Rapport final du projet n° ANR-05-NT05-3_45205

A. Identification

Programme – année	2005
Projet (acronyme)	REGLIS
Titre complet du projet	De la molécule à la cellule : développement, confrontation et intégration de modèles formels et de méthodes d'analyse
Coordinateur du projet (société/organisme)	Marie-France Sagot Équipe Baobab – Laboratoire de Biométrie et Biologie Évolutive, Université Lyon 1
Période du projet (date début – date fin)	15 décembre 2005 – 31 Mai 2009
Rapport confidentiel (OUI/NON)	Non
Date de fin de confidentialité	

Rédacteur de ce rapport

Civilité, prénom, nom	Mme Marie-France Sagot
Téléphone	04 72 44 82 38
Adresse électronique	Marie-France.Sagot@inria.fr
Date de rédaction	14/04/2009

B. Pour les projets partenariaux, liste des livrables et affectation éventuelle à chaque partenaire

Ce projet ne comportait qu'un seul partenaire.

C. Rapport factuel

C.1 **Tableau de résultats**

Nombres de publications

	International		France		Actions de diffusion		
	Articles acceptés dans revues à comité de lecture	Communications Internationales à comité de lecture	Articles France	Communications France	Articles vulgarisation	Conférences vulgarisation	Autres Séminaires invités
monopartenaire	27	13	0	4	1	0	34

Autres retombées

Nature	Commentaire
Brevets nationaux	
Brevets internationaux	
Autres (préciser en C.4)	Logiciels ou bases disponibles publiquement : 11

C.2 **Tableau de personnels**

	nombre de personnes employées en CDD sur le projet et financées par l'ANR	
	nombre	Mois-homme cumulés sur tous les partenaires depuis le début du projet
Doctorants		
Post-docs	2	7+ 1,15 mois
Ingénieur en CDD	5	30+31+2+1,5+24 mois
Stagiaires		
Autres		

Nom, prénom, qualification	Devenir des personnes employées en CDD sur le projet		
	emploi suite au projet		en recherche d'emploi
	chez les partenaires	ailleurs	
Ludovic Cottret, ingénieur, INRA Toulouse		CDD	
P.G.S. Fonseca, chercheur associé Instituto Superior Técnico, Lisbonne, Portugal		CDI	
Vincent Lacroix, postdoc CRG Barcelone puis Univ. Lyon 1 ; à partir du 01/10/2009, MCU Univ. Lyon 1		CDD CDI à partir 01/10/09	
Claire Lemaitre, postdoc Univ. Bordeaux 2		CDD	
Emmanuel Prestat, ingénieur, Centrale Lyon		CDD	
Patricia Thébault, MCU Univ. Bordeaux 2		CDI	
Amélie Véron, postdoc Univ. Lyon 1		CDD sur	

C.3 **Liste des publications et communications**

a. Publications

Revues internationales

V. Acuña, F. Chierichetti, **V. Lacroix**, A. Marchetti-Spaccamela, **M.-F. Sagot**, and L. Stougie. Modes and cuts in metabolic networks : Complexity and algorithms. Biosystems, 95:51-60, 2008.

S. S. Adi, **M. D. V. Braga**, C. Fernandes, C. Ferreira, F. Martinez, **M.-F. Sagot**, M. A. Stefanes, C. Tjandraatmadja, and Y. Wakabayashi. Repetition-free lcs with few reversals. In LAGOS, volume 30, pages 243–248. Electronic Notes in Discrete Mathematics, 2008.

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., 1:29, 2007.

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. IEEE/ACM Trans. Comput. Biology Bioinform., 5(3):348-356, 2008.

A. M. Carvalho, A. T. Freitas, A. L. Oliveira, and **M.-F. Sagot**. An efficient algorithm for the identification of structured motifs in dna promoter sequences. IEEE/ACM Trans. Comput. Biology Bioinform., 3(2) :126–140, 2006.

C. Chauve and **E. Tannier**. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. Plos Comp. Biology, 4:11, 2008.

M. Deloger, F.M. Cavalli, E. Lerat, C. Biémont, **M.-F. Sagot**, C. Vieira. Identification of expressed transposable element insertions in the sequenced genome of Drosophila melanogaster. Gene, 2009, in press.

Y. Diekmann, **M.-F. Sagot**, and **E. Tannier**. Evolution under reversals: Parsimony and conservation of common intervals. IEEE/ACM Trans. Comput. Biology Bioinform., 4(2):301–309, 2007.

P.G.S. Fonseca, **C. Gautier**, K. Guimarães, and **M.-F. Sagot**. Efficient representation and p-value computation for high order markov motifs. Bioinformatics, 24:160–166, 2008.

V. Lacroix, **L. Cottret**, **P. Thébault**, and **M.-F. Sagot**. An introduction to metabolic networks and their structural analysis. IEEE/ACM Trans. Comput. Biology Bioinform., 5(4):594-617, 2008.

V. Lacroix, C. G. Fernandes, and **M.-F. Sagot**. Motif search in graphs : Application to metabolic networks. IEEE/ACM Trans. Comput. Biology Bioinform., 3(4) :360–368, 2006.

