Agence Nationale de la Recherche ANR	**Fiche compte-rendu semestriel d'activité**	Date : 08/02/2008 Réf: ANR-05-NT05-3_45205 Nombre de pages : 6

No de Projet : ANR-05-NT05-3_45205

A. Identification

Projet (acronyme)	REGLIS
Coordinateur du projet (société/organisme)	Équipe Baobab – Laboratoire de Biométrie et Biologie Évolutive
Période du projet (date début – date fin contractuelle)	15 décembre 2005 – 15 décembre 2008

Rédacteur de ce rapport

civilité, prénom, nom	SAGOT Marie-France
téléphone	04-72-43-13-88
adresse électronique	Marie-France.Sagot@inria.fr

Compte rendu semestriel d'activité n°X/Y

Période faisant l'objet du rapport d'activité (date début – date fin)	15 Juillet 2007 - 15 Janvier 2008
Date de rédaction	8 Février 2008

B. Pour les projets partenariaux, rappel des livrables ou jalons alloués aux partenaires pour l'ensemble du projet

Ce projet ne comportait qu'un seul partenaire.

C. Retombées cumulées sur la durée du projet

Cette section rassemble des éléments cumulés qui seront suivis tout au long de l'avancée du projet et repris dans son bilan. Ils permettent d'apprécier l'impact du programme à différents niveaux. Cette section est constituée d'un tableau des publications, et d'une liste de résultats éventuellement plus qualitatifs.

Nombre de publications et de communications <u>cumulées</u> sur la durée du projet.

	International		France		Actions de diffusion		
	Articles acceptés dans des revues à comité de lecture	Communications Internationales à comité de lecture	Articles France	Communications France	Articles vulgarisation	Conférences vulgarisation	Autres Séminaires invités
monopartenaire	16	5	0	2	1		19

Liste des publications et communications relatives au projet et ne figurant pas dans les rapports antérieurs.

Articles acceptés dans des revues à comité de lecture

M. D. V. Braga, M.-F. Sagot, C. Scornavacca and **E. Tannier**. Exploring The Solution Space of Sorting by Reversals With Experiments and an Application to Evolution, to appear in *Transactions on Computational Biology and Bioinformatics*, 2008.

N. Mugnier, **L. Gueguen**, C. Vieira and C. Biémont. The heterochromatic copies of the LTR retrotransposons as a record of the genomic events that have shaped the Drosophila melanogaster genome, *Gene*, in press, 2008.

J. Allali, **M.-F. Sagot**. A multiple layer model to compare RNA secondary structures. *Software: Practice and Experience*, in press, 2008.

Articles acceptés dans des conférences internationales à comité de lecture

S. S. Adi, **M. D. V. Braga**, C. G. Fernandes, C. E. Ferreira, F. V. Martinez, **M.-F. Sagot**, M. A. Stefanes, C. Tjandraatmadja, Y. Wakabayashi. Repetition-free longest common subsequence. Latin-American Algorithms, Graphs and Optimization Symposium (LAGOS), 2007. To appear in *Electronic Notes in Discrete Mathematics*.

A. M. Carvalho, A. L. Oliveira and **M.-F. Sagot**, Efficient learning of Bayesian network classifiers: An extension to the TAN classifier, In M. A. Orgun and J. Thornton, editors, *Proceedings of the 20th Australian Joint Conference on Artificial Intelligence*, volume 4830 of *Lecture Notes in Artificial Intelligence*, pages 16-25. Springer-Verlag, 2007.

R. Lenne, C. Solnon, T. Stützle, **E. Tannier**, M. Birattari. Reactive Stochastic Local Search Algorithms for the Genomic Median Problem. accepted at *EvoCop*, proceedings to appear in *LNCS*, 2008.

Articles acceptés dans des conférences nationales à comité de lecture

L. Cottret, V. Acuña, H. Charles, **M.F. Sagot**. Recherche de précurseurs dans un réseau métabolique. Réseaux d'Interactions : Analyse, Modélisation, Simulation (RIAMS'08), 2008.

Séminaires présentés dans le cours du dernier semestre

Claire Lemaitre. Détection précise des points de cassure de réarrangement dans les génomes de mammifères. Alphy, 1er février 2008, Lyon.

Eric Tannier, "Reconstruction of mammalian ancestral genomes: the two pasimonies", séminaire du laboratoire G-SCOP, 8 novembre 2007, Grenoble.

Eric Tannier, "Reconstruction de génomes ancestraux: les deux parcimonies", séminaire "algorithmes" des équipes Bipop et Casys, 8 novembre 2007, Grenoble.

Eric Tannier, "Reconstruction de génomes ancestraux: les deux parcimonies", séminaire du laboratoire LIAFA, 22 janvier 2008, Paris.

Autres retombées :

Nature	Commentaire
Brevets nationaux	-
Brevets internationaux	-
Autres	Logiciels disponibles publiquement : 7

D. Eventuellement, résultat marquant du semestre écoulé

Le résultat le plus marquant du semestre écoulé est une méthode efficace de recherche d'ensembles de composés (appelés précurseurs) capables de synthétiser des composés « cibles » dans un réseau métabolique. Ce travail est en cours de rédaction.

