Agence Nationale de la Recherche ANR	**Fiche compte-rendu semestriel d'activité**	Date : 08/08/2008 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 prolongé jusqu'au 31 Mai 2009

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 Janvier 2008 - 15 Septembre 2008
Date de rédaction	30 Septembre 2008

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

Ce projet ne comporte 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	19	10	0	4	1	0	33

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

C. Lemaitre and **M.-F. Sagot**. A Small Trip in the Untranquil World of Genomes - A survey on the detection and analysis of genome rearrangement breakpoints. *Theoretical Computer Science*, 395:171-192, 2008 **(indiqué comme sous presse dans rapport précédent)**.

P. Peterlongo, N. Pisanti, F. Boyer, A. Pereira do Lago and **M.-F. Sagot**. Lossless filter for multiple repetitions, *J. Discrete Algorithms*, 6(2), 497-509, 2008, **(indiqué comme sous presse dans rapport précédent)**.

P. Peterlongo, J. Allali and **M.-F. Sagot**. Indexing Gapped-Factors using a tree, *Int. J. Found. Comput. Sci.*, 19:71-87, 2007, **(indiqué comme sous presse dans rapport précédent).**

N. Mugnier, **L. Guéguen**, 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*, 411:87-93, 2008 **(indiqué comme accepté dans rapport précédent).**

J. Allali, **M.-F. Sagot**. A multiple layer model to compare RNA secondary structures. *Software: Practice and Experience*, 38(8):775-792, 2008 **(indiqué sous presse dans rapport précédent).**

C. Lemaitre, E. Tannier, C. Gautier, M.-F. Sagot. Precise detection of rearrangement breakpoints in mammalian chromosomes. *BMC Bioinformatics*, 9:286-322, 2008.

V. Lacroix, L. Cottret, P. Thébault and **M.-F. Sagot**. An introduction to metabolic networks and their structural analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 5:360-368, 2008.

P. G. S. da Fonseca, C. Gautier, K. S. Guimarães and **M.-F. Sagot**. Efficient representation and P-value computation for high order Markov motifs. *Proceedings of the European Conference on Computation Biology (ECCB'08). Bioinformatics*, 24 :160-166, 2008.

V. Acuña, F. Chierichetti, **V. Lacroix**, A. Marchetti-Spaccamela, **M.-F. Sagot**, L. Stougie. Modes and cuts in metabolic networks: Complexity and algorithms. *BioSystems*, 2008, in press **(indiqué comme soumis dans rapport précédent).**

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. Electronic Notes in Discrete Mathematics,* vol. 30, pp. 243-248, 2008 **(indiqué comme sous presse dans rapport précédent)**.

R. Lenne, C. Solnon, T. Stützle, **E. Tannier**, M. Birattari. Reactive Stochastic Local Search Algorithms for the Genomic Median Problem. *EvoCop'08*, Lec*ture Notes in Computer Science*, 4972, 2008, 266-276 **(indiqué comme accepté dans le rapport précédent)**.

E. Tannier, C. Zheng and D. Sankoff. Multichromosomal genome median and halving problems. *Proceedings of WABI'08*, *Lecture Notes in BioInformatics*, vol. 5251, pages 1-13, 2008.

L. Cottret, P. V. Milreu, **V. Acuña**, F. V. Martinez, A. Marchetti Spaccamela, **M.-F. Sagot**, L. Stougie. Enumerating Precursor Sets of Target Metabolites in a Metabolic Network. *Proceedings of WABI'08*, *Lecture Notes in BioInformatics*, vol. 5251, pages 233-244, 2008.

S. Bérard, A. Château, C. Chauve, C. Paul, **E. Tannier**. Perfect DCJ rearrangements. In *Proceedings of RECOMB-CG Satellite Workshop, Lecture Notes in BioInformatics*, vol. 5267, pages 156-167, 2008.

Articles acceptés dans des conférences nationales à comité de lecture

F. Boyer, C. Chauve, **E. Tannier**. Prédiction de synténies dans le génome ancestral des amniotes. Jobim 2008, 2008.

Y.-P. Deniélou, F. Boyer, **M.-F. Sagot, A. Viari**. Recovering isofunctional genes: a synteny-based approach. Jobim 2008, 2008.

Séminaires présentés dans le cours du dernier semestre

L. Cottret. « Comparaison des réseaux métaboliques des organismes à génome réduit », Centre de Bioinformatique de Bordeaux, 29 Février 2008.

