Rapport sur le projet
ANR Blanc Numéro ANR-05-NT05-3_45205

Acronyme du projet : REGLIS
Titre : De la molécule à la cellule : développement, confrontation et intégration de modèles formels et de méthodes d'analyse
Responsable : Marie-France Sagot
Partenaires : BAOBAB-HELIX (désormais BAOBAB-BAMBOO), Université Claude-Bernard-Lyon 1 et INRIA Rhône-Alpes (seul partenaire)
Durée : 3 ans

1 Principaux résultats scientifiques obtenus

Les principaux résultats scientifiques obtenus sont détaillés selon les deux axes principaux que le projet contenait, à savoir l'inférence et la mise en évidence de régularités, d'une part au niveau génomique, et d'autre part au niveau des réseaux métaboliques et géniques. Un autre sujet, plus récemment abordé, vise à établir des liens entre ces deux axes. Ce sujet n'avait pas été mentionné dans le texte du projet au moment de sa soumission mais en est une suite naturelle. Il sera brièvement traité dans une troisième section. Seules les collaborations extérieures au Laboratoire de Biométrie et Biologie Évolutive de l'UCBL sont mentionnées.

Diverses perspectives au travail développé dans ce projet, certaines représentant des études déjà initiées et de nouvelles collaborations, seront évoquées en conclusion de ce rapport.

1.1 Inférence et la mise en évidence de régularités au niveau génomique

1.1.1 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 mammifères ont été étudiés ainsi que ceux d'un metathérien, d'un oiseau et de trois poissons. Des corrélations ont ainsi été mises en évidence entre la richesse en GC des isochores et certaines propriétés biologiques telles la longueur des introns, leur nombre, la densité en gènes en particulier compacts, la répartition en éléments répétés, etc. Les résultats de cette étude tendent à prouver que le phénomène de la conversion génique biaisée (BGC en anglais) ne serait pas présent chez un des poissons (le medaka). 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. Les conclusions en sont pour le moment que la BGC serait capable de maintenir des allèles récessives létales dans une population.

1.1.2 Répétitions

Les répétitions et leur distribution le long d'un génome sont un autre un élément important dans l'organisation chromosomique, et la recherche exhaustive de similarités multiples dans une séquence ou un ensemble de séquences est ainsi un problème fondamental en bioinformatique. Or la résolution exacte de ce problème souffre d'un temps d'exécution exponentiel, ce qui, en pratique, interdit l'exploitation de grosses masses de données. Nous avons alors travaillé sur un

algorithme de filtrage exact qui permet de supprimer rapidement des séquences en entrée de larges portions n'intervenant pas dans le résultat final. Une première version a été publiée [24, 23] et mise à disposition de la communauté via une interface web (`http://igm.univ-mlv.fr/~peterlon/ednimbus`). Ces résultats ont été étendus et parfois considérablement améliorés, en particulier en termes de sélectivité. Le nouvel algorithme, appelé Tuiuiu, a été décrit dans un article soumis à publication [25] et sera rendu publique prochainement. Ce travail a été fait en collaboration avec Pierre Peterlongo de l'INRIA de Rennes, Nadia Pisanti de l'Université de Pise en Italie et avec Gustavo A. T. Sakomoto et Alair Pereira do Lago de l'Université de São Paulo (USP) au Brésil.

1.1.3 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. Une étude des 29 gènes 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.

1.1.4 Contraintes environnementales sur la composition en bases des génomes

Une influence possible de l'environnement sur la composition globale en bases des génomes, par exemple en G+C, avait été proposée. Cela a fait l'objet de très nombreuses études, cependant souvent controversées. L'une des questions non tranchées concernait en particulier l'effet des UVs sur cette composition. Les lésions provoquées par les UVs sur l'ADN ont lieu de manière prépondérante sur les bases pyrimidiques adjacentes (C, T) et il est ainsi depuis longtemps considéré que ce mécanisme entraîne une forte pression de sélection sur la composition en bases.

Cette pression de sélection n'avait été que suggérée par la corrélation positive observée, chez les procaryotes, entre contenu en G+C et exposition aux UVs dans l'habitat naturel. Or, le contenu en G+C n'est qu'une mesure très indirecte de l'évitement des dinucléotides photo-sensibles. Nous avons donc re-étudié cette question en mesurant directement le contenu en dinucléotides de pyrimidine de chacun des génomes de procaryotes disponibles et de quelques virus marins. Nous avons pu montrer sur l'ensemble des données, et sur plusieurs exemples bien choisis, 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. Ces résultats ont fait l'objet d'une publication [22].

