Étude de l'évolution combinatoire des gènes par l'analyse de réseaux de similarité de séquence

Pierre-Alain Jachiet

Université Pierre et Marie Curie

Ecole doctorale Complexité du Vivant

Evolution Paris Seine - UMR 7138

Equipe Adaptation, Intégration, Réticulation et Evolution

Étude de l'évolution combinatoire des gènes par l'analyse de réseaux de similarité de séquence.

Par Pierre-Alain JACHIET

Thèse de doctorat en Biologie Évolutive

Dirigée par Eric BAPTESTE et Philippe LOPEZ

Présentée et soutenue publiquement le 02 Juillet 2014 devant :

Dr Eric **Bapteste** (CR CNRS, Université Pierre et Marie Curie)	Encadrant
Pr Didier **Casane** (PR, Université Paris Diderot)	Rapporteur
Dr Claudine **Devauchelle** (MC, Université Evry Val-d-Essonne)	Examinatrice
Pr François-Joseph **Lapointe** (PR, Université de Montréal)	Examinateur
Dr Philippe **Lopez** (MC, Université Pierre et Marie Curie)	Encadrant
Dr Laurent **Viennot** (DR INRIA, Université Paris Diderot)	Rapporteur
Pr Martin **Weigt** (PR, Université Pierre et Marie Curie)	Examinateur

Résumé

L'accumulation récente de données de séquences génomiques a montré que l'évolution des gènes n'est pas strictement arborescente. De nombreux processus évolutifs, comme l'exon shuffling, la fusion de gènes ou la recombinaison illégitime remodèlent les gènes, créant des structures composites, formées de parties dont les histoires évolutives sont différentes. Le développement de réseaux de similarité de séquences fournit un cadre analytique permettant d'étudier l'impact de ces processus sur l'évolution moléculaire, en structurant les relations de ressemblance entre séquences et en formalisant en termes de graphes la détection de gènes (triplets intransitifs) et de familles de gènes (cliques minimales séparatrices) composites. La taille des jeux de données actuels, de l'ordre de plusieurs millions de séquences, a également requis le développement de nouveaux outils et méthodes : parallélisation des comparaisons de séquences, visualisation de très grands réseaux par simplification en communautés de Louvain et identification de grands cycles. Appliquées à des jeux de données de génomes eucaryotes et viraux, ces méthodes ont démontré la présence de gènes composites dans tout le vivant et les éléments génétiques mobiles. En proportion, les gènes composites sont plus nombreux dans les génomes eucaryotes ; en nombre absolu, ils sont plus nombreux à être portés par des virus. Chez ces derniers, la distribution fonctionnelle des gènes composites est biaisée (enrichissement dans les familles essentielles pour la perpétuation du cycle viral), et les éléments des gènes composites trouvent même parfois leurs origines dans le matériel génétique de classes virales différentes. Plus généralement, l'étendue des processus combinatoires, en révélant des liens évolutionnaires autres que les liens d'homologie au sens fort, justifie une étude pluraliste des relations de similarité entre séquences.

Abstract

The recent accumulation of genomic sequence data has shown that gene evolution is not strictly tree-like. Many evolutionary processes, like exon shuffling, gene fusion or nonhomologous recombination remodel genes by creating composite structures that are made from parts with different evolutionary histories. The development of sequence similarity networks provides an analytical framework to study the impact of these processes on molecular evolution, by structuring the resemblance relationships between sequences and by formalizing, in terms of graph theory, the detection of composite genes (intransitive triplets) and gene families (clique minimal separators). The size of current data sets, typically several million sequences, has also required the development of new tools and methods: sequence comparison parallelization, large networks visualization with Louvain communities and large cycles identification. When applied to eukaryotic and viral genome data sets, these methods have shown that composite genes are found throughout cellular organisms and mobile genetic elements. Proportionally, composite genes are more numerous in eukaryotic genomes; in absolute number, they are more numerous in viruses. In the latter, composite genes functional distribution is biased (enrichment of genes families that are essential for the perpetuation of the viral cycle), and the various parts of composite genes sometimes even originate from the genetic material of different viral classes. More generally, the extent of combinatorial processes, by unravelling other evolutionary bonds than homology bonds in the strictest sense, legitimates a pluralistic study of similarity relationships between sequences.

Remerciements

Je tiens tout d'abord à remercier Éric Bapteste et Philippe Lopez pour m'avoir permis de réaliser cette thèse. Merci pour votre disponibilité, votre soutien et vos encouragements. Merci pour la liberté et la confiance que vous m'avez accordées. Merci Éric pour toutes nos conversations, qui m'ont beaucoup appris.

Je souhaite remercier tous les visiteurs du laboratoire, qu'ils y aient passé quelques heures, quelques jours ou plusieurs semaines : ce fut un plaisir d'apprendre, d'échanger ou de collaborer avec vous. Merci à Michel Habib, James McInerney, Philippe Colson, Lucie, Mathias, Judith, Ana, et bien d'autres.

Je remercie mes collègues, avec lesquels j'ai partagé la vie tranquille de notre camp retranché bioinformaticien, en bout de couloir avec vue sur la Seine. Merci Cédric pour ton sourire bienveillant et tes nouveaux rires retentissants. Merci Slim pour tes humeurs et ta bonne humeur. Merci Guifré pour les discussions de tous les instants. Merci Raphaël pour avoir repris le flambeau, à peine hier étais tu un sympathique arrivant que te voilà déjà demain le vétéran des doctorants.

Je remercie Gaëlle Boutin, Hervé Le Guyader et l'ancienne école doctorale *Diversité du vivant* pour l'organisation des journées de Roscoff, et de leurs fabuleux banquets de fruits de mers.

Merci à mes amis pour tous les moments vécus ensembles pendant ces trois années, des soirées crêpes du 98 aux virées en cabanes perdues dans les montagnes, des nuits de bricolage montreuilloises aux évasions sur des parois calcaires.

Merci à ma famille et à mes parents.

Merci à ma lutine préférée pour tout ce qu'elle est.

Sommaire

Chapitre 1 - Introduction 1
1.1. L'abondance des données moléculaires ouvre de nouvelles questions évolutives 2
1.2. La science des réseaux pour l'étude des données relationnelles 3
1.3. La comparaison de séquences 8
1.4. Similarité, homologie et convergence évolutive 14
1.5. Réseaux de similarité de séquences 20
1.6. Objectifs de cette thèse 33
Chapitre 2 - Identification de familles de gènes composites 35
2.1. Processus combinatoires d'évolution des gènes 36
2.2. Identification des processus combinatoires dans les réseaux de similarité de séquences 40
Chapitre 3 - Problème de l'homologie : extension du champ des ressemblances informatives pour les évolutionnistes 53
3.1. L'homologie en biologie 55
3.2. Famille de gènes homologues 56
Chapitre 4 - Application de la notion d'air de famille à des jeux de données de virus 79
4.1. La diversité des virus est immense et très peu connue 80
4.2. Différentes classifications des virus 80
4.3. Evolution des virus 83
4.4. Etude systématique des phénomènes combinatoires chez les virus 84
4.5. Enjeux nouveaux abordés lors de cette étude 85
Chapitre 5 - Conclusion 105
5.1. Analyse des réseaux de similarité de séquences 105
5.2. Étude de l'évolution combinatoire des gènes 110
5.3. Pour une démarche pluraliste en évolution 114
Chapitre 6 - Bibliographie 117

Table des figures

Figure 1-1: La révolution des nouvelles technologies de séquençage 2
Figure 1-2 : D'un tableau de données relationnelles à un réseau 4
Figure 1-3 : Exploration du réseau social de Roscoff selon différentes caractéristiques 5
Figure 1-4 : Les éléments d'un modèle de réseau 7
Figure 1-5 : Score d'alignement entre deux séquences 9
Figure 1-6 : Recherche d'alignements locaux avec BLAST 10
Figure 1-7 : Schéma de protéines des quatre grandes classes structurelles 15
Figure 1-8 : Masquage des régions de faible complexité 17
Figure 1-9 : Comparaison de séquences avec l'algorithme BLAST 21
Figure 1-10 : D'une sortie BLAST à un réseau de similarité de séquences 22
Figure 1-11 : Effet sur la connectivité d'un réseau du seuil sur la E-value 24
Figure 1-12 : Exploration graphique d'un réseau de similarité de séquences 25
Figure 1-13 : Apparition d'une composante connexe géante 26
Figure 1-14 : Homologie distante et partielle 28
Figure 1-15: Classification d'une protéine multi-domaines par l'algorithme GeneRAGE 30
Figure 1-16: Modèle d'évolution des familles multi-domaines 31
Figure 1-17 : La structure du voisinage des gènes reflète leurs relations évolutives 32
Figure 1-18 : Processus combinatoires considérés pour construire les familles homologues 33

Figure 2-1 : Delta du fleuve Léna en Sibérie 35
Figure 2-2 : Mécanismes de recombinaison homologue suite à une rupture double-brin 37
Figure 2-3 : Fusion de gène 38
Figure 2-4 : Phénomène d'exon-shuffling 39
Figure 2-5 : Composante connexe d'un réseau de similarité de séquences 41
Figure 2-6 : Patron de similarité formé par un gène fusionné dans un réseau 42
Figure 2-7 : Processus combinatoires et patron dans les réseaux 43

Figure 3-1 : Réseau des communautés de la GCC d'un réseau de gènes eucaryotes 53
Figure 3-2 : Réseau phylogénétique à plusieurs racines 60

Figure 4-1 : Diversité morphologique des virus 79
Figure 4-2 : Classification des virus proposée par David Baltimore 81
Figure 4-3 : Scénario d'apparition des lignées virales 83

Chapitre 1 - Introduction

1.1. L'abondance des données moléculaires ouvre de nouvelles questions évolutives 2

1.2. La science des réseaux pour l'étude des données relationnelles 3

1.2.1. L'exemple d'un réseau social 3

1.2.2. Les réseaux permettent de visualiser les données relationnelles 4

1.2.3. Les réseaux sont adaptés à une démarche exploratoire 6

1.2.4. Les réseaux permettent l'étude statistique des données relationnelles..... 6

1.2.5. Formalisation d'un modèle de réseau 7

1.3. La comparaison de séquences..... 8

1.3.1. Les scores d'alignements reposent sur un modèle d'évolution 8

1.3.2. BLAST est l'algorithme d'alignement de séquences de référence..... 9

1.3.3. Principes de fonctionnement de BLAST 10

1.3.4. Système de score employé par BLAST 11

1.3.5. Le choix d'un seuil sur la E-value n'a pas de sens statistique 12

1.4. Similarité, homologie et convergence évolutive 14

1.4.1. La convergence évolutive des séquences génétiques 14

1.4.2. La convergence évolutive ne concerne pas toutes les catégories structurelles de protéines 15

1.4.3. Détecter et masquer les régions sensibles à la convergence évolutive 16

1.4.4. Alternatives à la comparaison de séquences avec BLASTP..... 18

1.5. Réseaux de similarité de séquences..... 20

1.5.1. Construction d'un réseau de similarité de séquences..... 20

1.5.2. Visualisation d'un réseau de similarité de séquences 23

1.5.3. Algorithmes de réseaux pour former des familles homologues..... 27

1.6. Objectifs de cette thèse..... 33

1.1. L'abondance des données moléculaires ouvre de nouvelles questions évolutives

La quantité de séquences qui peut être exploitée en biologie évolutive a augmenté très rapidement au cours des quinze dernières années. Divers sauts technologiques ont spectaculairement fait chuter le coût et la complexité du séquençage de matériel génétique [Pareek et al., 2011]. A titre d'exemple, le séquençage du premier génome humain a duré 13 ans, de 1990 à 2003. Il a nécessité la contribution de 3.000 personnes au sein d'un consortium international, pour un budget d'environ 3 milliards de dollars. Aujourd'hui, le séquençage d'un génome humain nécessite moins d'une semaine, pour un coût inférieur à 5000 dollars [Soon et al., 2013] (Figure 1-1.A). Cette tendance se poursuit avec l'annonce de nouvelles ruptures technologiques (Ion Torrent, Nanopore) promettant de petits séquenceurs, de la taille d'un micro-onde voire d'une clé USB, toujours plus rapides, bon marché et simples d'utilisation. Le séquençage massif de matériel génétique est désormais une pratique courante, qui révolutionne de nombreux champs de la biologie.

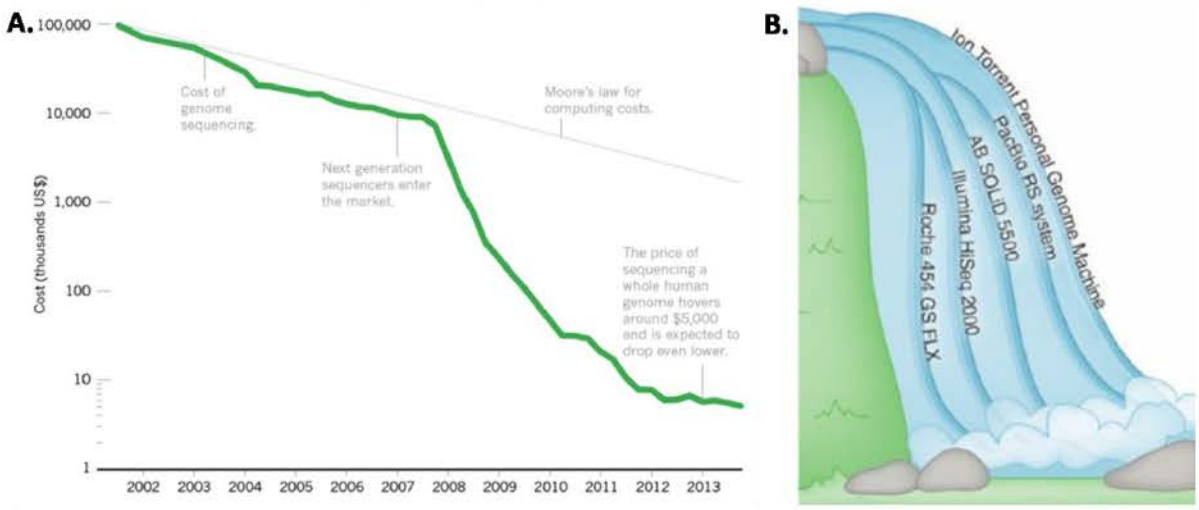


Figure 1-1: La révolution des nouvelles technologies de séquençage

A. Après la fin du Projet Génome Humain, le coût du séquençage a suivi pendant quelques années la loi de Moore, qui prédit un déclin exponentiel du coût des calculs informatiques. Après 2007, le coût du séquençage a chuté bien plus rapidement.
Check Hayden, Erika. "Technology: The $1,000 Genome." Nature. 2014

B. Les nouvelles plateformes de séquençage apportent un torrent de séquences, difficile à maîtriser, qui bouleverse de nombreux champs de la biologie.
Rusk, Nicole. "Torrents of Sequence" Nature Methods. 2011

Ce phénomène est souvent décrit comme le déferlement d'un torrent (Figure 1-1.B) ou d'un déluge de séquences, comme un flot puissant difficile à canaliser. En effet, l'analyse de l'information contenue dans de telles quantités de données pose beaucoup de difficultés pratiques. L'outil informatique est désormais indispensable à toutes les étapes, dès le séquençage des nucléotides et l'assemblage des fragments produits. La détection des gènes repose sur des modèles statistiques nécessairement automatisés. L'étude de ces gènes, une fois détectés, utilise ensuite des méthodes extrêmement intensives en calcul. C'est pourquoi cette thèse de biologie évolutive s'est concentrée sur le développement de nouvelles

approches *in silico* étendant l'exploitation de ces grands jeux de données moléculaires, fondées sur les réseaux, pour les raisons que nous allons maintenant exposer.

1.2. La science des réseaux pour l'étude des données relationnelles

Le mot 'réseau' est régulièrement employé dans le langage courant. On discute du déploiement d'un « réseau de télécommunications », de l'encombrement d'un « réseau de transport » ou de l'influence du « réseau social » d'un individu. Des approches de réseau se sont développées dans de nombreux domaines scientifiques depuis une vingtaine d'années, notamment en physique, biologie, informatique et économie, jusqu'à justifier la création du journal *Network Science* en 2013, publié par Cambridge University Press.

Quel est donc cet objet de réseau ? Et qu'apporte-t-il dans la compréhension de domaines si variés ? Pour l'introduire et en exposer les multiples intérêts, nous allons d'abord nous appuyer sur l'exemple des réseaux sociaux. Ces réseaux ont l'avantage d'être très intuitifs pour les animaux hautement sociaux que nous sommes. Ils sont employés depuis longtemps par les sociologues, qui ont développé de nombreux concepts et méthodes pour les analyser. Il s'agit d'un premier intérêt des réseaux : *ils permettent des échanges fructueux entre des domaines d'études très différents*.

1.2.1. L'exemple d'un réseau social

Prenons comme exemple une analyse de réseau effectuée sur le groupe de participants aux journées de l'école doctorale *Diversité du vivant*, lors de l'édition d'octobre 2012. Ce groupe comprenait des doctorants issus des différents laboratoires rattachés à l'école doctorale, et quelques encadrants. De nombreux membres du groupe n'avaient encore jamais eu l'occasion de se rencontrer, parce que leurs laboratoires sont dispersés géographiquement, et que la moitié des doctorants débutaient alors leur première année. Les encadrants de l'école doctorale et les doctorants en deuxième année, qui avaient participé aux journées de l'école doctorale un an auparavant, pouvaient en revanche connaître de nombreux participants. La structure sociale de ce groupe était donc *a priori* peu dense et hétérogène. Elle était amenée à se modifier fortement au cours des journées à venir et portait des informations sur le fonctionnement de l'école doctorale.

Pour collecter l'information sur la structure sociale de ce groupe, j'ai fait circuler dans le train menant à Roscoff une liste des participants, en demandant à chacun de noter qui il connaissait. Le résultat de cette enquête est un tableau (Figure 1-2.A) indiquant pour chaque personne la liste de ses connaissances. Ce tableau contient une description très simplifiée des relations sociales au sein du groupe, sans mesure de la relation entre individu. Malgré sa simplicité, ce tableau contient une information riche, qu'il paraît difficile d'analyser. Il est beaucoup plus facile d'appréhender cette information en la représentant par un réseau (Figure 1-2.B). Dans cette figure, les cercles (*nœuds* ou *sommets*) représentent les individus, et les traits (*arêtes*) représentent des relations de connaissance mutuelle entre individus. On

observe que les individus ont un nombre de connaissances (*degré*) très variable. On constate que le réseau est divisé en plusieurs parties déconnectées (*composantes **connexes***), avec une grande partie rassemblant l'essentiel des individus (*composante connexe géante*), une petite partie rassemblant deux individus ne se connaissant qu'entre eux, et deux individus isolés. On remarque que certains individus semblent plus *centraux*, tandis que d'autres sont en périphérie du réseau de connaissance.

A.

Personne	Liste des connaissances
1	5; 14
2	
3	22; 10; 15
4	5; 9; 11; 16; 19; 21
5	6; 9; 10; 11; 13; 15; 16; 17; 22; 26
6	5; 17; 26; 28
7	5; 11; 13; 17
8	15
9	4; 5; 11; 16; 21
10	

B.

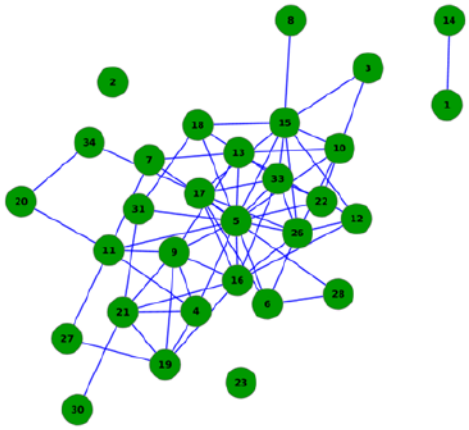


Figure 1-2 : D'un tableau de données relationnelles à un réseau
Les relations de connaissances entre participants aux journées de l'école doctorale à Roscoff sont représentées de façon intuitive par un réseau.

A. Tableau listant les connaissances de chaque individu (extrait)
B. Réseau de connaissance réciproque entre individus.

1.2.2. Les réseaux permettent de visualiser les données relationnelles

"Une image vaut mille mots" (Confucius)

Le dessin d'un réseau permet d'appréhender de façon intuitive des données relationnelles complexes. Il permet de *voir* un jeu de données, de distinguer rapidement les caractéristiques clés d'un phénomène, de construire des hypothèses. Les représentations graphiques sont plus intuitives que les tableaux de données parce que notre cerveau a davantage été sélectionné au cours de son évolution pour la représentation mentale des formes que pour le calcul matriciel. La richesse des formes d'un réseau sollicite à plein ses capacités visuelles. Les réseaux sont des objets mathématiques faciles d'accès, car les concepts, mesures et algorithmes de réseaux sont souvent imagés.

La figure 1-2 n'est qu'une étape préliminaire pour qui souhaiterait analyser en finesse la structure sociale du groupe de Roscoff. Ce dessin représente la *topologie* du réseau, c'est-à-dire les connexions entre sommets. Il est possible de mettre en valeur certains aspects de cette topologie en modifiant des aspects graphiques tels que la couleur, la taille ou la forme des nœuds et arêtes. Si l'on s'intéresse par exemple à la variation du nombre de connaissances des individus, on peut représenter le degré des nœuds par un code couleur (Figure 1-3.A). Cela fait ressortir certains individus isolés et d'autres très connectés. L'individu 5 est très fortement connecté avec 16 voisins. Il occupe donc une place importante dans ce réseau social, ce qui n'est pas surprenant étant donné qu'il s'agit de

Gaëlle Boutin, gestionnaire de l'école doctorale. On peut s'intéresser à des questions plus avancées, comme la circulation d'informations via les individus dans ce réseau. Une façon d'identifier les nœuds de communication clés est de représenter leur centralité d'intermédiarité (*betweenness*), à savoir la proportion de plus courts chemins entre individus dans le réseau qui passent par un sommet donné (Figure 1-3.B). En plus du sommet 5 qui est très central selon cette mesure, d'autres sommets plus périphériques ressortent car ils créent des liens vers des parties isolées du réseau.

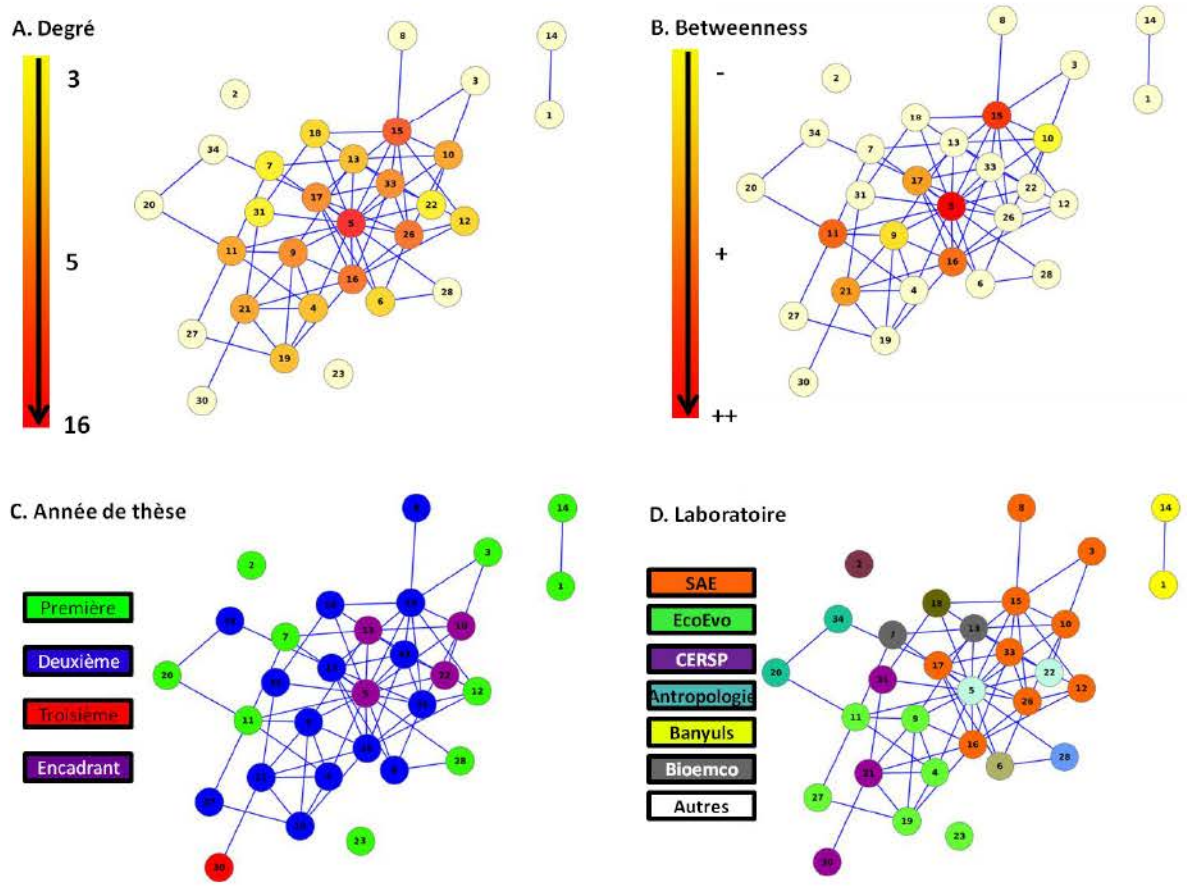


Figure 1-3 : Exploration du réseau social de Roscoff selon différentes caractéristiques

- **A.** Le *degré* d'un nœud indique le nombre de connaissances d'un individu, c'est-à-dire la taille de son réseau social local.
- **B.** La *betweenness* mesure la centralité d'intermédiarité, c'est-à-dire l'importance d'un individu pour les communications dans le réseau.
- **C.** La coloration par *année de thèse* des individus montre que ceux de première année sont plus périphériques.
- **D.** La coloration des individus en fonction de leur *laboratoire* d'origine permet d'étudier les relations au sein des laboratoires et entre laboratoires différents.

Les réseaux permettent également d'analyser des données transverses sur les individus et de les ajouter aux représentations graphiques. Dans la figure 1-3.C le réseau est coloré en fonction de l'année de thèse des individus; on observe ainsi une nette différence entre les anciens doctorants qui sont reliés par un tissu dense de connaissances, et les nouveaux doctorants qui sont plus en périphérie. Le réseau est coloré en fonction du laboratoire d'origine dans la figure 1-3.D. On peut ainsi observer que les laboratoires ont des tailles différentes, que certains sont isolés tandis que d'autres sont fortement connectés. On peut

aussi remarquer que certains laboratoires ne forment pas des cliques, c'est-à-dire que tous les membres du laboratoire ne sont pas connectés 2 à 2, peut-être en raison d'une dispersion géographique des équipes. Ces figures invitent donc à s'interroger sur la dynamique sociale et l'histoire de ce groupe.

1.2.3. Les réseaux sont adaptés à une démarche exploratoire

Ces quelques exemples montrent que la représentation de données par un réseau permet d'adopter naturellement une démarche *exploratoire*. La visualisation fait rapidement naître des hypothèses, guidée par des données qui n'avaient pas nécessairement été collectées pour y répondre. L'exploration est facilitée par la flexibilité du réseau, qui peut être associé à d'autres informations, voire redéfini différemment, pour approfondir les questions soulevées. Cette méthode exploratoire, distincte et complémentaire de la méthode hypothético-déductive [Kell, Oliver, 2004], est très pertinente pour s'intéresser à des objets nouveaux ou peu connus, tel que nous le ferons avec des séquences génétiques dans ce travail de thèse.

Les représentations graphiques sont des guides très utiles pour s'approprier les données, mais la démarche exploratoire ne doit pas s'y arrêter. Il serait hâtif de tirer des conclusions sur la structure du groupe de Roscoff à partir des dessins de la figure 1-3. Il est en effet difficile d'appréhender tous les détails de la structure d'un réseau en observant sa projection en 2 dimensions, ce qui est d'autant plus vrai que la taille du réseau augmente. Une solution serait de visualiser des parties restreintes du réseau, ou de comparer des projections selon plusieurs méthodes de placement des nœuds et arêtes, mais cette démarche est subjective, non systématique, et risque donc d'être biaisée.

Les doctorants de première année sur la figure 1-3.C sont ainsi décrits comme « périphériques », alors que certains ont pourtant un nombre élevé de connaissances (p. ex. les 11 et 12) : ils ne se retrouvent en périphérie que par le hasard de l'algorithme de projection. Pour aller au bout de la démarche, il s'agirait maintenant de formaliser en terme de réseau ce qu'est un nœud périphérique, en s'appuyant par exemple sur le riche corpus de concepts et d'algorithmes de la théorie des graphes. C'est dans ce cycle d'allers-retours entre représentations, construction d'hypothèses, formalisation et mesures de propriétés que la démarche exploratoire prend tout son intérêt.

1.2.4. Les réseaux permettent l'étude statistique des données relationnelles

A travers l'étude d'un groupe social, nous avons montré que les réseaux permettent de manipuler facilement des informations complexes, et qu'ils sont adaptés pour adopter une démarche exploratoire. Au-delà de ces aspects pratiques, le principal intérêt des réseaux est sans doute qu'ils permettent de poser un regard global sur les phénomènes, de les aborder comme des systèmes, à la différence des méthodes réductionnistes, plus classiques en

science. Les réseaux proposent en fait de procéder à un type nouveau de statistiques, qui traite la structure des données relationnelles [Brandes et al., 2013].

Les statistiques sont l'étude des données : elles s'intéressent à la collecte, au traitement, à l'interprétation et la présentation des données, afin de les rendre intelligibles. Les études statistiques habituelles sont conçues pour des ensembles sans structure, des unités ou des groupes indépendants. Puisque les phénomènes réels présentent généralement une structure, une grosse partie du travail de statistiques consiste à détecter les relations entre données, pour éliminer les biais de dépendance. Si l'on étudie par exemple l'association entre le nombre d'amis d'un doctorant et son année de thèse, on cherchera à échantillonner des doctorants qui ne sont pas amis les uns des autres. Certaines études statistiques classiques s'intéressent à des données relationnelles, comme par exemple la corrélation entre les âges de paires d'amis, mais elles considéreront alors des paires indépendantes, sans individu en commun.

Les réseaux proposent au contraire de s'intéresser à la structure des données relationnelles. Cette approche repose sur l'hypothèse, implicite mais très forte, que la structure des relations est fondamentalement importante. Elle pense cette structure comme une propriété émergente du système étudié, qui n'est pas résumée par la simple agrégation de ses éléments constitutifs. C'est cette conceptualisation qui permet, et nécessite, de nouvelles méthodes de pensée, une forme nouvelle d'inférence scientifique et le développement de nouveaux types de connaissances. Nous allons exploiter cet avantage en réalisant des réseaux de similarité de séquences.

