Rapport semestriel d'activité n°1 datant du 15/06/2006

A. Identification

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

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 décembre 2005 – 15 juin 2006
Rédacteur de ce rapport : nom	SAGOT Marie-France
téléphone	04-72-44-82-38
fax	04-72-43-13-88
adresse électronique	marie-france.sagot@inrialpes.fr
Date du rapport	15 juin 2006

B. Pour les projets multi-partenaires, rappel des tâches allouées par partenaire pour l'ensemble du projet (partir du planning généralement fourni dans le projet. Ce document est à remplir par le coordonnateur du projet à partir des informations fournies par les partenaires)

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

Eléments qualitatifs

C. Description des travaux effectués pour la période concernée et conformité de l'avancement aux prévisions (15 à 50 lignes maximum suivant le nombre de partenaires)

C1. Inférence et mise en évidence de régularités :

a. au niveau génomique

- Dans la perspective d'identifier des régularités au niveau génomique, nous nous sommes intéressés en particulier aux régions de ruptures de synténie (appelés points de cassure) dans les génomes de mammifères. Les travaux effectués jusqu'à présent ont consisté au développement d'une méthode de détection des points de cassure sur un génome. Cette méthode est basée sur l'analyse comparative de l'ordre des gènes orthologues et sur l'alignement de séquences intergéniques. L'analyse des séquences aux bornes des points de cassure détectés sera l'étape suivante.

- Reconstitutions de scénarios évolutifs par remaniements chromosomiques : exploration des solutions du tri des permutations par inversions. Le plus gros problème des méthodes combinatoires actuelles pour les reconstitutions de scénarios d'inversions est le très grand nombre de solutions qu'elles fournissent, sans critère pour les départager. Nous réduisons l'espace des solutions en ajoutant des contraintes de conservation de segments chromosomiques présents dans plusieurs espèces, et construisons des algorithmes proposant une solution optimale avec cette contrainte.

- Une vision générale des ET dans les génomes séquencés et annotés est disponible mais nous ne disposons que de peu d'informations concernant leur expression. Nous avons donc mené une étude sur l'expression des familles ETs présentes dans le génome de la drosophile à partir de données d'EST. S'agissant de séquences répétées, il a fallu établir une méthode d'attribution spécifique des EST aux copies d'ET. La difficulté à effectuer une telle attribution dans le cas de séquences répétées est bien connue, particulièrement dans le cadre de l'assemblage de séquences de génomes complets. Ce travail doit se poursuivre et conduire à moyen et long termes à une analyse du lien entre les éléments répétés (ETs et autres) et expression des gènes d'un organisme à un niveau global.

- Enfin, une analyse de l'influence possible de l'environnement sur la composition des génomes a été réalisée dans le cas particulier de l'exposition des bactéries aux UV et de leur contenu en dinucléotides. Cette analyse a apporté un élément supplémentaire dans la controverse concernant une telle influence. Dans le cas des UV, l'étude menée n'a pas révélé d'influence.

b. au niveau du réseau métabolique

Afin d'étudier les propriétés structurelles des réseaux métaboliques, plusieurs modèles peuvent être considérés. Dans un premier temps, nous avons choisi de modéliser un réseau métabolique par un graphe coloré non orienté dans lequel les noeuds correspondent aux réactions, les arêtes aux métabolites qui lient ces réactions et les couleurs aux classes de mécanismes réactionnels auxquelles les réactions appartiennent (EC number). Nous nous sommes appuyés sur ce modèle simple pour définir la notion de motif qui correspond à la notion de brique structurelle élémentaire d'un réseau métabolique. Ce travail préliminaire a soulevé de nouvelles questions ainsi que des idées d'amélioration énumérées à la suite et qui sont en cours de réalisation.

- Exploration de modèles de graphes aléatoires capables de capturer la structure d'un réseau métabolique dans le but de pouvoir étudier la sur-représentation des motifs dans un tel réseau.

- Modélisation de réseaux métaboliques par des hypergraphes permettant de représenter de manière plus réaliste les liens qui lient les métabolites entre eux et dès lors de dégager des propriétés structurelles plus fines.

- Rapprochement avec d'autres méthodes d'analyse des réseaux métaboliques basées sur une décomposition de la matrice stochiométrique et reformulation du problème de recherche de modes élémentaires dans ce cadre. Exploration de problèmes connexes et de leur complexité.

- Conception d'une nouvelle méthode de visualisation de graphes permettant de dessiner un réseau métabolique en prenant en compte sa structure en voies.

c. au niveau du réseau génique

Les travaux dans le cas de réseaux géniques ont débuté par les deux questions suivantes.