1.1.5 Distance de réarrangements entre génomes

Deux permutations de nombres entiers modélisent l'ordre des gènes dans deux génomes appartenant à des espèces distinctes. Les différences entre les deux ordres sont dus à des évènements évolutifs, dont un des plus courants est l'inversion de l'ordre d'un ensemble de gènes contigus. Pour expliquer les différences d'organisation entre les génomes et comprendre les contraintes qui pèsent sur ces organisations et les évènements qui les perturbent, des algorithmes qui permettent de reconstituer l'histoire des inversions qui ont différencié deux génomes ont été développés, y compris dans l'équipe.

Le principal problème de ce type d'algorithme 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 [4, 5].

Nous avons également travaillé à ajouter des contraintes biologiques pour limiter l'espace des solutions, et développé ainsi un algorithme qui transforme une permutation en une autre tout en préservant les groupes de gènes co-localisés dans les deux permutations [6, 12, 27]. Lorsque l'un des génomes est ancêtre de l'autre, il est plus intéressant biologiquement de prendre en compte dans le calcul des groupes de gènes co-localisés non pas seulement les deux génomes considérés mais également tous les ancêtres intermédiaires du génome actuel. Une première méthode dans ce sens est en phase finale de rédaction. Une autre contrainte intéressante à explorer dans le cas des bactéries concerne la symétrie observée des inversions autour de l'origine de réplication. D'autres types de contraintes ont également été considérées, par exemple pour prendre en compte la présence d'éléments dupliqués. Une première approche a été proposée qui considère ces dupliqués et correspond à la recherche d'une plus longue sous-séquence commune libre de répétitions [2]. Ce travail a été réalisé en collaboration avec les départements d'informatique de l'Université de São Paulo (USP) et de l'Université Fédérale de Mato Grosso do Sul (UFMS), toutes deux au Brésil.

1.1.6 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. Ces ruptures, appelées points de cassure, constituent des régions du génome qui ont potentiellement subi un ou plusieurs réarrangements génomiques. L'analyse fine de ces régions, au niveau génomique, pourrait permettre de mieux comprendre les causes et mécanismes de ces évènements évolutifs.

La détection précise des points de cassure constitue une étape préalable importante. Une étude approfondie des méthodes existantes dans le domaine a d'abord été effectuée [18], puis nous avons développé notre propre méthode pour répondre à nos besoins spécifiques de fiabilité et de précision des points de cassure. Cette méthode a été appliquée aux génomes entièrement séquencés des mammifères et a montré son efficacité, puisqu'elle permet d'obtenir une meilleur précision dans la définition des points de cassure que toutes les méthodes existantes à ce jour. Ce travail a fait l'objet d'une publication dans *BMC Bioinformatics* [19].

Cette méthode nous permet actuellement 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. Les résultats de ce travail sont en cours de rédaction. Ils sont réaliés avec l'équipe d'Alain Arnéodo de l'Institut François Jacob à l'ENS de Lyon. D'autres résultats préliminaires avaient également suggéré des différences de similarité

de séquence en particulier entre les séquences de points de cassure et des séquences non "cassées" du génome de l'homme proches ou plus éloignées sur le génome. Ces résultats feront l'objet d'une seconde publication.

Enfin, la 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 (publication en cours de soumission).

1.1.7 Reconstitution de génomes ancestraux

Deux méthodes de reconstruction de génomes ancestraux ont été développées, fondées chacune sur un principe différent : l'une, en collaboration avec le laboratoire d'informatique LIRIS de Lyon, construit l'ancêtre qui permet de minimiser un nombre d'évènements global qui peuvent expliquer les arrangements actuels de gènes [20] ; l'autre, développée en collaboration avec Cédric Chauve de l'université de Vancouver au Canada, rassemble des informations locales sur des morceaux de génomes conservés, et reconstruit un génome qui porte toutes ces informations (article soumis [9]). La reconstitution de génomes ancestraux a également motivé le calcul de médianes génomiques. Il s'agit, 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. Un travail a été accepté établissant la complexité théorique de plusieurs variantes de problèmes de médianes pour les cas où les génomes sont multi chromosomiques (comme c'est le cas de tous les vertébrés par exemple), et éventuellement dupliqués. Il a ainsi été établi qu'une variante du problème de la médiane devient polynomiale pour des génomes à plusieurs chromosomes, alors qu'elle est NP-complète pour un seul [29].