E. Description des travaux effectués et résultats obtenus pendant la période concernée. Conformité de l'avancement des travaux avec le plan initialement prévu. Prévision de travaux pour la (les) prochaine(s) période(s)

Travaux effectués et prévision

Inférence et mise en évidence de régularités au niveau génomique

Principaux résultats et prévisions :

Finalisation d'une méthode d'affinement des régions de cassures suite à un réarrangement et application à l'analyse comparative de divers génomes de vertébrés (homme, souris, rat, macaque, chimpanzé et chien). Un article sur la méthode avec application à l'homme, la souris et le chien est en cours de soumission, tandis que les 6 vertébrés ci-dessus ont été utilisés dans une étude de corrélation en cours des régions de cassure chez l'homme avec des sites de réplications prédits (par nos collaborateurs de l'ENS de Lyon du groupe d'A. Arnéodo). La corrélation avec d'autres types de structures génomiques est en cours d'analyse.

Poursuite de l'étude de l'activation de certains types d'éléments répétés, les éléments transposables chez *Drosophila melanogaster,* en utilisant des banques publiques d'EST. Il a été montré que seulement 70 familles d'éléments transposables chez la drosophile sont potentiellement exprimées et que parmi ces familles, seulement 202 copies sont en relation avec un unique EST. Une étude des 29 gènes (principalement impliqués dans des fonctions de signalisation, transcriptase inverse, développement et transport) insérés dans ces éléments transposables exprimés (62% LTR Retrotransposons / 31% Non LTR Retrotransposons / 7% DNA Transposons) a révélé que près de 60% sont insérés dans des introns, 20% en 5'UTR et 20% en 3'UTR. Un article est en cours de rédaction et l'analyse se poursuit maintenant à un niveau comparatif avec d'autres espèces de drosophile afin de détecter une possible signification évolutive.

Analyse exhaustive des miRNAs dans l'anophèle. Ce travail effectué avec le laboratoire de Eric Westhof à Strasbourg, a produit des résultats qui sont en cours d'analyse plus détaillée et de comparaison avec des résultats similaires obtenus sur la drosophile.

Application de l'algorithme de représentation de tous les scénarios optimaux d'inversion entre deux génomes présenté à une conférence dans le semestre précédent à l'étude de l'évolution des chromosomes humains X et Y, et des bactéries symbiotiques *Rickettsia*. Cette étude a permis de dégager un nouveau problème combinatoire très intéressant qui est celui de la génération aléatoire de permutations situées à une distance d'inversion donnée de la permutation identité. En parallèlle, une première méthode de recherche d'ordre des gènes dans les génomes ancestraux a été mise au point et l'article accepté à une conférence.

Inférence et mise en évidence de régularités au niveau des réseaux

Principaux résultats et prévisions :

Mise au point d'une méthode permettant d'identifier des ensembles de composés (appelés précurseurs) capables de synthétiser des composés « cibles » dans un réseau biologique. Après une formalisation algorithmique du problème, étude de sa complexité, et implantation, la méthode est en cours d'application au réseau métabolique de *Buchnera aphidicola*, en se focalisant dans un premier temps sur la recherche des précurseurs des acides aminés essentiels produits par la bactérie pour le puceron qui ne peut plus synthétiser ces acides aminés lui-même. Un article sur la méthode elle-même est en cours de rédaction. L'algorithme sera rendu disponible prochainement. Une analyse approfondie du réseau de *Buchnera* va se dérouler le long du prochain semestre.

L'algorithme d'inférence de motifs dans des réseaux métaboliques présent dans le logiciel MOTUS a été amélioré, donnant lieu à un première version qui est plus rapide de plusieurs ordres de magnitude que celle semi-naïve actuellement implantée dans MOTUS. Un article décrivant cet algorithme a été soumis et les travaux se poursuivent afin de traiter des motifs plus complexes. Parmi ceux-là, les motifs dont les occurrences induisent deux composantes connexes dictintes sont particulièrement intéressants pour une étude de l'évolution du réseau métabolique. À noter qu'une thèse, de Vincent Lacroix, sur ces motifs dans les réseaux métaboliques a été soutenue en Octobre 2007. Vincent Lacroix est actuellement postdoc dans l'équipe de Roderic Guigó à Barcelone sur le thème de l'épissage alternatif chez l'homme (dans le cadre du projet ENCODE) mais continue à collaborer avec nous sur les réseaux métaboliques.

De cette thèse, est issu également un article de revue sur l'analyse structurelle des réseaux métaboliques (acquisition des données, modélisation, analyse sur la base de mesures topologiques et évolution) qui a été récemment soumis.

Une autre thèse, de Paulo G. S. Fonseca, sera soutenue le 26 Février prochain sur le sujet de l'inférence de modules dans les réseaux génétiques en prennant en compte trois types d'informations simultanément : partage de motifs en séquence, données d'expression et évolution (conservation). Un premier article issu de cette thèse vient d'être soumis, et deux autres sont en préparation.