C. Lemaitre, **M. Braga**, **M.-F. Sagot**, **C. Gautier**, **E. Tannier**, G. Marais. Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes. Genome Biology and Evolution, 2009, in press.

C. Lemaitre and **M.-F. Sagot**. A small trip in the untranquil world of genomes. Theor. Comp. Sci., 395 :171–192, 2008.

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

C. Melodelima C, C. Gautier, The GC-heterogeneity of teleost fishes. BMC Genomics, 9:632, 2008.

C. Melodelima C, C. Gautier, and D. Piau, A markovian approach for the prediction of mouse isochores. J. Math. Biol., 55:353-364, 2007.

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, 19:19-36, 2008.

C. Melodelima, L. Guéguen, D. Piau**, C. Gautier**, A computational prediction of isochores based on hidden Markov models. Gene, 385:41-49, 2006.

N.D. Mendes, A.T. Freitas, **M.-F. Sagot**. Current tools for the identification of miRNA genes and their targets. Nucleic Acids Res, 2009, in press.

N. Pisanti, M. Crochemore, R. Grossi, and **M.-F. Sagot**. Bases of motifs for generating repeated patterns with wild cards. IEEE/ACM Trans. Comput. Biology Bioinform., 2(1):40–50, 2005.

L. Palmeira, **L. Guéguen**, and J. R. Lobry. UV-Targeted Dinucleotidess Are Not Depleted in Light-Exposed Prokaryotic Genomes. Molecular Biology and Evolution, 23(11):2214–2219, 2006.

P. Peterlongo, J. Allali and **M.-F. Sagot**. Indexing Gapped-Factors using a tree, Int. J. Found. Comput. Sci., 19:71-87, 2007.

P. Peterlongo, F. Boyer, A. P. do Lago, N. Pisanti, and **M.-F. Sagot**. Lossless filter for multiple repetitions. Journal of Discrete Algorithms, 6:497–509, 2008.

P. Peterlongo, G. A. T. Sakomoto, A. P. do Lago, N. Pisanti, and **M.-F. Sagot**. Lossless filter for multiple repeats with bounded edit distance. BMC Algorithms for Molecular Biology, 4:3, 2009.

C. Rezvoy, D. Charif, **L. Guéguen**, G. A. Marais, MareyMap: a R-based tool with graphical interface for estimating recombination rates. Bioinformatics, 23:2188-2189,2007.

S. Schbath, **V. Lacroix**, and **M.-F. Sagot**. Assessing the exceptionality of coloured motifs in networks. EURASIP J Bioinform Syst Biol., 2009, in press.

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

Conférences internationales avec actes (indexés dans Web of Science)

R. Bourqui, D. Auber, **V. Lacroix**, and F. Jourdan. Metabolic network visualization using a constraint planar graph drawing. In 10th conf. on Information Visualization, pages 489-496, 2006.

M. D. V. Braga, **M.-F. Sagot**, C. Scornavacca, and **E. Tannier**. The solution space of sorting by reversals. In ISBRA, volume 4463 of Lecture Notes in Computer Science, pages 293–304. Springer, 2007.

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

A. M. Carvalho, A. L. Oliveira, and **M.-F. Sagot**. Efficient learning of bayesian network classifiers. In Australian Conference on Artificial Intelligence, volume 4830 of Lecture Notes in Computer Science, pages 16–25. Springer, 2007.

L. Cottret, P. V. Milreu, **V. Acuña**, A. Marchetti-Spaccamela, F. V. Martinez, **M.-F. Sagot**, and L. Stougie. Enumerating precursor sets of target metabolites in a metabolic network. In WABI, volume 5251 of Lecture Notes in BioInformatics, subseries of Lecture Notes in Computer Science, pages 233–244. Springer, 2008.

Y.-P. Deniélou, F. Boyer, **M.-F. Sagot**, **A. Viari**. Multiple Alignment of Biological Networks: A Flexible Approach. Combinatorial Pattern Matching, Lecture Notes in Computer Science, 2009, in press.

V. Lacroix, C. G. Fernandes, and **M.-F. Sagot**. Reaction motifs in metabolic networks. In WABI, volume 3692 of Lecture Notes in Computer Science, pages 178–191. Springer, 2005.

R. Lenne, C. Solnon, T. Stützle, **E. Tannier**, and M. Birattari. Reactive stochastic local search algorithms for the genomic median problem. In EvoCOP, volume 4972 of Lecture Notes in Computer Science, pages 266–276. Springer, 2008.

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.

P. Peterlongo, N. Pisanti, F. Boyer, and **M.-F. Sagot**. Lossless filter for finding long multiple approximate repetitions using a new data structure, the bi-factor array. In SPIRE, volume 3772 of Lecture Notes in Computer Science, pages 179–190. Springer, 2005.

N. Pisanti, A. M. Carvalho, L. Marsan, and **M.-F. Sagot**. Risotto : Fast extraction of motifs with mismatches. In LATIN, volume 3887 of Lecture Notes in Computer Science, pages 757– 768. Springer, 2006.

M.-F. Sagot and **E. Tannier**. Perfect sorting by reversals. In COCOON, volume 3595 of

Lecture Notes in Computer Science, pages 42–51. Springer, 2005.

