Rapport semestriel d'activité n°3 datant du 15 juillet 2007

A. Identification

Programme – année	Programme non thématique 2005
Projet (acronyme)	REGLIS
Coordonnateur du projet (société/organisme - laboratoire ou entité de rattachement)	Équipe BAOBAB – Laboratoire de Biométrie et Biologie Évolutive

Référence convention/décision	NT05-3_45205
Période du projet (date début – date fin)	15 décembre 2005 – 15 décembre 2008
Période faisant l'objet du rapport d'activité (date début – date fin)	15 janvier 2007 – 15 juillet 2007
Rédacteur de ce rapport : nom	SAGOT Marie-France
téléphone	04-72-44-82-38
adresse électronique	marie-france.sagot@inria.fr
Date du rapport	15 juillet 2007

B. Pour les projets multi-partenaires, rappel des tâches allouées par partenaire pour l'ensemble du projet

Le projet REGLIS n'a qu'un seul partenaire, le partenaire principal (équipe BAOBAB-HELIX).

Eléments qualitatifs

C. Description des travaux effectués pour la période concernée et conformité de l'avancement aux prévisions

Nous n'indiquons dans le présent rapport que les travaux pertinents aux six derniers mois du projet. Certains sont la continuation de travaux initiés et déjà en partie introduits lors des deux rapports précédents. Nous avons essayé dans ce cas de ne présenter en détail que ce qui se réfère à la période couverte par ce rapport (15 Janvier à 15 Juillet 2007). Outre les travaux décrits ci-dessous, nous avons également avancé sur le problème de l'identification de petits ARNs, en particulier les microRNAs (thèse de Nuno Mendes) mais avons choisi de n'en parler que lors du prochain rapport, lorsque nous aurons pu discuter les résultats obtenus à ce jour avec nos collaborateurs (Eric Westhof et son équipe à l'Institut de Biologie Moléculaire et Cellulaire, Strasbourg). Il en sera de même en ce qui concerne la reprise de nos travaux sur la comparaison de structures d'ARNs avec Julien Allali (ex-doctorant de la coordonnatrice de REGLIS, MdC au Labri, Bordeaux), Claude Thermes et Yves d'Aubenton Carafa (CGM, Gif-sur-Yvette).

C1. Inférence et mise en évidence de régularités :

a. Au niveau génomique

1. **Amélioration des algorithmes de recherche de régularités dans les séquences nucléotidiques**

- **Filtrage de séquences en vue d'un alignement multiple :** La recherche exhaustive de similarités multiples dans une séquence ou un ensemble de séquences est un problème fondamental en génomique. Or la résolution exacte de ce problème souffre d'un temps d'exécution exponentiel, ce qui, en pratique, interdit l'exploitation de grosses masses de données. Fin 2006, Ed'Nimbus, une version opérationnelle d'une technique de filtrage permettant de supprimer rapidement des séquences en entrée de larges portions n'intervenant pas dans le résultat final, a été mise à disposition via une interface web (http://igm.univ-mlv.fr/~peterlon/ednimbus). À ce jour, cet algorithme, basé sur une publication acceptée depuis le dernier rapport dans "*Journal of Discrete Algorithms*", reçoit environ une vingtaine de requêtes hebdomadaires. Les résultats obtenus sont extrêmement prometteurs et peuvent être largement améliorés selon de nombreux axes de recherche. Ainsi, en collaboration avec Pierre Peterlongo de l'IRISA (Rennes), Alair Pereira do Lago et Gustavo Sakomoto (Université de São Paulo, Brésil), et Nadia Pisanti (Université de Pise, Italie) nous travaillons sur l'application d'une condition de filtrage plus contraignante, conduisant à de meilleurs résultats pour des temps de calculs similaires. Cette nouvelle approche, qui a conduit à un algorithme appelé TUIUIU, est en cours de finalisation (voir Section D. 1).

- **Isochores :** La structure en isochores des génomes de certains vertébrés a, depuis sa découverte en 1976, été un exemple privilégié de l'analyse des relations entre fonctionnement moléculaire du génome et information génétique. Définie au départ comme une régionalisation du contenu en C+G du génome, les isochores ont été associés au fil des années à de très nombreuses propriétés biologiques au point qu'il n'en existe plus réellement de définition consensuelle. Nous avons entrepris un re-examen de cette structure associant à la fois des approches aussi « objectives » que possible et des modélisations des relations entre propriétés. Ce travail repose sur une méthode de segmentation par HMM et a conduit à 3 publications en 2007 (voir « Livrables »).

- **Taux de recombinaison :** La comparaison de cartes physiques et génétiques est la méthode la plus employée pour estimer la variation du taux de recombinaison le long d'un chromosome. Elle utilise en amont une modélisation du processus d'occurrence des cross-overs (échange de brins d'ADN entre les branches des chromosomes d'une même paire lors de la méiose). Cependant ce modèle ne prend pas en compte les résultats concernant à la fois les processus moléculaires en cause (relation entre chiasma (entrecroisement de deux chromatides homologues), appariement des brins non homologues, et cross-over) et les chromosomes dans leur entier (nécessité d'un chiasma au moins par bras). Une modélisation plus réaliste est en cours et a déjà permis de retrouver une importante relation entre taux de recombinaison et longueur des bras des chromosomes.

2. **Recherche de régularités dans les régions de cassure suite à un réarrangement**

La comparaison de l'ordre des gènes entre plusieurs génomes permet l'identification des ruptures de synténie. Ces ruptures constituent des régions du génome qui ont potentiellement subi un ou plusieurs réarrangements génomiques. L'analyse fine de ces régions, au niveau génomique, pourrait permettre de mieux comprendre les causes et les mécanismes de ces évènements évolutifs au niveau de la cellule. L'identification de régularités dans ces régions, permettrait de localiser de manière systématique les réarrangements ou les régions susceptibles d'être réarrangées. Ce travail étant basé sur la localisation des gènes orthologues sur les deux génomes comparés, il dépend fortement de la qualité des assignations d'orthologie. Or, les méthodes d'assignations d'orthologie produisent des erreurs, notamment dues à l'existence de paralogues et de pseudo-gènes. Durant ces six derniers mois, nous avons ainsi développé une méthode automatique pour détecter les erreurs d'assignation d'orthologie.