Enfin, le travail sur les réseaux bayésiens en pharmacogénomique se poursuit avec des expériences *in silico* qui ont été effectuées utilisant la stratégie gloutonne de recherche associée au score de Dirichlet bayésien sur des données simulées dans un premier temps, puis sur des données publiées. Ce dernier est constitué de 72 échantillons (puces Affymetrix) de moelle osseuse (24) et sang périphérique (48) de patients atteints d'une leucémie dont le type (Lymphoïde ou Myéloïde) est connu. Dans un premier temps, la possibilité de gérer plusieurs milliers de variables sur le jeu d'entraînement (38 échantillons) a été testée. Le résultat renvoie à un problème de dimension qui a été

abordé pour le moment par deux types d'approches qui sont :
- tester le pouvoir discriminant de la méthode : nous avons ainsi réappris la structure mais en réduisant le nombre de variables aux 50 dont l'expression est corrélée au maximum avec le type de leucémie. On a pu par ce critère obtenir des résultats satisfaisants dans nos prédictions : une erreur de plus seulement est commise par rapport à une méthode antérieure, mais la notre a l'avantage de fournir des probabilités, permettant ainsi à un expert (en l'occurrence un médecin) de prendre des décisions plus avisées qu'avec de simples « oui ou non »;
- utiliser des connaissances biologiques comme *a priori*. Cette deuxième stratégie est en cours d'investigation.

F. Etat financier et ressources humaines (optionnel)

Bref descriptif de l'état de consommation des crédits

	Crédits consommés (en %)	Commentaire éventuel
Main d'œuvre (tous statuts confondus)	54%	Crédits non encore consommés mais déjà engagés : 85%
Equipement	36%	
Mission	37%	Crédits non encore consommés mais déjà engagés : 43%
Fonctionnement/prestations	100%	

Bilan des CDD cumulés depuis le début du projet

	nombre de personnes employées en CDD sur le projet et financées par l'ANR	
	nombre	personnes×mois cumulés sur tous les partenaires depuis le début du projet
Doctorants		
Post-doctorants	1	7 mois
Ingénieurs en CDD	4	35 mois à 100% + 10 mois à 50%
Stagiaires		
Autres		

G. Commentaires libres

Commentaire général à l'appréciation du coordinateur, sur l'état d'avancement du projet, les interactions entre les différents partenaires…

Le projet progresse de manière satisfaisante aussi bien au niveau du génome que des réseaux (métaboliques, génétiques). Plusieurs publications se trouvent actuellement en préparation et devraient être soumises dans les prochains mois. Tout en poursuivant ces travaux à chacun des deux niveaux de façon indépendante, nous abordons maintenant

des études à la fois théoriques (développement de méthodes combinatoires ou statistiques) et appliquées sur le rapport et l'interaction entre les deux niveaux. Plus particulièrement, nous allons commencer à aborder deux questions précises. La première porte sur l'impact de certains types de réarrangements génomiques, notamment la duplication complète, sur, initialement, le réseau métabolique. L'organisme qui sera choisi pour une telle étude sera celui de la paramécie avec lequel d'autres équipes du laboratoire travaillent également et dont l'annotation est disponible publiquement et est fiable. La deuxième question concerne quant à elle le lien entre organisation génomique et réseau, c'est-à-dire, entre la structuration d'un génome en régions relativement homogènes vis-à-vis de certaines caractéristiques (composition, présence de répétitions, régions de cassures suite à un réarrangement, densité en gènes et niveau de leur expression etc.) et les différents types de réseaux modélisant l'interaction entre les gènes ou autres éléments présents dans chacune de ces régions. Les organismes choisis varieront selon la disponibilité actuelle des données sur chacune de ces caractéristiques. Pour l'expression par exemple, nous avons commencé à travailler avec la levure. Toute méthode développée dans ce cadre devra et pourra bien évidemment être appliquée a d'autres organismes.

Facultatif : question(s) posée(s) à l'ANR…

Bien que le projet avance à un bon rythme conforme à ce que nous souhaitions, les diverses questions abordées soulèvent des problèmes nouveaux, soit méthodologiques, soit d'analyse, et nous aurions aimé ainsi pouvoir disposer de quelques mois de plus sur le projet. Comme nous avons par ailleurs été prudents sur nos dépenses, cela serait financièrement possible si nous pouvions disposer de l'argent qui nous reste non pas jusqu'à fin Novembre 2008 comme initialement prévu mais, idéalement, jusqu'à fin Mai 2009 si possible. Plus précisément, nous aimerions pouvoir prolonger les contrats de Emmanuel Prestat (pharmacogénomique) et de Paulo Fonseca (modules dans les réseaux génétiques) afin de leur permettre de mener jusqu'à terme le développement de logiciels complets et de valider ce développement et les analyses associées par des publications. Nous estimons que cela demande environ 15 mois encore. Nous aimerions savoir si ce prolongement de 6 mois est possible.