E. Tannier, C. Zheng, and D. Sankoff. Multichromosomal genome median and halving problems. In WABI, volume 5251 of Lecture Notes in BioInformatics, subseries of Lecture Notes in Computer Science, pages 1–13. Springer, 2008.

Chapitres de livre

E. Tannier, Sorting signed permutations by reversals (sorting sequence), The Encyclopedia of Algorithms, Springer, pages 860-863, 2008.

Conférences nationales avec actes

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

L. Cottret, **V. Acuña**, H. Charles, and **M.-F. Sagot**. Recherche de précurseurs dans un réseau métabolique. In J.-P. Comet, F. Quessette, and S. Vial, editors, RIAMS'07, 2007.

J.-J. Daudin, **V. Lacroix**, F. Picard, S. Robin, and **M.-F. Sagot**. Uncovering structure in biological networks. In Jobim 2006, Bordeaux, France, 2006.

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

Soumises

V. Lacroix, **L. Cottret**, O. Rogier, C. G. Fernandes, F. Jourdan, and **M.-F. Sagot**. Motus: a software and a webserver for the search and enumeration of node-labelled connected subgraphs in biological networks, 2008. submitted.

C. Lemaitre, L. Zaghloul, **M.-F. Sagot**, **C. Gautier**, A. Arnéodo, **E. Tannier**, B. Audit. Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation, 2009. submitted.

b. Communications

L. Palmeira. Séminaire invité au laboratoire d'Alain Arnéodo (Novembre 2006 – Lyon, France) : "*Theoretical approaches for the study of dinucleotide content in genomes*".

L. Palmeira. Séminaire invité au laboratoire d'Axel Meyer (Octobre 2006 – Constance, Allemagne) : « *Dinucleotides in evolution: analysis and modelling* » .

L. Palmeira. Séminaire invité au laboratoire d'Arndt von Haeseler (Novembre 2006 – Vienne, Autriche) : « *Neighboring-site dependencies in genome evolution: analysis and modelling* ».

L. Palmeira. Séminaire invité au laboratoire de Michael Laessig (Décembre 2006 – Cologne, Allemagne) : « *Theoretical approaches for the study of dinucleotide content in genomes* ».

E. Tannier. Séminaire invité en vue d'une collaboration avec Gianpaolo Oriolo, de l'Université de Rome Tor-Vergata (Septembre 2006) : « *Conservation and Rearrangements in Genomes* ».

C. Gautier and E. Prestat. Some bioinformatics tools for studying biodiversity (November 2006, Universidade Federal do Mato Grosso do Sul, Brazil).

C. Gautier and P. Gustavo. Séminaire invité à l'Université de São Paulo (Novembre 2006, Brésil) : « *Analysing gene expression* » (C. Gautier et P. Gustavo).

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

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

L. Palmeira. Models of DNA evolution with neighbor-dependent substitutions, *Otto Warburg Summer School and Workshop 2006* (September 2006 - Berlin, Germany).

L. Palmeira. Theoretical approaches for the study of dinucleotide content in genomes. *Theoretical Approaches for the Genome* (November 2006 - Annecy, France).

M.-F. Sagot. Conférence invitée, Federation of the European Societies in PLant Biology (FESPB) 2006. « Computational biology: Negative thoughts and (maybe) some positive actions », Palais des Congrès, Lyon, July 19, 2006.

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

E. Tannier, « Reconstruction of mammalian ancestral genomes: the two pasimonies », séminaire invité du laboratoire G-SCOP, 8 novembre 2007, Grenoble.

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

M.-F. Sagot. Séminaire invité au Department of Computer Science of the Technical University of Eindhoven, « A chatting tour through computational biology », Technical University of Eindhoven, the Netherlands, March 19, 2007

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

E. Tannier, « Reconstruction de génomes ancestraux: les deux parcimonies », séminaire invité du laboratoire LIAFA, 22 janvier 2008, Paris.

M.-F. Sagot. Séminaire invité au Department of Computer Science of the University of Rome "La Sapienza", « Open combinatorial / graph problems in computational biology », Rome, Italy, February 4, 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- March 27th - 28th , 2008.

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

E. Tannier, « Prédictions du passé des chromosomes », séminaire invité université libre de Bruxelles, 21 avril 2008.

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

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

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

L. Cottret. séminaire invité « 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.

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

C. Gautier. « Selection and randomness in evolution or selection of randomness in evolution? ». Conférence invitée à JOBIM, Nantes 2009.

C.4 **Liste des éléments de valorisation**

a. Logiciels et sites internet rendus publiques ou enrichis depuis Juin 2006

- BaobabLuna (Marília Braga) : implantation d'un algorithme qui permet l'exploration de l'espace des solutions du tri par inversions, avec un gain de temps important par rapport aux méthodes existantes.
  http://pbil.univ-lyon1.fr/software/luna/

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

- 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/

- Tuiuiu (Pierre Peterlongo avec A.P. do Lago et G. Sacomoto) : filtre de séquences qui peut être utilisé pour trouver de longues répétitions dans les séquences nucléotidiques, permet insertions et délétions
  http://mobyle.genouest.org/cgi-bin/MobylePortal/portal.py?form=tuiuiu
  (prototype)