1.2.5. Formalisation d'un modèle de réseau

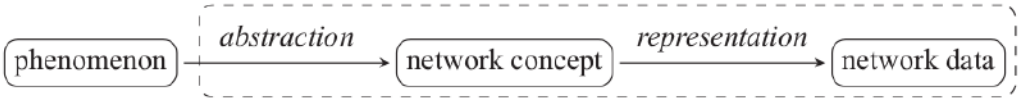


Figure 1-4 : Les éléments d'un modèle de réseau
Brandes, et al. « What is network science? » Network Science 2013

Pour généraliser, l'étude d'un système par un modèle de réseau nécessite deux étapes. La première consiste à *abstraire* le phénomène sous la forme d'un réseau, c'est-à-dire à définir des entités qui constitueront les nœuds et les relations entre ces entités qui constitueront les arêtes. Nous avons ainsi abstrait le système social de Roscoff sous la forme d'un *réseau de connaissance*. Cette abstraction réduit la réalité du phénomène. Elle en permet une étude générique, reproductible sur d'autres réseaux de connaissance. La seconde étape de modélisation consiste à *représenter* le phénomène sous la forme d'un réseau, c'est-à-dire à le mesurer d'une certaine façon pour obtenir l'objet réseau à proprement parler. Ces deux étapes sont résumées dans le schéma de la figure 1-4. Elles font parties intégrantes du modèle de réseau. Il est important de ne pas oublier ou négliger la façon dont elles ont été réalisées, pour donner du sens aux analyses effectuées ensuite. Lorsque l'on retourne au phénomène initial, la représentation permet une traduction correcte des résultats obtenus sur le réseau : un laboratoire densément connecté sera un

laboratoire dont la plupart les doctorants ont réciproquement déclarés qu'ils se connaissaient. L'abstraction rappelle la façon dont le phénomène a été pensé et modélisé : on a étudié les relations de connaissance, pas d'amitié. Voyons maintenant comment ce type de modèle s'applique aux données biologiques.

1.3. La comparaison de séquences

Lorsqu'un biologiste récupère un fichier de gènes issus d'un séquençage à haut-débit, il est comme un archéologue qui recevrait un lot de parchemins écrits dans une langue inconnue. S'il regarde une séquence isolée, il sera bien incapable d'en deviner la fonction. Il pourra identifier les mots du texte grâce à la correspondance quasi-universelle entre les codons de 3 acides nucléiques et les acides aminés. Il pourra délimiter certains paragraphes types, formés par des structures secondaires de protéines telles que les hélices alpha ou les feuillets béta. Mais le sens particulier, la fonction de la séquence, lui restera inconnue. Pour commencer à la déchiffrer, sa meilleure piste sera de comparer les nouvelles séquences à d'autres, dont les fonctions sont déjà identifiées. Si l'archéologue ne s'intéresse pas au sens des textes, mais qu'il cherche juste à les classer, à comprendre les liens qu'ils entretiennent entre eux et les mécanismes ayant produit leur diversité, il devra de même entreprendre un travail comparatif : entre plusieurs manuscrits du même lot ou entre manuscrits issus d'époques ou de sociétés différentes. De la même façon, la seule information accessible au biologiste de l'évolution vient de la comparaison des objets qu'il étudie. Une séquence seule ne porte pas d'information évolutive. L'étape préalable à toute analyse d'un jeu de séquences génétique est donc de les comparer, entre elles ou à des séquences de référence.

1.3.1. Les scores d'alignements reposent sur un modèle d'évolution

De nombreuses méthodes existent pour comparer les textes des séquences génétiques. On peut comparer des descriptions globales comme leur longueur, la fréquence de certains mots, la proportion des différentes lettres. Ces méthodes sont employées pour des questions particulières, mais l'on procède plus habituellement à des *alignements*, c'est-à-dire à des mises en correspondance entre les lettres le long des paires de séquences à comparer. L'hypothèse sous-jacente à un alignement est que les textes des séquences ont une origine ancestrale commune, et que leurs différences proviennent des générations de copies imparfaites qui les séparent. On imagine que certains mots ont pu être écrits différemment par les copistes ou remplacés par des synonymes, que des mots ou paragraphes ont pu être ajoutés ou supprimés. Pour rechercher les traces de ces liens ancestraux entre les textes observés aujourd'hui, on modélise leur évolution passée. On suppose qu'il existe des probabilités fixes que telle lettre ait été remplacée par telle autre, ou qu'un fragment de taille donnée ait été ajouté ou retiré. Ce modèle se traduit par un score que l'on attribue à l'alignement d'une paire de séquences : des correspondances entre fragments identiques font monter ce score, des correspondances entre lettres facilement interchangeables sont moins valorisées, tandis que des correspondances entre lettres très différentes, ou avec des trous insérés dans l'autre séquence font baisser le score. En

pratique les paramètres exacts des matrices de similarité entre lettres sont calibrés sur des jeux de séquences dont on connaît l'origine ancestrale commune. On peut utiliser différentes matrices calibrées selon différentes méthodes (p. ex. BLOSUM, PAM, WAG) et sur des jeux de divergence variable (p. ex. BLOSUM 45, 62, 80 pour des séquences avec moins de 45, 62 ou 80% de résidus identiques). La valeur négative d'un trou (*gap*) est souvent modélisée par une fonction affine de sa taille : un coût à l'ouverture, un coût à l'extension. *In fine*, on obtient une formule mathématique qui associe un score à l'alignement d'une paire de séquences.

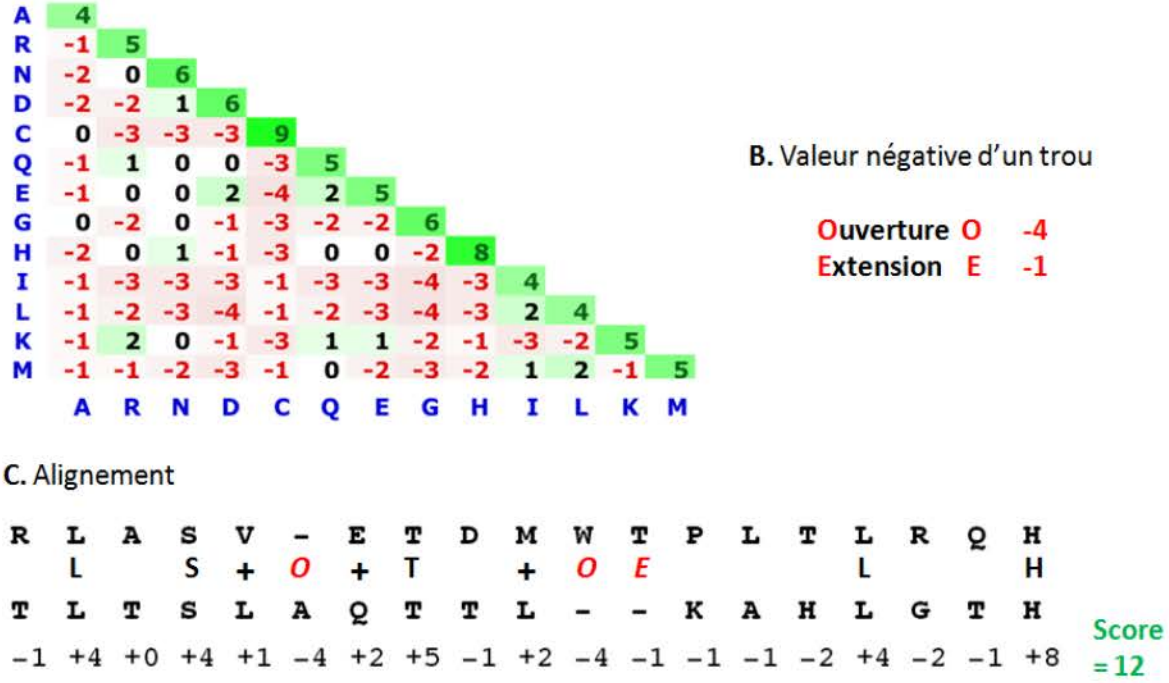


Figure 1-5 : Score d'alignement entre deux séquences
Adapté à partir de http://homepages.ulb.ac.be/~dgonze/TEACHING/stat_scores.pdf

A. Extrait de la matrice de substitution BLOSUM 62.
B. Coefficients de la fonction de coût d'un trou (*gap*)
C. Calcul du score brut d'un alignement par addition des scores de chaque position

1.3.2. BLAST est l'algorithme d'alignement de séquences de référence

On utilise ensuite des algorithmes pour chercher parmi l'ensemble des alignements possibles entre une paire de séquences celui qui a le meilleur score. Il s'agit là d'un problème d'optimisation difficile, car l'espace des correspondances entre séquences est très grand. Certains algorithmes [Needleman, Wunsch, 1970 ; Smith, Waterman, 1981] fournissent une solution exacte à ce problème (un alignement de score maximal), mais ils nécessitent beaucoup de calculs et sont donc relativement lents. Des algorithmes approximatifs beaucoup plus rapides ont étés développés, tel que FASTA [Pearson, 1990] et surtout BLAST (Basic Local Alignement Search Tool) [Altschul et al., 1990 ; Camacho et al., 2009]. Ces algorithmes reposent sur des heuristiques, c'est-à-dire sur des raccourcis intelligents pour effectuer la recherche plus rapidement.

L'algorithme et le programme BLAST ont été publiés en 1990, à une époque où le nombre de séquences disponibles commençait à augmenter rapidement grâce au perfectionnement des techniques de PCR et de séquençage. La puissance de calcul disponible à l'époque était alors bien plus faible qu'aujourd'hui. La capacité de BLAST à produire rapidement des alignements de bonne qualité lui a rapidement assuré un grand succès. Si la puissance de calcul des ordinateurs a beaucoup augmentée depuis (un ipad2 de 2011 est aussi puissant que le meilleur superordinateur de 1994 [Dongarra, Luszczek, 2011]), le nombre de séquences à comparer s'est également considérablement accru (cf. partie 1.1). L'intérêt pour des calculs rapides ne s'est donc pas démenti, et BLAST s'est instauré comme un standard. Cet algorithme est le plus utilisé en bioinformatique avec près de 50.000 citations pour l'article original. Il sera central dans la suite de notre travail. Nous allons donc en présenter quelques principes importants.

1.3.3. Principes de fonctionnement de BLAST

Tout d'abord BLAST est un algorithme d'alignement *local*. Il ne cherche pas à produire d'alignement complet des paires de séquences, mais à trouver des paires de régions ayant des scores d'alignement maximaux. Cela lui permet d'aligner des séquences qui ne sont similaires que sur une partie de leur longueur, soit que les parties restantes des séquences ont trop divergé, soit qu'elles n'ont pas la même origine ancestrale. Ce point sera très important pour notre étude de l'évolution combinatoire des gènes.

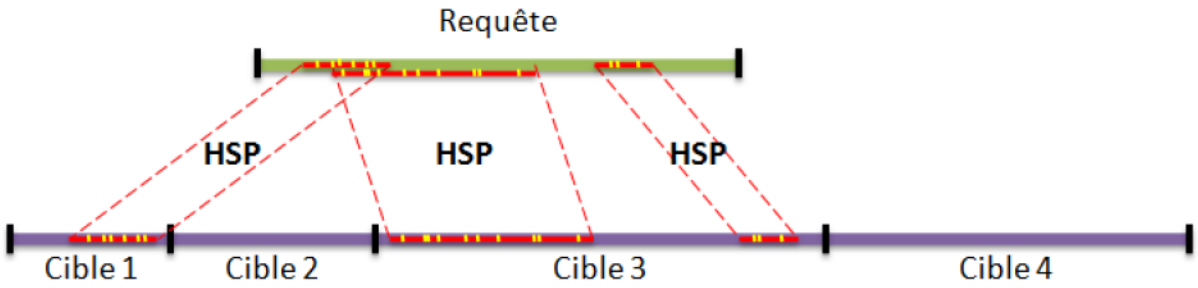


Figure 1-6 : Recherche d'alignements locaux avec BLAST
BLAST initialise des graines d'alignements locaux entre une séquence requête et des séquences cibles, à partir de petits mots de k-lettres (points jaunes). Il étend et joint ces alignements, jusqu'à obtenir des alignements localement optimaux (**HSP**), puis retourne ceux de scores supérieurs à un seuil défini par l'utilisateur.

Lors d'une recherche, BLAST distingue des séquences *requêtes* et des séquences *cibles*. Cette distinction a du sens pour un utilisateur qui cherche à aligner une séquence requête contre les séquences cibles d'une base de données. Si en revanche l'utilisateur a en tête une recherche symétrique, il doit réaliser que l'esprit de la recherche ne l'est pas et effectuer sa recherche dans les deux directions.

BLAST procède en deux étapes. La première étape consiste à initialiser des graines d'alignement à partir de mots de k lettres : par défaut k=3 pour les séquences d'acides aminés et k=11 pour les séquences d'acides nucléiques. Il identifie pour cela tous les mots de k lettres identiques ou très similaires entre la séquence requête et les séquences cibles. La

seconde étape consiste à étendre ces graines d'alignement, en mesurant l'augmentation ou la diminution du score de similarité. La recherche d'extension est stoppée lorsque le score décroit trop fortement, et l'alignement de meilleur score obtenu au cours de l'extension est conservé. Ces alignements localement optimaux sont appelés *HSP* (High-scoring Segment Pair) dans le jargon de BLAST (Figure 1-6). BLAST retourne en résultat l'ensemble des HSP dont le score (mesuré en terme de E-value, cf. ci-dessous) est meilleur qu'un seuil fixé par l'utilisateur. Deux séquences peuvent ainsi être alignées le long de plusieurs HSP distincts.

1.3.4. Système de score employé par BLAST

Le calcul du score brut d'alignement de BLAST **S** dépend de la matrice de similarité et du coût d'ouverture des gaps employés, tel qu'exposé dans la partie 1.3.1. Tel quel, il n'est pas comparable entre recherches et ne permet pas d'estimer la signification statistique d'un HSP. Ce score est donc normalisé en étudiant la distribution des scores de HSP entre des protéines aléatoires. Le modèle de protéines aléatoires employé consiste simplement à *tirer chaque acide aminé indépendamment, selon la distribution moyenne dans les séquences* [Altschul, Gish, 1996]. Les scores bruts des HSP entre de telles protéines aléatoires suivent alors une loi de distribution connue, caractérisée par la longueur des protéines **m** et **n** et les caractéristiques de la recherche[1]. Il est ainsi possible de calculer la probabilité d'obtenir un score au moins aussi bon « au hasard » (P-value), selon ce modèle de protéine aléatoire.

La qualité d'un HSP est généralement décrite par sa E-value **E**, définie comme le nombre de HSP avec un score au moins aussi bon que **S** attendus (**E**xpected) pour une recherche entre 2 protéines aléatoires de mêmes tailles. La E-value est donnée par la formule $\mathbf{E = Km*n\, e^{-\lambda S}}$. Elle augmente proportionnellement à la taille des séquences comparées, car dans une séquence aléatoire deux fois plus longue, on peut s'attendre à trouver 2 fois plus de HSP avec un score supérieur à S. La E-value décroit exponentiellement avec le score S. Cette décroissance exponentielle se comprend mieux en termes de probabilité[2]. Un HSP de score *S1+S2* correspond à deux HSP consécutifs de scores *S1* et *S2*, sa probabilité est donc le produit des probabilités de HSP de score *S1* et *S2*[3]. Cette multiplication des sous-scores explique la décroissance exponentielle de la E-value avec le score (car $e^{-\lambda S1} * e^{-\lambda S2} = e^{-\lambda(S1+S2)}$). En pratique, le score d'alignement est le déterminant principal de la E-value.

Une recherche d'alignements locaux par BLAST s'effectue habituellement contre une base de données *cible* comprenant plusieurs séquences. Pour calculer la E-value, BLAST considère la base de donnée comme une longue séquence de taille **n**. La E-value d'une HSP doit donc se comprendre dans son contexte, c'est-à-dire dans le cadre d'une recherche contre une base de donnée d'une certaine taille. Une autre transformation du score brut est

---

[1] Ces caractéristiques de recherche sont résumées par 2 valeurs de normalisation : **λ** pour le système de score ; **K** pour la taille de l'espace de recherche, qui varie avec la distribution moyenne d'acides aminés dans les séquences.

[2] La E-value est quasiment égale à la P-value pour des petits scores, cf. plus bas

[3] De la même façon : la probabilité d'obtenir 20 six consécutifs en lançant des dés, est le produit de la probabilité d'en obtenir 12 et de la probabilité d'en obtenir 8.

fournie par BLAST. Il s'agit du bit score[1] **S'** qui permet des comparaisons entre recherches effectuées avec différents systèmes de score, sans prendre en considération la longueur des séquences comparées. Ce score est exprimé dans une échelle logarithmique de base 2, d'où le nom *bit* qui réfère à une unité d'information. Une augmentation de 1 du bit score correspond à une division par 2 de la E-value.

Le bit-score est plus pratique que la E-value à plusieurs égards. Tout d'abord, il augmente avec la qualité d'une HSP tandis que la E-value décroit de façon contre-intuitive. Le bit score est généralement compris entre 0 et quelques centaines. Il s'approxime donc par un chiffre entier, tandis que la E-value s'exprime entre 0 et 1 et s'approxime par une puissance négative de 10. Si la variation exponentielle de la E-value avec le score à un sens statistique contre un modèle de protéine aléatoire, la variation du bit score est plus naturelle pour comparer la qualité de deux HSP. Enfin le bit score dépend uniquement de la qualité de la HSP, et pas de la longueur des séquences comparées, ce qui permet de comparer des HSP obtenus indifféremment contre des petites ou des grandes bases de données. De plus il n'est pas forcément pertinent de réduire la valeur attribuée à une HSP selon qu'elle ait été obtenue dans une petite ou dans une grande séquence. Malgré cela, l'usage dans la littérature est d'employer la E-value comme échelle de référence pour les recherches BLAST. On utilise parfois la transformation -log(E-value), qui a les premiers avantages du bit score, mais qui dépend de la longueur des séquences comparées.

1.3.5. Le choix d'un seuil sur la E-value n'a pas de sens statistique

La E-value (nombre attendu de HSP aléatoires de score >= S) ne devrait *a priori* pas être confondue avec la P-value (probabilité d'obtenir une HSP aléatoire de score >= S). Mais ces notions sont en fait très proches, et sont quasiment égales pour les petites valeurs qui nous intéressent (<0,1)[2]. Si l'on raisonnait en termes statistique habituels, on emploierait donc un seuil de 0,05 lors de l'alignement d'une séquence requête contre une base de données, pour limiter les chances d'obtenir des HSP dus à la pure variation aléatoire (1 chance sur 20). En pratique, les seuils employés sont beaucoup plus restrictifs. On considère rarement qu'un HSP est de bonne qualité lorsque sa E-value est supérieure à 10^{-5}. Des seuils encore plus restrictifs sont même souvent employés (10^{-10} voire 10^{-20}), et d'autres indicateurs de la qualité d'une HSP sont considérés en complément, en particulier le pourcentage de résidus identiques dans l'alignement.

Pourquoi utiliser des seuils de E-value si stricts ? Ce n'est pas pour éviter un biais de comparaisons multiples, qui ne produirait en théorie qu'une HSP aléatoire pour 1/(seuil E-value) séquences requêtes. D'ailleurs les E-value ne sont généralement pas corrigées en fonction du nombre de requêtes. La raison pratique est que la similarité entre deux séquences est déjà bien faible à l'œil pour une E-value de 10^{-5} : soit l'alignement

[1] Le bit score est définie par la formule $S' = (\lambda S - \ln K) / \ln 2$ (transformation affine du score). Il est lié à la E-value par la formule $S' = -\log_2 (E / (m*n))$.

[2] Car $P = 1 - e^{-E}$ et $x \simeq 1 - e^{-x}$ pour $|x| \ll 1$

correspondant est très court, soit il est plein de trous et a peu de résidus conservés. La raison théorique est que le modèle de protéines aléatoires employé pour calculer les scores BLAST n'est pas réaliste. Les protéines sont en réalité soumises à de nombreuses contraintes fonctionnelles, qui rendent impossibles la plupart des protéines aléatoires du modèle. Les acides aminés successifs ne sont pas indépendants car leurs propriétés physico-chimiques sont corrélées. Les protéines ont divers niveaux de structures, qui sont sélectionnés au cours des générations. L'espace que peuvent explorer les protéines réelles est donc beaucoup plus restreint que celui des protéines aléatoires du modèle de BLAST. Une paire de protéines réelles sans origine ancestrale commune peut donc potentiellement produire des HSP de faible E-value.

Il n'existe pas de modèle satisfaisant des protéines réelles, et donc pas d'arguments autres qu'empiriques pour choisir un seuil de E-value. Si l'on effectue une analyse à petite échelle, il est possible de privilégier la sensibilité de la recherche pour obtenir un maximum de résultats, quitte à éliminer les résultats douteux lors d'un examen individuel. En revanche lors d'analyses automatisées sur de grands jeux de données, il est gênant de ne pas avoir de critère de référence. Le seuil est souvent déterminé arbitrairement, par expérience ou pour les besoins de l'analyse subséquente, par exemple des HSP bien conservés pour réaliser des alignements multiples, ou des HSP en nombre raisonnable pour réaliser les calculs coûteux qui suivent.

1.4. Similarité, homologie et convergence évolutive

L'absence de référence claire pour interpréter un score d'E-value est liée à un problème traditionnel en évolution, qui est de déterminer uniquement à partir du texte de deux séquences si elles ont - ou non - une origine ancestrale commune. Cette question sera centrale dans notre analyse des réseaux de similarité de séquences. Deux séquences, ou deux régions de séquences, qui proviennent d'une même forme ancestrale par une succession de réplications, éventuellement imparfaites, sont dites homologues. L'idée initiale est qu'une similarité entre deux séquences est la trace d'une origine ancestrale commune. L'idée « un score similarité élevé indique une origine ancestrale commune » tient-elle ? La réponse courte à cette question est : non à cause de la convergence évolutive, mais l'on peut se prémunir des contre-exemples identifiés.

1.4.1. La convergence évolutive des séquences génétiques

De la même façon que 2 particules microscopiques dérivant selon un mouvement aléatoire dans un fluide peuvent être amenées par hasard à se trouver à petite distance l'une de l'autre, 2 séquences génétiques dérivant aléatoirement suite à des copies imparfaites dans l'espace des séquences peuvent être amenées à se ressembler. C'est précisément cette probabilité d'un niveau de ressemblance donné d'intervenir par le seul jeu du hasard qui est estimée par le modèle statistique de BLAST, corrigée par le fait que l'on regarde la proximité d'1 séquence à *N* autres et que pour une distribution moyenne donnée des lettres, l'espace exploré est plus réduit que l'ensemble des textes possibles. Pour rendre peu probable la considération de tels HSP similaires par le seul jeu du hasard, il suffit d'utiliser un seuil de probabilité faible. L'espace des séquences est extrêmement vaste (même après correction par la fréquence des acides aminés), ce qui explique que même une similarité faible entre deux séquences (par exemple d'E-value 10^{-5}) soit considérée comme extrêmement peu probable si elles sont parties à l'origine de régions indépendantes de l'espace des séquences. Deux protéines non-homologues ne peuvent donc pas devenir similaires si elles évoluent au sens du modèle de BLAST, c'est-à-dire par substitution aléatoires et indépendantes des sites.

Deux séquences sans origine ancestrale commune peuvent en revanche converger dans l'espace des séquences, et finir par se ressembler, si elles sont soumises à des biais évolutifs similaires. De tels biais existent au niveau de la réplication, lorsqu'un mécanisme moléculaire engendre les mêmes variations dans des séquences différentes. Le glissement de la polymérase peut par exemple induire des répétitions simples (p. ex. ATATATATAT), puis amplifier celles qui existent déjà. De tels biais existent également au niveau de la sélection des séquences. Des contraintes identiques sur la composition, la structure ou la fonction des protéines peuvent induire une sélection de certains variants, suffisamment importante pour que le texte de leurs séquences devienne en partie similaire au cours des générations. On parle de convergence évolutive. Il est difficile de distinguer en pratique les causes d'une convergence évolutive [Zhang, Kumar, 1997], qui ne sont d'ailleurs pas exclusives les unes des autres.

1.4.2. La convergence évolutive ne concerne pas toutes les catégories structurelles de protéines

La convergence évolutive de la structure primaire des séquences n'est pas un phénomène général. Son importance varie notamment fortement en fonction de la classe structurelle pour les séquences protéiques : fibreuses, membranaires, désordonnées ou globulaires [Wootton, 1994 ; Wong et al., 2010].

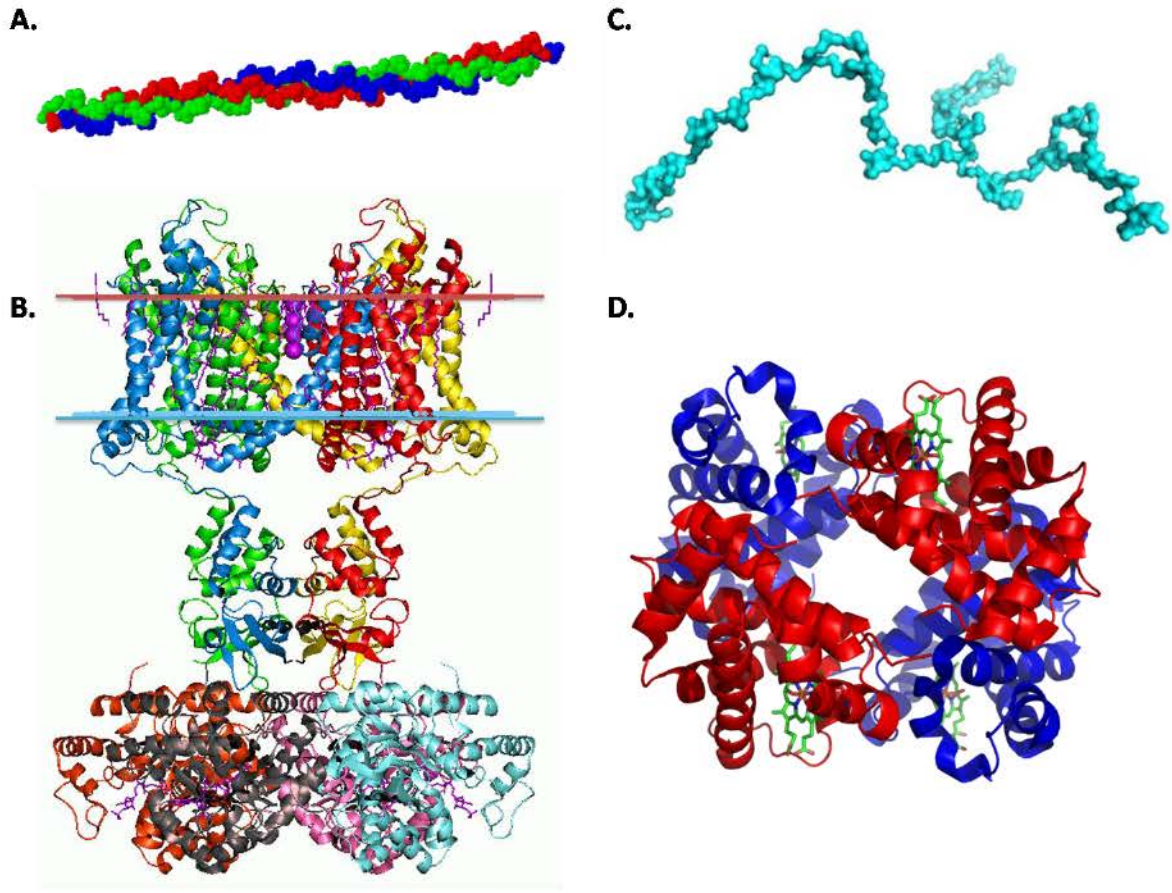


Figure 1-7 : Schéma de protéines des quatre grandes classes structurelles

- **A.** Fibreuse : Triple-hélice d'une molécule de collagène (source)
- **B.** Membranaire : Canal potassique (source). Les limites de la bicouche lipidiques sont indiquées par des lignes rouges et bleues
- **C.** Désordonnée : Juxtanodin (source) en solution
- **D.** Globulaire : Hémoglobine (source)

Les protéines fibreuses (scléroprotéines) sont de longues molécules en forme de filament. Elles sont insolubles dans l'eau et assurent des fonctions de structure dans les cellules. Ces protéines sont souvent répétitives, ce qui viole fortement l'hypothèse d'indépendance des sites du modèle aléatoire de protéine, et peut engendrer des scores de similarité élevés entre séquences non-homologues.

Les protéines membranaires sont liées à la membrane, soit en surface, soit en la traversant. Elles assurent diverses fonctions liées à leur localisation : transport transmembranaire, signalisation, adhésion cellulaire. Les régions intra-membranaires de ces protéines sont soumises à des pressions de sélection très spécifiques, du fait de leur

environnement chimique et de leur fonction, qui induisent fréquemment des convergences évolutives.

Les protéines intrinsèquement désordonnées sont caractérisées par l'absence de structure tridimensionnelle stable. Elles assurent notamment des fonctions de signalisation cellulaire et de reconnaissance moléculaire. Des régions intrinsèquement désordonnées existent également dans des protéines ordonnées, où elles servent de jonctions entre régions de structure globulaire ou membranaire. Les régions désordonnées des protéines sont souvent enrichies en quelques acides aminés. Certaines contiennent par exemple des segments riches en proline. Cet enrichissement ne respecte pas les proportions moyennes utilisées dans le modèle statistique de BLAST, ce qui réduit localement l'espace d'évolution des protéines et résulte en des scores de similarité élevés entre protéines non-homologues.

Par ailleurs, les protéines intrinsèquement désordonnées contiennent beaucoup de sites fonctionnels courts, comme les Short Linear Motifs (SLiM). Ces sites sont impliqués dans la reconnaissance ou l'attachement à d'autres molécules. Ils évoluent souvent de façon convergente pour effectuer des tâches similaires dans des protéines différentes. Leur fonction est directement codée dans leur séquence primaire, indépendamment de la structure tridimensionnelle de la protéine. Ils sont cependant de trop petite taille (3 à 11 acides aminés) pour être l'origine unique de HSP non-homologues de score élevé [Edwards et al., 2007].

Les protéines globulaires sont de forme sphéroïde et solubles dans l'eau. Cette classe très large comprend des protéines régulatrices, de transport, de structure, de signalisation et des enzymes. Les protéines globulaires ne semblent généralement pas présenter d'évolution convergente au niveau des séquences. On peut expliquer cela par l'absence des conditions identifiées dans les autres catégories structurelles de protéines. Le repliement tridimensionnel de la séquence d'acides aminés joue notamment un rôle très important dans la fonction de ces protéines. Ainsi malgré l'évolution convergente très bien documentée des sites actifs de certaines enzymes [Russell, 1998 ; Gherardini et al., 2007], ceux-ci se trouvent à des positions éloignées le long de la protéine. En plus de leur trop petit nombre, la convergence d'acides aminés éloignés ne résulte pas en des scores de similarité élevés. Plus généralement, une évolution convergente des structures tridimensionnelles n'implique pas la convergence des séquences primaires. Ainsi la notion d'évolution convergente dépend de la mesure de distance employée sur les objets étudiés.