Par ailleurs, nous étudions la corrélation de la position des points de cassure avec d'autres découvertes sur l'organisation du génome (en duplicons, en réplicons, etc.). Ce dernier point est le fruit d'une collaboration avec Alain Arnéodo, Benjamin Audit et Lamia Zaghloul, respectivement chercheurs et doctorante au Laboratoire Joliot-Curie de l'ENS de Lyon.

3. Étude des liens entre éléments répétés de différents types entre eux ainsi qu'avec d'autres facteurs, dont le niveau d'expression des gènes

Nous menons une large étude des liens entre éléments répétés de différents types, ainsi qu'entre ces éléments et d'autres facteurs, dont le niveau d'expression des gènes. Dans un premier temps, nous nous intéressons au niveau d'activation à l'échelle génomique de certains types d'éléments répétés, les éléments transposables. L'organisme choisi pour mener cette étude est *Drosophila melanogaster*. Ce travail a été effectué avec Cristina Vieira de l'équipe « Génomes et Populations » de notre laboratoire. Des données d'EST disponibles publiquement ont été utilisées.

Par ailleurs, les éléments transposables s'accumulent dans les régions hétérochromatiques, où ils sont progressivement dégradés par des processus mutationnels. Au sein d'une collaboration avec Nathalie Mugnier, Cristina Vieira et Christian Biémont de l'équipe « Génomes et Populations », une modélisation de l'évolution des copies de rétrotransposons LTR dans les régions hétérochromatiques des génomes de drosophiles a été effectuée. Nous avons également estimé, pour chaque classe d'âge, le nombre d'éléments transposables insérés dans le génome, à partir du nombre d'éléments trouvés par Repeat Masker.

4. Régularités et évolution : Reconstitution de scénarios évolutifs par remaniements chromosomiques

Les régularités dans l'organisation génomique observées dans différentes espèces nous permet d'essayer d'inférer des scénarios évolutifs de cette organisation. Étant donné leur interconnexion, nous avons choisi d'aborder en parallèle plusieurs questions relatives à de tels scénarios. Ces questions sont les suivantes.

- **Algorithme d'exploration de l'ensemble des scénarios évolutifs possibles**. Étant donnée l'organisation génomique actuelle de deux chromosomes (ou morceaux de chromosomes), nous disposons maintenant d'un algorithme permettant d'explorer (sans énumérer, ce qui prendrait trop de temps et de mémoire) l'ensemble des scénarios d'inversions expliquant les différences d'organisation entre les chromosomes.

- **Reconstitution de l'histoire moléculaire de la différentiation sexuelle.** En examinant les chromosomes actuels humains X et Y, nous tentons de découvrir les traces de leur histoire, fortement marquée par les remaniements, et en particulier par les inversions. Nous avons découvert plusieurs traces de ces inversions. Nous appliquons l'algorithme d'exploration mentionné au paragraphe précédent pour évaluer la vraisemblance d'hypothèses débattues par la communauté sur l'évolution du chromosome Y. Cet algorithme doit en outre prendre en compte le fait que les chromosomes X et Y comportent des duplications de gènes. Nous avons donc développé une méthode permettant de calculer une distance de réarrangements entre génomes comportant des familles de gènes. Ce dernier point est mené en collaboration avec des chercheurs mathématiciens de l'Université de São Paulo, Brésil et de l'Université Fédérale de Mato Grosso do Sul à Campo Grande, São Paulo.

- **Reconstitution de l'organisation des génomes de mammifères ancestraux.** L'utilité de cette reconstruction est d'étudier qualitativement et quantitativement les évènements évolutifs importants (du point de vue de l'organisation des molécules d'ADN) qui ont marqué l'histoire évolutive des mammifères. Après une large étude méthodologique des

publications récentes sur le sujet, nous espérons pouvoir proposer un modèle d'évolution qui concorde avec les données recueillies. Nous avons commencé par cette étude car, concernant les ancêtres communs aux mammifères, les différences dans les résultats publiés révèlent quelques écueils méthodologiques à éviter, ainsi que des possibilités d'études combinatoires sur les problèmes soulevés. Nous souhaitons construire des méthodes de reconstitution de génomes ancestraux qui intègreront ces études initiales, et les appliquer aux génomes de tétrapodes.

- **Application des méthodes de tri par inversion pour la compréhension de l'évolution de systèmes symbiotiques**. En collaboration avec Fabrice Vavre de l'équipe « Génétique et Évolution des Interactions Hôtes-Parasites », nous avons initié une étude de l'évolution des *Wolbachias*, une famille des bactéries intracellulaires qui forment un système symbiotique avec des insectes et des nématodes. Parmi ces bactéries, il y a au moins une lignée qui montre des caractéristiques bien différentes des autres (notamment son génome comporte plus d'éléments répétés et d'évènements de réarrangements, alors que les bactéries intracellulaires sont considérées comme plutôt conservées). Afin de mieux comprendre ce processus évolutif, étant donné qu'il n'existe pas encore suffisamment de données sur les *Wolbachias*, nous avons choisi comme référence les *Rickettsias*, qui sont évolutivement très proches des *Wolbachias*. L'enjeu est de tenter de faire le lien entre l'écologie du parasite, en particulier sa capacité à infester de nombreuses espèces d'hôte, et la structure de son génome.

5. Dépendances entre sites voisins dans les séquences d'ADN

Il a été montré que l'analyse de l'évolution des séquences nécessite l'incorporation de mécanismes faisant intervenir des dépendances entre sites voisins. Dans cette optique, un modèle de type markovien a été développé au niveau mathématique dans le cadre d'une collaboration avec Didier Piau de l'Université Joseph Fourier à Grenoble, et implanté sous forme d'un programme destiné à une prochaine large diffusion (voir section D et « livrables »). L'importance dans la reconstruction des arbres phylogénétiques de la prise en compte de ces dépendances de voisinages a été montrée (Leonor Palmeira, thèse soutenue le 13 Juillet 2007), justifiant ainsi pleinement ces développements méthodologiques. Les développements mathématiques et l'application bioinformatique font l'objet de deux articles en préparation.

b. Au niveau du réseau métabolique

1. Recherche de motifs dans les réseaux métaboliques

Au cours de cette période, nous avons utilisé intensivement notre outil de recherche de motifs en l'appliquant à notre source de données la plus fiable : le réseau métabolique *d'Escherichia coli* (reconstruit à partir de la base de données BioCyc). Les sous-réseaux identifiés comme étant des instances de motifs exceptionnels ont été systématiquement comparés aux structures connues d'opérons chez cet organisme. De manière générale, on trouve un enrichissement en opérons par rapport aux sous-réseaux qui ne sont pas des occurrences de motifs exceptionnels. Plus précisément, cet enrichissement est renforcé pour les motifs qui possèdent des occurrences non recouvrantes (*i.e.* toutes les occurrences ne sont pas dans la même zone du réseau). Nous avons également pu constater que cet enrichissement était renforcé pour les occurrences de motifs topologiques colorés (étiquetés). Ce résultat est à confirmer, il semble qu'il dépende de la taille de motif considéré.