- Fonction supplémentaire dans SeqinR version 1.0-6 (Leonor Palmeira avec Delphine Charif, Jean Lobry, Anamaria Necsulea) : Incorporation de statistiques non paramétriques pour l'étude de la sur- et sous-représentation de dinucléotides dans la bibliothèque SeqinR du logiciel R.
  http://cran.univ-lyon1.fr/web/packages/seqinr/index.html

- 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

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

- PSbR (Yoan Diekmann en collaboration avec Eric Tannier et M.-F. Sagot) : implantation d'un algorithme pour trier des permutations (de marqueurs) signée qui fournit une solution optimale qui respecte les intervalles communs si une telle solution existe.
  http://biomserv.univ-lyon1.fr/~tannier/PSbR/

- 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/

- Implémentation d'un algorithme pour énumérer tous les ensembles minimaux de précurseurs (Ludovic Cottret avec Paulo Milreu) : Sera rendu disponible prochainement via une interface web.

- Implémentation d'un algorithme pour obtenir tout les motifs de un taille pré-fixée dans une réseau métabolique (A. Vellozo) : développement et implémentation d'une structure de données pour stocker les motifs et sous-graphes connectés d'un réseau métabolique en utilisant moins de mémoire et de façon plus rapide que les autres programmes connus. Sera plus tard intégré à Motus.

b. Nouveaux partenariats et projets

Au delà des résultats spécifiques qui avaient été énoncés dans le texte soumis et qui ont pu en grande partie être traités de manière satisfaisante, ce projet avait indiqué un grand objectif général, qui était de renforcer et étendre un réseau de collaborations qui avait été en partie initié dans le cadre d'un projet précédent. Cet objectif a été globalement bien atteint. Le présent projet a ainsi déjà donné naissance à d'autres nouveaux, dont nous ne mentionnons brièvement que ceux déjà validés officiellement. Ces autres nouveaux projets portent sur des sujets qui sont la continuation de ceux de REGLIS notamment sur les ARNs (Projet ANR Blanc Brasero coordonné par Alain Denise du LRI de l'Université d'Orsay et impliquant aussi le Labri à Bordeaux, l'IGM à l'Université de Marne-la-Vallée et Sequoia de l'INRIA de Lille), les motifs dans les séquences (Projet Bioinformatique Israel-France et divers projets avec le Portugal) et les réseaux (Projet ANR Blanc NeMo coordonné par Stéphane Robin de l'AgroParisTech impliquant aussi l'INRA de Jouy et le Laboratoire Statistique et Génome de l'Université d'évry), ou de vastes extensions de ceux-ci, notamment à l'étude de l'organisation spatiale des chromosomes au sein des cellules et au rôle joué par cette organisation dans la régulation et la dynamique des génomes (Projet ARC-INRIA avec le MRC-Imperial College à Londres, Mistis de l'INRIA à Grenoble, l´INRA de Jouy et AgroParisTech, INRIA-LBBE), ainsi qu'à l'étude au niveau évolutif et réseaux de la relation hôte-parasite (Projet ANR-BBSRC MetNet4SysBio coordonné, côté Grande Bretagne originalement par Angela Douglas de l'Université de York et maintenant par Gavin Thomas, ainsi que côté français par Hubert Charles du Laboratoire BF2I de l'INSA de Lyon, Projet ANR Blanc Miri juste accepté avec BAOBAB-BAMBOO comme seul partenaire, et Équipe Associée Simbiosi INRIA avec Alberto Marchetti-Spaccamela de l'Université de Rome et Leen Stougie actuellement à l'Université Libre d'Amsterdam et au CWI, « Centrum voor Wiskunde en Informatica »).

D. Rapport scientifique

D.1 **Résumé consolidé**

Objectifs du projet

Ce projet, largement exploratoire, avait pour but, à travers une approche comparative à partir de données disponibles publiquement, d'arriver à une meilleure compréhension du réseau complexe d'interactions spatiales et temporelles entre les divers éléments composant le vivant (gènes, métabolites, etc.) et de ces derniers, individuellement ou collectivement, avec l'environnement. Une telle compréhension obligeait à adopter un point de vue à la fois local et global, statique et dynamique d'un organisme, allant du niveau moléculaire (génome) jusqu'à une vision générale du fonctionnement d'une cellule (réseaux biologiques). Plus spécifiquement, nous souhaitions nous intéresser à trois questions biologiques très générales : 1. existe-t-il des régularités, structurales ou fonctionnelles, dans la diversité qui est observée, qui pourraient représenter des indices d'une organisation profonde du vivant; 2. pouvons-nous identifier ces régularités de façon systématique et arriver ainsi à dégager un ordre dans le réseau complexe des interactions détectées; enfin, 3. comment ce réseau s'est-il mis en place au cours de l'évolution, pour accomplir quelles fonctions ? Les résultats attendus concernaient : 1. l'obtention de meilleurs formalismes de modélisation mathématique et algorithmes d'analyse; et 2. des éléments de réponse à la fois à des questions biologiques spécifiques (tests d'hypothèses) et plus générales (exploration systématique des données disponibles).