1.4.3. Détecter et masquer les régions sensibles à la convergence évolutive

Si l'on cherche à identifier uniquement les régions homologues entre séquences avec BLAST, comment se prémunir des scores élevés produite par évolution convergente ? Avant toute chose, il est possible de filtrer les HSP identifiées selon plusieurs critères : E-value, pourcentage d'identité, proportion de sites identiques, longueur des alignements locaux. Cela prémunit contre le fait que, globalement, les protéines sont plus similaires entre elles

que le suppose le modèle de BLAST, mais ne permet pas d'éviter les cas de forte évolution convergente présentés ci-dessus. La stratégie générique pour éviter ces cas consiste à identifier les régions concernées dans les séquences, puis à modifier la façon dont elles sont considérées par l'algorithme d'alignement.

Une majorité des cas de forte évolution convergente est le fait de régions enrichies en certaines lettres ou contenant des répétitions simples [Frith et al., 2010], par exemple AAACAAAAAGAA, ATATATATAT, PPCDPPPKPPP. Ces régions sont dites de faible complexité, car elles sont très redondantes et contiennent peu d'information au sens de l'entropie de Shannon [Shannon, 1948]. Différents programmes ont été développés pour identifier ces régions, qui fonctionnent sur les séquences nucléiques (DUST), protéiques (SEG), ou les deux : TRF [Benson, 1999], TANTAN [Frith, 2011a]. Les régions non globulaires des protéines qui évoluent de façon convergente ne sont pas toujours bien détectées par les méthodes précédentes, qui n'identifient que les cas les plus évidents de complexité faible. Elles échouent notamment à éliminer de nombreuses régions transmembranaires, les hélices coiled-coil, les segments de peptides signal, qui sont très propices à des similarités non-homologues. Il faut employer des outils de détection spécifiques à chaque cas pour les identifier.

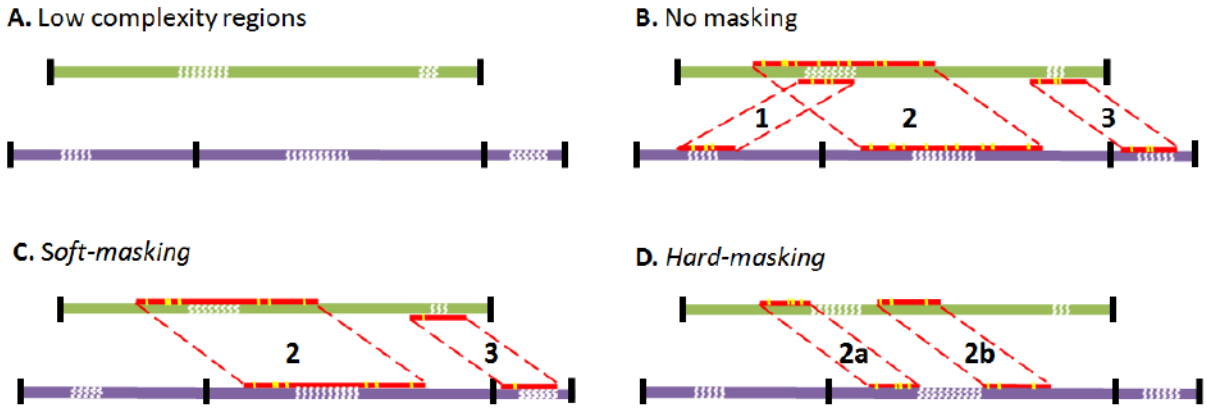


Figure 1-8 : Masquage des régions de faible complexité

- **A.** Identification des régions de faibles complexités (blanc), dans la séquence requête et les séquences cibles.
- **B.** Trois HSP sont identifiées sans masquage des régions de faible complexité.
- **C.** En appliquant un masquage doux, les graines d'alignements sont ignorées dans les régions de faible complexité. Le HSP 1 entre régions de faible complexité n'est alors pas trouvé. Les HSP 2 et 3 sont conservés par extension à partir de graines dans des régions voisines.
- **D.** En appliquant un masquage dur, les régions de faible complexité sont ignorées lors de l'alignement. Le HSP 3 est éliminé car il n'atteint pas un score suffisant. Deux HSP sont obtenus pour le HSP 2, de chaque côté de la région de faible complexité.

Les régions identifiées comme sujettes à convergence sont ensuite *masquées* par BLAST, selon deux stratégies générales. Le masquage fort (*hard-masking*) consiste à remplacer ces régions par des lettres joker (respectivement 'N' et 'X' pour les séquences nucléiques et protéiques) et à ne pas les prendre en compte lors des alignements. Le masquage doux (*soft-masking*) met ces régions en lettres minuscules. Elles sont alors ignorées lors de la phase d'initialisation des graines d'alignement par des mots de longueur fixe, mais considérées lors

de la phase d'extension des alignements. Les régions de faible complexité peuvent ainsi être incluses dans des alignements locaux, par extension depuis des zones de complexité plus élevée. Cela permet une meilleure détection des zones homologues et un score de similarité plus précis.

Cette méthode propose un bon compromis, mais n'évite pas certains alignements fallacieux, dominés par des régions de faible complexité augmentant considérablement le score. Un nouveau type de masquage (*gentle-masking*) a récemment été proposé [Frith, 2011b]. Comme le soft-masking, il autorise l'extension à travers des zones de faible complexité, mais il ne prend en compte que les résidus mésappariés dans ces zones : le score ne peut donc que y baisser. Les régions homologues sont ainsi mieux délimitées, tout en évitant une augmentation artificielle du score dans les régions de faible complexité. Cette procédure de masquage n'est pas encore implémentée par BLAST. Il est donc recommandé d'utiliser le hard-*masking* si l'on veut s'assurer de l'homologie le long des HSP, et d'appliquer ce masquage aux séquences requêtes et cibles. Par défaut, BLAST ne masque pas les séquences cibles. Il masque les séquences nucléiques requêtes avec DUST, selon le *hard-masking*. Sa procédure par défaut sur les séquences protéiques était auparavant la même en utilisant SEG. Les dernières versions utilisent désormais une autre approche par défaut, qui consiste à corriger localement les scores pour des biais de composition en acides aminés.

1.4.4. Alternatives à la comparaison de séquences avec BLASTP

Nous avons décrit une procédure générale pour comparer des séquences et éliminer les similarités qui ne proviendraient pas d'une origine ancestrale commune. Avant de s'interroger sur la façon dont nous allons exploiter ces relations entre séquences, il convient de préciser qu'il est parfois possible d'utiliser d'autres informations pour étudier leurs histoires communes. Les séquences que nous allons considérer sont des gènes, qui auront souvent été délimités dans des séquences plus longues (*contig*, génome). Ces gènes sont donc en relation physique avec d'autre gènes, de par leur succession le long d'un même support, dont ils partagent une part de l'histoire évolutive. Nous n'exploiterons pas directement cette information contextuelle, qui n'est d'ailleurs pas disponible si l'on étudie des données de transcriptomique ou de métagénomique.

Nous avons par ailleurs insisté sur le fait que nous comparions les *textes* des séquences génétiques. Il serait envisageable de comparer plutôt leur conformation tridimensionnelle, locale (structure secondaire) ou globale (structure tertiaire). La structure 3D d'une protéine est en effet bien mieux conservée que sa séquence primaire. Cependant il n'est pas possible d'acquérir directement de telles structures en grande quantité. Il est également difficile de prédire ces structures à partir des séquences (problème qui n'a d'ailleurs pas toujours une solution unique puisque les protéines peuvent changer de conformation 3D).

Il est en revanche courant d'étudier la structure en *domaines* des protéines. Cela consiste à identifier les domaines par des alignements locaux contre des domaines de référence, puis de comparer la composition en domaines des protéines. D'une part ce

problème est d'une nature un peu différente, similaire à l'étude de la composition en gènes des génomes, d'autre part il suppose un modèle d'organisation des protéines en domaines, dont nous allons nous passer car il recoupe l'objet de notre étude (évolution combinatoire au sein des gènes).

Finalement, il nous reste pour les séquences codant pour des protéines l'alternative de comparer les versions en acides nucléiques, ou les versions traduites en acides aminés. Nous choisirons classiquement de comparer les séquences protéiques, car elles sont mieux conservées que les séquences nucléiques, ce qui est préférable à la grande échelle évolutive à laquelle nous travaillerons. Elles permettent de retrouver des relations plus anciennes entre séquences. Nous utiliserons donc essentiellement le programme BLAST**P** de la suite BLAST.

Des programmes tels que PSI-BLAST permettent de détecter des relations entre séquences plus anciennes encore que celles détectées par BLASTP. PSI-BLAST identifie pour cela une famille de gènes similaires au gène requête, détermine un modèle statistique de cette famille, et s'en sert pour identifier de nouvelles séquences correspondants au modèle. Cette méthodologie repose donc sur la construction implicite de famille de gènes, ce qui recoupe l'objet de notre étude. Nous nous en tiendrons donc à la comparaison paire à paire des séquences par alignements locaux, afin de limiter les aprioris théoriques sur l'évolution des séquences.

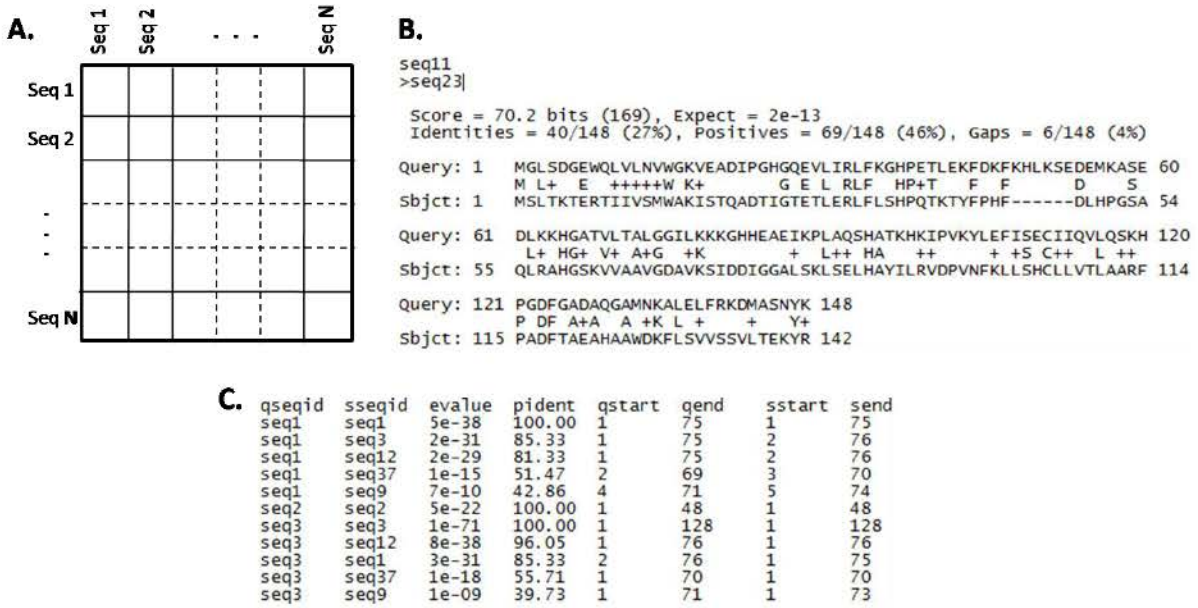
1.5. Réseaux de similarité de séquences

Lorsque l'on étudie un jeu de séquences génétiques, nous avons vu qu'une première étape est généralement de les comparer toutes 2 à 2 avec l'algorithme BLAST. Ce qui est fait ensuite dépend a priori des questions de recherche, mais dans une majorité de cas, un biologiste cherchera à rassembler ses séquences par groupes. S'il a par exemple procédé à un séquençage métagénomique d'ARN 16s pour estimer la diversité bactérienne dans un milieu, il classifiera les séquences obtenues en groupes de similarité élevée afin de définir des Unités Taxonomiques Opérationnelles. S'il cherche à étudier l'histoire évolutive des gènes, il produira des groupes de gènes homologues, et s'en servira pour réaliser des arbres phylogénétiques. S'il veut annoter fonctionnellement des séquences à partir de séquences de référence, il pourra former des groupes de gènes orthologues, et supposer que les gènes d'un même groupe ont des fonctions proches.

Dans chacun de ces cas, les groupes de séquences résument en un certain sens l'information de similarité. Ils fournissent une classification de référence pour la suite des analyses, sur laquelle on ne revient généralement pas. Les méthodes employées pour réaliser ces classifications ont donc un rôle central dans l'analyse des séquences. Pour chacun des types de regroupements, de nombreux programmes ont été développés. Cependant les algorithmes sous-jacents, et leurs paramètres, ne sont pas évidents à comprendre. De plus, il n'existe pas de consensus sur les meilleures méthodes à employer, et il est difficile pour un utilisateur de vérifier la qualité des regroupements effectués. Ainsi cette étape centrale est souvent réalisée de façon arbitraire, avec des programmes employés comme des boîtes noires, et sans possibilité de vérifier les résultats obtenus qui seront la base des analyses ultérieures. Nous allons introduire les réseaux de similarité de séquence en montrant qu'ils permettent notamment de mieux comprendre la construction des groupes de séquences, ce qui justifie d'ailleurs leur emploi par de nombreux biologistes.

1.5.1. Construction d'un réseau de similarité de séquences

Un réseau de similarité de séquences est un réseau dont les sommets sont des séquences, et dont les arêtes représentent une similarité entre paires de séquences. Ceci est l'abstraction d'un réseau de similarité de séquence (cf. figure 1-4 partie 1.2.5), qui en pratique peut être construit de plusieurs façons. La recherche de similarité est en général réalisée avec l'algorithme BLAST, en considérant le jeu de séquences à la fois comme requête et comme cible (Figure 1-9.A). BLAST renvoie alors tous les alignements locaux de similarité élevée qu'il trouve entre paires de séquences. On n'étudiera pas le détail des alignements obtenus (Figure 1-9.B), ce qui devient rapidement impossible dès que le nombre de séquences augmente, mais plutôt des descripteurs de ces alignements tels que la E-value et le pourcentage d'identité (Tableau 1). Typiquement, on utilisera la sortie tabulaire de BLAST (option "-outfmt 6"), dans laquelle chaque ligne décrit un HSP (Figure 1-9.C). Cette sortie peut déjà en tant que telle être interprétée comme un réseau, où chaque ligne est une arête entre une séquence cible et une séquence requête.



C.

qseqid	sseqid	evalue	pident	qstart	qend	sstart	send
seq1	seq1	5e-38	100.00	1	75	1	75
seq1	seq3	2e-31	85.33	1	75	2	76
seq1	seq12	2e-29	81.33	1	75	2	76
seq1	seq37	1e-15	51.47	2	69	3	70
seq1	seq9	7e-10	42.86	4	71	5	74
seq2	seq2	5e-22	100.00	1	48	1	48
seq3	seq3	1e-71	100.00	1	128	1	128
seq3	seq12	8e-38	96.05	1	76	1	76
seq3	seq1	3e-31	85.33	2	76	1	75
seq3	seq37	1e-18	55.71	1	70	1	70
seq3	seq9	1e-09	39.73	1	71	1	73

Figure 1-9 : Comparaison de séquences avec l'algorithme BLAST

A. Matrice schématique de la comparaison de toutes les séquences d'un jeu de données contre-elles mêmes. Les comparaisons sont effectuées dans les deux sens car le résultat de BLAST n'est pas symétrique.

B. Détail d'un fichier de sortie complet de BLAST avec les alignements

C. Détail d'un fichier de sortie BLAST tabulaire, avec uniquement quelques descripteurs par lignes

Descripteurs d'alignements locaux fournis par BLAST	Abbreviations
E-value	**evalue**
Bit score	bitscore
Longueur	length
Nombre d'appariements identiques	nident
Pourcentage d'appariements **ident**iques	**pident**
Début de l'alignement sur la séquence requête (**q**uery **start**)	**qstart**
Fin de l'alignement sur la séquence requête (**q**uery **end**)	**qend**
Début de l'alignement sur la séquence cible (**s**ubject **start**)	**sstart**
Fin de l'alignement sur la séquence cible (**s**ubject **end**)	**send**
Nombre d'appariements de score positifs	positive
Pourcentage d'appariements de score positifs	ppos
Nombre de mésappariements	mismatch
Nombre de trous ouverts	gapopen
Nombre de trous	gaps
Identifiant de la séquence requête (**q**uery **seq**uence **id**entification)	**qseqid**
Identifiant de la séquence cible (**s**ubject **seq**uence **id**entification)	**sseqid**

Tableau 1 : Descripteurs d'alignements locaux fournis par BLAST
Ce tableau présente une sélection de descripteurs d'alignements locaux fournis par BLAST, avec leur abréviation (l'aide de BLAST fourni une liste complète). Nous employons essentiellement les descripteurs dont l'abréviation est en gras.

La figure 1-10 est une représentation graphique du réseau d'une sortie brute de BLAST, et de sa conversion en un réseau de similarité de séquence. Dans le réseau brut, les arêtes

sont orientées, certaines arêtes pointent d'une séquence vers elle-même (boucle), et certaines arêtes parallèles ont les mêmes extrémités.

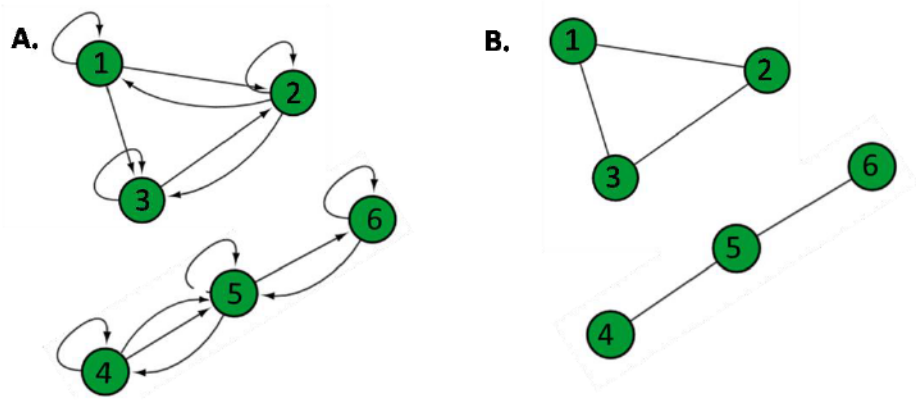


Figure 1-10 : D'une sortie BLAST à un réseau de similarité de séquences

A. Représentation directe d'une sortie BLAST tabulaire par un réseau, en considérant chaque ligne comme une arête du réseau. Il s'agit d'éliminer les boucles, et les arêtes multiples pour obtenir un réseau de similarité de séquences.

B. Réseau de similarité séquences correspondant après symétrisation, élimination des boucles, et élimination des arêtes multiples.

Les boucles apparaissent car l'on compare toutes les séquences à elles mêmes. Elles ne sont pas informatives et doivent être supprimées. Les arêtes sont orientées parce que la comparaison de BLAST n'est pas symétrique entre une séquence cible et une séquence requête. Les comparaisons effectuées dans les deux sens ne donnent pas exactement les mêmes résultats, en termes de limites des régions alignées, d'alignements, et de scores de similarité. Il est même possible qu'une arête soit présente dans un sens mais pas dans l'autre, si les E-value associées aux comparaisons se trouvent de part et d'autre du seuil limite donné en argument à BLAST. Cette asymétrie n'a pas de sens biologique, et il faut donc symétriser le réseau pour obtenir des arêtes non-orientées. Une façon naturelle de procéder est de conserver parmi les deux alignements entre deux séquences celui de meilleur score. On annote alors l'arête non-orientée par les descripteurs de cet alignement (pourcentage d'identité, longueur, etc.).

Enfin, il peut y avoir plusieurs HSP à des endroits distincts le long d'une paire de séquences, pour différentes raisons liées à l'évolution des séquences (rythmes de divergence variables, insertion de régions non-homologues) ou à l'algorithme BLAST (exclusion des régions de faible complexité). Diverses méthodes existent pour combiner ces résultats et se ramener à une unique arête. La plus simple, que nous allons employer par défaut, est de ne garder que le HSP de score maximal. Une méthode plus élaborée consiste à combiner les limites des différentes régions alignées en les représentants par une union d'intervalles, ou à combiner les scores des différentes HSP en un score global entre séquences.

Finalement, on retiendra la procédure suivante pour obtenir un réseau simple à partir d'une sortie BLAST : 1/ Éliminer les boucles (lignes avec une séquence en requête et en cible), 2/ Ne garder que l'arête (la ligne) de score maximal entre chaque paire de séquences. Plusieurs programmes permettent de réaliser cette procédure, tels que EGN [Halary et al.,

2013] ou Pythoscape [Barber, Babbitt, 2012]. Il est sinon possible de la programmer soi-même pour bien en maîtriser les étapes. Le réseau ainsi obtenu est une *représentation* (cf. figure 1-4, partie 1.2.5) quasiment directe de l'information de similarité produite par BLAST.

1.5.2. Visualisation d'un réseau de similarité de séquences

1.5.2.1. Distances dans la projection du réseau

Après avoir comparé des séquences avec BLAST et transformé le résultat en un réseau de similarité de séquence, la première chose que l'on souhaite généralement faire est de le visualiser. Parmi les logiciels de visualisation de réseaux, les plus pratiques sont pour nous Cytoscape [Shannon et al., 2003] et Gephi [Bastian et al., 2009]. Ces deux logiciels acceptent en entrée un réseau décrit par une liste d'arêtes tabulée, format de notre sortie BLAST épurée.

Diverses méthodes d'agencement (*layout*) des nœuds et des arêtes sont proposées par ces logiciels. Il est recommandé d'en essayer plusieurs, et de faire varier leurs paramètres pour obtenir différentes projections d'un réseau. Les layouts Force-Directed, Edge-Weighted Spring Embedded et Organic sont les plus pratiques dans Cytoscape, tandis que le layout ForceAtlas2 est rapide et facilement paramétrable dans Gephi. Le fonctionnement précis de ces layouts diffère, mais ils sont généralement décrits par la même analogie mécanique, d'un réseau comportant des poids sur les nœuds et des ressorts sur les arêtes qu'il s'agit de faire converger vers un état d'énergie minimale.

Certains layouts peuvent prendre en compte une valeur numérique associée aux arêtes, qui module la force de rappel des ressorts. On utilise dans ce cas le Bit score comme valeur numérique ou la transformée logarithmique de la E-value (*-log(E-value)*), plutôt que la E-value qui s'étale sur plusieurs centaines d'ordres de grandeurs. Si elle mérite d'être essayée et peut fournir des résultats intéressants, cette modulation du poids des arêtes n'est pas nécessaire, car la distance entre séquence se reflète déjà dans la topologie du réseau. Ce n'est pas une arête unique qui modifie fortement le placement des nœuds, mais la multiplicité des arêtes entre groupes de séquences similaires qui favorise l'émergence de structures dans la représentation finale.

La visualisation graphique d'un réseau met concrètement en pratique l'idée que l'on n'étudie pas des entités individuelles indépendantes, mais un ensemble structuré de relations. Ainsi la distance entre deux séquences dans une visualisation graphique ne représente *a priori* pas une distance évolutive portée par les arêtes, mais émerge de la prise en compte des similarités entre toutes les séquences. Un résultat frappant obtenu par Atkinson et ses collaborateurs [Atkinson et al., 2009] est que, même en exploitant uniquement la topologie du réseau, les distances 2D entre séquences dans une représentation graphique sont fortement corrélées aux distances en terme de E-value (coefficients de corrélation de 0.83 à 0.94 dans leurs analyses). Ils observent de plus que cette corrélation est peu sensible aux données manquantes, et qu'elle ne dépend pas de la taille du réseau mais davantage du seuil d'E-value employé. Ainsi la visualisation d'un réseau

de similarité de séquences permet effectivement de voir les distances entre séquences étudiées[1].

1.5.2.2. Effet des seuils de similarité sur la structure du réseau

La structure du réseau est fortement influencée par le niveau de similarité requis pour connecter les séquences. Le niveau de similarité de base est donné au moment de la comparaison BLAST, via le choix du seuil d'E-value maximal pour retourner un HSP. Il est possible de diminuer *a posteriori* ce seuil, et de filtrer les HSP sur d'autres valeurs, comme leur pourcentage d'identité ou leur longueur (exprimée en nombre de résidus ou en proportion des séquences couvertes). Ces filtres diminuent le nombre d'arêtes considérées (Figure 1-11) et la densité de connexions dans le réseau. De nouvelles structures locales apparaissent lorsqu'on ne conserve que les connexions entre séquences très similaires, jusqu'à éventuellement déconnecter certaines parties du réseau et créer de nouvelles composantes connexes.

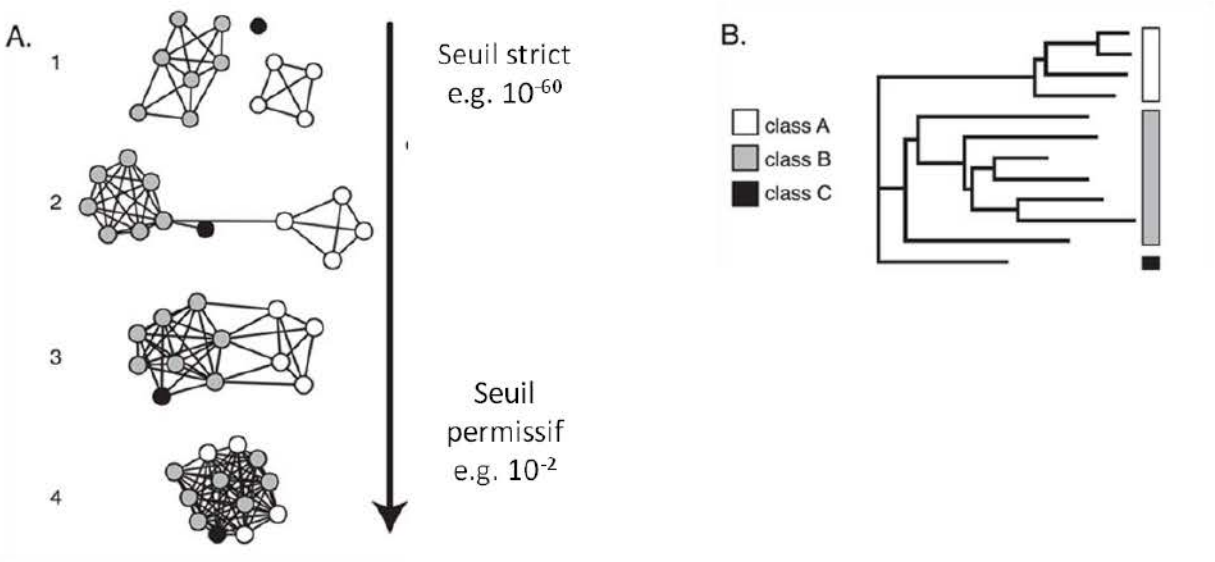


Figure 1-11 : Effet sur la connectivité d'un réseau du seuil sur la E-value
Adapté de Atkinson *et al.* 2009

- **A.** Un réseau de similarité de séquences créé sur trois classes simulées de protéines homologues est représenté à quatre seuils sur la E-value. À un seuil strict, le réseau se décompose en plusieurs composantes connexes; les séquences au sein d'une classe sont très similaires. À un seuil plus permissif, des similarités moins fortes sont ajoutées, et les groupes se connectent.
- **B.** Arbre phylogénétique simulé ayant produit les séquences des réseaux de **A.**

Il est intéressant de comparer les réseaux à différents seuils, en les construisant séquentiellement, ou faisant varier dynamiquement le seuil dans certains logiciels de visualisation (p. ex. en employant la dimension temporelle dans Gephi). Ces seuils peuvent avoir un sens biologique précis, si l'on ne s'intéresse par exemple qu'aux similarités >97 %,

[1] Un autre résultat obtenu par Atkinson et al. est que les distances entre séquences à l'intérieur d'un réseau (longueur du plus court chemin dans le réseau avec une distance de *-log(evalue)* sur les arêtes) sont fortement corrélées aux distances entre séquences dans un alignement multiple.

ou n'être que des limites arbitraires. Les analyses telles que la visualisation qui ne dépendent que de la topologie du réseau (i.e. sans considérer les valeurs portées par les arêtes) méritent d'être menées à différents seuils. Ces analyses multiples sont rapides car il suffit de masquer certaines arêtes sans recalculer le réseau, et facilitées lorsque les outils d'analyse incorporent des méthodes de seuillage.

1.5.2.3. Flexibilité de visualisation d'annotations

Un aspect très fructueux de la visualisation des réseaux, par rapport à leur analyse numérique, est l'exploration interactive des informations sur les séquences, en jouant sur la couleur, la taille et la forme des nœuds et des arêtes. On peut ainsi afficher les fonctions des protéines, leur génome ou leur environnement d'origine, la longueur des régions alignées, ou toute autre annotation spécifique aux données (Figure 1-12), et observer la formation ou non de patrons remarquables. Il est aussi possible de visualiser le résultat de calculs effectués sur le réseau, comme par exemple une classification en familles.

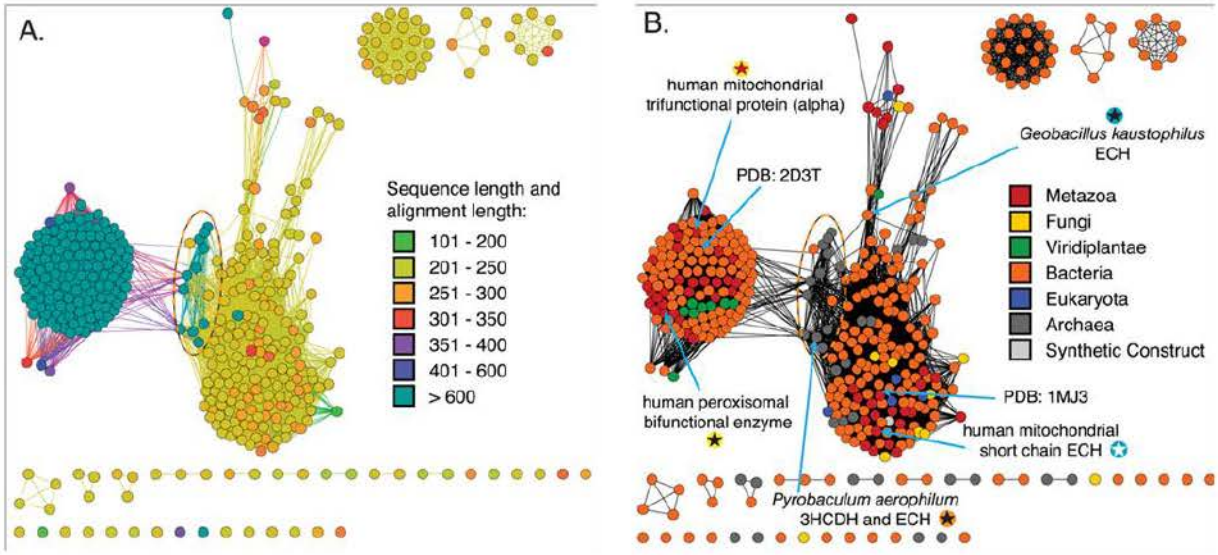


Figure 1-12 : Exploration graphique d'un réseau de similarité de séquences
Visualisation du réseau de similarité de 410 gènes de la famille enoyl-CoA hydratase, seuillé à une E-value de 10^{-50}

- **A.** Les nœuds sont colorés en fonction de la longueur des séquences, et les arêtes en fonction de la longueur de l'alignement.
- **B.** Les nœuds sont colorés en fonction du règne (Champignons, Metazoaires, Plantes vertes) ou du domaine (Bactéries, Eucaryotes, Archées) des organismes portant les gènes

Les réseaux permettent d'étudier des jeux de données de taille considérable. L'étape la plus coûteuse en calculs pour l'analyse d'un réseau est souvent la comparaison des séquences lors de sa création. Cette étape à une complexité proportionnelle au nombre de comparaisons, et donc proportionnelle au carré du nombre de séquences. Ce facteur quadratique correspond à un quadruplement du nombre de comparaisons pour un doublement du nombre de séquences. Cette étape initiale peut donc se révéler difficile avec des jeux de données toujours plus grands. Elle est de toute façon nécessaire et se parallélise aisément sur plusieurs ordinateurs (cf. discussion dans la chapitre 5). Au cours de ma thèse, j'ai ainsi couramment traité des jeux de données comprenant des millions de séquences,

issues de centaines de génomes, sans devoir recourir à des ressources externes à l'équipe ou à des superordinateurs.