Parallèlement, nos observations préliminaires nous ont également menés à développer de nouvelles fonctionnalités à notre algorithme de détection de motif, Motus. La notion d'exceptionnalité peut maintenant être gérée au niveau des groupes recouvrants d'occurrences (et non simplement au niveau du nombre d'occurrences). De plus, nous pouvons prendre en compte la notion de motif topologique coloré (grâce, pour l'instant, à l'interfaçage de Motus avec Nauty, un programme de

détection d'isomorphisme de graphes) qui semble prometteuse. Nous disposons donc désormais de plusieurs définitions emboîtées de la notion de motif.

De manière relativement orthogonale pour l'instant, nous avons également continué à travailler sur la notion de mode élémentaire. Un problème majeur que nous avions précédemment identifié est que le grand nombre de modes élémentaires croît très rapidement avec la taille du réseau, ce qui rend difficile leur interprétation. Nous avons donc commencé à proposer des méthodes de regroupement de ces modes élémentaires. Ce travail est réalisé en collaboration avec Leen Stougie (Université Technique de Eindhoven et CWI, Hollande), et de Alberto Marchetti-Spaccamela (Université « La Sapienza » de Rome, Italie).

2. Recherche de régularités au sein des réseaux métaboliques de bactéries en fonction de leur style de vie

Comme expliqué dans le précédent rapport, l'objectif ici est de modéliser le réseau métabolique de plusieurs bactéries libres, parasites ou mutualistes, endosymbiotes ou non, afin de distinguer des régularités propres aux différents styles de vie.

La première année du projet a été essentiellement destinée à se familiariser avec les méthodes et les outils de reconstruction de données métaboliques, et à développer de nouveaux outils de lecture et de formatage des données. À la date du dernier rapport, nous commencions tout juste la seconde étape qui était de comparer ces différents réseaux à plusieurs niveaux :

1. Au niveau structurel, en comparant certaines mesures classiquement appliquées sur les réseaux métaboliques : connectivité, modularité, diamètre...
2. Au niveau fonctionnel :
    1. Comparaison des motifs de réactions grâce au programme Motus ;
    2. Comparaison des capacités métaboliques propres à chaque organisme ;
    3. Comparaison de la robustesse des réseaux métaboliques.
3. Au niveau fonctionnel et structurel : en développant une méthode d'alignement de réseaux métaboliques qui prenne en compte les similarités locales à la fois structurelles et fonctionnelles. Le développement de cette méthode se fait en parallèle avec le développement d'une méthode efficace de comparaison de réactions biochimiques (dans le cadre d'un stage de Master M2 Pro en bioinformatique de l'Université de Rouen).

Ces perpectives sont toujours en cours mais ont connu cependant des développements significatifs depuis 6 mois (voir section D. 2.)

c. Au niveau du réseau génique

1. Utilisation des réseaux bayésiens à l'inférence des réseaux génétiques à partir de données d'expression : application en pharmacogénomique.

L'objectif est d'intégrer au sein d'un modèle probabiliste des données de natures différentes mais relatives au même système biologique. Les facilités de modélisation ainsi que la compréhension « intuitive » des modèles qu'elles offrent nous ont conduit à choisir des méthodes probabilistes sur les graphes, et en particulier les réseaux bayésiens. Les développements méthodologiques seront constamment confrontés à de « vrais » problèmes au sein d'une collaboration avec la plate-forme de pharmacogénomique de Lyon. Cette collaboration conduira à la modélisation simultanée de données cliniques, de réponses aux traitements et de puces à ADN (transcriptomique, CGH, polymorphisme), ceci dans le cadre d'études de plusieurs types de cancers (leucémie, cancer du sein, myélome multiple et cancer du poumon).

2. Inférence de modules d'expression à partir de données hétérogènes

Notre objectif ici est de traiter le problème de l'identification de modules de régulations, c'est-à-dire, de groupes de gènes co-régulés et de leurs régulateurs. Une importante originalité de notre travail est d'essayer d'identifier des modules qui sont évolutivement conservés. Du point de vue biologique, l'approche proposée s'appuie sur trois hypothèse principales :

- les gènes co-regulés sont sous le contrôle de protéines régulatrices (facteurs de transcriptions, notés TFs) communes et doivent donc présenter des patterns en séquence (motifs) communs dans leurs régions régulatrices, qui correspondent aux sites de fixation de ces TFs ;
- les gènes co-regulés répondent de manière coordonnée à certaines conditions environnementales ou de croissance, et doivent être co-exprimées sous ces conditions ;
- puisque les modules transcriptionnels sont censés être responsables d'importantes fonctions biologiques, ils sont soumis à une pression de sélection, et sont donc évolutivement conservés.

Nous avons ainsi défini le concept de métamodules de régulation transcriptionnelle (notés TRMMs) qui correspond à un groupe de gènes partageant des motifs de régulation et exhibant un comportement d'expression selon le contexte qui est cohérent et consistant quelles que soient les espèces. D'un point de vue méthodologique, nous constatons que l'incomplétude des données actuellement disponibles ainsi que leur haut niveau de bruit imposent des limitations sévères à la fiabilité des conclusions auxquelles nous pouvons arriver en analysant un seul type de données. C'est pour cela que nous proposons d'analyser simultanément des données expérimentales hétérogènes, particulièrement des séquences et des données d'expression de gènes provenant de diverses espèces.