Verrous ou points durs

Les verrous principaux du projet étaient liés à sa nature très exploratoire, initialement sur des questions biologiques extrêmement vastes et relativement vagues - existe-t-il des régularités, structurales ou fonctionnelles, dans la diversité qui est observée, qui pourraient représenter des indices d'une organisation profonde du vivant ? – dans un domaine qui en outre est fortement compétitif depuis quelques années. Les défis étaient donc de réussir à apporter des éléments de réponse à une question aussi large, en fait de dégager des voies de réponse ou d'exploration plus précises, tout en contribuant au principal objectif méthodologique qui consistait à obtenir de meilleurs formalismes de modélisation mathématique et algorithmes d'analyse.

Résultats majeurs

Il est clair que nous ne pouvions arriver aussi rapidement à une réponse nette et claire aux trois questions biologiques posées. Par contre, nous pensons avoir clairement rempli l'objectif de dégager de meilleurs formalismes de modélisation mathématique et algorithmes d'analyse, que ce soit au niveau génomique (lien entre recombinaison et organisation chromosomique, blocs de synténie et régions de cassure, génomes ancestraux, scénarios d'inversions) qu'à celui des réseaux biologiques (motifs, flux et ensembles minimaux de précurseurs, modules de régulation). À travers des tests d'hypothèse et une exploration systématique des données, nous avons aussi dégagé des régularités que nous sommes maintenant en mesure d'analyser selon différentes voies plus spécifiques.

Abstract

Objectives of the projet

This largely exploratory project had for objective, by using a comparative approach from publicly available data, to arrive at a better understanding ot the colpex network of spatial and temporal interactions between the different elements of a living system (genes, metabolites, etc.) and of the latter, individually or collectively, with the environment. Such an understanding obliged to adopt a vantage point at the same time local and global, static and dynamic of an organism, that goes from the molecular level (genome) to a general vision of the functioning of a celle (biological networks). More specifically, we wished to address three very general biological questions: 1. are there regularities, structural or functional, in the diversity of what is observed, that could represent indices of a deep organisation of living systems; 2. could we identify such regularities in a systematic fashion and thus uncover an order in the complex network of the detected interactions; 3. how has this network been set up in the course of evolution, to accomplish which function? The expected results concerned: 1. obtaining better formalisms of mathematical modelling and analytical algorithms; and 2. elements of response to biological questions both specific (hypothesis tests) and general (systematic exploration of the available data).

Limiting steps and hard issues

The main limiting steps of the project were related to its largely exploratory nature, initially around biological questions that were extremely vast and relatively vague – are there regularities, structural and functional in the observed diversity, that could represent indices of a deeper orgamisation of living systems? – in a field that has been strongly competitive since a few years. The main challenges were therefore to be able to provide some elements of response to a question so large, in fact to be able to arrive at more precise elements of response or exploration, while contributing to the main methodological objective that was to obtain better formalisms of mathematical modelling and analytical algorithms.

Major results

It is clear that we could not arrive as quickly at a clear response to the three biological questions asked. On the other hand, we believe that we clearly reached the objective of producing better formalisms of mathematical modelling and analytical algorithms, both at the genomic (link between recombination and chromosomal organisation, synteny blocks and break regions, ancestral genomes, inversion scenarii) and at the network level (motifs, fluxes and minimal precursor sets, regulation modules). Through hypotesis tests and a systematic exploration of the data, we also uncovered some regularities that we are now able to analyse in a more specific manner.

D.2 **Mémoire**

Introduction

Ce projet, largement exploratoire, avait pour but, à travers une approche comparative à partir de données disponibles publiquement, d'arriver à une meilleure compréhension du réseau complexe d'interactions spatiales et temporelles entre les divers éléments composant le vivant (gènes, métabolites, etc.) et de ces derniers avec l'environnement. Une telle compréhension obligeait à adopter un point de vue allant du niveau moléculaire (génome) jusqu'à une vision générale du fonctionnement d'une cellule (réseaux biologiques).
Plus spécifiquement, nous souhaitions nous intéresser à trois questions biologiques très générales : 1. existe-t-il des régularités, structurelles ou fonctionnelles, dans la diversité qui est observée, qui pourraient représenter des indices d'une organisation profonde du vivant; 2. pouvons-nous identifier ces régularités de façon systématique et arriver ainsi à dégager un ordre dans le réseau complexe des interactions détectées; enfin, 3. comment ce réseau s'est-il mis en place au cours de l'évolution, pour accomplir quelles fonctions ?
Les résultats attendus concernaient l'obtention de meilleurs formalismes de modélisation mathématique et algorithmes d'analyse, et des éléments de réponse à la fois à des questions biologiques spécifiques (tests d'hypothèses) et plus générales (exploration systématique des données).
Les principaux verrous du projet étaient liés à sa nature très exploratoire, initialement sur des questions biologiques extrêmement vastes et relativement vagues dans un domaine qui, en outre, est fortement compétitif. Les défis étaient donc de réussir à apporter des éléments de réponse à des questions aussi larges, en fait de dégager des voies de réponse et d'exploration plus précises, tout en contribuant au principal objectif méthodologique qui consistait à obtenir de meilleurs formalismes de modélisation mathématique et algorithmes d'analyse.
Il est clair que nous ne pouvions arriver aussi rapidement (3 ans) à une réponse nette et claire aux trois questions biologiques posées. Par contre, nous pensons avoir clairement rempli l'objectif de dégager de meilleurs formalismes de modélisation mathématique et algorithmes d'analyse, que ce soit au niveau génomique (lien entre recombinaison et organisation chromosomique, blocs de synténie et régions de cassure, génomes ancestraux, scénarios d'inversions) qu'à celui des réseaux biologiques (motifs, flux et ensembles minimaux de précurseurs, modules de régulation). À travers des tests d'hypothèse et une exploration systématique des données, nous avons aussi dégagé des régularités que nous allons maintenant analyser selon différentes voies plus spécifiques. Ces résultats sont synthétisés à la suite.