- Étude des méthodes informatiques pour l'inférence des modules transcriptionnels (i.e. clusters de genes co-regulés et leurs correspondants régulateurs) dans les réseaux de régulation à partir de données diverses (expression, séquence, etc.), et dans le cadre d'une approche comparative (exploitation simultanée de données issues de différents organismes).

- Développement (en cours) d'une méthode basée sur les réseaux bayésiens pour l'intégration d'informations issues de données de natures différentes (données cliniques, données de puces, expertise d'anatomopathologiste...) et leurs relations avec les réponses toxicologiques de traitements anti-cancéreux sur une population de patients. Ce projet, qui s'inscrit dans le domaine de la pharmacogénomique, est axé plus particulièrement sur la leucémie et le cancer du sein et fait l'objet d'une nouvelle collaboration avec le laboratoire de pharmacologie de la faculté de médecine de Grange Blanche à Lyon, grâce à laquelle nous aurons un retour d'expertise et des données originales pour tester et améliorer le modèle. L'approche actuelle est tournée vers la pharmacogénomique, cependant l'objectif global du projet est clairement méthodologique : développer un modèle probabiliste capable de lier les données phénotypiques d'un ensemble d'individus aux données d'expression, la méthode pourra ainsi s'étendre facilement à des domaines plus larges.

C2. Modélisation du réseau intégré

Pour le moment, nous nous sommes concentrés sur l'intégration des réseaux de régulation et métabolique. Le modèle biologique choisi a trait au shift diauxique de la levure (lequel se définit par le point de bascule entre la fermentation et la respiration). Avec l'idée de tendre vers la décomposition des réseaux biologiques en modules fonctionnels, nos travaux actuels ont porté sur l'intégration des données d'expression aux motifs répétés dans les réseaux métaboliques dont une définition et un algorithme de détection (appelé Motus) ont été proposés par Vincent Lacroix (voir section C1, item b ci-dessus pour la définition et section C3 ci-dessous).

C3. Développement de solutions informatiques

a. Acquisition et visualisation de données

- Développement d'un ensemble d'outils informatiques dédié au tri et au traitement des données métaboliques provenant de BioCyc.

- Etablissement de jeux de données en format SBML pour l'étude de réseaux métaboliques à partir des données de BioCyc.

- Développement d'une interface WEB pour le programme de recherche de motifs dans les réseaux métaboliques Motus, développé par Vincent Lacroix.

b. Implémentation des algorithmes

- Implémentation d'un algorithme de recherche de motifs dans un réseau métabolique. Cet algorithme est décrit dans un article initialement présenté à la conférence WABI en 2005, dont la version étendue a été acceptée pour publication dans la revue *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (sous presse, 2006).

- Implémentation d'un algorithme d'inférence de motifs dans un réseau métabolique

- Ces deux algorithmes ainsi que des méthodes élémentaires de calculs de propriétés de graphes sont disponibles dans Motus via une interface web, bientôt accessible publiquement.

D. Résultats obtenus pour la période concernée, dégager notamment les faits marquants (15 à 50 lignes maximum) *Décrire les résultats obtenus et préciser éventuellement les livrables déjà réalisés en interne au projet.*

Plusieurs résultats préliminaires ont été obtenus mais se trouvent en cours de finalisation et seront détaillés dans le prochain rapport, au bout d'une année de projet.

E. Difficultés rencontrées et solutions de remplacement envisagées (15 à 50 lignes maximum) *ex : impasse technique, abandon d'un partenaire ou d'un sous traitant, maîtrise des délais, maîtrise des budgets. Faut-il revoir le contenu du projet ? Faut-il revoir le calendrier du projet ?*

Les principales difficultés rencontrées en ces 6 premiers mois de projet ont porté, dans un cas, sur les données, et dans l'autre, sur à la fois les données et des problèmes conceptuels.

Le premier cas concerne les données d'orthologie entre génomes nécessaires pour la détection précise et l'analyse des régions de cassure dans un génome et, ainsi, pour arriver à une meilleure compréhension de la dynamique des génomes. Les données disponibles publiquement contiennent clairement des erreurs d'assignation d'orthologie qui compliquent et biaisent la recherche et l'étude de points de cassure. Cette difficulté n'invalide pas le problème mais va nous obliger à y ajouter une tâche supplémentaire qui va consister à revoir partiellement les méthodes de détection d'orthologie et à développer une approche propre.