d. Au niveau du réseau intégré

L'objectif est de proposer un cadre pour l'analyse des relations entre le génotype et le phénotype. Un des principaux objectifs étant la décomposition des réseaux biologiques en modules fonctionnels, nos travaux se concentrent sur l'analyse biostatistique des données d'expression en lien avec les réseaux métaboliques et de régulation. Nous nous sommes pour cela focalisés sur une étude de la répétition des motifs dans les réseaux métaboliques en corrélation avec l'expression et le mode de régulation des enzymes impliquées dans ces motifs. Les données d'expressions sont issues de plusieurs études de micro-arrays en accès libre sur la base de données SGD (Saccharomyces Genome Database http://www.yeastgenome.org) et les données de régulation sont fournies par Ana Teresa Freitas et Arlindo Oliveira du Technical University de Lisbonne, Portugal (récupérées à travers internet à l'URL : http://www.yeastract.com).

e. Relation entre dynamique des génomes et réseau métabolique

Nous avons initié une étude de l'impact de certains types de réarrangements sur le métabolisme d'un organisme. Dans un premier temps, nous nous intéressons au phénomène de duplication, en particulier aux duplications de génomes complets. L'organisme d'étude que nous avons adopté est celui de la paramécie. Le génome de la paramécie a subi au moins trois duplications complètes et il est relativement facile de distinguer les copies issues de chacune de ces duplications entre elles, ainsi que des copies issues de duplications localisées survenues de façon dynamique au cours de l'évolution. Une des hypothèses est que l'impact des dupliqués du premier type sur le métabolisme est tout à fait différent de l'impact que peuvent avoir les duplications dynamiques. Notre approche vise à la fois effectuer une étude exploratoire de ces différents types d'impacts, et essayer de tester des hypothèses biologiques très précises telles que : 1. les gènes dupliqués codant pour des enzymes et ayant conservé leur copie seraient proches dans le réseau, 2. ces gènes représenteraient préférentiellement les étapes initiales d'une voie de synthèse. Ces travaux sont effectués en collaboration avec Laurent Duret, Daniel Kahn et Jean-François Gout, chercheurs et doctorant respectivement dans l'équipe « Bioinformatique et Biologie Évolutive » de notre laboratoire.

D. Résultats obtenus pour la période concernée, dégager notamment les faits marquants

Nous ne présentons ici que les résultats nouveaux (publications, implantations, analyses bioinformatiques) obtenus depuis le dernier rapport d'activités (datant de Janvier 2007).

a. Au niveau génomique

1. Amélioration et extension des algorithmes de recherche de régularités dans les séquences nucléotidiques

- **Filtrage de séquences en vue d'un alignement multiple :** Les travaux sur Ed'Nimbus ont donné lieu à une publication acceptée depuis le dernier rapport dans la revue *Journal of Discrete Algorithms* (voir section « Livrables »). Un article présentant de nouvelles améliorations et extensions, ayant conduit à l'algorithme appelé TUIUIU, vient d'être soumis. Une implantation de TUIUIU sera rendue publiquement disponible prochainement.

- **Isochores** : Actuellement une méthode de partitionnement du génome fondée sur un apprentissage des propriétés a été développée sur des cas considérés comme exemplaires. Contrairement à toutes les méthodes précédentes de recherche des isochores, cette approche n'impose pas une définition *a priori* basée sur la fréquence C+G. De plus, une extension intégrant un raisonnement de type analyse comparative a récemment pu élargir le spectre des espèces ayant des isochores aux poissons. Plus récemment, nous nous sommes investi dans l'analyse et la modélisation de l'impact de la recombinaison génétique sur la structure en isochores.

- **Taux de recombinaison :** Une implémentation en R et Tcl/tk des cartes de Marey (graphe croisant distance physique et distance génétique) ainsi que des estimations du taux de recombinaison a été publiée dans *Bioinformatics* et est disponible sur la plate-forme PRABI à l'adresse http://pbil.univ-lyon1.fr/software/mareymap/.

2. Recherche de régularités dans les régions de cassure

Le filtre mis au point a été appliqué sur les gènes orthologues entre l'homme et le chimpanzé, et la qualité des assignations d'orthologie obtenues a été évaluée en comparant les régions de synténie correspondantes avec celles provenant d'alignement de génomes complets. Ainsi, avec les gènes orthologues, on retrouve toutes les régions de cassure détectées par les alignements de génome complet, plus d'autres non détectées. Ces données sont donc validées et seront utilisées pour l'analyse des régions de cassure.

Par ailleurs, un article initial portant sur l'étude de la corrélation de la position des points de cassure avec les réplicons connus ou inférés chez certains mammifères (principalement homme) est en préparation. Les résultats obtenus semblent orienter le modèle d'évolution par remaniements des génomes (un sujet encore très controversé) dans le sens de zones du génome plus propices aux cassures.

3. Étude des liens entre éléments répétés de différents types entre eux ainsi qu'avec d'autres facteurs, dont le niveau d'expression des gènes

Les résultats obtenus jusqu'à présent indiquent que 23 familles d'éléments transposables (ET) ne semblent pas exprimées dans *Drosophila melanogaster* alors que 70 le sont. Parmi ces 70, 33 familles possèdent des copies d'ET qui ont été associées à un unique EST (202 copies au total). De plus, il a été détecté un nombre de copies exprimées significativement plus grand dans le chromosome X par rapport aux autres chromosomes. Nous poursuivons actuellement les analyses

par un examen des 20kb flanquant les copies ETs n'ayant été associées qu'avec un seul EST. Cette analyse utilise les informations disponibles sur les sites de FlyBase et de FruitFly pour positionner ces copies relativement à d'autres caractéristiques du génome (euchromatine, hétérochromatine, centromère, gène, autre ET etc.), calculer la densité en gènes et le contenu en C+G de leurs régions flanquantes et vérifier si l'expression des gènes dans ces régions est corrélée avec celle de la copie, et, enfin, vérifier s'il y a sur-expression de ces ETs dans le chromosome X.

Le travail sur l'évolution des copies de rétrotransposons LTR dans les régions hétérochromatiques des génomes de drosophiles a permis de déceler une vague récente d'insertion de tels éléments dans le génome de *Drosophila melanogaster*. Ce travail a été soumis à *Molecular Biology and Evolution.*

4. Reconstitution de scénarios évolutifs par remaniements chromosomiques

- **Exploration des solutions du tri par inversions**. Nous avons développé un algorithme qui permet l'exploration de l'espace des solutions du tri par inversions, avec un gain de temps très important par rapport aux méthodes existantes. L'article qui décrit cette méthode a été accepté depuis le précédent rapport à la conférence ISBRA (qui a eu lieu en Mai 2007). Par ailleurs, nous avons été invités à soumettre une version longue de cet article à *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

- **Tri par inversion en présence d'éléments dupliqués**. Une première approche a été proposée qui correspond à la recherche d'une plus longue sous-séquence commune libre de répétitions. La complexité du problème a été étudiée montrant que ce dernier est APX-difficile (il est difficile d'obtenir même une approximation de la solution). Pour des instances pas trop grandes (comme cela est le cas, par exemple, des chromosomes X et Y), une formulation en termes de programmation linéaire par nombres entiers a été implantée. Un article a été soumis.