1. Inférence et la mise en évidence de régularités au niveau génomique

1.1 Impact de la recombinaison sur l'apparition et l'évolution des structures génomiques

Dans le but de modéliser l'impact de la recombinaison sur l'apparition et l'évolution des structures génomiques, notamment les isochores, plusieurs génomes de vertébrés ont été analysés. Les résultats de cette étude, qui indiquent que le phénomène de la conversion génique biaisé (BCG) ne serait pas présent chez certains poissons, s'appuient en particulier sur des travaux innovants de détection de structures tels les isochores (Melodelima 2006, Melodelima 2007 2 fois, Melodelima 2008 2 fois). Un premier modèle mathématique de l'impact du BGC dans une population de taille infinie a également été construit et des simulations de cet impact réalisées dans une population de taille finie (publication en préparation).

1.2 Répétitions

Les répétitions et leur distribution le long d'un génome sont un autre élément important dans l'organisation génomique. Or la résolution exacte de ce problème interdit actuellement en pratique l'exploitation de grosses masses de données. Nous avons alors développé un algorithme de filtrage sans perte permettant de supprimer rapidement des séquences en entrée de larges portions n'intervenant pas dans le résultat final (Peterlongo 2007 et 2008, http://igm.univ-mlv.fr/~peterlon/ednimbus). Ces résultats ont depuis été étendus et améliorés, en particulier en termes de sélectivité (algorithme Tuiuiu, Peterlongo 2009).

1.3 Éléments transposables

Parmi les répétitions les plus importantes d'un génome, tant par leur nombre que par leur fonction éventuelle, se trouvent les éléments transposables. Nous avons voulu ainsi étudier l'activation de ces éléments transposables chez Drosophila melanogaster en utilisant des banques publiques d'EST. Nous avons pu montrer que seulement 70 familles d'éléments transposables chez la drosophile sont potentiellement exprimées dont seulement 202 copies sont en relation avec un unique EST. Un pourcentage significativement plus élevé de copies exprimées a par ailleurs été observé sur le chromosome X par rapport aux autres (Deloger 2009)

1.4 Contraintes environnementales sur la composition en bases des génomes

Une influence possible de l'environnement sur la composition globale en bases des génomes avait été proposée mais était controversée. L'une des questions en ouvert concernait en particulier l'effet des UVs sur une telle composition, notamment le contenu en G+C. Nous avons re-étudié cette question et avons pu montrer qu'il n'existe pas de relation entre contenu en dinucléotides de pyridimine et exposition aux UVs dans l'habitat naturel. Ceci suggère que les génomes procaryotes et viraux n'ont pas développé un évitement des dinucléotides de pyrimidine comme stratégie évolutive pour contrer l'effet délétère des UVs, mais probablement des stratégies de protection et de réparation de l'ADN (Palmeira 2006).

1.5 Analyse des régions de cassure et de leur distribution

La comparaison de l'ordre des gènes entre plusieurs génomes permet l'identification des ruptures de synténie, appelées points ou régions de cassure. L'analyse fine de ces régions pourrait permettre de mieux en comprendre les mécanismes. Leur détection précise est ainsi importante. Une étude approfondie des méthodes existantes dans le domaine a été effectuée (Lemaitre et Sagot 2008), puis nous avons développé notre propre méthode qui a montré son efficacité puisqu'elle permet d'obtenir une meilleure précision dans la définition des régions de cassure que toutes les méthodes existantes à ce jour (Lemaitre et al. 2008). Nous avons étudié les corrélations entre ces régions sur le génome humain avec d'autres structures génomiques, tels les isochores et la densité en gènes (article soumis). La méthode a également été utilisée afin d'identifier des duplications aux bornes d'inversions sur le chromosome Y, renforçant ainsi l'hypothèse que ce chromosome a divergé du X par des inversions successives responsables de l'arrêt de la recombinaison entre les deux chromosomes (Lemaitre 2009).

1.6 Distance de réarrangements entre génomes

Pour expliquer les différences d'organisation entre les génomes et comprendre les contraintes qui pèsent sur cette organisation et les évènements qui la perturbent, des algorithmes permettant de reconstituer l'histoire des réarrangements entre génomes ont été développés (Tannier 2007). Leur principal problème est cependant de ne donner qu'une seule solution alors que souvent il en existe de nombreuses équivalentes. Nous

avons donc développé un algorithme qui permet de compter le nombre de solutions optimales au tri par inversions, de les regrouper par classes d'équivalence, et d'énumérer toutes les classes sans énumérer toutes les solutions (Braga 2007, Braga 2008). Nous avons également travaillé à ajouter des contraintes biologiques qui limitent parfois l'espace des solutions : groupes de gènes co-localisés dans les deux génomes (Berard 2008, Diekmann 2007, Sagot 2005) et dans les ancêtres intermédiaires du génome actuel (article soumis), symétrie observée des inversions autour de l'origine de réplication (cas des procaryotes, article en cours de rédaction), présence d'éléments dupliqués (Adi 2007, article soumis).