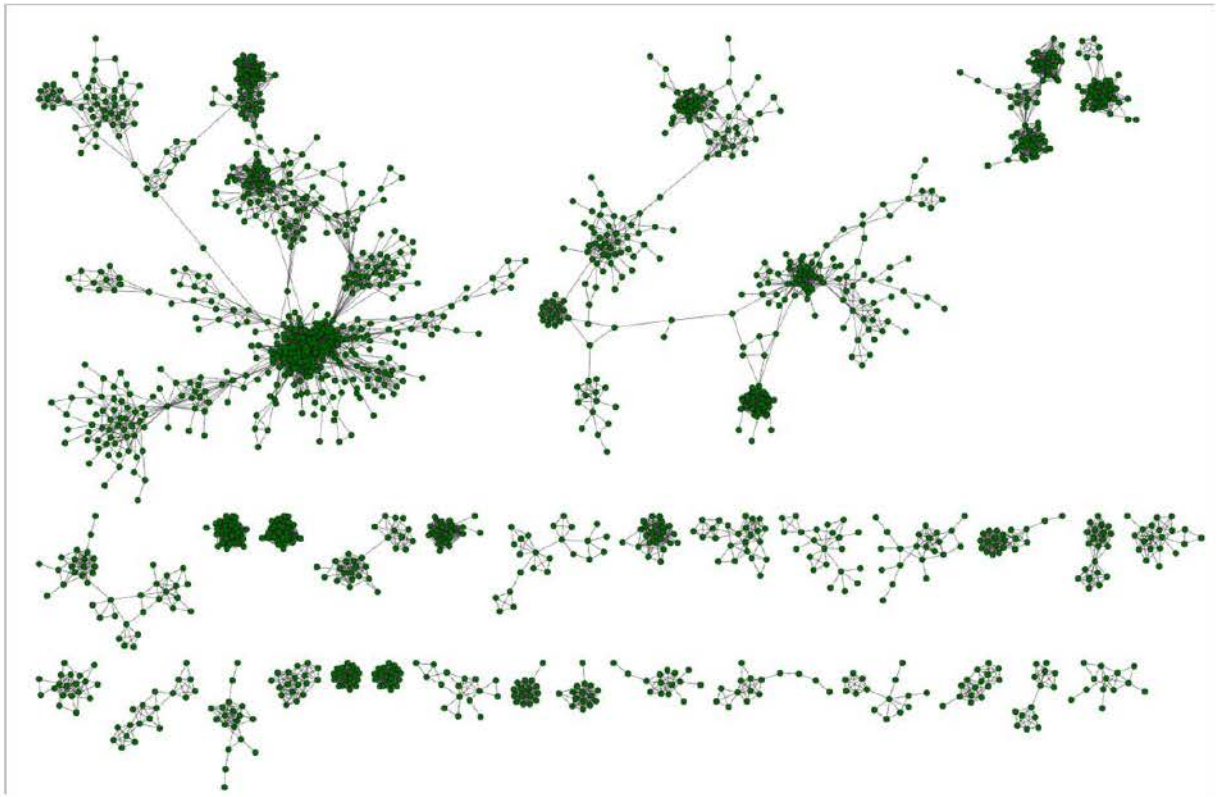


Figure 1-13 : Apparition d'une composante connexe géante
Extrait d'un petit réseau de similarité de séquence, dans lequel apparaissent deux grandes composantes connexes. L'ajout de nouvelles séquences à ce jeu de données a tendance à relier les composantes connexes existantes, jusqu'à former une composante connexe géante qui agglomère la majorité des nœuds et arêtes du réseau.

Une difficulté apparaît cependant pour la visualisation de ces grands réseaux de similarité de séquences. Un de leur trait majeur est en effet la présence d'une composante connexe géante (GCC) qui apparaît avec l'augmentation du nombre de séquences (Figure 1-13)[1]. Selon la diversité des séquences étudiées, cette composante peut en agglomérer plus de la moitié, et contenir plus de 90 % des arêtes du réseau. Il devient alors rapidement impossible de créer en mémoire les dizaines de millions d'objets nécessaires à la visualisation interactive de ces GCC. Nous proposerons une solution à ce problème dans l'article du chapitre 3 (cf. figure 5 de l'article), en divisant le GCC en régions densément connectées, puis en procédant à une visualisation à deux échelles : le réseau des liens entre communautés d'une part, les sous-réseaux internes à ces communautés d'autre part. Cette double échelle permet une visualisation de grande ampleur des relations entre séquences, mais souffre d'une relation moins dynamique, moins intuitive et moins directe avec les données. Il devient alors parfois préférable d'employer une approche exploratoire sur un

[1] Une composante connexe géante apparait dans de nombreux réseaux réels de grande taille [Newman et al., 2001]

sous-ensemble du jeu de données, pour développer des hypothèses et concevoir des mesures que l'on appliquera ensuite sur tout le réseau.

La visualisation d'un réseau de séquences est un moyen rapide d'observer les relations de similarité qu'elles entretiennent, mais ce n'est pas le seul intérêt des réseaux pour les évolutionnistes.

1.5.3. Algorithmes de réseaux pour former des familles homologues

La construction de groupe de gènes homologues est une démarche fréquente, notamment dans l'optique de construire des alignements multiples puis des arbres phylogénétiques. Les différents algorithmes utilisés pour former des groupes de séquences homologues à partir d'une comparaison BLAST s'expriment naturellement à partir d'un réseau de similarité. Les réseaux ont parfois été utilisés pour *concevoir* ces algorithmes, mais sont souvent peu exploités pour les *expliquer*, et rarement utilisés pour *visualiser leurs résultats*. Nous proposons d'illustrer l'intérêt des réseaux de similarité de séquences sur ce problème classique. Nous présenterons ainsi les difficultés que rencontrent ces algorithmes, qui seront notre point de départ pour étudier les phénomènes combinatoires d'évolution des séquences.

1.5.3.1. Les réseaux permettent de construire des familles homologues divergentes

Une hypothèse centrale des méthodes de construction de familles homologues est qu'*une similarité entre deux séquences implique leur homologie*, c'est-à-dire qu'elles ont une origine ancestrale commune. Cette hypothèse est raisonnable dans la mesure où l'on emploie des seuils de similarité suffisamment élevés sur la E-value et le pourcentage d'identité, et que l'on masque les régions propices à la convergence évolutive lors des alignements. En revanche la réciproque de cette hypothèse est fausse : certaines séquences homologues ont trop divergé depuis leur séquence ancestrale commune et ne sont plus similaires. De telles séquences homologues mais non similaires sont appelées des *homologues distants*. Ainsi, deux séquences voisines dans un réseau de similarité de séquences peuvent être considérées homologues ; tandis que deux séquences qui ne sont pas voisines peuvent – ou non – être homologues.

Il est possible de retrouver le lien d'homologie entre des homologues distants, lorsqu'ils sont similaires aux mêmes séquences homologues intermédiaires. Cela correspond dans le réseau à considérer que les voisins (homologues) d'une séquence intermédiaire sont homologues entre eux (Figure 1-14.B). En reproduisant successivement ce principe, on peut retrouver les homologues distants d'une séquence via plusieurs homologues intermédiaires. Un algorithme naturel pour former des familles homologues serait donc de considérer les composantes connexes du réseau de similarité de séquences. Deux séquences sont regroupées dans une même famille si et seulement si elles sont connectées dans le réseau, directement ou par un chemin de similarité. Construire de telles familles homologues divergentes permet d'étudier l'évolution biologique à une plus grande échelle de temps.

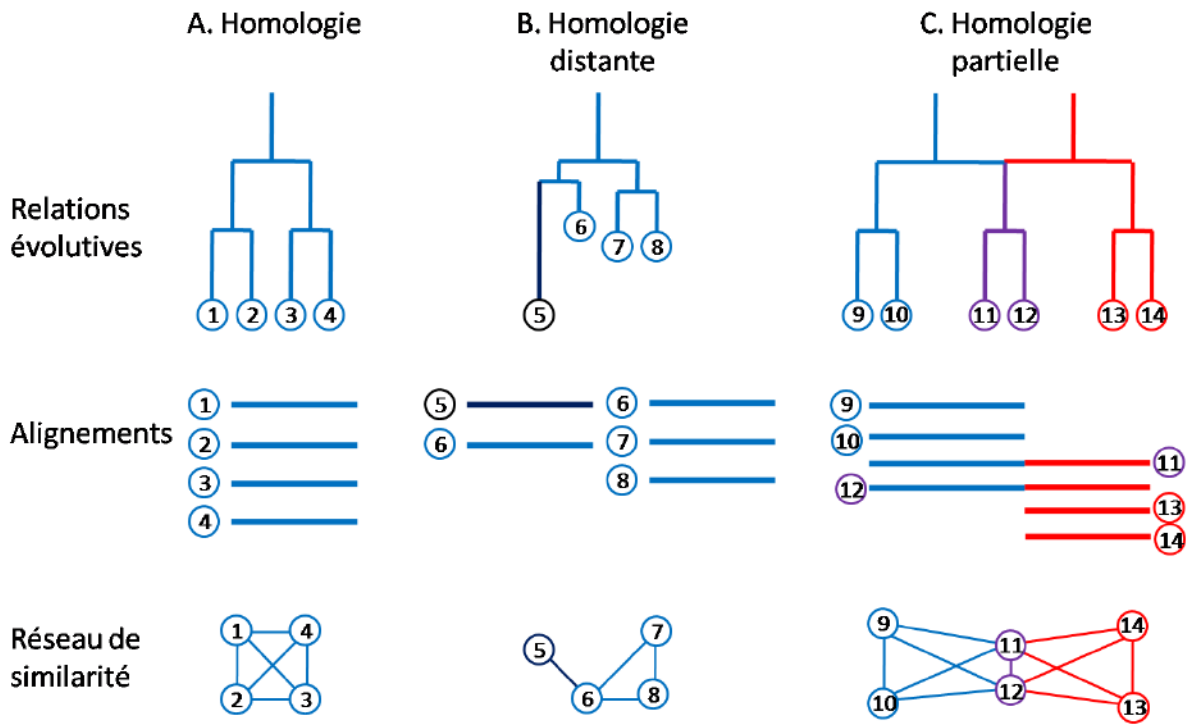


Figure 1-14 : Homologie distante et partielle

A. Cas idéal d'un groupe de séquences ayant peu divergée depuis une origine ancestrale commune. Les séquences sont toutes similaires deux à deux. Elles sont homologues sur l'ensemble de leur longueur.

B. Homologie distante. La séquence 5 a beaucoup divergée des autres séquences. Elle est homologue des séquences 7 et 8 mais ne leur est plus similaire. Il est possible de retrouver ces relations d'homologies distantes grâce à la séquence 6 intermédiaire.

C. Homologie partielle. La séquence 11 est partiellement homologue des séquences 9 et 13, mais ces séquences ne sont pas homologues entre elles.

1.5.3.2. Le problème de l'homologie partielle

L'algorithme de construction de familles homologues proposé ci-dessus, appelé couramment *simple linkage*, construit en pratique des familles avec de nombreuses séquences non-homologues [Liu, Rost, 2004]. Ce problème découle du problème de l'homologie, que nous discuterons plus en détail dans le chapitre 2 de ce manuscrit, mais dont nous pouvons déjà présenter les conséquences pratiques.

On considère classiquement que des séquences homologues le sont sur l'ensemble de leur longueur, par descendance depuis une séquence ancestrale commune. L'algorithme de *simple linkage* repose sur l'idée que l'homologie entre séquences est une relation transitive : si A et B ont une origine ancestrale commune, et si B et C ont une origine ancestrale commune, alors on peut inférer que A et C ont une origine ancestrale commune.

Suite à des processus combinatoires d'évolution des gènes (cf. partie 2.1), certaines séquences ne sont homologues que sur une partie de leur longueur. Or l'homologie partielle entre séquences n'est pas une relation transitive : si B est homologue avec A le long d'une région, et homologue avec C le long d'une région différente, alors rien ne permet d'affirmer

que A et C sont homologues (Figure 1-14.C). De tels cas de figures sont à l'origine du regroupement de gènes non-homologues par l'algorithme de *simple linkage*.

Un alignement local entre deux séquences ne permet que d'inférer leur homologie partielle, le long des régions alignées. La difficulté de la construction de familles homologues est de retrouver les homologues distants, malgré l'incertitude sur les relations d'homologie partielles. Différentes catégories de méthodes ont été proposées pour résoudre ce problème.

1.5.3.3. 1ère solution : Ne garder que les alignements locaux couvrant l'essentiel des séquences

Une réponse naturelle au problème de l'homologie partielle serait de ne s'intéresser qu'aux alignements locaux couvrant entièrement les séquences : deux séquences qui sont similaires - et s'alignent - sur toute leur longueur sont entièrement homologues. Un tel critère éliminerait les relations d'homologie partielles, mais aussi de nombreuses relations entre homologues complets. Il ne permettrait que de construire des familles homologues très conservées. Le rythme d'évolution est en effet souvent hétérogène le long des séquences : certaines régions sont conservées sur un grand nombre de générations, tandis que d'autres divergent rapidement.

Des critères moins stricts sont souvent employés, par exemple sur une proportion minimale des séquences couvertes par les alignements de 60 ou 80%. Si ces critères permettent d'étudier des relations entre homologues légèrement divergents, ils éliminent toujours les relations plus distantes et le problème de l'homologie partielle commence à se poser. Des tests empiriques ont montré que ce critère n'est pas un moyen fiable d'améliorer la reconstruction de familles homologues [Song et al., 2008].

1.5.3.4. 2ème solution : étudier les protéines multi-domaines

Le phénomène d'homologie partielle remet en cause la conception des protéines comme des unités évolutives indivisibles. Une conception alternative est de considérer une protéine comme un assemblage de modules évolutifs indépendants, appelés domaines protéiques. Un *domaine protéique* est ici défini comme une sous-unité *évolutive* stable, qui peut être combiné avec d'autres modules pour former une grande diversité de protéines dîtes multi-domaines. Plusieurs approches utilisent ce modèle de protéine pour éviter le problème d'homologie partielle lors de la construction de familles homologues. Elles diffèrent dans leur méthodologie et dans leur conception de l'impact du phénomène d'homologie partielle.

1.5.3.4.a. La piste de GeneRAGE : étudier les familles mono-domaines

Une première approche consiste à s'intéresser aux protéines constituées d'un unique domaine évolutif, en considérant que les protéines multi-domaines sont des assemblages de protéines mono-domaines (Figure 1-15.A). Selon ce modèle, une protéine multi-domaine est similaire à plusieurs familles mono-domaines, et les connecte dans le réseau (Figure 1-15.B).

L'algorithme GeneRAGE [Enright, Ouzounis, 2000] propose de détecter les protéines multi-domaines[1], de les retirer du réseau, pour pouvoir ensuite reconstruire les familles mono-domaines sans les agglomérer.

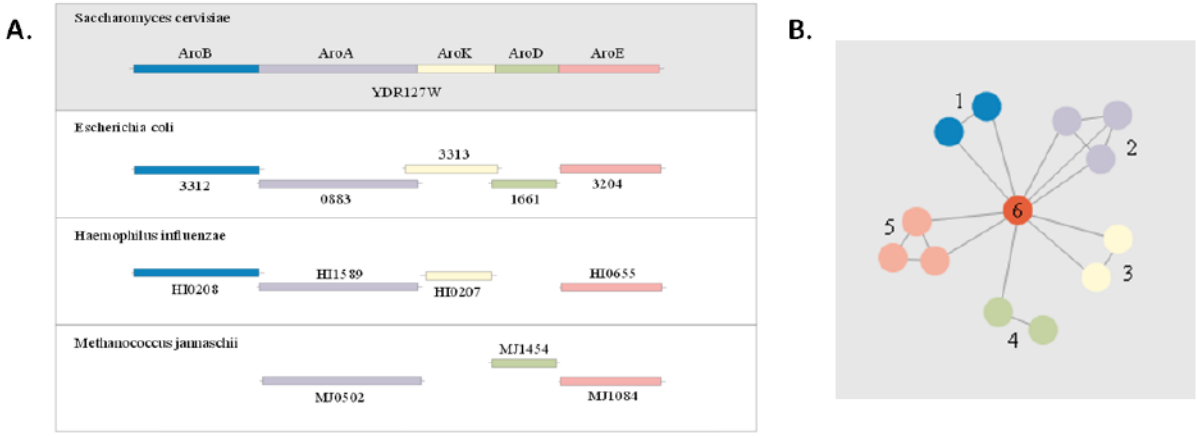


Figure 1-15: Classification d'une protéine multi-domaines par l'algorithme GeneRAGE
Adapté de (Enright, Ouzounis 2000).

A. Représentation du cluster *aro* dans plusieurs espèces. Chez la levure Saccharomyces cervisiae le gène YDR127W code pour une protéine multifonctionnelle, tandis que dans les autres espèces les enzymes correspondantes sont codées par d'autres gènes, non nécessairement voisin sur les génomes

B. Résultat du groupement par GeneRAGE des gènes en différentes familles homologues représenté sur un réseau de similarité de séquences. Le gène 6 est la protéine multifonctionnelle YDR127W.

L'algorithme GeneRAGE ne fonctionne pas bien en pratique, car il repose sur des hypothèses erronées concernant l'évolution des protéines. Il suppose implicitement que les protéines multi-domaines sont rares, qu'elles sont que du bruit qu'il faut éliminer pour étudier les séquences évoluant « convenablement », sans combinaison. Or les protéines multi-domaines représentent jusqu'à 80% des séquences chez les eucaryotes, et jusqu'à 60% chez les procaryotes [Han et al., 2007].

De plus, GeneRAGE suppose qu'une protéine multi-domaine peut être identifiée comme une concaténation de protéines mono-domaines. Or ces protéines mono-domaines peuvent ne pas être présentes dans le jeu de données étudié, voire ne pas exister puisque les fusions ne sont qu'un processus combinatoire parmi d'autres (p. ex. le brassage de domaine, cf. partie 2.1). Les protéines multi-domaines ne seraient alors pas détectées par GeneRAGE.

1.5.3.4.b. La piste de Neighborhood Correlation : étudier les familles multi-domaines

Dans la panoplie des méthodes pour construire des familles homologues, l'approche proposée par Song *et al.* (2008) se situe à l'opposée de celle de GeneRAGE. Ces auteurs prennent explicitement en compte la présence d'une proportion élevée de protéines multi-domaines dans les génomes, et souhaitent identifier indifféremment des familles mono-

[1] GeneRAGE identifie les protéines multi-domaines en étudiant la structure du réseau, de façon similaire à la méthode de détection des *séquences composites* que nous présenterons dans le chapitre 2.

domaines et multi-domaines. Ils proposent pour cela un modèle théorique d'évolution des protéines multi-domaines, dans lequel une famille est caractérisée par une partie centrale « identitaire », héritée à une position fixe du génome, sur laquelle viennent s'insérer des domaines « accessoires » (Figure 1-16.A). Deux protéines sont considérées homologues si elles partagent cette partie identitaire. Deux protéines similaire le long d'un domaine accessoire sont considérées non-homologues (Figure 1-16.B).

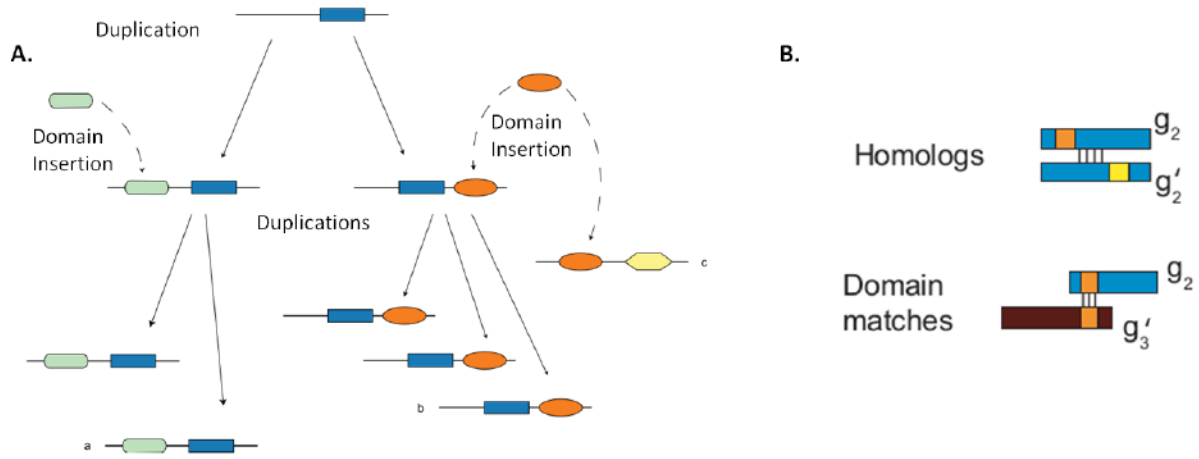


Figure 1-16: Modèle d'évolution des familles multi-domaines

A. Evolution d'une famille multi-domaines hypothétique par duplications et insertion de domaines à une position génomique.

B. Selon ce modèle d'évolution, il est possible de distinguer deux protéines homologues de la même famille qui ont des domaines insérés différents, et deux protéines non-homologues qui sont similaires sur des domaines insérés identiques.

La seule donnée de la similarité entre les deux protéines ne suffit pas à distinguer leur relation, c'est pourquoi Song et ses collaborateurs proposent d'étudier la structure locale du réseau de similarité, et plus particulièrement le voisinage commun de ces protéines. Deux protéines qui ont une majorité de voisins en commun sont probablement homologues sur la partie centrale identitaire ; tandis que deux protéines qui ont une faible proportion de voisins en commun ne partagent sans doute qu'un domaine accessoire (Figure 1-17). Cette intuition est mesurée par l'indice de corrélation de voisinage (*Neighborhood Correlation*), qui défini un nouveau score entre séquences. Ce score permet de définir un nouveau réseau dans lequel les voisins sont homologues sur des parties identitaires. Ce réseau a une structure en communautés (boules) plus forte que le réseau de similarité initial, mais les composantes connexes doivent cependant toujours être découpées pour séparer les protéines non-homologues [Joseph, Durand, 2009].

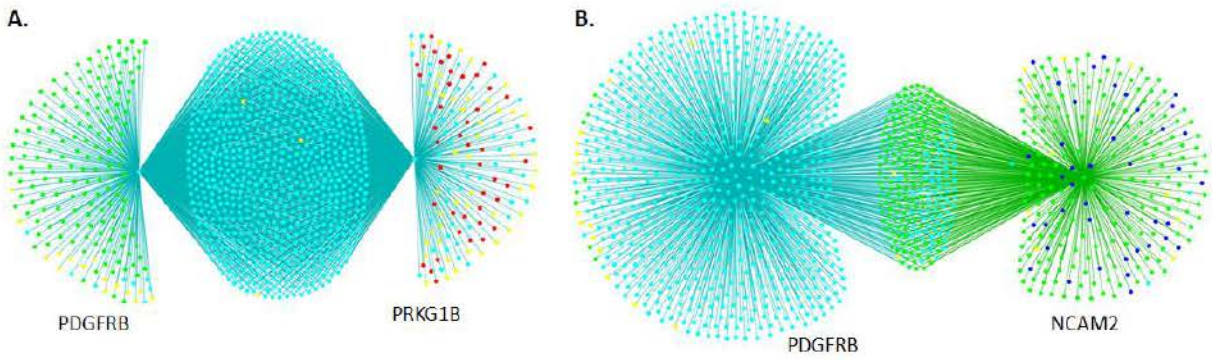


Figure 1-17 : La structure du voisinage des gènes reflète leurs relations évolutives

A. Les kinases PDGFRB et PRKG1B sont homologues et partagent la majorité de leurs voisins. Leur voisinage commun contient 779 gènes, essentiellement des kinases (turquoise). Leurs voisins non partagés (183 et 142 respectivement) sont majoritairement dus à des similarités sur les domaines *Ig* (vert) et *cNMP* (rouge).

B. Les gènes PDGFRB et NCAM2 ont peu de voisins non partagés. Ils ont 232 voisins en commun, avec lesquels ils partagent le même domaine, tandis que PDGFRB a 730 voisins uniques et que NCAM2 en a 240.

Les deux méthodes exposées ci-dessus intègrent l'existence de phénomènes d'évolution combinatoire, en employant la notion de domaine, mais avec une conception bien différente de l'évolution des protéines multi-domaines. GeneRAGE considère que la création de protéines multi-domaines est rare et concatène de protéines entières. Neighborhood Correlation considère que les protéines multi-domaines forment des familles, qui sont modifiées par des l'insertion/délétion éventuellement fréquentes de domaines accessoires.

1.5.3.5. Limites des modèles de construction de familles homologues

Plus généralement, les méthodes de construction de familles homologues peuvent être classées d'après leur conception de l'évolution combinatoire des protéines selon deux axes (Figure 1-18). Le premier axe indique la fréquence à laquelle on envisage les phénomènes de combinaison. Le second axe indique l' « équilibre » des phénomènes de combinaison considérés entre les deux fragments. On dira qu'une combinaison est déséquilibrée lorsqu'une protéine principale est modifiée par un fragment accessoire. On dira qu'elle est au contraire équilibrée, lorsque les deux fragments ont le même statut (p. ex. fusion de deux gènes) et qu'on ne peut pas attribuer une ascendance préférentielle à l'un des deux.

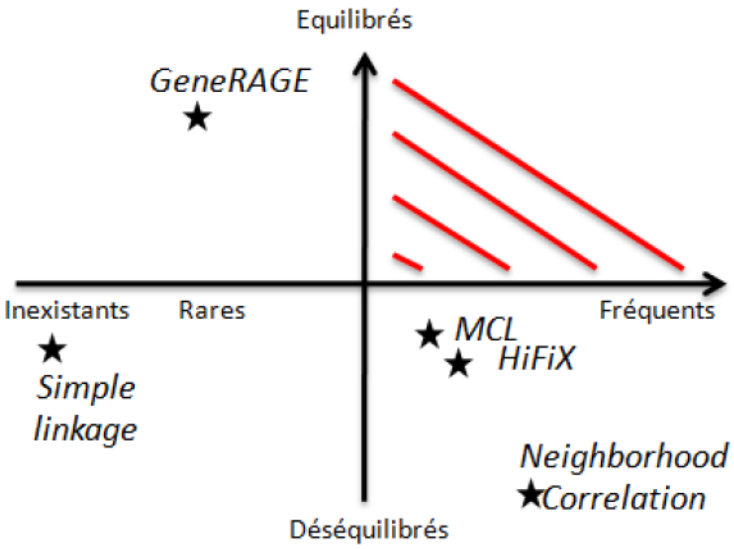


Figure 1-18 : Processus combinatoires considérés pour construire les familles homologues
Les méthodes de construction de familles homologues peuvent être classées selon la façon dont elles considèrent les processus combinatoires d'évolution des séquences.
L'axe horizontal indique la fréquence des phénomènes combinatoires considérés, d'inexistant à fréquent. L'axe vertical indique si les combinaisons considérées font participer des fragments de séquences de façon déséquilibré ou équilibrée. Il n'existe pas de modèle de famille homologue et de méthode de construction considérant des phénomènes combinatoires fréquents et équilibrés

Les méthodes présentées jusqu'à présent permettent d'envisager la classification en familles homologues lorsque les phénomènes combinatoires sont soit rares, soit déséquilibrés. Or l'expérience montre qu'en général ces conditions ne sont pas respectées. Il semble difficile de concevoir une notion robuste de famille homologue en présence de phénomènes de combinaison fréquents et équilibrés. Organiser les données moléculaires en familles constitue donc une avancée pratique pour comprendre l'évolution de la vie sur terre, tout en soulevant des questions théoriques encore mal maîtrisées que nous allons aborder dans ce travail.

1.6. Objectifs de cette thèse

Au vu du grand intérêt de l'étude des données moléculaires par des méthodes de réseaux, je me suis concentré sur des applications fondamentales de ces méthodes pour faire progresser les analyses évolutionnaires.

En premier lieu (chapitre 2), j'ai démontré que l'utilisation de graphes de similarité de séquences permettait non seulement l'identification de familles de gènes homologues, mais aussi de familles de gènes composites, constituées sur la base de matériel génétique provenant de plusieurs familles de gènes. J'ai analysé la distribution de ces gènes composites dans le vivant et les éléments génétiques mobiles, ainsi que leurs fonctions. J'ai confirmé que les génomes eucaryotes contiennent les proportions de gènes composites les plus

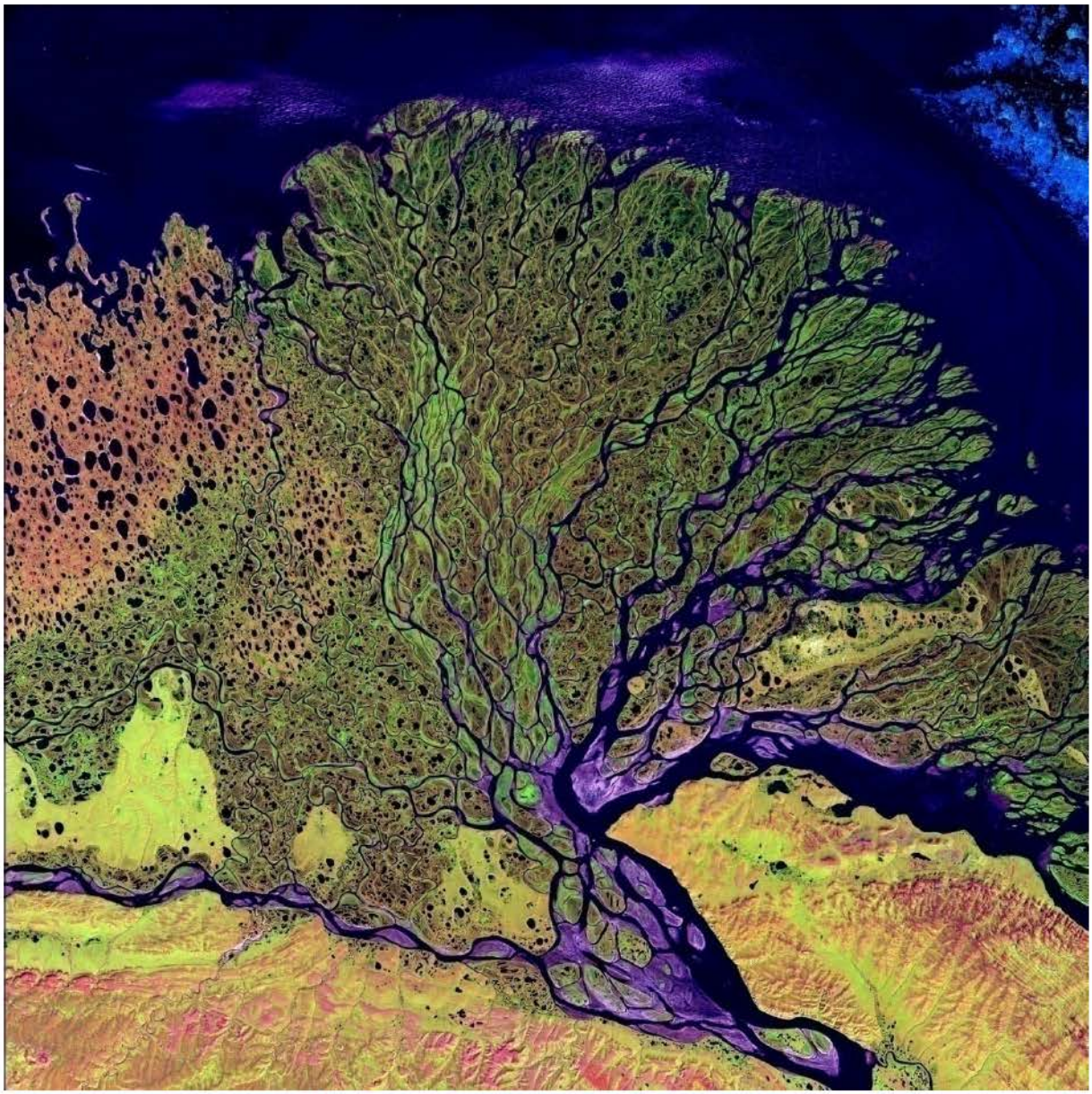
élevées dans le vivant. Mais surtout, cette démarche m'a conduit à établir l'existence de relations complexes et évolutivement informatives entre les séquences.