La seconde principale difficulté concerne la modélisation de réseaux de signalisation qui peuvent être vus comme des réseaux intégrés plus riches que les réseaux génique+métabolique. La difficulté ici porte à la fois sur la quantité et surtout la fiabilité relativement faibles des données disponibles pour la modélisation, et sur la meilleure façon de représenter ensemble des données de natures diverses. Cette partie du projet concernait toutefois un plus long terme et nous avons ainsi choisi de nous concentrer dans un premier temps sur le problème de l'intégration de données géniques (essentiellement données d'expression) et métaboliques, et cela à partir des motifs répétés détectés par l'algorithme Motus dans les réseaux métaboliques (voir section C3).

F. Livrables externes réalisés (15 à 50 lignes maximum)

Pour les articles et communications écrites, préciser s'il s'agit d'articles dans des revues à comité de lecture / d'ouvrages ou chapitres d'ouvrage / d'articles dans d'autres revues / de communications dans des colloques ou des congrès / de dépôt de brevet… Référencer selon les normes habituelles. Mentionner également s'ils peuvent ou non faire l'objet de communications externes par l'ANR et son unité support

Indiquer, *Le cas échéant*, les thèses démarrées, en cours et/ou soutenues en relation directe avec le projet :

Préciser le titre, date de soutenance (prévue ou réelle), soutien financier, devenir des étudiants pour les thèses soutenues

Ainsi qu'indiqué en section C3, une interface web du logiciel Motus sera bientôt mis à disponibilité du milieu académique (l'interface existe déjà mais n'a pas encore été rendue accessible car elle en cours de test dans l'équipe). La suite de routines ayant permis une première identification des éléments transposables actifs dans la drosophile (applicable aux génomes d'autres organismes eucaryotes) sera également, une fois nettoyée, rendue disponible sur demande.

Par ailleurs, les articles suivants ont été acceptés ou soumis entre le 15 Décembre 2005 et le 15 Juin 2006 (seuls sont indiqués les articles de l'équipe BAOBAB ayant directement à voir avec le projet).

V. Lacroix, C. G. Fernandes and M.-F. Sagot. The Motif Search Problem in Graphs: Application to Metabolic Networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, sous presse.

C. Lemaitre and M.-F. Sagot. A Small Trip in the Untranquil World of Genomes. soumis à *Theoretical Computer Science*, 2006.

L. Palmeira, L. Guéguen and J. Lobry. UV-targeted dinucleotides are not depleted in light-exposed Prokaryotic genomes. *Mol. Biol. Evol*., sous presse.

G. Autres commentaires

Aucun pour le moment.

Eléments quantitatifs

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

(*préciser la date, le lieu, l'objet, le nombre des participants)*

Une rencontre a été organisée par BAOBAB à Bertinoro, Italie (http://www.inrialpes.fr/helix/people/sagot/bertinoro2006/bertinoro.html) qui a permis de prendre contact avec d'autres chercheurs dont certains pourraient devenir de futurs collaborateurs. La rencontre était financée par les propres invités. Quatre membres de BAOBAB y ont participé. Les frais d'avion étaient déjà couverts par des fonds précédant le début du projet REGLIS, et nous n'avons pas eu, en tant qu'organisateurs, à payer de logement sur place (centre universitaire de Bologne).

Par ailleurs, deux membres de BAOBAB ont effectué une visite à Rome dans le cadre d'une collaboration qui débute sur le problème du calcul de modes élémentaires dans un réseau métabolique. La collaboration concerne un chercheur informaticien de l'Université de Rome, Alberto Marchetti-Spaccamela, et un chercheur informaticien de l'Université de Eindhoven en Hollande, Leen Stougie. Nous étions invités.

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

Nous n'avons encore effectué aucune dépense de fonctionnement ou d'équipement sur ces 6 premiers mois. Des dépenses de fonctionnement sont prévues déjà pour les prochains 6 mois et seront détaillées dans le prochain rapport. Les dépenses d'équipement sont prévues dès l'arrivée du postdoc, prévue pour la fin de 2006 (le recrutement d'un ingénieur ayant été prioritaire pour nous – voir section K).

Partenaire	Fonct. (Keuros)	Equip. nature	Equip. (Keuros)
Baobab			
Total projet			

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

Nota : sans objet pour les programmes « Blanc » et « JCJC »

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 d'étude**	**01/05/2006**	**12 mois**

Ce contrat sera renouvellé à son échéance d'un an.