1.7 Reconstitution de génomes ancestraux

Deux méthodes de reconstruction de génomes ancestraux ont été développées, fondées chacune sur un principe différent (Lenne 2008, Chauve 2008). La reconstitution de génomes ancestraux a également motivé le calcul de médianes génomiques permettant de reconstituer l'organisation des chromosomes des ancêtres en même temps que les évènements qui expliquent les différences entre les génomes actuels (Tannier 2008).

1.8 Vers l'épigénétique

Les chromosomes eucaryotes forment des contacts à large échelle qui apparaissent conservés dans des cellules d'un même tissu à un même stade du développement et semblent être liés à la régulation et à certains réarrangements génomiques. Les données issues des techniques de détection de ces contacts entre chromosomes (4 et 5C) sont bruitées et nécessitent une analyse statistique et algorithmique spécifiques que nous avons commencé à effectuer en collaboration avec une biologiste du MRC à Londres. En parallèle, une simulation de données de type 4 et 5C sous diverses conditions initiales de cellules artificielles est en train de nous permettre d'évaluer la capacité de ces techniques à représenter les contacts spécifiques présents dans les cellules simulées, et ainsi de préparer les nouvelles techniques (confidentielles) mises au point par nos collaborateurs du MRC (en préparation).

2. Processus et réseaux

2.1 Régulation et motifs en séquence et en structure

Nos travaux sur la détection de motifs en séquence (dans le cas de l'ADN), à partir de modèles appris de ces motifs ou ab initio, se sont poursuivis au long de ces années (Carvalho 2006, Carvalho 2007, Pisanti 2006), et, depuis environ un an, ont été étendus aux ARNs, notamment les microARNs (Mendes 2009).

2.2 Reconstruction de réseaux métaboliques

La prise en main d'outils déjà existants et le développement de nos propres routines (librairie ParseBioNet, http://biomserv.univ-lyon1.fr/baobab/parsebionet/) nous ont permis de produire, filtrer et traiter nous-mêmes les données sur lesquelles nous travaillons avec nos méthodes. Ces données ont été rendues publiques (http://pbil.univ-lyon1.fr/software/symbiocyc/, article soumis). Nous avons également contribué à l'élaboration d'un outil de visualisation de réseaux métaboliques par une approche innovante qui permet de voir le réseau dans son ensemble tout en respectant au maximum le découpage en voies métaboliques (Bourqui 2007).

2.3 Motifs et modules dans les réseaux

L'étude des réseaux biologiques est en pleine expansion, avec notamment l'apparition de nouvelles méthodes d'analyse de graphes. La notion de motif topologique a ainsi été proposée (équipe d'Uri Alon) et a permis de mettre en évidence la présence de boucles de rétroaction dans les réseaux génétiques. Dans le contexte des réseaux métaboliques

cependant, on peut constater que deux sous-réseaux ayant même topologie peuvent avoir des fonctions très différentes. Nous avons donc proposé une nouvelle définition, celle de motif coloré. À la différence d'un motif topologique, un motif coloré n'a pas de structure spécifique. Il représente plutôt un ensemble de mécanismes réactionnels encodés par les étiquettes des noeuds du réseau et formant un sous-graphe connexe. Cette définition est également intéressante dans le cas de réseaux d'interactions où le niveau de bruit (arêtes non fiables) est élevé, les étiquettes dénotant alors des caractéristiques fonctionnelles. La recherche de motifs colorés constituait un problème original en théorie des graphes dont nous avons caractérisé la complexité et proposé un algorithme, Motus (http://pbil.univ-lyon1.fr/software/motus), pour le résoudre (Lacroix 2005 et 2006). Enfin, des méthodes pour l'évaluation statistique de motifs topologiques ou colorés respectivement ont été proposées (Schbath 2008).

2.4 Inférence de modules dans les réseaux de régulation

Les motifs peuvent être vus comme des modules de petite taille bien que la notion de module soit probablement plus large et sans lien avec celle de « quelque chose » qui se répète. La notion de module est ainsi plus souvent associée à celle de « quelque chose », par exemple un sous-réseau, capable de fonctionner de façon quasiment indépendante de son contexte. Nous nous sommes intéressés au problème de l'inférence de modules dans le cas de réseaux génétiques et en prenant en compte trois types d'informations simultanément : le partage de motifs en séquence, des données d'expression et l'évolution (conservation). Un premier article a été accepté sur ce sujet (Fonseca 2008) et un deuxième est en cours de rédaction.

2.5 Inférence de modules dans les réseaux intégrés

Nos travaux ont également porté sur l'analyse des données d'expression en lien avec les réseaux métaboliques dans une étude des relations entre le génotype et le phénotype. Nous nous sommes pour cela focalisés sur une analyse des sous-graphes connexes dans les réseaux métaboliques en corrélation avec l'expression et le mode de régulation des enzymes impliquées. Cette analyse (réalisée sur la levure) a révélé que les gènes enzymatiques induisant un sous-graphe connexe dans le réseau partagent des mêmes sites de régulation et patrons d'expression significativement plus souvent qu'attendu (article en préparation).