Dans un deuxième temps (chapitre 3), j'ai démontré, en collaboration avec des évolutionnistes moléculaires, que ces relations complexes justifient la définition de nouveaux termes pour décrire l'évolution des gènes de façon plus complète et satisfaisante. En particulier, j'ai contribué à introduire les notions « d'homologie arborescente », « d'homologie réticulée » et « d'air de familles » en biologie moléculaire. Ces travaux établissent que, faute d'un modèle de graphe suffisamment inclusif, les évolutionnistes évacuent un phénomène évolutif crucial de leurs analyses, l'origine multiple des adaptations moléculaires, quand ils se concentrent sur les adaptations survenant au sein d'une seule lignée de gènes.

Pour illustrer leur importance quantitative et mieux comprendre les règles de l'évolution de nouveaux gènes par des mécanismes d'introgression se déroulant au delà des frontières des familles de gènes, j'ai finalement choisi d'étudier la composition de gènes chez les entités biologiques les plus nombreuses de la planète : les virus (chapitre 4). J'ai démontré que toutes les classes de virus disposent de gènes composites, qui se retrouvent surreprésentés dans les fonctions biologiques importantes pour l'interaction des virus avec leurs hôtes cellulaires et pour la réplication des virus. Par conséquent, en nombre absolu, les éléments génétiques mobiles portent le plus grand nombre de gènes composites, qu'ils font circuler entre différents hôtes cellulaires, contribuant à la création d'un grand réseau d'échange de matériel génétique. Les frontières au sein de ce réseau sont encore plus complexes à établir qu'on ne l'imaginait puisque l'évolution de gènes composites sur la base de l'association d'informations génétiques présentes dans des virus à ADN d'une part et dans des virus à ARN d'autre part, mise en évidence par mes travaux, illustre que la composition de gènes ne s'effectue pas seulement en dehors des frontières des familles de gènes mais aussi parfois entre lignées de virus très distinctes.

D'une manière générale, toutes ces analyses encouragent l'extension de l'étude des séquences au delà des modèles évolutionnaires classiques pour mieux tenir compte de la multiplicité des origines des adaptations moléculaires dans le monde vivant.

Chapitre 2 - Identification de familles de gènes composites



© NASA/USGS EROS Data Center Satellite Systems Branch

Figure 2-1 : Delta du fleuve Léna en Sibérie
L'image de ce réseau fluvial illustre la façon dont les gènes évoluent : ils divergent par accumulation de mutations, puis se rejoignent, fusionnent ou échangent des fragments, avant de diverger à nouveau.
Adapté de l'article : Selosse, MA « L'évolution par fusion », Pour la Science 2011

2.1. Processus combinatoires d'évolution des gènes 36
2.1.1. Définition 36
2.1.2. Recombinaison homologue 36
2.1.3. Recombinaison non-homologue 37
2.2. Identification des processus combinatoires dans les réseaux de similarité de séquences 40
2.2.1. Motivations 40
2.2.2. Un gène fusionné forme un patron intransitif dans un réseau 41
2.2.3. Les gènes composites sont au centre des patrons intransitifs produits par des processus combinatoires 42
2.2.4. Elimination des patrons intransitifs causés par l'homologie distante 43
2.2.5. Formalisation de la recherche de gènes composite dans un réseau de similarité de séquences 44

2.1. Processus combinatoires d'évolution des gènes

2.1.1. Définition

On désigne par processus combinatoires d'évolution des gènes la production de nouveaux gènes à partir de la combinaison de plusieurs séquences physiquement distinctes. Ces processus contrastent avec les modèles où les séquences sont considérées comme des réplicateurs clonaux, divergeant progressivement en accumulant des mutations ponctuelles (substitution, insertion, délétion). Les processus combinatoires sont des exceptions à ce modèle évolutif, à côté des autres exceptions que sont l'incorporation d'une région non-codante (*origination de novo*) et les modifications non ponctuelles internes à une séquence (p. ex. duplication locale de grands fragments). Ils sont donc sources de problèmes pour la construction de phylogénies et autant que possible éliminés lorsque l'on étudie l'évolution des gènes au moyen d'arbres. Il s'agit donc de développer des concepts et des approches alternatives à l'approche phylogénétique classique pour étudier l'évolution combinatoire des gènes.

2.1.2. Recombinaison homologue

La recombinaison homologue est une première catégorie de processus combinatoire (Figure 2-2). Il s'agit de l'échange de séquences nucléotidiques entre deux molécules d'ADN similaires ou identiques. Ce phénomène est présent dans tous les organismes vivants, où il sert des fonctions variées. Il intervient notamment pour réparer des ruptures double-brin de l'ADN, et pour créer de la diversité génétique au sein des génomes, lors de la méiose chez les eucaryotes ou lors du transfert de gènes homologues chez les bactéries. Ce mécanisme

combinatoire est comparable à une forme de sexe au niveau des gènes [Minshull, Willem Stemmer, 1999], dans le sens où il ne crée pas de chimère, mais mélange des entités qui sont identiques à des variations locales près.

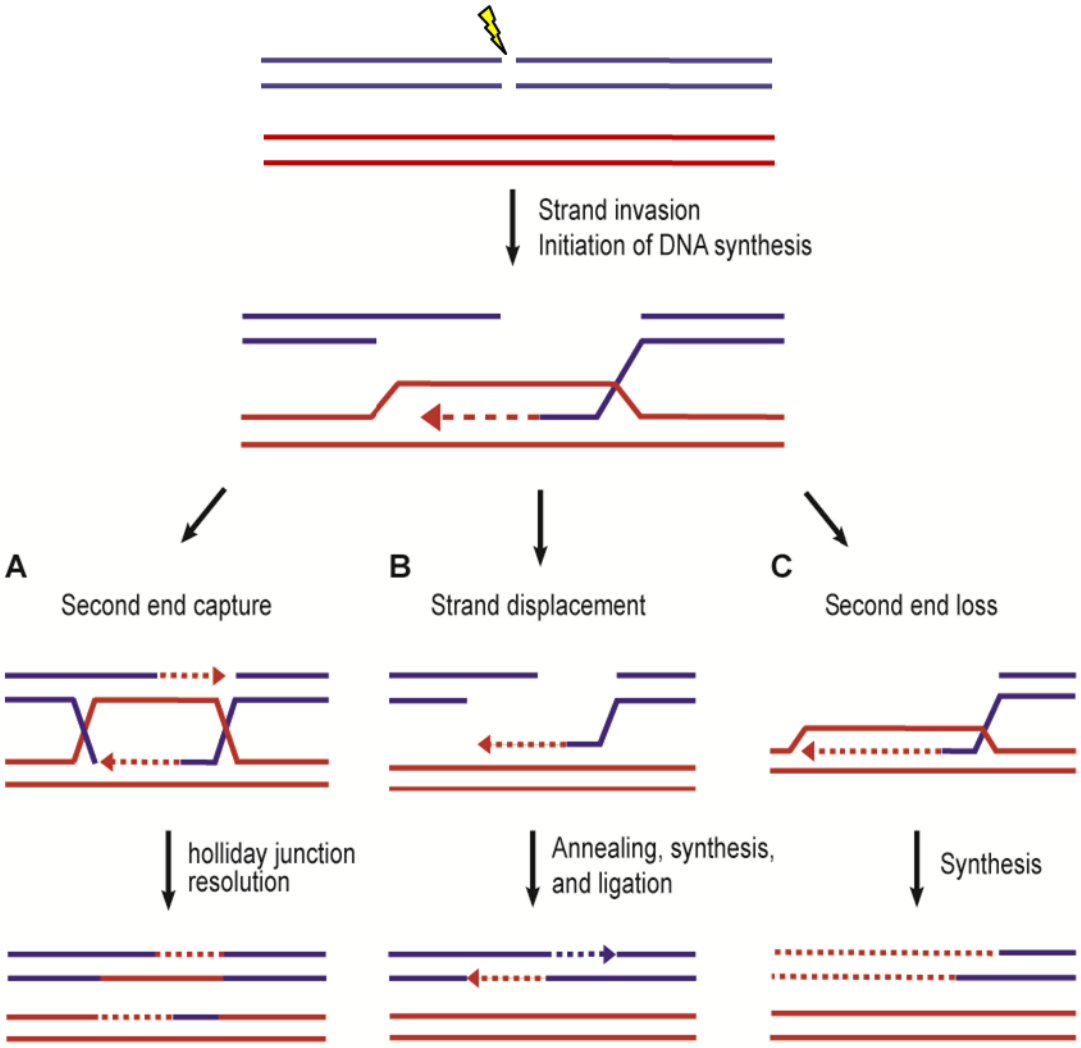


Figure 2-2 : Mécanismes de recombinaison homologue suite à une rupture double-brin
Une rupture double-brin de l'ADN (bleu) peut être réparée à partir d'une séquence homologue, selon différents mécanismes de recombinaison. La séquence réparée contient au final un fragment, plus ou moins long, provenant de la séquence homologue.
Adapté de Barlow *et al.* "Timing Is Everything: Cell Cycle Control of Rad52." Cell Division, 2010

2.1.3. Recombinaison non-homologue

Les autres processus combinatoires, qui ne consistent pas en un remplacement d'un fragment de séquence par un fragment homologue, sont désignés par le terme général de recombinaison non-homologue. Ce type de recombinaison est parfois appelée recombinaison illégitime, ce qui sous-entend qu'il s'agit d'un phénomène anormal, rare et souvent délétère. A plus large échelle évolutive, la recombinaison non-homologue est cependant un phénomène important, à l'origine de nombreuses nouvelles familles de gènes [Bornberg-Bauer et al., 2010].

De nombreux types de recombinaison non-homologue ont été nommés, par des auteurs qui abordaient ces questions selon des approches variées. Certains des termes employés désignent des mécanismes très précis, tandis que d'autres décrivent des phénomènes beaucoup plus généraux, avec un recoupement entre les différentes réalités décrites.

Au niveau moléculaire, il existe de nombreux mécanismes qui peuvent aboutir à la combinaison de séquences non-homologues. Sans rentrer dans le détail de ces mécanismes, on peut dire que la recombinaison non-homologue est favorisée par la présence de régions similaires entre les séquences, de fragments répétitifs, voire de sites spécifiques de recombinaison [Bapteste, 2013]. Elle peut intervenir par exemple lors d'un crossing-over inégal entre chromosomes, lors du déplacement d'éléments transposables au sein d'un génome ou lors d'une réparation incorrecte d'une rupture double brin de l'ADN.

2.1.3.1. Fusion de gènes

La fusion de gènes est un cas particulier fréquent de recombinaison non-homologue [Pasek et al., 2006 ; Zhou et al., 2008], qui produit un gène contenant l'intégralité de deux séquences distinctes mises bout à bout. Elle intervient souvent entre gènes physiquement proches le long du génome, par le biais d'une mutation ponctuelle entraînant la disparition d'un codon stop, ou d'une délétion de la région intergénique (Figure 2-3). Les fusions de gènes créent de nouvelles fonctions [Long, 2000], améliorent l'activité catalytique ou permettent une co-régulation des protéines dont les gènes ont été fusionnées.

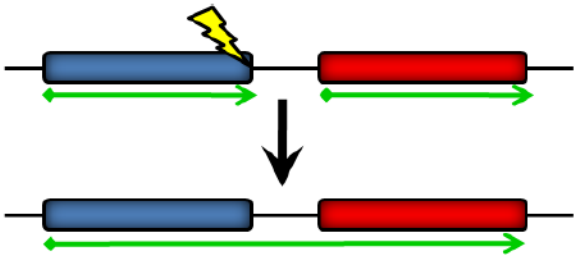


Figure 2-3 : Fusion de gène
Suite à une mutation ponctuelle, les gènes adjacents bleu et rouge forment un nouveau gène fusionné

2.1.3.2. Brassage d'exons

La structure en intron-exon des gènes eucaryotes favorise la recombinaison non-homologue [Gilbert, 1978]. Les introns sont les régions des séquences génétiques qui sont excisées lors de la phase d'épissage de l'ARN, et qui ne sont jamais présentes dans l'ARN mature. Les introns permettent de produire plusieurs protéines différentes à partir d'une même séquence, grâce à une combinatoire des exons retenus lors de l'épissage alternatif de l'ARN. A l'échelle évolutive, les introns sont des sites privilégiés de recombinaison non-homologue. Ils contiennent en effet des segments répétés favorables à la recombinaison, ce qui permet de combiner les exons sans toucher à leurs séquences (Figure 2-4). Ce phénomène, appelé brassage des exons, ou exon-shuffling, est très efficace pour créer de

nouvelles protéines, car les exons forment souvent des sous-unités fonctionnelles et structurelles indépendantes dans les protéines.

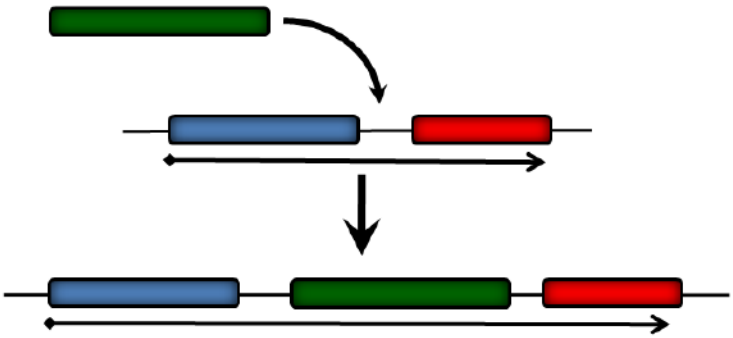


Figure 2-4 : Phénomène d'exon-shuffling
L'insertion d'un exon (vert) dans l'intron d'un gène, produit un nouveau gène avec trois exons.

2.1.3.3. Modèle d'organisation des protéines en domaines

La constatation que les protéines ne sont pas des unités évolutives indivisibles a conduit à introduire la notion de domaines protéiques, définis comme des modules évolutifs indépendants[1]. En pratique, la délimitation de domaines évolutifs passe par l'analyse des régions similaires entre de nombreuses séquences. Plusieurs bases de données de domaines protéiques ont été construites (InterPro, Pfam, PROSITE, ProDom, SMART, NCBI Conserved Domain Database, SUPERFAMILY) [Forslund, Sonnhammer, 2012] qui diffèrent par leur exhaustivité, par leur méthode de construction - notamment par l'intervention manuelle d'experts -, et par les annotations fonctionnelles ou structurelles qu'elles proposent. Ces bases de données permettent d'identifier les domaines présents dans chaque protéine, puis d'étudier leur histoire et les règles de leurs combinaisons. Un résultat frappant est que la majorité des protéines comprennent plusieurs domaines (4/5 chez les eucaryotes, 2/3 chez les procaryotes) [Han et al., 2007]. La majorité des protéines a donc été le produit de phénomènes combinatoires au cours de leur évolution. La description en termes de domaines s'affranchit de l'étude précise des mécanismes moléculaires, ainsi que de l'identification des partenaires de la combinaison. Il arrive que les limites des domaines et des exons correspondent, un domaine correspondant souvent plusieurs exons, auquel cas les notions de brassage d'exons et de brassage de domaines coïncident. Il est notamment proposé que la proportion élevée de séquences introniques chez les métazoaires a permis la création d'un grand nombre de protéines multi-domaines, qui ont joué un rôle important dans l'émergence de la multicellularité [Patthy, 1999].

Le modèle d'organisation des gènes en domaines permet d'étudier l'évolution de sous-unités évolutives des gènes et de leurs combinaisons. Il ne peut cependant pas rendre compte de tous les processus combinatoires au niveau des gènes. Le mouvement de

[1] La notion de domaine a aussi été introduite pour décrire d'autres notions de modules protéiques indépendants, en termes de fonction, de structure, ou de repliement. Ces notions sont liées, et les méthodes d'identifications propres à chaque définition recouvrent souvent des régions similaires.

fragments de séquences ne respecte pas nécessairement la jonction de domaines évolutifs cohérents : l'insertion d'une séquence dans une autre peut par exemple correspondre à la jonction des structures 3D correspondantes [Russell, 1994]. Il est bien souvent difficile de déterminer des limites nettes aux domaines, et de nombreux gènes ne sont couvert que partiellement par des domaines. Le modèle de protéines évoluant par combinaison de modules appelé domaines présente donc des limites.

2.2. Identification des processus combinatoires dans les réseaux de similarité de séquences

Nous avons présenté en introduction l'intérêt des réseaux de similarité pour étudier les phénomènes classiques (non-combinatoires) d'évolution des séquences. Ils sont rapides à construire. Ils permettent de mieux maitriser la construction des familles de gènes homologues, à la base des analyses phylogénétiques. Ils permettent d'explorer avec une bonne approximation les relations évolutives au sein des familles de gènes, même lorsque ces familles sont trop grandes, ou trop divergentes, pour reconstruire des arbres phylogénétiques [Atkinson et al., 2009]. Les réseaux ne sont cependant qu'une représentation pertinente des données relationnelles, et ne sont pas contraints par le cadre théorique de la phylogénie. Ils permettent d'étudier les relations évolutives issues de phénomènes combinatoires, difficiles à concevoir et à analyser à partir des modèles arborescents classiques. Cette possibilité est à l'origine de l'intérêt pour les réseaux de mes directeurs de thèse, qui ont d'abord étudié les phénomènes combinatoires à l'échelle des génomes, au moyen de réseaux de partage de gènes [Halary et al., 2009]. Ils ont ensuite construit des réseaux de similarité à l'échelle des gènes, et découvert que leur structure est marquée par des phénomènes combinatoires. Mon travail de thèse a consisté à suivre cette piste de recherche, c'est-à-dire à étudier l'évolution combinatoire des gènes par l'analyse des réseaux de similarité de séquence.

2.2.1. Motivations

Plusieurs auteurs avaient déjà employé les informations de similarité entre séquences pour détecter des phénomènes combinatoires et plus particulièrement pour identifier des gènes fusionnés. Ainsi, deux articles parus indépendamment en 1999 dans les journaux Nature et Science [Enright et al., 1999 ; Marcotte et al., 1999], visaient à améliorer l'annotation des gènes en identifiant les gènes fusionnés. Leur hypothèse était que si deux gènes fusionnent préférentiellement s'ils interagissent dans des voies cellulaires, détecter un gène fusionné permet d'inférer l'interaction entre ses composants, telle une pierre de rosette moléculaire (« Rosetta Stone »). Une autre motivation pour détecter les gènes fusionnés est d'améliorer la construction de familles homologues [Enright, Ouzounis, 2000]. Enfin, certains auteurs se sont attachés plus spécifiquement à l'étude et à la visualisation des réseaux de similarité de séquence (Figure 2-5) [Adai et al., 2004]. Dans ce contexte, l'étude des gènes fusionnés permet de mieux comprendre la structure de ces réseaux, puisque les

gènes fusionnés introduisent des ponts dans les graphes. Mon travail s'inscrit dans la lignée de cette dernière catégorie d'auteurs. Notre objectif est d'améliorer la compréhension fondamentale de la structure du réseau des similarités entre gènes, et d'éclaircir des questions telles que la nature de la composante connexe géante qui relie une majorité des séquences dans un réseau. Nous ajoutons une perspective et des questionnements évolutifs, qui n'avaient pas été abordé directement par les approches précédentes.

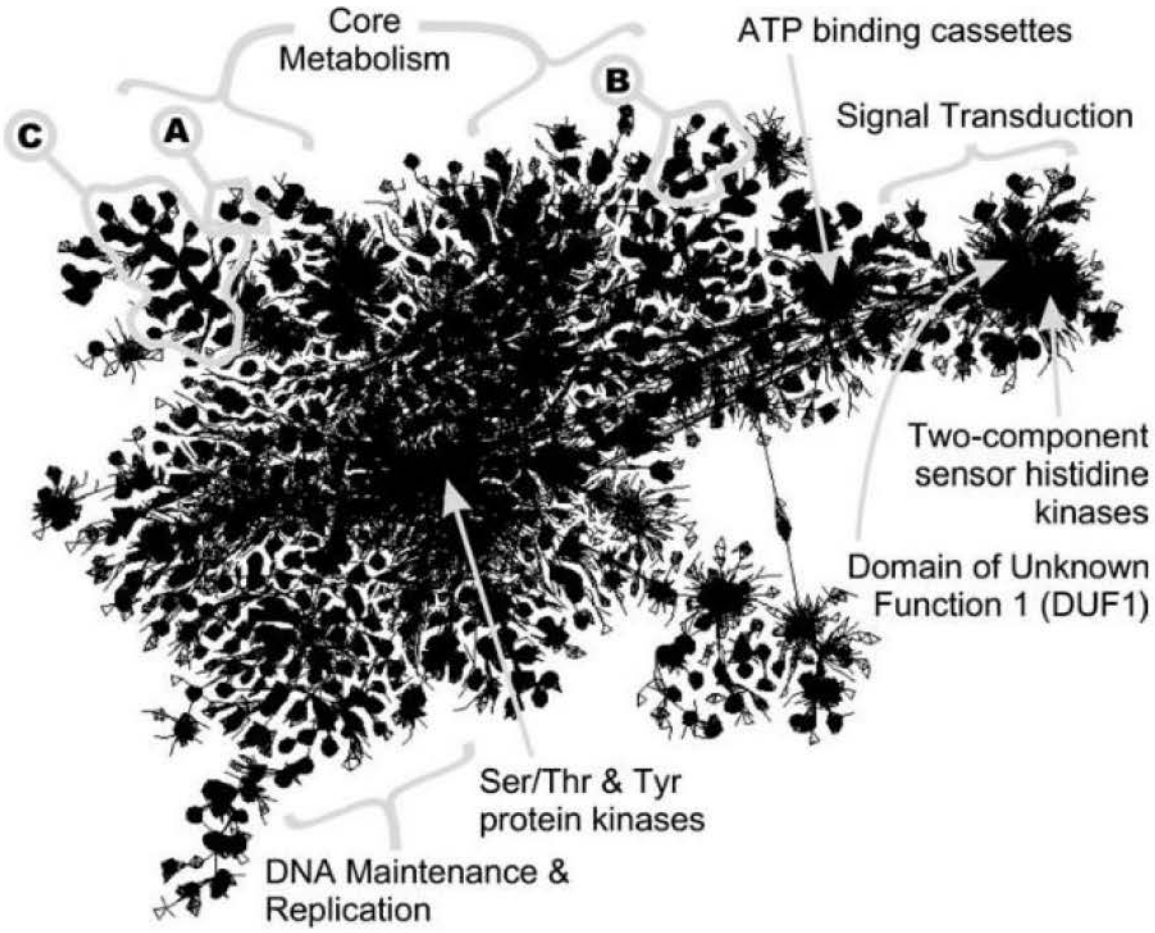


Figure 2-5 : Composante connexe d'un réseau de similarité de séquences
Source : Adai *et al.* "LGL: Creating a Map of Protein Function with an Algorithm for Visualizing Very Large Biological Networks." Journal of Molecular Biology (2004)

2.2.2. Un gène fusionné forme un patron intransitif dans un réseau

D'un point de vue méthodologique, les méthodes antérieures pour détecter des gènes fusionnés fonctionnent toutes selon le même principe. L'idée centrale est que si deux gènes A et B fusionnent pour former un nouveau gène C, ces trois gènes forment une relation de similarité non-transitive (Figure 2-6). A est similaire à C, C est similaire à B, mais A et B ne sont pas similaires. Ce patron intransitif est remarquable. Il fournit une procédure très simple pour identifier les gènes fusionnés : (i) énumérer tous les triplets de gènes, (ii) garder ceux qui ne sont pas transitifs. On infère alors que le gène C au centre d'un tel triplet est une fusion entre les gènes A et B.

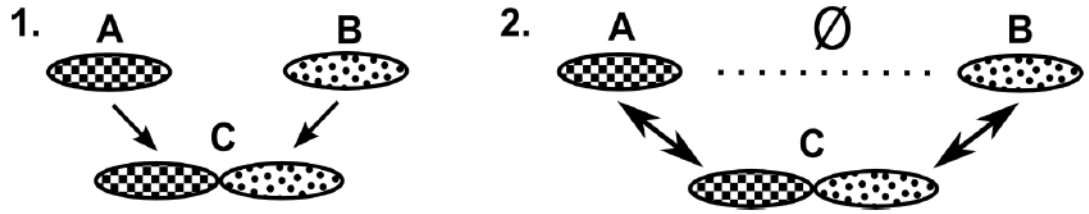


Figure 2-6 : Patron de similarité formé par un gène fusionné dans un réseau

1. Fusion des gènes A et B pour former le gène C
2. Les gènes A, C et B forment un patron intransitif dans le réseau de similarité de séquence

2.2.3. Les gènes composites sont au centre des patrons intransitifs produits par des processus combinatoires

Dans la suite, nous appellerons « composants » les gènes A et B, et « composite » le gène central C. Il s'agit de notre définition d'un **gène composite** : un gène **au centre d'un patron de similarité intransitif,** lorsque **ce patron est le résultat d'un processus combinatoire**. Nous choisissons d'employer le terme « composite » plutôt que le terme « fusionné » plus souvent utilisé parce que d'autres processus combinatoires que les fusions (combinaison complète de séquences) produisent des patrons intransitifs (Figure 2-7.A). Les combinaisons de fragments de séquences, notamment étudiées sous les désignations de brassage de domaines (domain-shuffling) ou brassage d'exons (exon shuffling), produisent le même patron, de même que les fissions (création de deux gènes indépendants par division d'un gène ancestral). Ces fissions sont bien plus rares [Pasek et al., 2006]. Nous les considérons comme des processus combinatoires car elles ne sont pas directement discernables des fusions dans un réseau de similarité. De plus les fissions ne sont souvent possibles et viables que pour des gènes déjà modulaires, et illustrent le fait que différentes combinaisons de domaines (ici des partitions de domaines) contribuent à l'évolution des gènes.

Notre définition d'un gène composite est fondée sur l'information de similarité entre séquence, qui est symétrique et ne permet pas déterminer l'orientation temporelle des événements. Composite ne signifie donc pas résultat d'une combinaison, mais composé de fragments, fragments qui existent et fonctionnent indépendamment dans des gènes composants. En ce sens, les fissions sont tout autant originales que les fusions par rapport au modèle d'un gène évoluant comme une unité indivisible. Par ailleurs, notons que les gènes produits par une combinaison de gènes homologues, ou partiellement homologues, forment des patrons de similarité transitifs, et ne sont donc pas considérés ici comme composites (Figure 2-7.B).

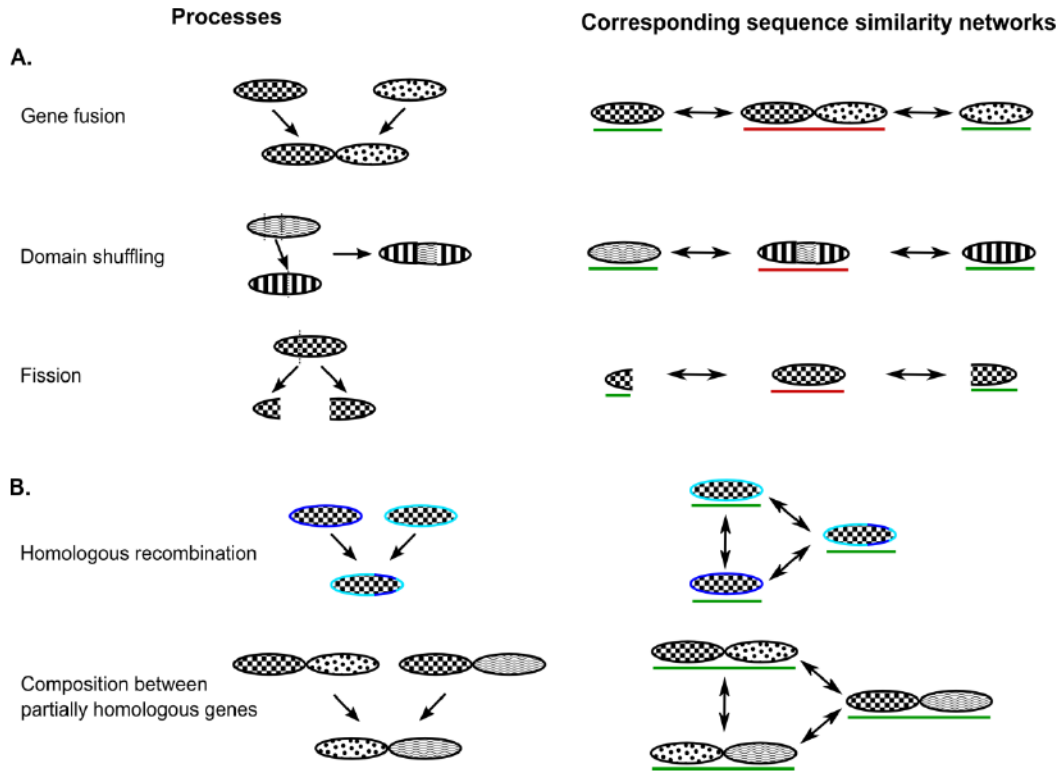


Figure 2-7 : Processus combinatoires et patron dans les réseaux

Gauche. Différents processus combinatoires d'évolution des gènes sont illustrés. Une forme de dessin identique indique une homologie entre séquences.