Une première application a été faite à l'étude de l'histoire évolutive des vertébrés visant estimer l'organisation des chromosomes d'espèces disparues dont on connaît des descendants. Des génomes ancestraux pour les mammifères boréoeuthériens et pour les amniotes ont ainsi été proposés, qui prennent en compte également les génomes dupliqués malgré la difficulté de retrouver des synténies dans ce cas. Le génome proto-mammifère est très proche des propositions déjà parues dans la littérature, tandis que le génome proto-amniote valide certaines associations synténiques proposées dans des études prospectives récentes, encore largement divergentes.

1.2 Inférence et la mise en évidence de régularités au niveau des réseaux

1.2.1 Régulation et motifs en séquence et en structure

Le travail précédent s'appuie fortement sur la capacité à détecter les motifs en séquence (dans le cas de l'ADN) et en structure (dans le cas de l'ARN, notamment des petits ARNs), à partir de modèles appris de ces motifs ou *ab initio*. Des travaux dans ce sens, en cours depuis quelques années avant la soumission du projet et son démarrage, se sont poursuivis au long de cess trois dernières années. Ils ont particulièrement bénéficié de certaines collaborations non françaises au sein du projet, notamment avec le groupe d'Arlindo Oliveira et Ana Teresa Freitas à l'Instituto Superior Técnico de Lisbonne au Portugal, celui de Roberto Grossi et Nadia Pisanti à l'Université de Pise en Italie, et, beaucoup plus récemment, avec le groupe de Yves van de Peer en Belgique (postdoc de Y. van de Peer actuellement en visite étendue à Lyon). Les travaux avec le groupe

portugais ont donné lieu à un certain nombre de publications sur les motifs ADN [7, 8, 21, 26], et, depuis quelques mois, sur les microARNs. Concernant ce dernier, un article de revue en cours de soumission et un travail est en cours sur la détection de miARNs dans le génome de *Anopheles gambiae* dans une collaboration avec Eric Westhof de l'Institut de Biologie Moléculaire et Cellulaire de Strasbourg. Ce travail sera poursuivi également avec Ana Tereza Vasconcelos du Laboratório Nacional de Computação Científica (LNCC) de Petrópolis, Brésil, qui participe du séquençage d'une nouvelle souche de l'anophèle (*Anopheles darlingi*).

1.2.2 Reconstruction, bases de données et visualisation de réseaux métaboliques

Reconstruction et base de données Durant ce projet, nous avons acquis une certaine expertise sur la reconstruction des données métaboliques. La prise en main d'outils déjà existants (comme les Pathway Tools de BioCyc `http://bioinformatics.ai.sri.com/ptools/`) et le développement de nos propres routines (librairie ParseBioNet `http://biomserv.univ-lyon1.fr/baobab/parsebionet/`) nous a permis de produire, filtrer et traiter nous mêmes les données que nos méthodes utilisent. Ces données ont été rendues publiques notamment via la construction du site web SymbioCyc (`http://pbil.univ-lyon1.fr/software/symbiocyc/`. Un article présentant cette base est en préparation.

Visualisation Nous avons contribué à l'élaboration d'une méthode et d'un outil de visualisation de graphes métaboliques complets. L'approche utilisée, qui permet de visualiser le graphe dans son ensemble tout en respectant au maximum le découpage en voies métaboliques, est très innovante [3]. Nous avons collaboré pour cela avec des spécialistes en visualisation de structures complexes : Fabien Jourdan de l'INRA de Toulouse, David Auber et Romain Bourqui du LABRI à Bordeaux.

1.2.3 Analyse et évolution de réseaux biochimiques

Motifs colorés dans les réseaux métaboliques L'étude des réseaux biologiques a été en pleine expansion ces dernières années, avec en particulier l'apparition de nouvelles méthodes d'analyse de graphes. L'équipe de Uri Alon a notamment proposé la notion de motif topologique, qui a été appliquée au réseau de régulation de la transcription de la bactérie *Escherichia coli*, mettant en évidence la présence de boucles de rétroaction, utiles pour caractériser le comportement dynamique du réseau.