2.6 Réseaux dans un contexte de pharmacogénomique

Les réseaux de régulation ont également été étudiés dans le contexte de la pharmacogénomique. Dans ce cas, des réseaux bayésiens sont utilisés comme modèles afin de trouver une signature de gènes discriminant certaines maladies à partir de données de transcriptome. Nous avons mis en évidence l'avantage d'initialiser la procédure d'apprentissage avec un graphe (obtenu à partir de la matrice de corrélations) qui ne décrit pas des interactions aussi finement qu'un réseau bayésien, mais qui est fiable, très rapide et simple à calculer, même pour plus de 4000 variables (Prestat 2008, article en préparation).

2.7 Analyse structurale générale

Dans le but de définir les spécificités du réseau métabolique sélectionnées au cours de l'évolution, nous avons réalisé la comparaison topologique de 35 réseaux métaboliques de bactéries au style de vie différents (parasites / mutualistes intracellulaires, libres, fixatrices d'azote). De façon surprenante, seuls 31 composés et une réaction sont communs à tous les organismes dans le réseau des petites molécules. Même en prenant en compte les défauts d'annotation, toujours présents, ces chiffres demeurent très faibles, ce qui pourrait remettre en question la notion de métabolisme minimal, en tout cas pour ce qui concerne certains symbiotes (article en préparation).

2.8 Capacité de réseaux métaboliques : Modes élémentaires

Un modèle simple de graphe peut être suffisant pour concevoir et appliquer des méthodes telles que la recherche de motifs mais trouve ses limites lorsqu'on veut aller plus loin dans l'analyse des résultats obtenus. Une extension naturelle consiste à représenter un réseau par un hypergraphe qui permet alors de capturer de manière plus réaliste les liens entre métabolites, et dès lors de dégager des propriétés structurelles plus fines. Cela permet en outre un rapprochement intéressant avec d'autres méthodes d'analyse des réseaux métaboliques qui s'appuyent sur une décomposition de la matrice stoichiométrique (modèles basés sur les contraintes). Nous nous sommes ainsi intéressés au problème de la recherche de modes élémentaires dans un réseau métabolique et à des questions connexes (Acuña 2008).

2.9 Capacité de réseaux métaboliques : Précurseurs

Afin d'identifier les interactions métaboliques entre un organisme et son environnement, nous avons développé une méthode, Pitufo, pour identifier les ensembles minimaux de métabolites (que nous appelons précurseurs) suffisants pour la synthèse de métabolites cibles (Cottret 2008). Cette méthode est la première exacte permettant de 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 aphidicola* pour y identifier les précurseurs des acides aminés, métabolites centraux dans la fonction symbiotique de la bactérie, a permis de retrouver des résultats connus dans la littérature, et découvrir de nouveaux (article en préparation).

Conclusion et perspectives

Au delà des résultats spécifiques qui avaient été énoncés dans le texte soumis et qui ont pu en grande partie être traités de manière satisfaisante, ce projet avait indiqué un grand objectif général, qui était de renforcer et étendre un réseau de collaborations qui avait été en partie initié dans le cadre d'un projet précédent. Cet objectif a été globalement bien atteint. Le présent projet a ainsi déjà donné naissance à d'autres nouveaux, dont nous ne mentionnons brièvement que ceux déjà validés officiellement. Ces autres nouveaux projets portent sur des sujets qui sont la continuation de ceux de REGLIS notamment sur les ARNs (Projet ANR Blanc Brasero coordonné par Alain Denise du LRI de l'Université d'Orsay et impliquant aussi le Labri à Bordeaux, l'IGM à l'Université de Marne-la-Vallée et Sequoia de l'INRIA de Lille), les motifs dans les séquences (Projet Bioinformatique Israel-France et divers projets avec le Portugal) et les réseaux (Projet ANR Blanc NeMo coordonné par Stéphane Robin de l'AgroParisTech impliquant aussi l'INRA de Jouy et le Laboratoire Statistique et Génome de l'Université d'évry), ou de vastes extensions de ceux-ci, notamment à l'étude de l'organisation spatiale des chromosomes au sein des cellules et au rôle joué par cette organisation dans la régulation et la dynamique des génomes (Projet ARC-INRIA avec le MRC-Imperial College à Londres, Mistis de l'INRIA à Grenoble, l´INRA de Jouy et AgroParisTech, INRIA-LBBE), ainsi qu'à l'étude au niveau évolutif et réseaux de la relation hôte-parasite (Projet ANR-BBSRC MetNet4SysBio coordonné, côté Grande Bretagne originalement par Angela Douglas de l'Université de York et maintenant par Gavin Thomas, ainsi que côté français par Hubert Charles du Laboratoire BF2I de l'INSA de Lyon, Projet ANR Blanc Miri juste accepté avec BAOBAB-BAMBOO comme seul partenaire, et Équipe Associée Simbiosi INRIA avec Alberto Marchetti-Spaccamela de l'Université de Rome et Leen Stougie actuellement à l'Université Libre d'Amsterdam et au CWI, « Centrum voor Wiskunde en Informatica »).