Droite. Les patrons formés par ces processus dans les réseaux de similarité de gènes. Les gènes soulignés en rouge sont au centre de patrons intransitifs

A. Les processus combinatoires entre gènes sans homologie forment des patrons intransitifs. Ils définissent des gènes composites.

B. Les processus combinatoires entre gènes homologues ou partialement homologues forment des patrons transitifs. Ils ne définissent pas des gènes composites.

2.2.4. Elimination des patrons intransitifs causés par l'homologie distante

S'il est simple d'énumérer les gènes satisfaisants à la première partie de la définition d'un gène composite (au centre d'un patron de similarité intransitif), il faut ensuite s'assurer qu'ils vérifient la seconde (que ce patron est le résultat d'un processus combinatoire). En effet, un patron intransitif peut aussi être formé par des homologues distants, qui ont trop divergé pour être directement similaires, mais sont similaires à un troisième homologue intermédiaire conservé (cf. 1.5.3.1). L'homologie distante est très fréquente. Elle créé de nombreux patrons intransitifs qui impactent fortement la structure des réseaux [Bapteste et al., 2012]. La reconstruction de famille homologue utilise d'ailleurs ces patrons intransitifs pour regrouper les homologues distants. La distinction de la cause d'un patron intransitif, entre homologie distante et processus combinatoire, est donc une question centrale pour améliorer les analyses évolutionnaires basées sur les relations de similarité.

De la même manière qu'il est difficile de s'assurer que deux séquences sont homologues partiellement et pas complètement, il est difficile de s'assurer qu'un patron intransitif n'est pas causé par de l'homologie distante. Tout au plus peut-on chercher à éliminer les cas

d'homologies distantes les plus flagrants, ce qui a été fait essentiellement de deux manières dans la littérature.

Un premier filtre consiste à ne garder que les gènes composites qui sont similaires aux gènes composants le long de régions non chevauchantes. Ce test élimine de nombreux faux positifs, ce qui montre l'importance du phénomène d'homologie distante. Il n'est cependant pas suffisant, car deux homologues distants peuvent n'être que faiblement similaires à un même homologue intermédiaire, et s'aligner le long de petites régions différentes.

Un deuxième filtre consiste à vérifier avec une méthode plus sensible si les séquences présumées composantes ne sont pas homologues. Certains auteurs ont eu pour cela recours à l'algorithme d'alignements Smith-Waterman, plus précis que BLAST. D'autres ont employé BLAST avec un seuil de similarité plus sensible (E-value plus grande notamment). Ce qui est important ici n'est pas tant la sensibilité absolue des méthodes de recherche de similarité, mais la différence de sensibilité entre la méthode initiale qui a permis d'identifier les triplets intransitifs et la méthode employée ensuite pour vérifier l'absence de similarité entre les composants. Il est en effet peu probable que deux séquences homologues distantes soient fortement similaires à une séquence intermédiaire, mais très dissimilaires entre eux. Deux objets peuvent difficilement être très proche d'un troisième, mais très lointains l'un de l'autre. Ou plus précisément cela est possible pour les séquences composites, mais nécessite une divergence très hétérogène (dans des régions différentes) des homologues distants par rapport à la séquence homologue intermédiaire.

2.2.5. Formalisation de la recherche de gènes composite dans un réseau de similarité de séquences

La détection de gènes composites a été entreprise par de nombreux auteurs selon les mêmes principes généraux décrits ci-dessus, mais en ré-implémentant à chaque fois le procédé, sans s'appuyer explicitement sur les initiatives précédentes. Les scripts et programmes employés n'ont en général pas été publiés avec les articles. Dans l'optique d'étudier systématiquement les gènes composites, j'ai implémenté le programme générique FusedTriplets, distribué sous licence libre GPL. FusedTriplets m'a également permis de comparer le principe des méthodes antérieures à une nouvelle méthode de détection de familles de gènes composites, et non pas seulement de composites isolés, que j'ai développé en collaboration avec le docteur en informatique Romain Pogorelcnik dans l'article suivant.

BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 7 2013, pages 837–844
doi:10.1093/bioinformatics/btt049

Sequence analysis

Advance Access publication January 30, 2013

MosaicFinder: identification of fused gene families in sequence similarity networks

Pierre-Alain Jachiet[1,†], Romain Pogorelcnik[2,†,*], Anne Berry[2], Philippe Lopez[1] and Eric Bapteste[1]

[1]UMR CNRS 7138 Systématique, Adaptation, Evolution, Université Pierre et Marie Curie, 75005 Paris, France and [2]LIMOS, Ensemble Scientifique des Cezeaux, 63173 AUBIERE, France

Associate Editor: Mario Albrecht

ABSTRACT

Motivation: Gene fusion is an important evolutionary process. It can yield valuable information to infer the interactions and functions of proteins. Fused genes have been identified as non-transitive patterns of similarity in triplets of genes. To be computationally tractable, this approach usually imposes an *a priori* distinction between a dataset in which fused genes are searched for, and a dataset that may have provided genetic material for fusion. This reduces the 'genetic space' in which fusion can be discovered, as only a subset of triplets of genes is investigated. Moreover, this approach may have a high–false-positive rate, and it does not identify gene families descending from a common fusion event.

Results: We represent similarities between sequences as a network. This leads to an efficient formulation of previous methods of fused gene identification, which we implemented in the Python program *FusedTriplets*. Furthermore, we propose a new characterization of families of fused genes, as clique minimal separators of the sequence similarity network. This well-studied graph topology provides a robust and fast method of detection, well suited for automatic analyses of big datasets. We implemented this method in the C++ program *MosaicFinder*, which additionally uses local alignments to discard false-positive candidates and indicates potential fusion points. The grouping into families will help distinguish sequencing or prediction errors from real biological fusions, and it will yield additional insight into the function and history of fused genes.

Availability: *FusedTriplets* and *MosaicFinder* are published under the GPL license and are freely available with their source code at this address: http://sourceforge.net/projects/mosaicfinder.

Contact: pogorelc@isima.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 30, 2012; revised on January 4, 2013; accepted on January 27, 2013

1 INTRODUCTION

1.1 Biological and evolutionary motivation for studying gene fusion

Fused genes, which result from the fusion of previously separate genes, or parts of their sequences, are key evolutionary entities (Patthy, 2003). It is generally believed that such events are rare, and that the resulting genes are often deleterious. However, fused genes can also encounter evolutionary success (Rogers and Hartl, 2012). Gene fusions have been reported in the three domains of life, e.g. in (hyper)thermophilic Archaea (Rodrigues *et al.*, 2007), in bacteria (Nie *et al.*, 2011; Pasek *et al.*, 2006) and in Eukaryotes (Durrens *et al.*, 2008; Ekman *et al.*, 2007; Zhou *et al.*, 2008). As a matter of fact, it has been estimated that two-fifths of the prokaryotic genes and more than two-thirds of the eukaryotic genes are composed of several domains (Han *et al.*, 2007), which have been likely combined through fusion events. In the latter taxa, gene fusions have been particularly well documented in animals (Buljan *et al.*, 2010; Marsh and Teichmann, 2010). In particular, it was shown that domain rearrangements occurred in 35.9% of gene families within the *Drosophila* clade (Wu *et al.*, 2011), significantly affecting processes of signalling and development. More problematically, in humans, gene fusions were reported to play a role in cancer. These fusions notably concerned relatively large and conserved genes (Narsing *et al.*, 2009) and members of the rapidly accelerated fibrosarcoma family of protein kinases, recently identified as characteristic aberrations of the most common tumours of the central nervous system in children (Lawson *et al.*, 2011). The widespread occurrence of gene fusion is notably explained by the fact that gene fusion can lead to new functions (Long, 2000). For instance, in the ciliate unicellular eukaryote *Tetrahymena thermophila*, gene fusions contributed to the evolution of processes, such as phospholipid synthesis, nuclear export and surface antigen generation (Salim *et al.*, 2011). Likewise, a gene fusion occurred in the early history of fungi, resulting in cellobiose dehydrogenases involved in the degradation of cellulose and lignin (Zamocky *et al.*, 2004). Later, the fungi *Candida albicans* benefited from the fusion of the 5′ domain of ALS5 (agglutinin-like sequence 5) to the tandem repeat region and 3′ domain of ALS1 producing an original ALS protein, likely involved in the adhesion to host and abiotic surfaces (Zhao *et al.*, 2011). As some of these fused genes increased the fitness of their carrier, they were maintained in genomes and gave rise to new gene families. Thus, various

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

fused globins have been reported, occasionally supplanting the parental gene form, as was the case for the fusion of β/Δ globin, before the radiation of *Paenungulata* (the clade containing elephants, dugongs and manatees and hyraxes) (Opazo *et al.*, 2009). Similarly, some of the toxins exploited by sea anemones to paralyse their preys have evolved by gene fusion, as they improved the transcript stability and secretion of these toxins (Moran *et al.*, 2009). From an evolutionary perspective, fused gene families can convey useful information about the history of life. They can provide valuable markers for phylogenetic analysis. It was suggested that these slow and rare events could be informative for reconstructing the phylogeny of plants (Nakamura *et al.*, 2007). Moreover, Stechmann and Cavalier-Smith (2002) used a derived gene fusion to propose a rooting of the eukaryotic tree. However, convergences and lateral transfers of gene fusions have also been reported, e.g. both in eukaryotes (Abdelnoor *et al.*, 2006; Aleshin *et al.*, 2007; Makiuchi *et al.*, 2007) and in diverse bacterial phyla, where gene fusions of polyamine biosynthetic enzymes *S*-adenosylmethionine decarboxylase (AdoMetDC, speD) and aminopropyltransferase (speE) orthologues, catalysing *de novo* diamine to triamine formation (Green *et al.*, 2011), and fused genes involved in histidine biosynthesis have been laterally transferred (Fani *et al.*, 2007 and so forth), suggesting that fused genes should only be used as phylogenetic markers with great care (Waller *et al.*, 2006). Finally, from a functional perspective, fused gene families serve as precious 'Rosetta stones' (Adai *et al.*, 2004) for the identification of potential protein–protein interactions and metabolic or regulatory networks (Enright *et al.*, 1999; Marcotte *et al.*, 1999). Our purpose in this article is to propose a new method for finding fused genes and to group them into families, which yields additional insight into the function and history of these genes.

1.2 Fused gene detection: state of the art

All current *in silico* methods for finding fused genes are based on sequence similarities (Durrens *et al.*, 2008; Enright *et al.*, 1999; Marcotte *et al.*, 1999; Rogers *et al.*, 2009; Salim *et al.*, 2011; Snel *et al.*, 2000; Suhre, 2004). The idea is that a fused gene (or *composite gene*) is similar to two *component* genes, which are not pairwise similar and align on disjoint parts of the fused gene (Fig. 1). In the rest of the article, we will use these terms of *composite* and *component* genes, as proposed in Enright *et al.* (1999). We will designate as a *fused triplet* a triplet of genes that exhibits this non-transitive pattern of similarity. Many variations around this idea have been implemented to identify composite genes and their components since the Marcotte *et al.* and the Enright *et al.* 1999 articles. They encounter four types of issues.

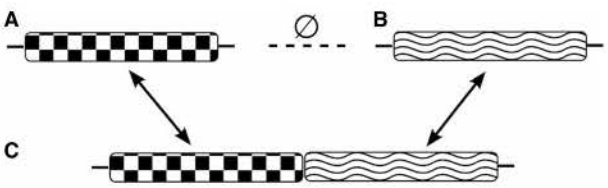


Fig. 1. Composite (fused) gene C and its two components A and B. A and B are similar to disjoint parts of C. A and B are dissimilar

First, the number of fused triplets rapidly becomes enormous for big datasets. Previous authors usually distinguished *a priori* between a query dataset (genome), within which composite genes were searched for, and a reference dataset (genomes, Clusters of orthologous groups of proteins (COGs)), in which components could be found. This greatly reduced the number of candidate triplets, with the drawback that some triplets are missed, as only a subset is investigated.

Second, some triplets may not result from a fusion but from distant homologues, i.e. a pair of homologous sequences that display no similarity at the sequence level, but that are both similar to a third intermediate sequence (Park *et al.*, 1997). Two types of tests are usually performed to exclude those false positives. The first test cross-checks that component genes are not similar, either with the same algorithm at a more permissive threshold [most of the time a higher Basic Local Alignment Search Tool (BLAST) E-value (Yanai *et al.*, 2001)] or with a more accurate algorithm (Enright *et al.*, 1999) such as Smith–Waterman (Smith and Waterman, 1981). The second test checks whether component genes align along non-overlapping regions of the candidate composite genes (Enright and Ouzounis, 2001; Yanai *et al.*, 2001). These controls eliminate many false positives.

Third, strongly supported fused triplets may result from sequencing or prediction errors (Pasek *et al.*, 2006), if a gene is artificially split into two separate genes, or if two adjacent genes are artificially fused into a single one. As those errors are presumably random and rare, a control is to identify other occurrences of candidate composite and component genes in closely related genomes.

A fourth and central issue is the grouping of identified component and composite genes *a posteriori* into gene families descending from a common fusion event. This grouping is necessary to count evolutionary events and to perform general functional analyses. First, if one could group composite triplets descending from a common fusion event, it would summarize the information contained in this enormous number of triplets of genes into fewer triplets of gene families, and, therefore, avoid long post-analyses of the results. This is, however, far from obvious and computationally challenging. Second, grouping into families would reduce the risk of distant homologies, as the absence of similarity between any pair of genes from two component families is much more robust than the absence of similarity between two component genes. Third, potentially artefactual composite or component genes would be easily identified, as they are the only representatives of their family (trivial family of size one).

A proxy to achieve a grouping into families has been to map genes on pre-existing family classifications (Suhre, 2004; Yanai *et al.*, 2001), usually Clusters of orthologous groups of proteins (COG)/Clusters of orthologous groups of eukaryotic proteins (Tatusov *et al.*, 2003). This is only partially satisfactory, as by definition, families of composite genes do not match a single COG and, therefore, are overlooked by that approach. Novel gene families (e.g. environmental) that have not been associated to a COG family are likewise difficult to detect. Alternatively, Enright and Ouzounis (2000) grouped composite genes into families by simple linkage. This is straightforward, as similarity between sequences is already computed to look for fused triplets. But simple linkage will aggregate unrelated composite genes if multiple fusion events have occurred in the history of some

genes (Supplementary Fig. S1). Moreover, this method does not allow reconstructing component gene families and their relation with composite families.

1.3 Fused gene detection: our approach

We propose to explicitly represent similarity between DNA or protein sequences (hereafter called genes) as a network. Sequence similarity networks were first proposed in a study conducted by Tatusov *et al.* (1997) and used for larger scale studies in the study conducted by Enright *et al.* (2002). This approach enables to apply efficient graph theory concepts and tools to mine similarity information (Atkinson *et al.*, 2009; Halary *et al.*, 2009; Song *et al.*, 2008; Tordai *et al.*, 2005). We propose a new characterization of families of composite genes, with a robust and fast method of detection, well-suited for the automatic analysis of large datasets, without using an *a priori* distinction on the datasets from which families of composite genes may be identified.

We also unify the existing methods for composite gene detection by transposing them into a sequence similarity network. This enables us to compare our new tool called MosaicFinder with the existing gene-centred approach, which we call FusedTriplets. MosaicFinder not only directly groups composite and component gene families but also reduces the risk of outputting a large number of false positives. In such searches, questions of macro-evolution should be addressed with a carefully selected dataset (e.g. introducing sequences from genomes that are representatives from the many taxonomical groups under comparison).

2 METHODS

2.1 Preliminary notions

A *graph* $G = (V, E)$ is a set of vertices V and a set of edges E that link some pairs of vertices together (our graphs are undirected). Sequence similarity networks are graphs with sequences (or genes) as vertices, connected by edges when they are found to be similar by a pairwise comparison method (Smith and Waterman, 1981), BLAST (Altschul *et al.*, 1990) and BLAST-Like Alignment Tool (Kent, 2002). Two vertices are *adjacent* if they are linked by an edge, i.e. two sequences are adjacent if they are similar. The *neighbourhood* of a vertex x is the set $N(x)$ of vertices that are adjacent to x (x not included). Given a subset X of vertices, we will call *common neighbourhood* of X, denoted $CN(X)$, the intersection of the neighbourhoods of all the vertices of X [i.e. $CN(X) = \bigcap_{x \in X} N(x)$]. Hence, the common neighbourhood of a set of genes F (e.g. a gene family) is the set of sequences in the dataset that are similar to every sequence of F, sequences of F excluded. A *clique* (also called *complete* subgraph) is a set of pairwise adjacent vertices. A set of genes F is a clique if for every pair of sequence (u, v) in F, u and v are similar. It usually means that sequences in F have a conserved homologous region in common. A graph is *connected* if there is a path between any pair of vertices. A *connected component* is a maximal connected subgraph. Note that two sequences in the same connected component may not have any homologous region in common (Fig. 2). A *separator* is a set of vertices whose removal increases the number of connected components. A *clique separator* is a separator that is a clique. A *clique minimal separator* (which we will shorten to CMS) is a clique separator, which is minimal for the separation of two given vertices (the reader is referred to Berry *et al.*, 2010 for graph definitions and details on CMSs).

Typically, a sequence similarity network can be reconstructed for a large dataset by connecting genes that are related in a BLAST (Altschul *et al.*, 1990) search, with an *E-value* score better than a user-defined threshold. Sequence similarity networks are graphs with sequences (or genes) as vertices, directly connected by edges when they show a similarity greater than a user-defined threshold. For a given comparison between two sequences, the alignment, score and E-value are not symmetric. They can vary depending on which sequence is used as the query. The network is symmetrized by considering the best match of each pairwise comparison. As the greatest asymmetry is found in the better-scoring comparisons [i.e. at a much more stringent threshold than the ones used for network reconstruction (Atkinson *et al.*, 2009)], this procedure does not impact the topology. Thus, the structure of this network captures much of the history of gene evolution: not only classical divergence by point mutations but also recombinations, fusions and fission events (Adai *et al.*, 2004). Conserved families of genes with a single common ancestor are all connected to each other in a connected component of the graph (unless they evolved beyond recognition by BLAST). They form cliques of sequences in the network, which are aligned over most of their length. Divergent families will form less densely connected groups of vertices because the common ancestry between some pairs of their genes is less frequently detected.

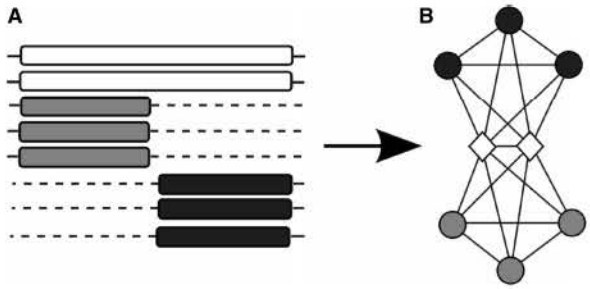


Fig. 2. (**A**) Multiple alignment of composite genes (white) and component genes (grey and black). (**B**) Similarity network of those genes. The white vertices form a composite gene family. They are a clique minimal separator of the network. The black vertices and the grey vertices form two separate component families

2.2 FusedTriplets: implementation of the gene-centred method

As explained in the introduction and by Figure 1, a composite gene is characterized in the similarity network by connecting two component genes that are not adjacent, and which are similar to disjoint parts of the composite gene. This leads to the following steps to identify fused triplets: enumeration of all non-transitive triplets of genes and cross-check of the absence of similarity between component genes and test of their alignment overlap along the composite gene.

We cross-check the absence of similarity between component genes in the same way as Yanai *et al.* (2001). It simply consists of testing whether they remain dissimilar at a more permissive threshold than the one used for triplet enumeration. For example, if one considered component and composite genes similar if $E-value \leq 1e-10$, one can test whether component genes have an $E-value \leq 1e-5$ to make sure that they are dissimilar [see Atkinson *et al.* (2009) for the effect of the threshold on similarity network topology]. This method presents the great advantage that it requires no further computation, as it only uses the similarity network information.

We reject triplets whose component gene alignment along composite gene overlap by >20 amino acids (as in Yanai *et al.*, 2001). This small overlap is allowed because BLAST alignments tend to extend slightly beyond homologous regions.

We implemented these steps in the Python script FusedTriplets.

2.3 MosaicFinder

When exploring a large dataset including several genomes, there may be several representatives of a given fusion event, which we would want to group into composite and component gene families. Rather than doing this grouping *a posteriori* from the results of a gene-centred approach, an interesting prospect is to identify those families directly in the similarity network (Fig. 2).

A family of composite genes has the particularity to link otherwise non-connected groups of nodes. The characteristic non-transitive pattern of composite genes extends to families. We propose to characterize a composite gene family as a CMS of the sequence similarity network (Berry *et al.*, 2010). A composite gene family is a *separator*, as its removal disconnects component gene families. It is *minimal*, as every composite gene is similar to the components. The additional condition that the separator is a *clique* describes the requirement that the family of composite gene is conserved. It should be noted that a composite gene that is the only representative of its family will be identified, as a clique minimal separator of size one.

For all these reasons, CMS is a good model to identify composite gene families. In addition, CMSs present several interesting properties: the number of CMSs bounded by the number of vertices, and an exact polynomial-time algorithm exists to identify them. MosaicFinder works in several consecutive steps, which are detailed later in the text.

STEP 1: Construction of the similarity network

MosaicFinder takes the result as input of all-against-all BLAST comparisons between the sequences under study, in the form of a simple flat-table, including information about the region that aligns between pairs of sequences (BLAST qstart, qend, sstart, send). To determine whether two sequences are similar, MosaicFinder relies on a pair of similarity scores, the 'E-value' and 'percentage of identity' of these two sequences. The results are then represented as an undirected network $G = (V, E)$, where V is the set of sequences, and edge is $(u, v) \in E$ if the similarity score S_{uv} or S_{vu} is higher than a user-defined threshold.

STEP 2: Identification of fused gene families

A graph algorithm (Berry *et al.*, 2010) is then applied to find clique minimal separators in this network and to propose candidate families of composite genes. This is the central and longest step of MosaicFinder.

STEP 3: Identification of component families

The component families are then identified by disconnections in the common neighbourhood of each CMS. This common neighbourhood does not contain the nodes from the separator. It, therefore, contains several connected components, which we defined as component families. Figure 3 illustrates this process.

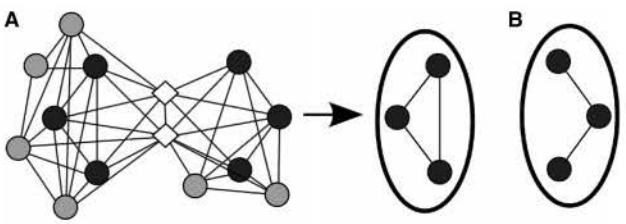


Fig. 3. (A) White nodes are a clique minimal separator. Black nodes are its common neighbourhood. Note that some grey nodes may be connected to the separator but not be in its *common* neighbourhood. (B) The subgraph of the CMS common neighbourhood contains two connected components, which define its component families. Note that component families are not required to be fully connected

STEP 4: Cross-checking similarity between component families (optional)

MosaicFinder optionally tests component families for distant homology. A simple way to verify the absence of any similarity between component families is to test whether they remain disconnected with a more permissive threshold of identity than the threshold used to identify the CMS. It should be stressed that the *gap* between the two thresholds, and not the absolute value of the permissive one, ensures that component families are not distant homologs. This test is optional because a disconnection between two component families is already robust. Furthermore, raising the BLAST score can increase the risk of detecting false positives, especially for an E-value beyond 1e-3 (Fokkens *et al.*, 2010) and for large datasets.

STEP 5: Use of alignments to eliminate false positives

Undetected distant homologues in the common neighbourhood of a CMS may still lead to an overestimation of the number of component families and composite gene families. MosaicFinder further tests for the presence of such distant homologues using information about the regions that align in BLAST comparison between vertices from the CMS and vertices from their common neighbourhood. There is a false positive when the different component families align in the same region of a candidate composite gene because such a significant overlap in an alignment suggests that homology between sequences of different component families was undetected. As different genes from a component family F may align to slightly different parts of a potential composite gene v, we used the median alignment of F along v, defined as follows. The start position of the median alignment of F along v is the median over all start positions of alignments of F family members along gene v, and the end is similarly the median over all end positions. MosaicFinder rejects a candidate composite gene if the median alignment of different component families overlaps on >20 (by default) amino acids. This small overlap is allowed because BLAST extends alignments as far as possible, and small non-homologous flanking regions may artefactually align. Otherwise, the composite gene is accepted, and a 'fusion point' is calculated as the middle point between the median alignments of each component families.

STEP 6: Output

MosaicFinder outputs a table of genes and gene families involved in fusion events. This table indicates the fusion event that genes are involved in, and their groupings into composite or component families. It additionally indicates a fusion point for composite genes.

3 RESULTS

We implemented MosaicFinder and FusedTriplets, which we used to compare the detection of composite gene families with the existing methods for detecting composite genes. As there exists no large manually curated database of composite genes to use as a test bed, we simulated the evolution of composite genes and composite gene families to test the accuracy of MosaicFinder.

We also ran tests on real databases, but we have less information on the validity of our (or other) methods in this context. We focused our attention on the number of composite genes detected.

3.1 Test of MosaicFinder on simulated composite gene families

We simulated the evolution of component and composite gene families under various evolutionary circumstances to test and compare the sensitivity and specificity of MosaicFinder and FusedTriplets in their detection of composite genes (Fig. 4).

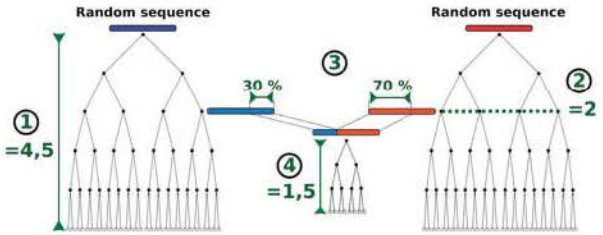


Fig. 4. Simulation of the evolution of component and composite gene families. Two random initial sequences are evolved along five-level perfect binary trees to produce two component gene families. Total length of both trees is scaled by the same evolutionary rate (parameter 1). A pair of sequences evolving along these trees is chosen at the same distance from root (parameter 2, *fusion level*). A given percentage (parameter 3) of the first sequence is fused with a fragment of the second sequence to create a new composite sequence of the same length. This sequence is evolved along a perfect binary tree with five fusion levels, scaled by a given evolutionary rate (parameter 4) to produce the composite gene families. In this figure, component families are divergent (tree length $\approx$ 5 MNSS), whereas composite family is conserved (tree length $\leq$ 2)

We used Seq-Gen (Rambaut and Grass, 1997) to simulate the evolution of component families, under the Whelan and Goldman model of amino acid substitution and a site-specific rate heterogeneity following a continuous gamma distribution ($\alpha = 1$). Ancestral sequences of 300 amino acids were generated randomly for each component family. These sequences were then evolved along perfect (complete) binary trees with five levels, i.e. symmetric and balanced trees with $2^5 = 32$ leaves at the fifth level, resulting in component families with 32 genes. We explored the effect of gene family divergence on composite gene detection under the hypothesis that the more divergent gene families are, the harder they are to detect. We produced gene families with different degrees of divergence as follows. We scaled these ultrametric phylogenetic trees with Seq-Gen (option -d) so that the total length of a tree can be measured as the distance from the root to any of the leaves in units of mean number of substitutions per site (MNSS). Typically, a tree of length 2 MNSS resulted in conserved families with all pairs of sequences presenting an E-value of $\leq$1e-10 (therefore, corresponding gene families forming cliques in a gene network reconstructed at that threshold). By contrast, trees of length 5 MNSS resulted in divergent families in which homology between many pairs of sequences was no longer detectable by BLAST. In these rapidly evolving trees, 99% of all pairs of maximally distant sequences presented an E-value of $\leq$1e-10 and 90% an E-value of $\leq$1e-5. To cover the range from highly conserved to highly divergent gene families, we explored 14 evolutionary rates, from 0.5 to 7 with a step of 0.5 (parameter 1). We generated simulated fusion events from a pair of component families evolving at the same evolutionary rate. A pair of sequences evolving along these trees was chosen at the same distance from the tree root [*fusion level* from 0 to 5 (parameter 2)]. We used this pair of sequences to create a novel 300 amino acids composite sequence made of 10–50% of the first sequence fused with 90–50% of the second sequence (parameter 3). This ancestral composite sequence was then evolved along a third perfect binary tree with five fusion levels, so that genes from composite and component gene families had undergone the same number of diversification events starting from ancestral component sequences (Fig. 4). The composite family was evolved at the same 14 evolutionary rates (parameter 4) that were used for the component families, thereby producing highly conserved to highly divergent composite families. For recent fusion events (*fusion level* = 0), the composite sequence was left unmodified. This protocol was repeated 10 times for each combination of the four parameters. We, therefore, simulated $10 * 14 * 6 * 5 * 14 = 58.800$ fusion events.

3.2 Result on simulation

For each simulated fusion event, we compared all pairs of genes from this dataset with BLASTp (E-value of $\leq$1e-5). We searched the resulting similarity network with MosaicFinder, FusedTriplets and FusedTriplets_E10, i.e. FusedTriplets with a more stringent 1e-10 E-value threshold and a cross-check of the absence of similarity between component genes/families at the original 1e-5 E-value threshold. The main explanatory parameters for composite gene detection results are evolutionary rates of component and composite gene families. We analysed the proportion of edges between composite and component genes that were recovered in the network, for various combinations of evolutionary rates (Supplementary Fig. S2.1). As expected, the majority of connections between fast evolving (>3 MNSS) component and composite families were lost, which defines an 'evolutionary zone' within which both MosaicFinder and FusedTriplets will work best. We compared the three methods with respect to the proportions of detected false positives (component genes that were erroneously identified as composite genes) and the proportions of detected true positives (composite genes that were correctly identified). In our simulations, all methods returned few false positives ($\leq$5%) for all combinations of evolutionary rates. However, FusedTriplets displayed higher proportions of false positives than FusedTriplets_E10 and MosaicFinder, this latter seemed as the method which is the least prone to outputting false positives (Supplementary Fig. S2.2). Such false positives seem to arise for particular combinations of evolutionary rates, leading to triplets in which there is a composite gene, located at one of the extremities of the triplet and two component genes. This topology is obtained when the intermediate component gene (i.e. the false positive) is connected on the one hand to its homologue (the other component gene) because of some sequence similarity that is still detectable for a given region of their sequences, and on the other hand to the composite gene via a different region of its sequence. As gene networks based on real data often connect sequences through partial regions of similarity, the analysis of real data may result in the detection of such false positives. These results strongly suggest that it is generally a good idea to use two thresholds with different stringencies in the detection of candidate composite genes, in analyses with FusedTriplet. Regarding the detection of true positives, there seemed to be 'evolutionary zones' in which all three methods recovered significant proportions of composite genes in our simulations (Supplementary Fig. S2.3.A). However, on closer examination, within these zones, the methods performed differently. First, we compared FusedTriplets with MosaicFinder (Supplementary

Fig. S2.3.B). Logically, MosaicFinder returned less true positives than FusedTriplets because MosaicFinder cannot detect candidate composite genes that are not also proposed by FusedTriplets. However, FusedTriplet_E10 (less sensitive than FusedTriplets to false positives, as described earlier in the text) is less efficient for detecting composite genes than MosaicFinder. Therefore, using MosaicFinder to analyse large datasets seems as a good trend. Overall, MosaicFinder is more robust than FusedTriplet, as it produces almost no false positives and successfully detects composite genes and groups them into families. We also investigated how other parameters (percentage of fused material from the component genes, fusion levels) affected the detection of false positives and true positives by these three methods (Supplementary Fig. S3). We observed that composite genes simulated in more recent events were more frequently detected than composite genes simulated in older events by all methods, and especially by MosaicFinder (Supplementary Fig. S3.3). Likewise, composite genes simulated in more balanced fusion events (e.g. when composite genes received fragments of similar sizes from the component genes) were more frequently detected than composite genes simulated in less balanced fusion events by all methods (Supplementary Fig. S3.5). This was expected because it is harder for any method to detect similarity between composite genes and component genes on shorter fragments, but FusedTriplet_E10 was more affected by this problem than the other methods. Regarding the fusion points, we find that MosaicFinder accordingly estimates the position of the fusion points. In all, 94% of the computed fusion points are <5 amino acids away from the *true* fusion point, and 99% <16 amino acids away. This variation is due to the imprecision of BLASTp alignments. Those numbers validate *a posteriori* the 20 amino acids overlap allowed between component families on composite genes.