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

(*Il peut s'agir de ressources financières, ressources humaines, allocations de recherche,…)*

Conjointement avec une autre équipe du Lab. de Biométrie et Biologie Évolutive (équipe « Génomes et Populations » de Christian Biémont), nous avons déposé une demande de bourse BDI CNRS qui a été acceptée. L'étudiant recruté sur cette bourse sera co-dirigé par Cristina Vieira de l'équipe « Génomes et Populations » et par Marie-France Sagot, coordonnatrice de ce projet. Le sujet portera sur la relation entre éléments répétés (en particulier éléments transposables) et régulation globale des gènes chez les eucaryotes. Il concernera dans une première étape le développement de méthodes de détection systématique d'éléments répétés dans un génome eucaryote.

M. *Le cas échéant*, modalités d'utilisation du complément de financement « pôles de compétitivité » (15 lignes maximum)

Rappel : ceci ne s'applique pas aux entreprises, mais seulement aux laboratoires publics et autres structures non soumises à l'encadrement communautaire des aides d'Etat à la R&D. Le complément de financement est destiné à couvrir des frais supplémentaires liés à la participation aux activités du pôle : ingénierie de projets partenariaux publics-privés, recherche de partenaires ; valorisation de la recherche ; relations inter-pôles et internationales…

N. CADRE RESERVE AU COORDONNATEUR DU PROJET (15 à 50 lignes maximum)

Commentaire général sur l'état d'avancement du projet, les interactions entre les différents partenaires, les efforts particuliers en matière d'interdisciplinarité, l'ouverture internationale, etc.

Ce projet comporte comme difficulté principale le fait qu'il exige, afin d'essayer de répondre à la question très générale « est-ce que la structure des systèmes vivants est simple, ou simplifiable en des principes généraux, ou bien est-ce que la vie est faite essentiellement d'exceptions? » d'avancer sur plusieurs fronts à la fois dans la recherche de régularités de natures diverses au niveau à la fois du génome et du réseau d'interactions. Dans ses six premiers mois, nous avons choisi d'avancer sur ces différents sujets simultanément mais en restant dans un premier abord proches de résultats que nous avions déjà et que nous avons essayé d'étendre. Le projet avance de façon satisfaisante et si nous sommes seuls partenaires, son financement nous permet d'établir des collaborations internationales importantes pour l'avenir. Ces collaborations nouvelles concernent pour le moment des chercheurs en bioinformatique venant de l'informatique ou des mathématiques des universités de Rome (Italie), Eindhoven (Hollande), Tel Aviv (Israel), Toulouse et Bordeaux ainsi que des chercheurs biologistes des universités de Bordeaux et York (UK). Ces collaborations concernent plus particulièrement le volant « réseaux » du projet. Des collaborations portant sur le volet génomique et dynamique des génomes sont encore à renforcer. Le projet compte bien sûr toujours sur les nombreuses collaborations multidisciplinaires nationales et internationales déjà existantes avant le dépôt de ce projet, notamment au niveau international avec le Portugal, l'Italie, le Brésil, Israel, et l'Angleterre. Outre ces collaborations, il était important pour nous également, lors de ces 6 premiers mois, de commencer à établir des jeux de données fiables et des interfaces aux débuts de méthodes que nous développions, ce que nous avons fait. Ces interfaces, dotées notamment d'outils de visualisation, sont importantes à la fois pour nous et pour améliorer la qualité et l'efficacité de notre dialogue avec la communauté des bioinformaticiens et des biologistes. Il est prévu que les méthodes interfacées soient mises à la disposition de la communauté mais dans un premier temps nous souhaitons, d'abord les tester plus exhaustivement, ensuite en bénéficier nous mêmes afin de tester nos modèles et éventuellement les revoir de façon également un peu exhaustive. Notre priorité au cours de la deuxième année de projet sera de réaliser ces tests des modèles. C'est à ce moment que nous prévoyons le recrutement du post-doctorant prévu dans le projet.

CADRE RESERVE A l'USAR

Nom du coordinateur scientifique de l'USAR :
Date :

Glossaire

Livrable : tout composant matérialisant le résultat de la prestation de réalisation. Toute production émise par le titulaire au cours du projet : document, courrier revêtant un caractère officiel , module de code logiciel, dossiers de tests, application intégrée, objet, dispositif…

Livrable interne : réalisé au sein du programme et non communiqué à l'extérieur du programme.

Livrable externe : élément diffusé ou livré hors de la communauté du projet de recherche..

Faits marquants : élément non nécessairement quantifiable mais significatif pour le projet.