Dans le contexte des réseaux métaboliques cependant, on peut constater que deux sous-réseaux ayant la même topologie peuvent avoir des fonctions très différentes. Nous avons donc proposé une nouvelle définition de motif, celle de motif coloré. À la différence d'un motif topologique, un motif coloré n'a pas de structure spécifique mais représente plutôt un ensemble de caractéristiques fonctionnelles encodées par les étiquettes des noeuds du réseau.

Formellement, un réseau métabolique est modélisé par un graphe coloré et un motif est défini comme un multi-ensemble de couleurs (une couleur correspond ici à un mécanisme réactionnel, symbolisé par un numéro EC). Une occurrence d'un motif est définie comme un ensemble de noeuds connectés et colorés par les couleurs du motif.

La recherche de motifs colorés constitue un problème original en théorie des graphes. Nous avons donc tout d'abord caractérisé sa complexité [16, 17] puis proposé un algorithme pour le résoudre [16, 17]. Ce travail a conduit également au développement d'un logiciel, MOTUS, utilisable en ligne

de commande et à travers le web à l'adresse `http://pbil.univ-lyon1.fr/software/motus/`. Un article sur le logiciel MOTUS a été soumis à publication [14].

Motifs colorés dans les réseaux d'interaction en général La recherche et l'énumération de motifs est actuellement poursuivie dans les réseaux métaboliques (afin de permettre le traitement de plus longs motifs) mais aussi dans ceux d'interaction protéine-protéine, ou dans tout autre réseau où il s'avère important de prendre en compte la valeur des étiquettes des noeuds dans la détection de motifs alors que la topologie précise du motif n'est pas toujours d'intérêt ou est à négliger dans un premier temps à cause du haut taux d'erreurs sur les interactions détectées. La définition de motifs colorés en général pose également la question de leur évaluation statistique. Des travaux sont en cours sur ce sujet impliquant l'équipe coordonnatrice du projet et une équipe de l'INRA à Jouy (publication soumise [28]).

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, si tant est qu'une telle notion existe de façon claire et unique, soit probablement plus large et indépendante de celle de "quelque chose" qui se répète, au moins au sein d'un même organisme (réseau). La notion de module est ainsi plus souvent associée à celle de "quelque chose", par exemple un sous-réseau, qui est capable de fonctionner de manière "quasiment" indépendante de son contexte. Nous nous sommes intéressés au sujet 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 cette année sur ce sujet [13] et un deuxième est en cours de rédaction.

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 et de régulation dans l'objectif de proposer un cadre pour l'étude des relations entre le génotype et le phénotype. Nous nous sommes pour cela focalisés sur une étude de la répétition de motifs dans les réseaux métaboliques en corrélation avec l'expression et le mode de régulation des enzymes impliquées dans ces motifs. Les données d'expressions sont issues de plusieurs études de micro-arrays en accès libre sur la base de données SGD (Saccharomyces Genome Database `http://www.yeastgenome.org`) et les données de régulation sont fournies par nos collaborateurs Ana Teresa Freitas et Arlindo Oliveira de l'Instituto Superior Técnico de Lisbonne, Portugal (`http://www.yeastract.com`). Cette étude a révélé que les gènes enzymatiques proches sur le réseau métabolique (induisant un sous-graphe connexe) partagent de mêmes sites de régulation et patrons d'expression significativement plus souvent qu'attendu. Un article présentant ces résultats est en préparation.

Réseaux dans un contexte de pharmacogénomique Les réseaux de régulation ont également été étudiés dans le contexte de la pharcogénomique. Dans ce cas, des réseaux bayésiens sont utilisés comme modèles afin de trouver une signature de gènes discriminant les leucémies aiguës lymphoïdes des myéloïdes, ainsi que pour la recherche de réseaux de régulation dans les cas du cancer du sein, à partir de données de puces transcriptome. Dans cette exploration des réseaux bayésiens appliqués à des problématiques de cancérologie, nous nous sommes servi des différentes facettes de cette méthode manipulant des concepts à la fois complexes (de probabilités conditionnelles)

et instinctifs (le raisonnement à partir du graphe de dépendances). Nous avons mis en évidence l'avantage d'initialiser la procédure d'apprentissage de structure avec un graphe qui certes ne décrit pas des interactions aussi finement qu'un réseau bayésien, mais qui peut aussi s'apparenter à un graphe d'interactions (obtenu à partir de la matrice de corrélations) et surtout qui est très rapide et simple à calculer, même pour plus de 4000 variables. Des résultats initiaux de ce travail ont été présentés (oralement et à travers un poster) à la 9ème Conférence Internationale sur la Science des Systèmes de Santé qui s'est déroulé à Lyon en Septembre 2008.