- **Reconstitution de l'organisation des génomes de mammifères ancestraux.** Une étude méthodologique des publications récentes sur le sujet de la reconstitution de l'organisation de mammifères ancestraux a été soumise à une revue depuis le précédent rapport. Cette étude suggère clairement que les différences observées entre ces méthodes sont dues, non au type de données utilisées (cytogénétiques versus génomiques) comme cela est suggéré ou affirmé dans de nombreux articles récemment parus, mais plutôt à l'algorithme utilisé pour un même ensemble d'organismes. Cet article recense quelques biais méthodologiques à éviter dans toute méthode de construction de génome ancestral.

5. Dépendances entre sites voisins dans les séquences d'ADN

Deux articles vont bientôt être soumis :

- Le premier (comme une *Application Notes* pour *Bioinformatics)* présentera un module Python pour la simulation de l'évolution de séquences le long d'un arbre sous des modèles avec dépendances entre sites. L'implantation est assez flexible pour intégrer des dépendances à courte échelle (effet des CpG), et des dépendances à plus large échelle (effet de mots plus longs, contrainte d'usage du code). Elle permet d'étudier par simulations de Monte-Carlo des modèles que l'on ne peut pas étudier analytiquement : dynamique générale du modèle, atteinte d'un état stationnaire, comportement à l'approche de cet état, distribution stationnaire.
- Le deuxième article concerne la mise en évidence d'une faible robustesse des méthodes d'inférence phylogénétique face à l'écart à l'hypothèse d'indépendance entre sites.

b. Au niveau du réseau métabolique

1. Recherche de motifs dans les réseaux métaboliques

Il est désormais possible de visualiser les occurrences d'un motif (grâce à une applet java développée par notre collaborateur Fabien Jourdan sur l'interface web de l'algorithme Motus disponible en accès restreint à http://pbil.univ-lyon1.fr/software/motus en utilisant le login : baobab et le mot de passe : baobab, et bientôt accessible publiquement).

De nouvelles fonctionnalités ont été implantés dans Motus, notamment la notion de regroupement des occurrences selon la topologie ou selon l'ordre des couleurs. De même, de nouveaux formats de sortie ont été développés pour faciliter l'analyse en post-traitement.

Un premier résultat concernant l'interface entre métabolisme et génome a été obtenu. En effet, la structure du réseau métabolique en motifs semble pouvoir être mise en correspondance dans certains cas avec la structure du génome en opérons. Ce résultat est encore à confirmer pour différentes tailles de motif.

Un algorithme de regroupement de modes élémentaires a été implanté et est en train d'être utilisé, par nous uniquement pour le moment, afin d'étudier ces modes.

2. Recherche de régularités au sein des réseaux métaboliques de bactéries en fonction de leur style de vie

Les analyses comparatives que nous voulons effectuer sur les réseaux métaboliques de différentes bactéries en fonction de leur style de vie dépendent essentiellement de la qualité de la reconstruction des données métaboliques, et cette dernière dépend à son tour fortement de la qualité de l'annotation des fonctions assignées à chaque gène. Dans la première partie du projet, les données métaboliques pour chaque organisme étaient reconstruites à partir des annotations des gènes telles qu'elles avaient été assignées par l'équipe en charge du séquençage. Ces annotations sont le plus souvent le fruit de méthodes automatiques, utilisant les informations d'orthologie entre séquences protéiques.

Afin de diminuer les biais dus à la qualité de ces annotations, nous avons décidé de changer la source des annotations et de développer les outils nécessaires pour pouvoir utiliser les données générées par HAMAP (High-quality Automated and Manual Annotation of Microbial Proteomes), système développé par l'équipe de Swiss-Prot et basé sur des annotations manuelles de protéines, capable d'annoter semi-automatiquement les protéines codées par un organisme, et de prendre en compte les mises à jour des annotations. La mise à jour de nos bases est actuellement en cours, mais environ la moitié est dorénavant construite à partir des données HAMAP. Le lien entre les données HAMAP (système dédié aux protéines) et les données génomiques se font à travers le site integr8, hébergé à l'EBI.

Dans le précédent rapport, nous mentionnions la naissance de SymBioCyc (http://biomserv.univ-lyon1.fr/baobab/symbiocyc), portail Web donnant accès aux différentes reconstructions métaboliques générées par les Pathway-Tools et aux analyses effectuées sur ces données. SymBioCyc est dorénavant accessible publiquement en version beta et intègre les reconstructions à partir des données HAMAP pour la moitié des organismes présents. Les réseaux métaboliques reconstruits sont représentés sous forme de 3 différentes sortes de graphes : le graphe des composés (dirigé ou non) dont les noeuds sont les métabolites et les arêtes les réactions, le graphe des réactions (dirigé ou non) dont les noeuds sont les réactions et les arêtes les composés, et le graphe biparti qui possède deux types de noeuds : les composés et les réactions. Ces graphes sont téléchargeables sous formes de fichiers facilement utilisables pour visualiser ou analyser le graphe avec les outils dédiés.

L'ensemble des données présentes dans SymBioCyc a été généré par les fonctions de la librairie Java parseBioNet. Celle-ci intègre également les outils statistiques sur les graphes de la librairie Java Jung, ce qui permet les mesures classiques sur les graphes présents dans SymBioCyc. Ces

mesures, tels que le degré moyen des noeuds, les mesures de centralité et de connectivité, vont nous permettre de caractériser les graphes présents dans SymBioCyc et nous donner des éléments de comparaison entre eux.