L. Cottret. « Functionality and comparison: two cases of using graph topology for investigating metabolic networks ». Université Fédérale de Mato Grosso do Sul, Campo Grande, Brésil, 18 Mars 2008.

L. Cottret. « Using graph topology to find precursors of target compounds in a metabolic network ». Rencontres Stic AmSud à Valparaiso, Chili, 28 Mars 2008.

C. Lemaitre. « A method to detect precisely rearrangement breakpoints in mammalian genomes ». Dynamics of genomes, workshop Valparaiso, Chili, 27-28 mars 2008.

F. Picard. « Assessing the exceptionality of Network motifs ». Workshop on Modeling of Genetic Regulatory and Metabolic Networks, Valparaíso Complex Systems Institute –ISCV, Valparaíso, Chile, ISCV, 28 Mars 2008.

F. Picard. « Assessing the exceptionality of Network motifs ». Séminaire du Laboratoire Bordelais de Recherche en Informatique (LABRI), Bordeaux, Mars 2008.

E. Tannier, « Prédictions du passé des chromosomes », Université Libre de Bruxelles, 21 Avril 2008.

E. Tannier, « Prédictions of ancestral chromosomes », MIEP 2008, Saint-Martin de Londres, 10 Mai 2008.

L. Cottret. « Comparaison de réseaux métaboliques ». Laboratoire Bordelais de Recherche en Informatique. 15 Mai 2008.

F. Picard. « Linear models for the joint analysis of multiple array-CGH profiles », Emerging statistical challenges in Genomes and Translational Research, Banff 2008, Canada, 1-6 Juin 2008.

L. Cottret. « Metabolic network reconstruction from genomic annotations ». Université de Glasgow. 5 Juin 2008.

M.-F. Sagot. « Extracting modules and motifs from networks ». Université de Glasgow. 5 Juin 2008.

E. Tannier, « Comparaison de l'organisation des génomes en blocs de synténie et réplicons », GTGC'08, 3 Juillet 2008.

E. Prestat. « Modeling pharmacogenomics data with Bayesian Networks ». 9ème Conférence Internationale sur la Science des Systèmes de Santé, 3 Septembre 2008. Cette présentation était accompagnée d'un poster.

Autres retombées :

Nature	Commentaire
Brevets nationaux	-
Brevets internationaux	-
Autres	Logiciels ou bases disponibles publiquement : 11

D. Eventuellement, résultat marquant du semestre écoulé

Les résultats les plus marquant du semestre écoulé concernent : 1. la mise au point de la méthode de détection des points de cassure plus précise que toutes celles existant actuellement et son application à l'étude de l'organisation des génomes de mammifères ; 2. une analyse comparative des réseaux métaboliques de bactéries libres et symbiotiques, et plus particulièrement des endocytobiotes, qui n'avait jamais encore été réalisée de façon aussi large et systématique ; et 3. une application de la méthode de calcul de tous les scénarios d'inversions génomiques à l'étude de certaines classes de symbiotes. Ces trois travaux (partie théorique et applications biologiques) seront, chacun, défendus dans le cadre d'une thèse d'ici Novembre 2008.

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)

Inférence et mise en évidence de régularités au niveau génomique

Principaux résultats et prévisions :

L'étude de limpact de la recombinaison sur l'apparition et l'évolution des structures génomiques, notamment les isochores (longues régions d'ADN (>>300 kb), ayant une composition relativement homogène en bases (taux de Guanine et Cytosine homogène) et des limites également bien marquées), a bien avancé. Plusieurs génomes de mammifères ont ainsi été étudiés. Des corrélations ont été mises en évidence entre la richesse en GC des isochores et certaines propriétés biologiques telles la longueur des introns, leur nombre, la compacité des gènes, la densité en gènes, la répartition des éléments répétés, etc. Nous n'avons pas retrouvé ces corrélations chez certains poissons (medaka et danio) et certaines corrélations y apparaissaient même inversées. Six séquences de medaka de Japon ont pu être récupérées et comparées. Grâce à cette comparaison, nous sommes en train de démontrer que le phénomène de la conversion génique biaisé (BGC en anglais) ne serait pas présent chez cette espèce. Un modèle mathématique pour l'impact de la BGC dans une population de taille infinie a également été construit et des simulations de cet impact ont été réalisées dans une population de taille finie. Les conclusions de ces premières modélisations montrent que le phénomène de la BGC serait capable de maintenir dans une population des allèles récessives létales.