3.3 Biological results

Our analyses of a real large dataset (591.439 sequences from the three domains of life and from mobile genetic elements, such as viruses and plasmids) with MosaicFinder extended our knowledge on the evolution of composite gene families. First, it showed that all types of genomes, whether they come from cellular organisms or from their mobile genetic elements, are concerned by the process of gene fusion (Supplementary Fig. S4). Eukaryotic genomes are significantly much more affected by this process than prokaryotic genomes and genomes of mobile elements; however, when the focus of the analysis is limited to the evolution of prokaryotes and their mobile genetic elements, these latter, in particular the plasmids, can be showed to be critically involved in that process. An excess of families of composite genes are found on plasmids, suggesting that these important vessels of DNA mobility are involved in the creation and/or the distribution of composite genes. This conclusion is consistent with the literature that claims that genomic evolution cannot be accurately described without taking the role of these infracellular entities into account (Bapteste and Burian, 2010; Bapteste *et al.*, 2012; Halary *et al.*, 2009). Moreover, our implementation allowed us to study the triplets centred on composite genes detected by MosaicFinder (Supplementary Fig. S5), offering an additional way to investigate the respective contribution of

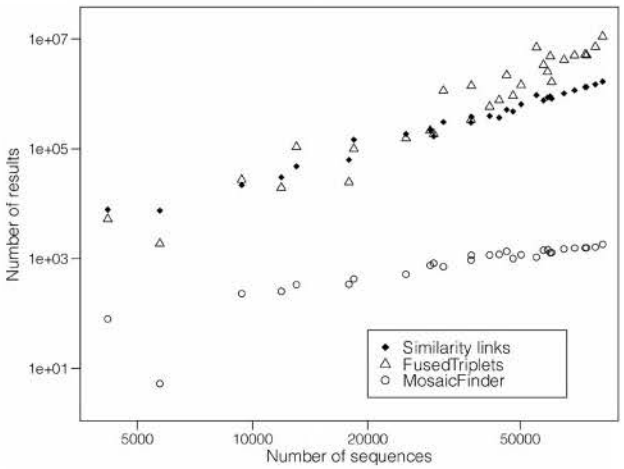


Fig. 5. Comparison of the number of edges between sequences, the number of identified composite families by MosaicFinder and the number of identified composite triplets by FusedTriplets (logarithmic scales)

cellular entities and mobile genetic elements to the process of gene fusion. When sequences from all genomes are analysed, 53% of these triplets only connect genes from cellular organisms; yet, when the focus is on the genomes of prokaryotes and mobile genetic elements, this proportion logically drops to 42% (because of the removal of intra-eukaryotic fusions). This result is consistent with our previous observations and is remarkable because it means that although a small majority of gene fusions apparently exclusively involves the genetic material from cellular organisms (when eukaryotic genomes are taken into consideration), a very large fraction of gene fusion events detected in that dataset have possibly involved the contribution of at least one mobile genetic element. An analysis of these triplets at a finer scale further suggests that mobile genetic elements were likely providers of some DNA for up to 39% of these fusions (which can be deduced from the sum of the percentages of triplets in which at least one sequence of mobile genetic element is at least at one of the extremities of the triplets), and that mobile genetic elements were possibly carriers of composite genes for up to 20% of these fusions (as can be deduced from the sum of the percentages of the per cent of triplets in which the composite sequence is carried by a mobile genetic element). When only genomes of prokaryotes and mobile genetic elements are considered, mobile genetic elements seem to act as providers of DNA/carriers of composite genes in up to 48/25% of the gene fusions, respectively. The functions of these composite genes are described in Supplementary Figure S6.

We also built datasets of various sizes, composed of 1–30 prokaryotic complete genomes, to search for composite and component genes. Figure 5 presents the number of composite gene triplets and triplets of families output by FusedTriplets and MosaicFinder, respectively, compared with the number of edges between sequences. FusedTriplets outputs an enormous amount of fused triplets (up to 11 millions), whereas MosaicFinder outputs only up to 1821 fusion events. FusedTriplets outputs up to 5339 potential composite genes for the biggest dataset (note that a given composite gene can correspond to many different triplets), whereas MosaicFinder finds 2490 unique composite genes. Most composite families (1313)

identified by MosaicFinder contain only one sequence. Of the remaining 508 fusion events, for 349 composite gene families, we could only detect one representative sequence of one of the two component families involved in the event. Thus, 159 detected fusion events (amounting to 985 composite genes) involved composite and component gene families with more than one sequence. These numbers show the great number of potentially misleading fusion events, and the interest of MosaicFinder is to identify them.

4 DISCUSSION

We proposed a new characterization of families of composite genes, as clique minimal separators in sequence similarity networks, and implemented this method into the C++ program MosaicFinder. We showed that on simulated data, MosaicFinder identifies conserved composite gene families well. Even if MosaicFinder was not designed to do so, it also identifies the evolutionary conserved fraction of composite genes from divergent families. In cases where divergent genes have evolved too much to show similarity to both component families, MosaicFinder proves to have a very low false positive rate.

We show that MosaicFinder gives good results quickly, with the advantage that genes are grouped into families, thus avoiding the extra work of regrouping the composite genes after they are output. Moreover, this information may be visualized as an annotated graph using Cytoscape (Shannon *et al.*, 2003). Supplementary Figure S7 gives an example.

According to our results from Section 3, MosaicFinder generates few false positives. Since in the real dataset constructed with 30 complete prokaryote genomes, MosaicFinder detected the impressive rate of one fusion gene of 33 genes, we can conjecture that in real data, there are in fact many composite genes.

Future work consists in breaking up long cycles with a local approach, as long cycles may mask fusion families by connecting component families indirectly (Supplementary Fig. S8).

5 SOFTWARE

MosaicFinder is based on the graph theoretic tool of clique separator decomposition. MosaicFinder is reliable for studying fusion events for phylogenetic research as well as for functional biology. The program has been developed in C++. FusedTriplets is a Python script that generalizes previous approaches to find composite genes, based on sequence similarity network abstraction. Both programs are freely available with their source code at this address http://sourceforge.net/projects/mosaicfinder/.

ACKNOWLEDGEMENT

The authors thank Michel Habib for suggesting using clique minimal separators.

Funding: P.A.J. was funded by an AMX PhD grant, R.P. and A.B. were funded by the French Agency for Research (DEFIS program TODO) and the ANR-09-EMER-010.

Conflict of Interest: none declared.

REFERENCES

Abdelnoor,R.V. *et al.* (2006) Mitochondrial genome dynamics in plants and animals: convergent gene fusions of a MutS homologue. *J. Mol. Evol.*, **63**, 165–173.

Adai,A. *et al.* (2004) LGL: creating a map of protein function with an algorithm for visualizing very large biological networks. *J. Mol. Biol.*, **340**, 179–190.

Aleshin,V.V. *et al.* (2007) Do we need many genes for phylogenetic inference? *Biochemistry (Mosc.)*, **72**, 1313–1323.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Atkinson,H.J. *et al.* (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS ONE*, **4**, e4345.

Bapteste,E. and Burian,R.M. (2010) On the need for integrative phylogenomics, and some steps toward its creation. *Biol. Philos.*, **25**, 711–736.

Bapteste,E. *et al.* (2012) Evolutionary analyses of non-genealogical bonds produced by introgressive descent. *Proc. Natl Acad. Sci. USA*, **109**, 18266–18272.

Berry,A. *et al.* (2010) An introduction to clique minimal separator decomposition. *Algorithms*, **3**, 197–215.

Buljan,M. *et al.* (2010) Quantifying the mechanisms of domain gain in animal proteins. *Genome Biol.*, **11**, R74.

Durrens,P. *et al.* (2008) Fusion and fission of genes define a metric between fungal genomes. *PLoS Comput. Biol.*, **4**, e1000200.

Ekman,D. *et al.* (2007) Quantification of the elevated rate of domain rearrangements in metazoa. *J. Mol. Biol.*, **372**, 1337–1348.

Enright,A. and Ouzounis,C. (2001) Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol.*, **2**, 10034.

Enright,A.J. and Ouzounis,C.A. (2000) GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics*, **16**, 451–457.

Enright,A.J. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Enright,A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.

Fani,R. *et al.* (2007) The role of gene fusions in the evolution of metabolic pathways: the histidine biosynthesis case. *BMC Evol. Biol*, **7 (Suppl. 2)**, S4.

Fokkens,L. *et al.* (2010) Enrichment of homologs in insignificant BLAST hits by co-complex network alignment. *BMC Bioinformatics*, **11**, 86.

Green,R. *et al.* (2011) Independent evolutionary origins of functional polyamine biosynthetic enzyme fusions catalysing de novo diamine to triamine formation. *Mol. Microbiol.*, **81**, 1109–1124.

Halary,S. *et al.* (2009) Network analyses structure genetic diversity in independent genetic worlds. *Proc. Natl Acad. Sci. USA*, **107**, 127–132.

Han,J.-H. *et al.* (2007) The folding and evolution of multidomain proteins. *Nat. Rev. Mol. Cell Biol.*, **8**, 319–330.

Kent,W.J. (2002) BLATThe BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Lawson,A.R.J. *et al.* (2011) RAF gene fusion breakpoints in pediatric brain tumors are characterized by significant enrichment of sequence microhomology. *Genome Res.*, **21**, 505–514.

Long,M. (2000) A new function evolved from gene fusion. *Genome Res.*, **10**, 1655–1657.

Makiuchi,T. *et al.* (2007) Occurrence of multiple, independent gene fusion events for the fifth and sixth enzymes of pyrimidine biosynthesis in different eukaryotic groups. *Gene*, **394**, 78–86.

Marcotte,E.M. *et al.* (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753.

Marsh,J.A. and Teichmann,S.A. (2010) How do proteins gain new domains? *Genome Biol*, **11**, 126.

Moran,Y. *et al.* (2009) Fusion and retrotransposition events in the evolution of the sea anemone *Anemonia viridis* neurotoxin genes. *J. Mol. Evol.*, **69**, 115–124.

Nakamura,Y. *et al.* (2007) Rate and polarity of gene fusion and fission in *Oryza sativa* and *Arabidopsis thaliana*. *Mol. Biol. Evol.*, **24**, 110–121.

Narsing,S. *et al.* (2009) Genes that contribute to cancer fusion genes are large and evolutionarily conserved. *Cancer Genet. Cytogenet*, **191**, 78–84.

Nie,Y. *et al.* (2011) Two novel alkane hydroxylase-rubredoxin fusion genes isolated from a dietzia bacterium and the functions of fused rubredoxin domains in long-chain n-alkane degradation. *Appl. Environ. Microbiol.*, **77**, 7279–7288.

Opazo,J.C. *et al.* (2009) Origin and ascendancy of a chimeric fusion gene: the beta/delta-globin gene of paenungulate mammals. *Mol. Biol. Evol.*, **26**, 1469–1478.

Park,J. *et al.* (1997) Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.*, **273**, 349–354.

Pasek,S. *et al.* (2006) Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics*, **22**, 1418–1423.

Patthy,L. (2003) Modular assembly of genes and the evolution of new functions. *Genetica*, **118**, 217–231.

Rambaut,A. and Grass,N.C. (1997) Seq-gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, **13**, 235–238.

Rivals,I. *et al.* (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.

Rodrigues,M.V. *et al.* (2007) Bifunctional CTP: inositol-1-phosphate cytidylyltransferase/CDP-inositol: inositol-1-phosphate transferase, the key enzyme for di-myo-inositol-phosphate synthesis in several (hyper) thermophiles. *J. Bacteriol.*, **189**, 5405–5412.

Rogers,R.L. and Hartl,D.L. (2012) Chimeric genes as a source of rapid evolution in *Drosophila melanogaster*. *Mol. Biol. Evol.*, **29**, 517–529.

Rogers,R.L. *et al.* (2009) Formation and longevity of chimeric and duplicate genes in *Drosophila melanogaster*. *Genetics*, **181**, 313–322.

Salim,H.M.W. *et al.* (2011) deFuser/detection of fused genes in eukaryotic genomes using gene deFuser: analysis of the *Tetrahymena thermophila* genome. *BMC Bioinformatics*, **12**, 279.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Snel,B.B. *et al.* (2000) Genome evolution-gene fusion versus gene fission. *Trends Genet.*, **16**, 9–11.

Song,N. *et al.* (2008) Sequence similarity network reveals common ancestry of multidomain proteins. *PLoS Comput. Biol.*, **4**, e1000063.

Stechmann,A. and Cavalier-Smith,T. (2002) Rooting the eukaryote tree by using a derived gene fusion. *Science*, **297**, 89–91.

Suhre,K. (2004) FusionDB: a database for in-depth analysis of prokaryotic gene fusion events. *Nucleic Acids Res.*, **32**, 273D–276D.

Tatusov,R.L. *et al.* (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Tatusov,R.L. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.

Tordai,H. *et al.* (2005) Modules, multidomain proteins and organismic complexity. *FEBS J.*, **272**, 5064–5078.

Waller,R.F. *et al.* (2006) Lateral gene transfer of a multigene region from cyanobacteria to dinoflagellates resulting in a novel plastid-targeted fusion protein. *Mol. Biol. Evol.*, **23**, 1437–1443.

Wu,Y.-C. *et al.* (2011) Evolution at the subgene level: domain rearrangements in the *Drosophila* phylogeny. *Mol. Biol. Evol.*, **29**, 689–705.

Yanai,I. *et al.* (2001) Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc. Natl Acad. Sci. USA*, **98**, 7940–7945.

Zamocky,M. *et al.* (2004) Ancestral gene fusion in cellobiose dehydrogenases reflects a specific evolution of GMC oxidoreductases in fungi. *Gene*, **338**, 1–14.

Zhao,X. *et al.* (2011) ALS51, a newly discovered gene in the Candida albicans ALS family, created by intergenic recombination: analysis of the gene and protein, and implications for evolution of microbial gene families. *FEMS Immunol. Med. Microbiol*, **61**, 245–257.

Zhou,Q. *et al.* (2008) On the origin of new genes in *Drosophila*. *Genome Res.*, **18**, 1446–1455.

Chapitre 3 - Problème de l'homologie : extension du champ des ressemblances informatives pour les évolutionnistes

Figure 3-1 : Réseau des communautés de la GCC d'un réseau de gènes eucaryotes
La couleur des communautés indique leur proportion de gènes composites de vert foncé (0%) à violet (100%)

3.1. L'homologie en biologie 55
3.1.1. La méthode comparative nécessite de mettre en correspondance des caractères entre organismes 55
3.1.2. Introduction du terme homologie 55
3.1.3. Transition à une ontologie évolutionnaire 56
3.2. Famille de gènes homologues 56
3.2.1. Bases ontologiques du concept de familles de gènes homologues 57
3.2.2. Difficultés posés par les processus combinatoires 57
3.2.3. Adaptation des modèles d'évolution des familles de gènes 58
3.2.4. Propositions de l'article pour tenir compte de ces difficultés 59

Le concept d'homologie s'est transformé au cours des siècles, pour s'adapter à de nouvelles conceptions du monde vivant, de nouveaux objets d'études, et de nouvelles données. Aujourd'hui, la diversité du vivant est majoritairement étudiée selon une perspective évolutive, à partir d'informations génétiques. Deux gènes sont dits homologues s'ils « possèdent *une* origine évolutive commune ». Cette définition est généralement interprétée selon une conception arborescente de l'évolution, pour signifier que deux gènes homologues descendent d'un même gène ancestral, par une succession de copies imparfaites. Elle permet de définir une famille de gène de façon analogue aux taxons d'organismes, comme l'ensemble des descendants d'un gène ancestral.

Les phénomènes d'évolution non-arborescents ne sont pas bien décrits, voire pas considérés par ce concept d'homologie. L'article présenté dans ce chapitre propose de réinterpréter le concept d'homologie selon une conception élargie de l'évolution, qui tient compte des processus combinatoires. Il propose d'étendre la relation d'homologie aux gènes qui n'ont qu'une origine ancestrale commune sur une partie de leur séquence. Il propose aussi de décrire des airs de familles entre les gènes qui partagent des ressemblances partielles. Il discute des perspectives de recherche qu'ouvre cette interprétation, et incite au développement de nouvelles méthodes d'analyses de l'évolution des séquences.

Ce travail est le fruit d'une collaboration entre plusieurs groupes de recherche. Ma principale contribution est présentée dans la partie « *Case 4* », autour de la figure 5. Cette figure expose l'importance des phénomènes combinatoires dans l'évolution des gènes eucaryotes. Elle permet de visualiser l'étendue et la diversité des relations évolutives entre séquences, habituellement découpées en îlots lors d'une interprétation strictement arborescente de l'homologie. Elle montre enfin l'intérêt méthodologique des réseaux pour analyser la diversité des gènes. Une version de cette figure a servie d'illustration à la couverture du journal Molecular Biology and Evolution, dans l'édition de mars 2014 où est paru cet article.

3.1. L'homologie en biologie

3.1.1. La méthode comparative nécessite de mettre en correspondance des caractères entre organismes

Une méthode générique en biologie, pour étudier la diversité des organismes et comprendre leur fonctionnement, consiste à les comparer les uns aux autres. Historiquement les chercheurs ont d'abord comparé la morphologie des organismes, en la décrivant par la délimitation de caractères tels que les organes. Avec l'accès aux séquences génétiques, ils comparent aujourd'hui massivement des gènes. Si l'on évoquera principalement ces deux types de caractères par la suite, il en existe bien d'autres, car chaque discipline délimite et compare des caractères spécifiques : des tissus en histologie, des réactions chimiques en biochimie, ou encore des comportements sociaux en éthologie.

Une étape préalable à toute comparaison entre organismes consiste à mettre en correspondance les caractères par lesquels on les décrit, c'est-à-dire de décider quel caractère est comparé à quel autre. Cette étape est moins évidente qu'il ne pourrait paraître. Faut-il par exemple comparer les ailes du papillon Moro sphinx à celles du colibri à ventre blanc ? Toutes deux permettent à ces espèces de voler, mais leurs structures sont difficilement comparables. A l'inverse, faut-il comparer les ailes du colibri et les nageoires pectorales du dauphin de l'Araguaia ? Ces organes ont des fonctions et des formes très différentes, mais leur organisation structurelle est similaire. La réponse à ces questions dépend en fait du regard l'on porte sur les organismes. Elle varie entre un scientifique qui étudie la motricité des animaux et un autre qui étudie leur organisation structurelle ; chacun compare les caractères qui se correspondent dans le cadre théorique de son étude.

3.1.2. Introduction du terme homologie

Le terme homologie signifie littéralement 'la qualité de correspondre dans le cadre d'un discours'[1]. Il est employé dans une déclinaison de ce sens général par de nombreuses sciences. En biologie, il a été introduit au milieu du XIX^{ème} siècle par Richard Owen, grand spécialiste d'anatomie et notamment d'ostéologie (étude des os). Owen s'intéressait à l'organisation du corps des animaux. Il étudiait la position structurelle des organes pour les mettre en correspondance entre organismes. A la suite de nombreux prédécesseurs botanistes et zoologistes, cette démarche systématique se révéla fructueuse pour produire des classifications cohérentes de la diversité des animaux, contrairement à des comparaisons d'organes basées sur leur forme générale ou leur fonction. Richard Owen formalisa cette méthode en théorisant l'idée que les organismes d'un même groupe (p. ex. vertébrés [Owen, 1848] partagent un même plan d'organisation structurelle, qu'il appela archétype. C'est en ce sens qu'il introduisit la notion d'homologie, pour qualifier les mêmes organes dans l'archétype chez différents organismes, indépendamment de leur forme ou de

[1] Le mot Homologie est construit à partir des racines grecques homos ('identique', 'similaire') et logos ('parole', 'discours').

leur fonction[1]. Cette dernière précision de la définition est importante. Il s'agit pour Owen de distinguer les organes homologues - éventuellement très différents en apparence comme peuvent l'être une aile et une nageoire -, des organes analogues, c'est-à-dire similaires en forme ou en fonction mais sans équivalence dans un plan d'organisation. Comparer les ailes d'un papillon à celles d'un oiseau n'a pas de sens dans le cadre de réflexion de Richard Owen. La distinction entre homologues et analogues avait pour objectif de lever cette source de confusion courante.

3.1.3. Transition à une ontologie évolutionnaire

La théorie d'archétype développée par Owen fournit une base ontologique au concept d'homologie, c'est-à-dire qu'elle donne un sens à la relation ainsi posée entre organes. Lorsque Darwin introduit la théorie de l'évolution, il emploie les résultats de la morphologie comme une preuve de la descendance avec modification. Il assimile la notion abstraite d'archétype à l'existence d'un ancêtre commun : "On my theory, unity of type is explained by unity of descent.". Darwin remplace l'explication essentialiste des patrons observés par une explication historique. Ce faisant, il transforme le sens attribué à l'homologie. Deux organes ne sont plus homologues s'ils correspondent aux mêmes parties de l'archétype, mais s'ils dérivent du même organe chez un ancêtre commun. Cependant, les critères pour mettre en correspondance (« *to homologize* » en anglais) les caractères ne sont pas modifiés, la position structurelle d'un organe prime sur sa fonction ou sa forme. Pour notre propos futur, il est intéressant de noter que la base ontologique du concept d'homologie a été radicalement transformée, tout en réemployant les mêmes méthodes pratiques de détermination.

3.2. Famille de gènes homologues

La théorie de l'évolution est devenue un cadre conceptuel central en biologie[2], employé par de nombreuses disciplines pour développer des analyses comparatives. Tout en conservant le sens de « partager une origine évolutive commune », le concept d'homologie a servi à développer une diversité de méthodes pour traiter d'autres caractères que les caractères morphologiques. Ainsi, l'homologie entre gènes est déterminée en alignant leurs séquences, rarement en comparant leurs positions relatives dans les génomes comme ce qu'il se pratique pour les organes. Cet alignement repose explicitement sur un modèle d'évolution des gènes, à la fois théorique (accumulation de mutations ponctuelles) et empirique (détermination des valeurs des matrices de similarité).

Bien souvent, l'étape préliminaire à une étude comparative n'est pas de mettre en correspondance 2 à 2 les gènes homologues, mais de construire des *familles* de gènes

[1] Richard Owen définie ainsi la notion d'homologie : "the same organ in different animals under every variety of form and function."

[2] Selon Dobzhansky « Rien n'a de sens en biologie si ce n'est à la lumière de l'évolution »

homologues. La classification en familles permet en effet d'organiser la diversité des gènes, d'étudier sa distribution dans les organismes, de modéliser son évolution. La construction de familles de gènes homologues permet éventuellement d'inférer indirectement l'homologie entre des séquences trop divergentes pour s'aligner. Il s'agit de l'étape préalable à toute analyse phylogénétique.

3.2.1. Bases ontologiques du concept de familles de gènes homologues

La partition des gènes en familles a pour objectif de refléter une partition naturelle, due à leurs relations évolutives[1]. Une famille homologue regroupe « *l'ensemble des descendants d'un ancêtre commun* ». Les gènes d'une même famille sont donc homologues 2 à 2, tandis que deux familles différentes ne partagent pas de gènes homologues. Le concept de *famille homologue* est plus fort que celui d'*homologie ;* il suppose que la relation d'homologie entre gènes est transitive. La base ontologique de ce concept de famille est un modèle arborescent d'évolution des séquences. Ce modèle est directement hérité de la représentation arborescente de l'évolution des organismes, de leur classification phylogénétique pour lesquels les gènes servent de marqueurs. Nous appellerons Strong Tree Thinking (STT) ce mode de pensée arborescent, le terme « strong » soulignant sa vocation d'universalité.

Le concept de *famille homologue* est également compatible avec un modèle d'évolution plus réticulé, qui autorise les recombinaisons homologues au sein des familles de gènes. Ce modèle suppose que les séquences évoluent selon une architecture générale arborescente, à laquelle viennent s'ajouter quelques échanges de segments homologues. Nous appellerons Phylogenetic Network Thinking (PNT) ce mode de pensée, pour référer au réseau phylogénétique employé pour reconstituer l'histoire de ces familles. Notons en passant qu'un modèle de population serait également compatible avec la partition des gènes en familles homologues, tout en faisant disparaître la référence à un ancêtre commun. Un tel modèle de population, qui pourrait être adapté pour décrire les relations de parenté entre séquences, n'est pas employé à plus large l'échelle évolutive ; de la même façon que l'on passe d'un modèle de population à un modèle arborescent pour décrire les relations de parenté entre organismes sexués [O'HARA, 1997]. Par conséquent l'identification de groupements exclusifs de séquences similaire ne s'explique pas forcément par un processus d'évolution arborescent.

3.2.2. Difficultés posés par les processus combinatoires

La diversité des processus responsables des ressemblances entre séquences d'une part, et l'étendue des changements que ces processus induisent dans les molécules d'autre part, ont une conséquence bien connue. L'application stricte du concept de famille homologue développé ci-dessus rencontre des difficultés pratiques, à commencer par celui de l'identification de séquences homologues au sens fort (STT). Puisque deux séquences

[1] Miele "Proteins can be naturally classified into families of homologous sequences that derive from a common ancestor."

similaires sont homologues, un réseau de similarité de séquences peut être considéré comme un réseau d'homologie entre séquences, dans lequel deux séquences reliées par une arête sont homologues, tandis que deux séquences non reliées peuvent être ou non-homologues. Pour détecter des séquences homologues en appliquant le concept strict de famille homologue, qui implique notamment la transitivité de l'homologie, il s'agirait de choisir les composantes connexes d'un réseau de similarité de séquences pour construire des familles homologues. Or le résultat ainsi obtenu ne serait pas satisfaisant, car il agglomérerait la majorité des séquences comprises dans la composante connexe qui apparaît dans les réseaux de similarité comprenant suffisamment de séquences ; les séquences ainsi agglomérées par de nombreuses relations d'homologie intermédiaires n'ayant bien souvent aucune trace d'une origine ancestrale commune exclusive, que ce soit dans leur similarité, leur structure, leur positions dans les génomes, ou leur fonction.

Ces difficultés pratiques proviennent de l'absence de prise en compte des phénomènes combinatoires (autres que la recombinaison homologue) dans le concept standard de familles homologues. Ces phénomènes produisent en effet des relations d'homologie partielle entre séquences, qui ne sont pas transitives et ne permettent donc pas une partition de l'espace des séquences en familles homologues distinctes. Une autre perspective permet de mieux comprendre et décrire ces « agrégations » de séquences dans les réseaux de similarité. On peut en effet considérer que si une séquence a plusieurs origines évolutives différentes, elle descend de plusieurs ancêtres différents et ne peut être attachée à une unique famille de gène.

3.2.3. Adaptation des modèles d'évolution des familles de gènes

Pour répondre à ces difficultés, les auteurs ajoutent généralement des événements d'acquisition de domaines *accessoires* par les gènes à un modèle d'évolution par ailleurs majoritairement arborescent [Song et al., 2008 ; Miele et al., 2012]. Cette conception des processus combinatoires permet de toujours attribuer une origine principale aux gènes, à laquelle vient parfois s'ajouter une origine auxiliaire. Elle préserve la définition d'une famille homologue comme l'ensemble des descendants (principaux) d'une séquence ancestrale commune. Les relations d'homologie partielle sont alors considérées comme « artefactuelles », produites par les domaines présents dans plusieurs familles de gènes qualifiés d'ubiquitaires (promiscuous), qui créent un phénomène de « chaînage de domaine » (domain-linkage).

En pratique, une approche permet apparemment de retrouver des familles homologues dans les agrégats de sommets des réseaux de gènes, en dépit du partage de domaines ubiquitaires. Les familles sont délimitées en découpant le réseau de similarité de séquences en groupes de sommets 'homogènes', en employant l'un des nombreux algorithmes de clustering de graphe existants [Fortunato, 2010]. Ces algorithmes sont généralement développés selon des principes indépendants d'un modèle d'étude particulier. Ils peuvent donc être employés indifféremment par différentes les disciplines (informatique, physique, science sociale, ingénierie, biologie). Le choix d'un algorithme pour un problème donné

reposera donc sur la qualité empirique des groupes produits et sur des caractéristiques générales telles que la possibilité de faire varier la résolution du découpage ou de ne pas imposer le nombre de communautés (clusters) au préalable. La facilité d'emploi des implémentations existantes joue également un rôle important dans le choix d'un algorithme, qui sera d'autant plus utilisé qu'il est facile à installer, simple et rapide. L'algorithme de clustering le plus utilisé pour construire des familles homologues est MCL [Enright et al., 2002], inventé à partir d'un modèle de flux dans le réseau. Cet algorithme doit autant son succès à la qualité des familles produites, difficilement mesurable objectivement, qu'au fait d'avoir été le premier algorithme de ce type spécifiquement appliqué à cette question et accompagné d'une implémentation efficace et facile d'emploi. Au cours de ma thèse, j'ai souvent employé un autre algorithme de clustering appelé 'méthode de Louvain'. Cet algorithme est très utilisé pour l'analyse de réseaux sociaux, car il est simple, bien implémenté, et extrêmement rapide sur de grands jeux de données. Miele et ses collaborateurs ont d'ailleurs développé en 2012 une nouvelle méthode de construction de familles homologues à partir de la méthode de Louvain [Miele et al., 2012].

Les méthodes pratiques de construction de familles de gènes semblent donc dépasser le problème des phénomènes combinatoires sans les modéliser explicitement. Un modèle d'évolution arborescent (STT ou PNT) avec acquisition de domaines accessoires est souvent invoqué pour ne garder que certaines arêtes fiables du réseau, entre séquences qui s'alignent sur une grande proportion de leur longueur et sont donc de la même famille (voire partie 1.5.3.3 1ère solution : Ne garder que les alignements locaux couvrant l'essentiel des séquences). Cette procédure élimine beaucoup de relations, sans pour autant produire un réseau de similarité qui se divise naturellement en familles homologues. Song et ses collaborateurs n'ont de même pas obtenu de résultat décisif en éliminant les relations présumées entre domaines accessoires, grâce à un indice de réseau qui intègre toutes les informations locales de similarités entre séquences. Ces procédures peuvent être employées comme filtres préalables, mais nécessitent toujours l'emploi d'algorithmes de clustering. C'est pourquoi en définitive, les familles sont délimitées en dégageant un signal statistique majoritaire parmi la complexité des relations de similarité entre séquences, par des méthodes largement indépendantes de toute conception précise des mécanismes d'évolution des gènes.