Afin de progresser dans l'analyse des réseaux métaboliques des endosymbiotes en particulier, nous avons par ailleurs développé des méthodes permettant de détecter les composés « essentiels » non produits par le réseau métabolique de la bactérie. Nous appelons ici composés essentiels, les composés qui devraient être produit par la bactérie, sur la base d'expériences physiologiques. Cette méthode aura deux utilisations dans notre cas : 1. identifier certaines erreurs dans la reconstruction métabolique, et 2. proposer des composés qui pourraient être fournis par l'hôte. Le laboratoire « Biologie Fonctionnelle, Insectes et Interactions (BF2I) » de l'unité mixte INSA-INRA de Lyon possède quelques données sur les transports possibles entre *Buchnera aphidicola* et le puceron, nous pourrons ainsi confronter nos résultats à leurs données. Cette méthode a été intégrée à la librairie parseBioNet et est actuellement en cours de test.

c. Au niveau du réseau génique : Inférence de modules d'expression à partir de données hétérogènes

Le travail sur l'inférence de modules d'expression à partir de données hétérogènes est effectué dans le cadre d'une thèse en co-tutelle avec un professeur brésilien de l'Université Fédérale de Pernambuco, Brésil. Les thèses brésiliennes doivent être pré-soutenues au plus tard un an avant la défense finale. Cette pré-soutenance implique la rédaction d'un manuscrit comportant une revue bibliographique, une formalisation détaillée du problème qui sera défendu, et une description des résultats attendus. La pré-soutenance a été réalisée avec succès en Février 2007. La thèse elle-même sera défendue avant Mars 2008.

Un programme, appelé MeMEx (« MetaModules Explorer ») est en cours d'implantation. Une première version fonctionnelle est attendue à la fin de l'été ou au début de l'automne. MeMEx est un logiciel convivial en Java/Swing qui sera mis à la disposition de la communauté. Il servira de plateforme pour une étude incluant quatre organismes pour lesquels des données étendues sont disponibles et qui ont déjà été utilisés dans des études comparatives précédentes.

d. Au niveau du réseau intégré

Nos études ont porté sur l'analyse de la corrélation entre la décomposition du réseau métabolique de la levure en motifs de taille k et les données d'expression obtenues à partir de différentes conditions environnementales. Pour chacune de ces conditions, les analyses ont été répétées fournissant une liste de gènes dont l'expression a présenté une modification (augmentation ou diminution) par rapport à l'état normal de la levure (aucune modification de l'environnement). À partir de ces données, des tests statistiques d'autocorrélation sur les graphes et de simulations ont été mis en oeuvre. Les premiers résultats montrent que les enzymes sous- ou sur-exprimées sont significativement plus souvent connectées dans le réseau que les enzymes dont l'expression ne varie pas. Ces résultats se généralisent pour des groupes d'enzymes dont le nombre peut varier de 2 à 6. Un second volet de l'analyse intègre les réseaux de régulation par le biais des facteurs de transcriptions régulant les enzymes. Les analyses statistiques effectuées (tests de permutation) ont mis également en évidence l'existence d'une corrélation significative entre la connexion topologique des enzymes et la régulation transcriptionnelle. L'étape suivante consistera à élargir cette étude avec l'idée de définir le seuil critique de connexions entre enzymes afin de proposer une première définition de module fonctionnellement significatif.

e. Relation entre dynamique des génomes et réseau métabolique

Le réseau de la la paramécie a été reconstruit (en utilisant la méthode Priam développée par Daniel Kahn et nos propres outils comme parseBioNet), première étape de l'étude de l'impact de certains types de réarrangements sur le métabolisme d'un organisme. Nous avons débuté maintenant l'étape

suivante, qui consiste initialement à tester des hypothèses telles que : 1. les gènes dupliqués codant pour des enzymes et ayant conservé leur copie seraient proches dans le réseau, 2. ces gènes représenteraient préférentiellement les étapes initiales d'une voie de synthèse.

E. Difficultés rencontrées et solutions de remplacement envisagées

Aucune difficulté majeure n'est à signaler outre le problème de la récolte de données dans certains cas ainsi qu'il l'avait été indiqué dans le rapport précédent.

F. Livrables externes réalisés

a. Articles acceptés, sous presse ou soumis entre Janvier 2007 et Juillet 2007

1. Articles dans des revues internationales

Y. Diekmann, M-F. Sagot, and E. Tannier, Evolution under reversals: parsimony and conservation of common intervals, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(3):301-309, 2007 (indiqué sous presse dans le rapport précédent).

E.Tannier, A. Bergeron, M.-F. Sagot, Advances on Sorting by Reversals*, Discrete Applied Mathematics*, 155(6-7):881-888, 2007 (indiqué sous presse dans le rapport précédent).

P. Peterlongo, N. Pisanti, F. Boyer, A. Pereira do Lago and M.-F. Sagot. Lossless filter for multiple repetitions, *J. Discrete Algorithms*, 2007, sous presse (indiqué comme soumis dans le rapport précédent).

P. Peterlongo, J. Allali and M.-F. Sagot. The Gapped-Factor Tree, *Int. J. Found. Comput. Sci.*, 2007, sous presse (indiqué comme soumis dans le rapport précédent).

R. Bourqui, L. Cottret, V. Lacroix, D. Auber, P. Mary, M.-F. Sagot, and F. Jourdan. Metabolic network visualization eliminating node redundance and preserving metabolic pathways, *BMC Syst. Biol*. 1(1):29, 2007.

C. Melodelima C, C. Gautier, and D. Piau, A markovian approach for the prediction of mouse isochores. *J. Math. Biol.,* 2007, sous presse.

C. Melodelima, L. Guéguen, C. Gautier, D. Piau, A Markovian approach for the analysis of the gene structure. *International Journal of Foundations of Computer Science*, 2007, sous presse.

C. Rezvoy, D. Charif, L. Guéguen, G. A. Marais, MareyMap: a R-based tool with graphical interface for estimating recombination rates. *Bioinformatics*, 2007, sous presse.

M.-F. Sagot, E. Tannier, Mammalian ancestral genome reconstructions: a small discourse on the method, 2007, soumis.

F. Chierichetti, V. Lacroix, A. Marchetti-Spaccamela, M.-F. Sagot, L. Stougie. Modes and Cuts in Metabolic Networks: Complexity and Algorithms, soumis (SPOR-Report 2007-01, Technische Universiteit Eindhoven, 2007, http://www.win.tue.nl/bs/spor/2007-01.pdf).

2. Articles dans des conférences internationales à comité de lecture

C. Melodelima, L. Guéguen, D. Piau, C. Gautier, Segmentation of the chimpanzee genome using a HMM model. *Lecture Notes in BioInformatics*, vol. 4414, pages 251-262, 2007.

