triad) ; cette librairie peut être alors utilisée pour un crible fonctionnel contre n'importe quel membre de la famille basé sur la sélection d'un bi-ligand de très haute affinité qui pourra être utilisé comme inhibiteur.

La définition des protéines cibles est basée sur une approche BioIT structurale où de nouveaux membres (appartenant éventuellement à une grande sous-famille, tyrosine, sérine-thréonine kinase…) d'une de ces deux familles de référence sont recherchés de manière croisée entre les banques génomiques relatives au microorganisme visé et la PDB. L'analyse structurale in silico utilisée permet de « classer » la protéine, conduisant soit un crible direct avec une librairie existante soit, après analyse RMN des sites actifs de la protéine, à la synthése d'une nouvelle librairie adaptée à cette cible.

La société ne fait pas état de développement bioinformatique stricto sensu, car la valeur ajoutée n'est pas dans le coté prédictif mais dans la synthèse de librairies de composants sur une base combinatoire.

[5] Triad Therapeutics, Inc., 9381 Judicial Drive, San Diego, CA 92121 www.triadthera.com

3 - Cengent. (Activité de **Geneformatics** absorbée en 2003).[6]

Comme Triad, cette société est engagée dans les phases en amont de la recherche de nouvelles molécules comme médicaments potentiels. Ils vont de la définition de la cible jusqu'au criblage de composants, avec une assistance BioIT à chaque étape beaucoup plus importante que chez Triad. En relation également avec de grandes entreprises pharmaceutiques, ils ont focalisé leur activité sur la famille des protéines de type tyrosine phosphatases (PTP). Ces protéines régulent des voies de signalisation et sont impliquées dans le diabète, l'obésité et le développement tumoral.

Geneformatics a ainsi développé une série d'outils informatiques. Le software FFFm (« Fuzzy Functionnal Form ») permet d'identifier sur une base structurale des domaines fonctionnels putatifs de la famille PTP à partir des séquences génomiques ou de cDNA. Cette première sélection a une valeur prédictive faible, sur la base des acides aminés importants du site actif le software « Touchstone Signature » permet d'affiner cette sélection initiale et a conduit à l'identification d'une centaine de PTP humaines qui n'avaient pas été préalablement identifiées avec les outils informatiques publics. Enfin un crible virtuel de composants pharmacologiques peut être mené et complété par un criblage expérimental sur les protéines cibles synthétisées et purifiées avec une évaluation de l'interaction par RMN.

Région de San Francisco

4 - Entelos[7]

La société Entelos à Foster City dans le Silicon Valley travaille comme Genomatica dans le domaine de la biologie des systèmes. Cette société a été créée en 1996 par M. Braidman avec 45 Million de $, elle emploie 70 personnes, moitié biologistes et moitié informaticiens avec un minimum de 4 personnes par projet. Elle possède un cluster de 200 processeurs pour faire des simulations et n'a aucune activité « expérimentale ».

Grâce à sa plate-forme « Physiolab » Entelos développe des modèles et des simulations de différentes pathologies humaines à toutes les échelles possibles, du groupe de patients jusqu'au niveau cellulaire et moléculaire.

Entelos travaille en collaboration avec des compagnies pharmaceutiques sur leurs pathologies d'intérêt : Astro-Zeneca, Aventis (Asthme), Bayer, Organen, Johnson-Johnson, Pfizer ou l'« American Diabetes Association ». En juillet 2003, Entelos a reçu avec son partenaire Pfizer le prix « BioITWorld's Best Practices Award ».

Ce travail inclut la collecte et l'intégration (souvent totalement manuelle) de toutes les données bibliographiques, biologiques et cliniques (les connaissances au sens global) de la pathologie, la modélisation en circuits grâce à des outils mathématiques des voies métaboliques pouvant être impliquées dans la maladie en est un aspect important. L'approche va du général au particulier (top-down) ; l'objectif final est de construire un modèle capable de reproduire l'ensemble des données collectées.

[6] Cengent Therapeutics, Inc., 10929 Technology Place, San Diego, CA 92127 www.cengent.com
[7] Entelos, 110 Marsh Drive, Foster City, CA 94404 www.entelos.com

L’objectif est bien entendu de réduire les temps et les coûts de la recherche des entreprises pharmaceutiques et cela à différents niveaux. Cette plate-forme contribue en amont à la définition de nouveaux médicaments par la recherche de gènes cibles ; la force prédictive des modèles élaborés permet, par exemple, d’identifier des voies métaboliques à haute incidence vis à vis de la maladie, d’en exclure d’autres et de rationaliser ainsi l’approche en réduisant les points d’intérêt à quelques protéines ou enzymes dans certaines voies métaboliques et en minimisant in fine le nombre d’expériences à mettre en oeuvre (ex avec Pfizer, rôle du PD4, 4 voies sur 26 ont été caractérisées alors qu’elles n’avaient pas été détectées par des études in vitro préalables ; l’étude a nécessité 20000 simulations). En aval, la réalisation d’essais pré-cliniques et cliniques sur des groupes de patients virtuels a permis de réduire le nombre de tests à effectuer de manière sensible.