Analyse structurale générale Dans le but de définir les spécificités du réseau métabolique sélectionnées au cours de l'évolution, 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. Nous souhaitons ensuite déterminer si d'autres propriétés du réseau 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. Ce travail fera l'effet d'une publication avant la fin du projet. Il est réalisé en collaboration avec Hubert Charles et Yvan Rahbé de l'équipe BF2I de l'INSA de Lyon.

Revue des analyses structurales de réseaux métaboliques Un article de revue sur l'analyse structurale des réseaux métaboliques – diverses questions et approches proposées, leurs limites et réalisme – a été accepté à publication récemment [15].

1.2.4 Capacité de réseaux métaboliques

Flux et 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 pousser plus loin 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 qui associent les métabolites entre eux, et dès lors de dégager des propriétés structurelles plus fines. Cela permet en outre un rapprochement inéressant avec d'autres méthodes d'analyses des réseaux métaboliques basées 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. Ce travail est réalisé en collaboration 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"). Il a déjà donné lieu à une publication [1].

Recherche de précurseurs Afin d'identifier les interactions métaboliques entre un organisme et son environnement, nous avons développé une méthode permettant d'identifier les ensembles

minimaux de métabolites (que nous appelons précurseurs) suffisants pour la synthèse de métabolites cibles. Ces travaux ont été le fruit d'une collaboration entre l'équipe coordonnatrice et 3 autres, à l'Université de Campo Grande do Sul au Brésil (Fabio V. Martinez et Paulo V. Milreu), l'Université "La Sapienza" de Rome (Alberto Marchetti-Spaccamela) et l'Université Libre d'Amsterdam (Leen Stougie). Un article sur la méthodologie est déjà publié [11, 10]. La méthode, appelée PITUFO, sera rendue disponible prochainement.

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* 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 aphidicola*. Ce travail est réalisé en collaboration avec Hubert Charles et Yvan Rahbé de l'équipe BF2I de l'INSA de Lyon.

1.3 Entre génomes et réseaux

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.

2 Sélection de publications

Une sélection des publications issues de ce projet sont indiquées à la fin du document.

3 Conclusion et futurs développements

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, lÍNRA 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, et Projet ANR Blanc Miri juste accepté avec BAOBAB-BAMBOO comme seul partenaire).

Références

[1] V. Acuna, F. Chierichetti, V. Lacroix, A. Marchetti-Spaccamela, M.-F. Sagot, and L. Stougie. Modes and cuts in metabolic networks : Complexity and algorithms. *Biosystems*, 2008.

[2] 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.

[3] 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.

[4] 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.

[5] 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.*, in press, 2008.

[6] 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.

[7] 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.

[8] 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.

[9] C. Chauve and E. Tannier. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. submitted, 2008.

[10] L. Cottret, P. V. Milreu, V. Acu na, 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.

[11] L. Cottret, V. Acu na, 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.

[12] 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.

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

[14] 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.

[15] 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 :360–368, 2008.

[16] 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.

[17] 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.

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

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

[20] 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.

[21] N. P., 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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. submitted, 2008.

[26] 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.

[27] 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.

[28] S. Schbath, V. Lacroix, and M.-F. Sagot. Assessing the exceptionality of coloured motifs in networks. 2008.

[29] 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.

Rapport fait le 28 Septembre 2008

Marie-France Sagot
Directrice de Recherche INRIA
Équipe-projet BAMBOO, INRIA Grenoble Rhône-Alpes &
Équipe BAOBAB, Lab. de Biométrie et Biologie Évolutive (LBBE)
Univ. Lyon I
43 bd du 11 Novembre 1918
69622 Villeurbanne cedex
Courriel : Marie-France.Sagot@inria.fr