M. Braga, M-F. Sagot, C. Scornavacca, E. Tannier, The solution space of sorting by reversals, *Lecture Notes in BioInformatics*, vol. 4463, 2007 (indiqué comme soumis dans le rapport précédent).

S. S. Adi, M. D.V. Braga, C. G. Fernandes, C. E. Ferreira, F. Viduani Martinez, M.-F. Sagot, M. A. Stefanes, C. Tjandraatmadja, Y.Wakabayashi, Repetition-free LCS, 2007, soumis.

b. Communications orales dans des conférences, colloques ou rencontres

E. Tannier, Genome rearrangements: the two parsimonies, présentation invitée au « Minisymposium on computational biology at the CANADAM 2007 conference », 27-31 Mai, 2007.

c. Logiciels et sites internet rendus publiques ou enrichis depuis Janvier 2007

- Ed'Nimbus (Pierre Peterlongo) : filtre de séquences qui peut être utilisé pour trouver de longues répétitions dans les séquences nucléotidiques
  http://igm.univ-mlv.fr/~peterlon/officiel/ednimbus/

- Fonctions supplémentaires dans SeqinR version 1.1-1 (Leonor Palmeira avec Delphine Charif, Jean Lobry, Anamaria Necsulea).
  http://cran.univ-lyon1.fr/src/contrib/Descriptions/seqinr.html

- Comparaison de cartes physiques et génétiques (Laurent Guéguen avec Clément Rezvoy, Delphine Charif et Gabriel Marais).
  http://pbil.univ-lyon1.fr/software/mareymap/

- Motus (Vincent Lacroix en collaboration avec Odile Rogier du PRABI) : recherche et inférence de motifs dans les réseaux métaboliques. Disponible à certains chercheurs externes au projet, bientôt disponible à toute la communauté. Diverses fonctionnalités ont été ajoutées depuis le dernier rapport (visualisation des occurrences, regroupement des occurrences par topologie ou ordre des couleurs, formats de sortie)
  http://pbil.univ-lyon1.fr/software/motus

- SymBioCyc (Ludovic Cottret) : site web incluant les reconstructions métaboliques et les principales caractéristiques de celles-ci pour 13 bactéries libres, parasites et mutualistes.
  http://biomserv.univ-lyon1.fr/baobab/symbiocyc/

- parseBioNet (Ludovic Cottret) : librairie Java permettant le filtre et l'analyse des réseaux métaboliques.
  http://biomserv.univ-lyon1.fr/baobab/parsebionet/

3. Thèses démarrées, en cours et/ou soutenues en relation directe avec le projet

Les thèses suivantes sont en relation directe avec le projet :

Nom du doctorant	Date de début de la thèse	Section(s) décrivant sujet principal
Vicente Acuña	Septembre 2006 (Lyon)	b1 (modes élémentaires) et e
Marília Braga	Septembre 2005 (Lyon)	a4
Yves-Pol Deniélou	Septembre 2006 (Grenoble)	réseaux et réarrangements
Marc Deloger	Septembre 2006 (Lyon)	a3
Vincent Lacroix	Septembre 2004 (Lyon)	b1 et e
Claire Lemaitre	Septembre 2005 (Lyon)	a2, a4 et e
Nuno Mendes	Janvier 2007 (Lyon)	motifs ARNs
Leonor Palmeira	Septembre 2004 (Lyon)	a5

La thèse de Leonor Palmeira a été soutenue le 13 Juillet 2007. Celle de Vincent Lacroix sera soutenue le 26 Octobre 2007.

Par ailleurs, Alexandra Popa, en stage M2 avec BAOBAB depuis Septembre 2006, a obtenu une bourse thèse de l'École Doctorale E2M2 de l'Université Lyon I. Elle travaillera sur les deux derniers points de la section (C)a1.

G. Autres commentaires

Les personnes suivantes ont été recrutées sur des fonds de l'ANR :

Ludovic Cottret, recruté comme ingénieur en Juin 2006 : sujet principal décrit dans les sections b2 et e.

Emmanuel Prestat, recruté comme ingénieur en Décembre 2006 : sujet principal décrit dans la section c1.

Patricia Thébault, recrutée comme postdoc en Février 2007 : sujet principal décrit dans la section d.

Eléments quantitatifs

H. Liste des réunions/séminaires/colloques organisés durant la période et des missions à l'étranger

a. Missions, séminaires et visites en France

Séminaire de Vincent Lacroix au laboratoire Physiologie Mitochondriale de Jean-Pierre Mazat (Février 2007 à Bordeaux, France) : « Motif Search in Metabolic Networks ».

Réunion du groupe de travail en génomique comparative (GTGC), le 9 juillet 2007 à Marseille (réunion satellite de la conférence francophone de bioinformatique Jobim 2007), co-organisée par Eric Tannier.

Visite de Stéphane Robin et Sophie Schbath à Lyon le 16 Juillet 2007 pour travailler sur un projet commun avec Ana Pombo (Londres) autour de l'organisation spatiale de la chromatine et son rôle possible dans la régulation des gènes et les réarrangements génomiques.

b. Missions, séminaires et visites à/de l'étranger

Séminaire de Vincent Lacroix au laboratoire de Roderic Guigó (Janvier 2007 à Barcelone, Espagne) : « Motif Search in Metabolic Networks ».

Visite de Claire Lemaitre, Ludovic Cottret, Vicente Acuña, Marília Braga et Marie-France Sagot à Lisbonne, Portugal, du 12 au 15 Février 2007 pour assister au « Workshop on Evolution, Co-evolution and Epigenetics » co-organisé par Marie-France Sagot.

Visite de Vincent Lacroix, Claire Lemaitre et Marie-France Sagot en Mars 2007 (du 1er au 3 Mars) dans le laboratoire de Ana Pombo à Londres pour travailler sur l'organisation spatiale de la chromatine (aspects expérimentaux).

Participation de Christelle Melo de Lima à la Conférence Bioinformatique à Berlin, du 10 au 15 Mars 2007 pour y présenter l'article : C. Melodelima, L. Guéguen, D. Piau, C. Gautier, Segmentation of the chimpanzee genome using a HMM model, *Lecture Notes in BioInformatics*, vol. 4414, pages 251-262, 2007.