Un cadre méthodologique a été mis au point qui permet d'estimer l'organisation des chromosomes des espèces disparues dont on connaît des descendants, c'est-à-dire de reconstruire des génomes ancestraux de vertébrés. Ce cadre a permis entre autres de proposer des génomes ancestraux pour les mammifères boréoeuthériens et pour les amniotes, et permet de prendre en compte les génomes dupliqués malgré la difficulté de retrouver des synténies dans ce cas.

La recherche de génomes ancestraux motive le problème du calcul de médianes qui consiste, à partir de génomes de trois espèces ou plus, de reconstituer l'organisation des chromosomes de leurs ancêtres en même temps que les évènements évolutifs qui expliquent les différences entre les génomes. Nous avons progressé dans l'étude algorithmique des médianes en établissant la complexité théorique de plusieurs variantes de problèmes de médianes pour les cas où les génomes sont multi chromosomiques (comme, par exemple, tous les vertébrés), et éventuellement dupliqués.

Les calculs de médianes génomiques font une utilisation intense des calculs de distances par réarrangements génomiques. La complexité du calcul de distance se ressent donc sur l'efficacité des méthodes de calcul des configurations ancestrales. Nous avons progressé dans le calcul des scénarios de réarrangements qui préservent les groupes de gènes co-localisés dans deux génomes pour une variante de la distance d'inversion appelée DCJ.

Une méthode de détection des points de cassure dans les génomes de mammifères a été publiée depuis le dernier rapport. Grâce à cette méthode, nous disposons de points de cassure plus précis. Cela a permis d'étudier les corrélations entre les positions des points de cassure sur le génome humain avec d'autres structures génomiques, telles que l'organisation en domaines de réplication, les isochores et la densité en gènes.

Cette méthode a également été utilisée dans le cadre d'un travail sur l'évolution des chromosomes X et Y chez l'homme. Elle a permis d'identifier des duplications inversées aux bornes d'inversions sur le chromosome Y, renforçant ainsi l'hypothèse que le chromosome Y a divergé du chromosome X par des inversions successives responsables de l'arrêt de la recombinaison entre ces deux chromosomes. Ce résultat fait partie d'un travail plus vaste sur l'analyse des scénarios de réarrangements entre ces deux chromosomes, grâce à une méthode de tri par inversions développée dans l'équipe.

Cette méthode qui permet, pour des distances d'inversions relativement petites (jusqu'à 20 environ), de calculer tous les scénarios d'inversions et de les représenter de façon compacte, a été étendue afin de prendre en compte des contraintes biologiques supplémentaires importantes lorsque l'on connaît les génomes qui sont ancestraux ou lorsque les organismes étudiés sont des bactéries. Dans le premier cas, la préservation des groupes de gènes co-localisés entre deux génomes prend en compte non pas seulement ces deux génomes (dont l'un est ancêtre de l'autre) mais également tous les ancêtres intermédiaires du génome actuel, tandis que dans le second cas, seules les inversions symétriques par rapport au début de la réplication sont considérées. La méthode est actuellement appliquée à l'étude de l'évolution des Rickettsies et des Mycobactéries.

Un système automatique pour l'analyse des données de séquençage de petits ARNs a été développé qui permet d'identifier la localisation la plus probable dans le génome de petites séquences supposées être celles de microARNs et identifiées lors du séquençage, et aussi l'établissement de plusieurs critères de filtrage pour isoler parmi ces positions identifiées celles qui sont de bons candidats à être des précurseurs de microARNs. Ce système inclut un outil de visualisation web des résultats filtrés selon l'ensemble de critères choisis. En utilisant ce système sur des données obtenues du laboratoire d'Eric Westhof (IBMC/CNRS, Strasbourg) concernant le moustique *Anopheles gambiae*, nous avons pu identifier 36 bons candidats (ceux qui passaient tous les filtres) dont 3 sont des microARNs déjà connus.

Inférence et mise en évidence de régularités au niveau des réseaux

Principaux résultats et prévisions :

La méthode de détection des précurseurs abordée lors du dernier rapport a fait l'objet d'un article accepté. Ce projet a été le fruit d'une collaboration internationale. Cette méthode est la première exacte pour trouver tous les ensembles de précurseurs d'un ensemble de métabolites cibles. De plus, pour la première fois, une attention toute particulière a été donnée au traitement des cycles. L'application au réseau de l'endocytobiote *Buchnera* est en cours. Les précurseurs des acides aminés, métabolites centraux dans la fonction symbiotique de la bactérie, ont été calculés. Mis à part des résultats connus dans la littérature, de nouveaux ont été identifiés, beaucoup plus difficiles à trouver en réalisant une étude gène à gène ou par comparaison avec les voies de synthèse classiquement définies. Nous allons étendre cette analyse également aux précurseurs des cofacteurs dont les voies de synthèse sont encore mal établies chez *Buchnera*.