Visite de Vincent Lacroix, Vicente Acuña et Marie-France Sagot rendue en Mars 2007 (du 13 au 16 Mars) à l'Université de Eindhoven (Pays-Bas) pour travailler avec Leen Stougie et Alberto Marchetti-Spaccamela sur l'énumération et le regroupement de modes élémentaires.

Visite de Roded Sharan, de l'université de Tel-Aviv, Israel, à Lyon en Mars 2007 (du 19 au 21 Mars) pour dégager des pistes de collaborations entre nos deux équipes sur l'analyse des réseaux biologiques.

Visite de Gabriel Valiente, de l'université technique de Catalogne, à Lyon en Mars 2007 (du 27 au 29 Mars) pour travailler sur la comparaison de réseaux métaboliques.

Participation de Marília Braga à la conférence ISBRA 2007 à Atlanta du 4 au 11 Mai 2007 pour y présenter l'article : M. Braga, M-F. Sagot, C. Scornavacca, E. Tannier, The solution space of sorting by reversals, proceedings of ISBRA'07, Lecture Notes in BioInformatics, vol. 4463, 2007.

Participation de Céline Scornavacca au « Minisymposium on computational biology at the CANADAM 2007 conference », Calgary, Canada du 27 Mai au 8 Juin 2007 et présentation de l'article mentionné ci-dessus.

Visite de Eric Tannier à l'Université Simon Fraser à Vanvouver en Mai 2007.

I. Par rubrique et par partenaire, établir la consommation des dépenses financées par l'ANR, depuis le démarrage du projet.

Partenaire	Fonct. (Keuros)	Equip. nature	Equip. (Keuros)
UCBL	81,07		
Total projet	81,07		

J. *Le cas échéant et pour les programmes thématiques*, préciser les travaux réalisés par les partenaires étrangers associés au projet sans aide de l'ANR

K. Liste des personnels recrutés en CDD par des établissements publics dans le cadre du projet sur l'aide allouée par l'ANR

Nom	Prénom	Qualifications	Date de recrutement	Durée du contrat (en mois)
Cottret	Ludovic	Ingénieur	01/05/2006	28 mois
Prestat	Emmanuel	Ingénieur	01/12/2006	24 mois
Thébault	Patricia	Postdoc	01/02/2007	7 mois
Fonseca	Paulo	Ingénieur	01/02/2007	10 * 0.50 mois

L. *Le cas échéant*, indiquer les différents types d'aides complémentaires obtenues grâce à ce projet.

M. *Le cas échéant*, modalités d'utilisation du complément de financement « pôles de compétitivité »

N. CADRE RESERVE AU COORDONNATEUR DU PROJET

Relativement au dernier rapport datant du 15 Janvier 2007, nous avons poursuivi nos travaux sur la recherche de régularités de natures diverses à tous les niveaux, du génome aux réseaux, ainsi qu'à l'interface entre ces niveaux. Ce qui a caractérisé plus particulièrement cette dernière période de six mois par rapport au précédent rapport a été tout d'abord des avancées en ce qui concerne à la fois l'organisation et la dynamique des génomes, avec des résultats soumis ou qui feront l'effet d'une soumission à publication dans les prochains mois. Nous avons également développé une nouvelle collaboration (avec l'équipe d'Alain Arnéodo de l'ENS de Lyon) qui devrait nous permettre d'établir de façon plus précise les liens entre la structure d'un génome (en réplicons, duplicons, etc.) et les divers réarrangements que ce dernier subi au cours de l'évolution. Nous espérons ainsi pouvoir également revoir et approfondir notre compréhension du modèle d'évolution par remaniements. Grâce à une amélioration de la modélisation du processus d'occurrence des cross-overs, notamment en prennant en compte les résultats concernant à la fois les processus moléculaires en cause (relation entre chiasma, appariement des brins non homologues, et cross-over) et les chromosomes dans leur entier (nécessité d'un chiasma au moins par bras), nous avons également mis à jour une importante relation entre taux de recombinaison et longueur des bras des chromosomes.

Une autre avancée originale par rapport au rapport précédent concerne l'inférence de modules d'expression à partir de données hétérogènes (motifs en séquence, puces et évolution). La méthode MeMEx qui sera rendue disponible avant le prochain rapport représente un logiciel complet dont chaque différent volet s'attaque à un problème original en soi : développement d'une nouvelle structure de données pour la prise en compte de dépendances entre sites à l'intérieur d'un motif en séquence ; nouvel algorithme de détection de bi-clusters dans des données d'expression ; phylogénie de modules dans les réseaux géniques.

Des progrès significatifs ont également été réalisés en ce qui concerne l'analyse du réseau intégré, où une corrélation importante a été détectée entre, d'un côté, l'ensemble de réactions connectées dans le réseau métabolique de la levure et, d'un autre côté, la sous/sur-expression de gènes et le partage de TFs. De même, une corrélation positive a été observée, dans le cas de *Escherichia coli*, entre les opérons et les motifs ayant plusieurs paquets non recouvrant d'occurrences qui sont détectés par le logiciel Motus, qui a par ailleurs été enrichi en fonctionnalités nouvelles au cours de ces derniers mois.

Concernant les jeux de données pour l'étude des réseaux métaboliques, notamment de bactéries, nous avions mentionné dans notre précédent rapport les difficultés que rencontrions encore dans certains cas pour en obtenir qui soient fiables, ou suffisamment complètes. Au vu de cela, nous avons décidé de changer la source des annotations et de construire les outils nécessaires pour pouvoir utiliser les données engendrées par HAMAP (High-quality Automated and Manual Annotation of Microbial Proteomes), système développé par l'équipe de Swiss-Prot et basé sur des annotations manuelles de protéines, capable d'annoter semi-automatiquement les protéines codées par un organisme et de prendre en compte les mises à jour des annotations. Ce sont ces données qui se trouvent désormais, pour la moitié des organismes, à la base des reconstructions que nous avons rendues accessibles publiquement en version beta via le portail Web SymBioCyc. Ce souci mis partiellement de côté, nous concentrons désormais notre attention sur l'analyse comparative de ces réseaux en fonction du mode de vie des organismes, et sur le lien entre certaines classes de remaniements génomiques, essentiellement duplications pour le moment, et réseau métabolique.

Enfin, nous avons continué à rendre les méthodes que nous développons et leurs améliorations / extensions disponibles à travers des interfaces Web.

CADRE RESERVE A l'USAR

Nom du coordinateur scientifique de l'USAR :
Date :