Dans le but de définir les spécificités du réseau métabolique sélectionnées au cours de l'évolution des différents styles de vie, nous sommes en train de réaliser la comparaison topologique de 35 réseaux métaboliques de bactéries au style de vie différents (parasites / mutualistes intracellulaires, libres, fixatrices d'azote). Notre première idée a été de comparer les intersections des ensembles de composés et de réactions des différents organismes. De façon surprenante, dans le réseau des petites molécules, seuls 31 composés et une réaction sont communs à tous les organismes. Même en prenant en compte les défauts d'annotation, toujours présents, ces nombres demeurent faibles, ce qui pourrait remettre en question la notion de métabolisme minimal, spécialement lorsqu'on étudie des bactéries au métabolisme réduit de par leur nature de symbiote obligatoire.La seconde idée, en cours, est de déterminer si les propriétés dans le graphe des noeuds eux-mêmes peuvent être reliés aux modes de vie des bactéries. Pour cela, nous mesurons le degré et la centralité des noeuds dans les graphes de réaction et dans

les graphes de composés et déterminons si certains composés ont un « comportement » différent selon le mode de vie considéré.

Les principaux résultats quant à l'identification de modules dans les réseaux de régulation transcriptionnelle concernent le développement de méthodes informatiques pour la caractérisation des ensembles de gènes co-régulés selon trois critères de modularité, à savoir, le partage de motifs de séquence communs dans leurs régions régulatrices, l'existence de profils de d'expression cohérents sous un ensemble de conditions expérimentales, et la conservation évolutive. Ces critères reposent sur trois postulats biologiques: le premier suppose que des gènes co-régulés partagent des motifs correspondant à des sites de liaison de protéines régulatrices communes, le deuxième considère que les gènes co-régules agissent ensemble et ont donc besoin d'être exprimés de façon coordonnée, et le troisième prétend que les modules de co-régulation doivent être soumis à une forte pression sélective puisqu'ils sont sensés être responsables de fonctions biologiques importantes. Nous proposons des mesures objectives de modularité basés sur chacun de ces critères, ainsi que des algorithmes capables de les calculer à partir de données expérimentales de nature diverse disponibles dans des bases publiques, telles que les données de séquence, des données d'expression, et des données phylogénétiques.

Organisme uni-cellulaire séparant les fonctions somatiques et germinales dans des noyaux différents, la paramécie présente à la fois des caractéristiques simples et complexes qui en font un excellent sujet d'études. D'une part, suite au séquençage de son génome, trois duplications complètes du génome ont été détectées chez la paramécie. D'autre part, une annotation automatique des gènes basée sur l'homologie de leur séquence par rapport à des enzymes connues a permis la reconstitution d'une estimation du réseau métabolique de la paramécie.

Nous étudions l'influence de la structuration des gènes de la paramécie en réseau métabolique sur la rétention de dupliqués. En particulier, il semble que le taux de rétention en dupliqués du voisinage dans le réseau métabolique influence la rétention de dupliqués chez les enzymes catalysant une réaction donnée. En effet, les réactions catalysées par des paires d'enzymes issues des duplications complètes du génome ont tendance à former des aggrégats dans le réseau métabolique de la paramécie.

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)	(voir engagés dans colonne de droite)	Crédits non encore consommés mais déjà engagés : 93%
Equipement	-	Erreur assignation lors précédent rapport, reporté sur fonctionnement
Mission	91%	
Fonctionnement/prestations	78%	

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	51 mois à 100% + 19 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 continue de progresser de manière satisfaisante aussi bien au niveau du génome que des réseaux (métaboliques, génétiques), avec également des publications à un rythme régulier. Surtout, les travaux théoriques et le développement des divers algorithmes auquel ces travaux ont conduit commencent à produire des résultats biologiques également intéressants, et posent de multiples nouvelles questions.

Deux plus précises avaient déjà été mentionnées lors du dernier rapport. Elles concernaient l'impact de certains types de réarrangements génomiques, notamment la duplication complète, sur le réseau métabolique, et 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. Une étude de la première question est bien avancée en utilisant comme modèle le génome de la paramécie. La deuxième question se projette progressivement vers une vision en 3D des chromosomes à l'intérieur des noyaux dans une collaboration avec Ana Pombo du MRC Imperial College de Londres.

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