3.2.4. Propositions de l'article pour tenir compte de ces difficultés

Le concept de famille homologue est utile pour étudier l'évolution des séquences génétiques et de leurs fonctions, mais il ne peut s'accommoder que de phénomènes combinatoires accessoires et rares dans son inspiration arborescente (STT, PNT). La partition en familles homologues élimine les nombreuses relations de similarité, évolutivement informatives, entre gènes de familles différentes. Toute une partie de l'histoire évolutive réelle des séquences devient donc invisible à la plupart des analyses qui nécessitent un tel découpage *a priori*. Pour dépasser ce biais systématique, l'article présenté ci-dessous propose de développer une ontologie alternative des notions d'homologie et de famille de

gènes, qui permet d'étudier davantage de relations évolutives, dont ces phénomènes combinatoires habituellement occultés. Après avoir exposé les intérêts de cette perspective, il l'illustre par différents cas d'études et appelle à développer des modèles mathématiques alternatifs, tels que des réseaux phylogénétiques à plusieurs racines (Figure 3-2).

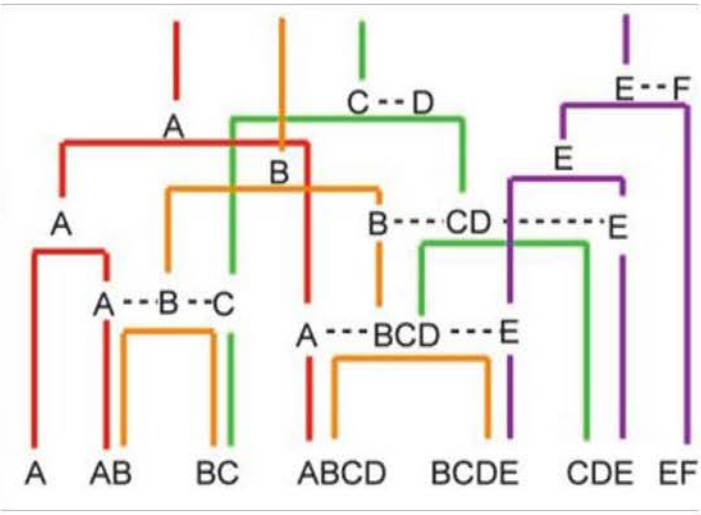


Figure 3-2 : Réseau phylogénétique à plusieurs racines
Evolution d'un ensemble de gènes à partir de 4 gènes ancestraux portant 6 domaines.
Schéma du résultat d'un algorithme de clustering de graphe proposé par Uchiyama, Ikua. : "Hierarchical Clustering Algorithm for Comprehensive Orthologous-Domain Classification in Multiple Genomes." Nucleic Acids Research, 2006

L'article propose d'employer le concept d'air de famille (family ressemblance) pour décrire la nature des groupes de séquences formés par les méthodes de clustering [Bapteste et al., 2012]. Ce concept proposé par le philosophe Wittgenstein [Wittgenstein, 1968] permet en effet de s'affranchir de modèles trop incomplets parce que trop restrictifs de l'évolution des séquences, tout en permettant de délimiter des familles et de tester des hypothèses à propos de l'évolution d'une diversité de séquences. Nous proposons ainsi la conjecture suivante : « Deux protéines non-homologues, mais qui partagent un air de famille – c'est-à-dire qui sont proches dans le réseau de similarité de séquences – seront plus similaires en terme de fonction, de localisation cellulaire, ou autre caractéristique, que des protéines qui ne partagent pas d'air de famille. »

Une autre proposition est d'analyser directement l'évolution des séquences à partir du réseau de similarité, suivant ainsi un mouvement typique de la « network science » initié dans d'autres disciplines. Cette approche est rendue possible grâce au traitement informatique de grandes quantités d'informations. De nombreuses questions évolutives peuvent ainsi être abordée directement, sans recours préalable à un arbre des séquences issues d'un seul ancêtre commun, et donc sans restriction a priori des processus évolutifs générant la diversité des séquences. Sur le long terme, seule la pratique des différentes approches permettra de mesurer leurs intérêts respectifs pour les études de biologie évolutive. On peut cependant déjà constater que les questions abordées par ces différentes approches ne seront assurément pas les mêmes. Ce constat plaide pour une démarche

pluraliste pour exposer une plus grande diversité de phénomènes (processus et patrons) évolutifs.

Ma contribution a été significative pour parvenir à ces résultats. En bref, j'ai développé de nouveaux scripts permettant d'automatiser la construction et l'analyse des centralités et de la structure de composantes connexes géantes de tailles croissantes (plusieurs millions de nœuds et centaine de millions d'arêtes), grâce à des méthodes de parallélisation et de décomposition du réseau. J'ai aussi développé des stratégies améliorant la visualisation de ces graphes (leurs communautés, leur cycles). En outre, j'ai caractérisé de nouveaux patrons par rapport à mon travail précédent, qui signent la présence de composites multiples (cf. figure 1 de l'article présenté dans le chapitre 4). Ces outils ont permis de montrer la prévalence et la complexité des relations entre séquences, et de l'importance de modèles non-arborescents, même chez les organismes eucaryotes, pour l'étude desquels, contrairement aux procaryotes, on s'appuie presque systématiquement sur des arbres et une logique STT.

A Pluralistic Account of Homology: Adapting the Models to the Data

Leanne S. Haggerty,[1] Pierre-Alain Jachiet,[2] William P. Hanage,[3] David A. Fitzpatrick,[1] Philippe Lopez,[2] Mary J. O'Connell,[4] Davide Pisani,[1,5] Mark Wilkinson,[6] Eric Bapteste,[2] and James O. McInerney*[,1,3]

[1]Bioinformatics and Molecular Evolution Unit, Department of Biology, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland

[2]Unité Mixte de Recherche 7138 Systématique, Adaptation, Evolution, Université Pierre et Marie Curie, Paris, France

[3]Center for Communicable Disease Dynamics, Harvard School of Public Health, Boston, MA

[4]Bioinformatics and Molecular Evolution Group, School of Biotechnology, Dublin City University, Glasnevin, Dublin, Ireland

[5]School of Biological Sciences and School of Earth Sciences, University of Bristol, Bristol, United Kingdom

[6]Department of Life Sciences, The Natural History Museum, Cromwell Road, London, United Kingdom

***Corresponding author:** E-mail: james.o.mcinerney@nuim.ie.

Associate editor: David Irwin

Abstract

Defining homologous genes is important in many evolutionary studies but raises obvious issues. Some of these issues are conceptual and stem from our assumptions of how a gene evolves, others are practical, and depend on the algorithmic decisions implemented in existing software. Therefore, to make progress in the study of homology, both ontological and epistemological questions must be considered. In particular, defining homologous genes cannot be solely addressed under the classic assumptions of strong tree thinking, according to which genes evolve in a strictly tree-like fashion of vertical descent and divergence and the problems of homology detection are primarily methodological. Gene homology could also be considered under a different perspective where genes evolve as "public goods," subjected to various introgressive processes. In this latter case, defining homologous genes becomes a matter of designing models suited to the actual complexity of the data and how such complexity arises, rather than trying to fit genetic data to some a priori tree-like evolutionary model, a practice that inevitably results in the loss of much information. Here we show how important aspects of the problems raised by homology detection methods can be overcome when even more fundamental roots of these problems are addressed by analyzing public goods thinking evolutionary processes through which genes have frequently originated. This kind of thinking acknowledges distinct types of homologs, characterized by distinct patterns, in phylogenetic and nonphylogenetic unrooted or multirooted networks. In addition, we define "family resemblances" to include genes that are related through intermediate relatives, thereby placing notions of homology in the broader context of evolutionary relationships. We conclude by presenting some payoffs of adopting such a pluralistic account of homology and family relationship, which expands the scope of evolutionary analyses beyond the traditional, yet relatively narrow focus allowed by a strong tree-thinking view on gene evolution.

Key words: **homology, network, comparative genomics, epaktolog, ortholog, paralog.**

The meaning of scientific terms cannot and should not remain fixed forever by the priority of the original definition. This is simply because our experience constantly outruns our terminology.

—Theodosius Dobzhansky (Dobzhansky 1955)

Defining Gene Families: A Central Complex Task in Evolutionary Studies

Homology is acknowledged as an elusive concept, and yet it is central to comparative evolutionary biology, underpins phylogeny reconstruction (Felsenstein 2004) and developmental biology (Brigandt 2003), and is used extensively in ethology and psychology (Ereshefsky 2007). On the one hand, we have ontological concepts of homology, and on the other hand, practical homology definitions and the relationship between these theoretical and operational issues is a neglected area of evolutionary biology. In this manuscript, we explore a plurality of ontological bases for understanding homology in macromolecular sequences, and by extension, we explore concepts and definitions of gene family. The ontology—the study of what objects exist and how they relate to one another—is an important aspect of enquiry that is generally addressed before any practical effort to apply this ontology. We contend that a tree-thinking perspective has strongly influenced consideration of what the ontological basis of homology might be and has needlessly and unhelpfully constrained understanding through the notion that homologs fit into neat genealogical families that have evolved their differences according to some underlying phylogenetic tree.

It has long been recognized that sequence evolution is not tree-like, in particular because of domain shuffling (Enright et al. 1999; Marcotte et al. 1999; Portugaly et al. 2006). It has also long been recognized that this non-tree–like evolution results in a network of sequence relationships (Sonnhammer and Kahn 1994; Park et al. 1997; Enright and Ouzounis 2000; Heger and Holm 2003; Ingolfsson and Yona 2008; Song et al. 2008). However, for an almost equally long period of time, it has been assumed that the right way to process this network was to carve it into homologous parts by clustering (Tatusov et al. 1997; Enright and Ouzounis 2000; Yona et al. 2000). Relevant clusters have generally been considered to be gene families with all members presenting full homology with one another. Smaller relevant clusters have also been proposed by identifying homologous domains, for example, families of sequences presenting homology over their entire length but frequently of smaller size than entire genes (Sonnhammer and Kahn 1994; Park and Teichmann 1998; Apic, Gough, Teichmann 2001b; Wuchty 2001; Enright et al. 2003; Song et al. 2008). Both of these relatively local perspectives on sequence relationships are familiar to most biologists.

Consequently, the task of defining gene families has been generally delegated to software programs that search for clusters or communities of phylogenetically related sequences. Increasingly, with genomic data sets of genuinely enormous sizes, the problem is considered best handled by such programs. And yet, the practice of placing genes into discrete gene families seems somehow at odds with the existence of domain databases (Corpet et al. 2000; Majumdar et al. 2009) that clearly demonstrate the pervasive influence of non-tree-like processes in molecular evolution (Levitt 2009). We propose not to carve up this network but to analyze its local (Sasson et al. 2003; Atkinson et al. 2009) and global structure (Adai et al. 2004).

In the last 20 years, as public repositories of macromolecular data have been greatly expanded, it has become increasingly apparent that a tree-thinking perspective on molecular evolution, while useful in many situations, is inadequate in a broader context and is far short of universality (because for instance, many, perhaps most genomes do not evolve solely in a tree-like fashion). We address the fundamental meanings of homology and its processual causes because, without precise insights into their meanings, we can only design algorithms or methods of defining homologies and families that carry caveats about the kinds of homologies that are being prioritized.

Without wishing to be critical of the useful and important work of others, it nonetheless seems unavoidable that we must take examples from the literature to provide some context. The TribeMCL (Enright et al. 2003) approach to defining gene families illustrates the problem quite clearly. In the manuscript describing the algorithm, analysis of a database of 311,257 proteins is reported. Depending on the settings of the software, 82,692 "families" could be identified, or 75,635 families or 60,934 families, with the entire automated process taking ~14 h on a large computing cluster (Enright et al. 2003). In this case, the "concept" of family was not explicitly explored at an ontological level (though it built on a general understanding of gene family at that time); therefore, the "definition" of a family was an operational one, based on a setting in a software programme instead of exploring evolutionary history and whether it might be simple or complex. In this case, family definition is a uniformly applied rule where one software option fits all. Here, we suggest that alternatives to such simple approaches are desirable, though perhaps more difficult to achieve. Similarly, while we stress that the TribeMCL approach has proved to be of enormous benefit, we argue that many important evolutionary events and types of family relationship can be missed if this kind of approach is the only one that is taken.

A number of points should be made at this stage before getting to the main argument of the article. In this article, we specifically wish to discuss homology in the context of genes and other genetic components, such as promoters and sub-gene elements—what we term genetic goods (McInerney et al. 2011). For such data, the notion of "homolog" and "gene family" has been written about extensively, but there is still no universally agreed consensus on what either of these terms mean (Duret et al. 1994; Natale et al. 2000; Perriere et al. 2000; Tatusov et al. 2000; Dessimoz et al. 2012; Miele et al. 2012). Additionally, there are significant technical limitations for the detection of homologies. Certain cutoffs are imposed on any analysis, which leads to de facto homologies being missed because the sequences no longer manifest a level of similarity that is greater than expected by random chance. Despite ambitious efforts to reduce this complication, it is likely that large-scale underdetection of homologs is still a problem (Weston et al. 2004; Noble et al. 2005). The argument therefore might be made that every sequence is possibly homologous to every other sequence. That is to say, all extant molecular sequences can trace their ancestry to a single nucleotide that has evolved by duplication and mutation. This idea is not better than the alternative hypothesis that they do not all share common ancestry, because terminal transferase enzymes that exist can generate DNA sequences in a template-independent manner (Greider and Blackburn 1985). Nonetheless, fundamental limitations for software programs do not mean we cannot make progress in understanding homology concepts and improve gene family classification. Acknowledging a plurality of concepts will enhance practical gene family classifications. In particular, we wish to acknowledge that homologies are embedded within a wider set of relationships that we call "family resemblances," and this is fundamentally different to the traditional notion of homology.

Homology Concepts and Homology Definitions

The notion of homology has a long and rich history, starting from before DNA was discovered. In 1868, Owen (1868) wrote a now classic book summarizing his ideas on homologies of the vertebrate skeleton. Owen did not have an evolutionary explanation for homology and interpreted the homologies that he inferred as variants on some kind of "archetype"—an ideal form of the organ that was constructed

by a creator. In his book, Owen (1868) declared that there were three different kinds of homology. Special homology describes when two organs had the same connection to the body and performed the same function. This meant that the pectoral fin of a porpoise was homologous to the pectoral fin of a fish, even though they were manifestly different otherwise. General homology referred to morphological features or parts of features that were of "the same organ" under every variety of form and function. Finally, serial homology referred to organs that were repeated on the body—bristles on the legs of a fly for instance. Owen's chief reason for writing this book seems to have arisen from his frustration with his fellow scientists using the word "analog" when they meant homolog.

Since then, within the field of morphology, the concept of homology has been subject to substantial debate, much of which can be seen as reflecting tensions between qualitative comparative anatomy and quantitative phylogenetics on the one hand, and causal and acausal accounts on the other. Thus, there have been proposals to synonymize homology with the cladistic concept of synapomorphy and accounts of "biological homology" (Mindell and Meyer 2001) that seek to accommodate new data from developmental biology on patterning and differential gene expression by explicating the notion of homology in terms of shared developmental pathways. The importance of ontogeny notwithstanding, of particular conceptual interest, is the notion of genetic piracy (Roth 1988) in which homology of some morphological character persists despite the genetic basis of the trait changing more or less completely over evolutionary time. These other debates illustrate how new data and new understandings of evolution often necessitate new usage of terms and clarification of concepts and models.

When similar frustrations arose almost 140 years after Owen's work, a collection of prestigious scientists felt the need to clarify the meaning of the word homology in molecular sequence data (Reeck et al. 1987). Interestingly, this clarification did not entertain the notion that different types of homology may be required to handle molecular data, possibly because to a certain extent, there was a general consensus on the ontological concept of homology (corresponding to Owen's general homology) though a lack of consensus on the practical identification of homologs. A reading of the literature today would corroborate the feeling that the practical level seems to be the one at which the problems of "defining" homologous genes lies, though in fact, the problems have much deeper ontological roots.

Walter Fitch commented that "homology [. . .] is indivisible" (Fitch 2000). This sentiment is often used in the teaching of evolutionary biology classes and indeed is often quoted. However, Fitch (2000) also allowed for chimeric genes as one exception to this general model. Thus, he wrote:

> If the domain that is homologous to the low-density lipoprotein receptor constitutes 20% of enterokinase, then enterokinase is only 20% homologous to that lipoprotein receptor, irrespective of its percent identity. If at the same time, this common domain were half of the lipoprotein receptor, the receptor would be 50% homologous to the enterokinase. The homologies are not the same in both directions if the proteins are of unequal length! This is the only situation where "percent homology" has a legitimate meaning and, even there, it is dangerous and better called, as Hillis has suggested, partial homology.

In Fitch's view, saying that two proteins were homologous along part of their length was fraught with the potential for misinterpretation. Therefore, the phrase "partial homology" needs to be used with care and should only mean that "this part (X%) of sequence 1 is homologous to that part (Y%) of sequence 2." In this case, some parts of sequences 1 and 2 do have a common ancestor, but we are implicitly acknowledging that their last common ancestor is not also a common ancestor of sequences 1 and 2 in their entirety. It would be a mistake to consider such a change in phrasing merely as a matter of rhetoric. Reeck et al. (1987) pointed out that a precise definition of homology would indeed be "an unimportant semantic issue" if it did not "interfere with our thinking about evolutionary relationships." At that time, in the late 1980s, the problem stemmed from the common interchanging of the words "similarity" with "homology" (e.g., saying that two sequences were 80% homologous when the authors really meant that they were 80% similar in sequence). Reeck et al. offered the solution that "homology should mean 'possessing a common evolutionary origin' and in the vast majority of reports should have no other meaning." Accordingly, Fitch later offered the opinion that homology was "[. . .] an abstraction, in that it is a relationship, common ancestry [. . .]" (Fitch 2000). This last point, we feel, is particularly important.

Thus, the consensus among molecular biologists became that similarity was defined as quantitative by comparing the sequences in question, but that homology was qualitative—sequences are homologs or they are not. In fact, the majority of the literature from that time to present day suggests that homology is a term that specifically refers to genes or proteins that manifest significant sequence similarity along the majority of their length. Databases such as homologene (http://www.ncbi.nlm.nih.gov/homologene, last accessed December 10, 2013) and COG (http://www.ncbi.nlm.nih.gov/COG/, last accessed December 10, 2013) only contain genes that are allowed to be in one family. Although we do not deny that database entries of such sequences are likely or certain to be homologs, sole focus on those kinds of evolving entities (entries that trace their heredity to a single common ancestor) and the heuristic of requiring homologs to manifest near- or full-length significant sequence similarity has clearly resulted in biases and information loss, as has been demonstrated (Sonnhammer and Kahn 1994; Park et al. 1997; Enright and Ouzounis 2000; Heger and Holm 2003; Ingolfsson and Yona 2008; Song et al. 2008). Even if we had a universally agreed definition of the gene (Epp 1997), it remains much more complicated to decide what might be a gene family.

Gene length can vary from dozens of nucleotides (the shortest human gene is 252 nucleotides in length) to several hundreds of thousands of nucleotides. Genes evolve by point mutation, legitimate and illegitimate recombination, exon shuffling, fusion, fission, invasion by selfish mobile elements, domain replacement, and so forth. Is a gene that has a transposon inserted into the middle no longer considered to be a member of this family? If a gene loses an exon and is now quite different in length from other members, then is it no longer considered to be a member of this family? In other words, our current knowledge of the diversity of evolutionary processes means that the generally agreed upon concept of homology needs revision and clarification, and other concepts such as family resemblance need to be introduced.

Recently, there has been an increased focus on the problems that domain shuffling in particular has created for efforts to distinguish orthologs and paralogs from sequences that appear to be orthologous and paralogous, when in fact they are not. Strictly speaking, two genes are orthologous when they are found in different species and can trace a direct lineage back to a single genomic locus in a common ancestor. It can be expected that the sequence in this common ancestor was not significantly different in domain architecture to the orthologs we observe today—though it is not clear how different is too different. Paralogs can trace their most recent common ancestor to a duplication event, again with the expectation that the most recent common ancestor will have had a similar structure. However, in the event that two genes or proteins look similar because they have been independently assembled through domain shuffling, they will not fulfill these criteria. In such cases, the word "Epaktolog" has been suggested to reflect similarity that is a consequence of independently "imported" domains (Nagy, Bányai, et al. 2011; Nagy, Szláma, et al. 2011). Specifically, the authors "[. . .] refer to proteins that are related to each other only through acquisition of the same type of mobile domains as epaktologs" (Nagy, Bányai, et al. 2011). This is an important consideration, and to date we do not have a rigorous analysis of known proteins to understand the extent to which similar proteins are in fact epaktologs and not orthologs or paralogs. However, we argue here that there are additional important relationships beyond those found in epaktologs (see later).

The most widely used method of allocating genes to a gene family is the Markov Clustering Algorithm (MCL) (Enright et al. 2002), which simulates flow through a network of sequence similarity and cuts the network at those places where flow is most restricted. A sequence similarity network is composed of nodes and edges, with the nodes representing gene or protein sequences and the edges representing some measure of similarity between the sequences. In practice, only "significant" levels of sequence similarity are represented at all, and these significant similarities are likely to represent homologous relationships because they are too similar to have arisen by random chance. Examples of such networks are given in figures 2, 4, and 5 and will be discussed later in this article. The idea behind the clustering approaches such as MCL is that unimportant relationships as defined by small, common, promiscuous domains can be safely deleted, leaving the more important relationships, and these can be used to define families. This approach is hugely successful, garnering well in excess of 1,500 citations at the time of writing. The authors have been careful to say that this method should be used with care, and indeed, appropriate usage of MCL for conservative analyses of particular kinds of homologs is expected to result in few if any errors. However, an ontological premise for this method is that a gene can only belong to one homologous family—the method explicitly does not allow a gene to belong to more than one family. This is because it is assumed that either there are "natural" discrete families and the relative strength of association between a gene and its family will emerge from the analysis or that some relationships are more important than others and the minor relationships can be dismissed as relatively unimportant. Although the philosophy of the approach (clearly influenced by the underlying assumption that gene evolution might be tree-like and takes place independently in different families) has not been explored extensively in the literature, we will argue that the effect of this algorithm is to principally enforce a tree-based viewpoint on gene families. This introduces persistent issues in homology definition that can best be overcome by first adopting more realistic starting assumptions on how genes evolve, second by adopting new concepts of homology, and third by adjusting our methods accordingly.

Defining Homologs Meets Different Kinds of Problems

The lack of agreement in how to define homologs (Fitch 2000; Enright et al. 2003; Li et al. 2003; Wong and Ragan 2008; Majumdar et al. 2009; Dessimoz et al. 2012; Miele et al. 2012) reflects the historical ideas concerning homology and the attempt to fit notions that were developed for one purpose (morphological systematics and comparative anatomy) to data that are only obliquely related to this purpose. The first evolutionary character matrices (Abel 1910; Tillyard 1919) were rectangular consisting of M rows $\times$ N columns. Most phylogenetic software programs today require such rectangular matrices, and if the sequence data do not fit into a matrix, then the user has two choices—either add characters to represent "missing" data or prune the data until it becomes rectangular (Capella-Gutierrez et al. 2009). Therefore, there is an implicit assumption that data matrices should look like this and an explicit requirement that the data is made to look this way. Given that discussions of the pruned parts of alignments rarely make their way into the final manuscript, we have no clear idea how often these nonconforming data sets arise as a result of introgression and gene family membership that involves more than one family.

Additionally, focusing on different aspects of sequence relationships, that is, the homology of entireties or of parts, leads to different inferences of relationships and, consequently, to a lack of consensus. The reality is of course that different parts of a gene sequence might have different histories, so an honest appraisal of homology might require a more radical view of homology than is traditional. Recently,

Song et al. (2008) offered a good example of this when they asserted the restrictive caveat that homologous genes must be descended from a common ancestor that had the same multidomain structure as contemporary sequences. Two genes that share a single domain and whose common ancestor had quite a different structure are not considered to be homologous in their model. The distinction between the two different kinds of evolutionary trajectory is of course important; however, it does seem to confuse the notion of homology being the concept of relationship through common ancestry, irrespective of how subsequent introgressive events have changed the overall domain neighbourhood. It is quite likely that what Song et al. (2008) call domain sharing but not homology is what Fitch (2000) and Hillis (1994) would call partial homology. Though it is perfectly reasonable to say that convergently remodeled proteins with similar structures cannot be true orthologs or paralogs, they are homologs, nonetheless.

Three Homology Models

In terms of homology concept and delineating homology groupings, a fundamental problem lies in the a priori model that we apply to our approach. Here we define three sets of models, and we discuss how these models can affect notions of homology. First, we have "strong tree thinking" (STT). This perspective sees that the important, perhaps only, relationships are those that have arisen along a diversifying phylogenetic tree, and events such as residue substitution and small indel events account for the changes between sequences. A phylogenetic tree, we emphasize, allows no introgressive events (Bapteste et al. 2012). STT is useful when analyzing sets of homologs that have a tree-like history and is generally seen in the analysis of nonrecombining orthologs to determine species relationships (Doherty et al. 2012) or nonrecombining paralogs to understand duplication events (e.g., Feuda et al. 2012). Next, we define "phylogenetic network thinking" (PNT) where legitimate recombination events are allowed, and these turn a phylogenetic tree into a phylogenetic web (Huson and Scornavacca 2011) relating closely related sequences without affecting homology relationships. PNT is extremely useful for analyzing legitimate recombination (Huson and Bryant 2006) and understanding incongruence in gene or genome histories. Finally, we have "goods thinking" (GT) that sees evolutionary history as being characterized by the vertical and horizontal transmission of genetic goods, allowing introgressive evolutionary events (e.g., legitimate and illegitimate recombination events, fusion, fission, etc.) and depicting relationships between sequences in a more pluralistic manner (McInerney et al. 2011; Bapteste et al. 2012). GT is the least conservative perspective and is the main focus of this manuscript. Its biological implications are potentially huge because it has been proposed that introgression of domains has resulted in the evolution of various signaling systems (Apic and Russell 2010) and a correlation has been suggested between the prevalence of proteins with multidomain architectures and organismal complexity (Apic, Gough, Teichmann 2001a). Indeed, a modest increase in number of domains allows for numerous novel genetic interactions, thus a small increase in genes sharing goods could be largely sufficient to construct complex hosts (Koonin et al. 2002).

Going back to Reeck et al. (1987) important definition according to which "homology should mean 'possessing a common evolutionary origin' and in the vast majority of reports should have no other meaning," we want to stress that a fundamental issue stems from the interpretation of the word "a" in the quoted sentence. Traditionally, evolutionary biologists have used the word "a" in the STT sense (O'Hara 1997) or the PNT sense and judge that it means "one." For both of these perspectives, the definition of homology can only mean that homologs must trace back to a single common ancestor without gene remodeling by sharing of DNA from other lineages. According to these perspectives, the community of descent that unites complete genes with complete genes corresponds to the objects such as the branches on phylogenetic trees or networks when these structures have been constructed from genes that are homologous along their entire length (Li et al. 2003) and where the genes have not been remodeled by illegitimate recombination throughout their history. This is probably the most commonly understood definition of homology, and it is certainly the focus of many software tools and algorithmic developments. Embracing this perspective (STT/PNT homology concept), a standard operational criterion (STT/PNT homology definition) for homology is, for instance, that homology extends for at least, say, 70% or 90% of the length of the two genes being examined.

However, if we interpret a in GT sense; McInerney et al. 2011), "a common ancestor" means "at least one" ancestor in common with other proteins. Then, our concept of homology is quite different and allows us to analyze a greater number of evolutionary events and relationships, though we must be much more careful about what we say about these evolving entities. So far, this GT perspective has not been explored much. The concept of homology has usually been described in terms of just the STT/PNT viewpoint—rather than the GT viewpoint—and software and databases have been geared toward the analysis of homologs defined under the aegis of the STT/PNT concept.

Instead of the traditional, narrower view of homology, we advocate that the pluralistic account of evolutionary processes and thus a pluralistic interpretation of the term "a" in Reecks et al.'s definition is now scientifically most fruitful, because it results in definitions of GT-style homologs and family resemblances that can encompass a greater variety of our empirical observations on sequence structures and is a better fit to our observations on the processes responsible for sharing of genetic "parts," at the molecular level, in evolution. Indeed, STT/PNT expectations for how homologs should look have resulted in practical definitions of homology that have often restricted how we have viewed gene, genome, and protein evolution, have affected the software and databases that have been developed to analyze genomic data, have affected the ways in which we think we should analyze macromolecular sequence data and may have frequently succeeded in blinding us to many crucial evolutionary events. For

instance, a recent publication (Miele et al. 2012) that deals with finding homologs from an exhaustive comparison of all macromolecular sequences in a data set against all other sequences in that data set starts with an opening line in the "motivation" section of the abstract thusly: "Proteins can be naturally classified into families of homologous sequences that derive from a common ancestor." The manuscript then goes on to describe a very promising method for clustering protein sequences into groups that manifest extensive similarity along almost their entire length. This clearly is a STT/PNT view of protein evolution, where a family is defined in explicit, though narrow terms and all segments of sequence are expected to descend from one same common ancestor, not many ancestors. While being a completely legitimate way of thinking about some protein relationships, the complexity of the majority of data falls outside this narrow framework, and we advocate that additional homology concepts can provide an augmented view of protein evolution.

Homology Concepts That Do Not Assume Tree-Like Evolution

Is there a homology concept that fits the data better than the STT/PNT concept and could it conceivably reduce the likelihood of overly restrictive and potentially incorrect inferences occurring? We think that the most efficient way to ameliorate the risk of error and to really account for evolutionary relationships between sequences is to realize where the most fundamental problem lies. Most algorithms would run quickly if genetic data had genuinely evolved in a tree-like way. In fact, no sophisticated algorithm would be necessary at all, as the gene families could be easily parsed from an all-versus-all gene similarity search and, assuming the search was sensitive enough, they would naturally fit into their respective families. However, real data have experienced more complex evolutionary processes (Nagy, Bányai, et al. 2011; Nagy, Szláma, 2011).

We propose that methods for defining homologous genes (gene families) that require homology to extend along most of the sequence (Miele et al. 2012) might be described by the search for "tribes" of proteins. We choose the word tribes, because this is the original meaning for the word phylogeny (from the Greek *Phylos* meaning "tribe" and *Genis* meaning "origin"; Sapp 2009). Therefore, such tribes of sequences are likely to be amenable to phylogenetic tree or network construction using standard software currently available (Felsenstein 2004; Huson and Scornavacca 2011). We note that this fits well with the objective of such programs as TribeMCL (Enright et al. 2003).

In continuing with the etymology of the word phylogeny, we wish to point out, however, that tribes are known to split and merge with other tribes, to subsume, and to be subsumed. Although analyzing homology along the entire length of a sequence is somewhat akin to a tribal origin analysis (a phylogenetic analysis of that tribe), it is by no means the only way that we can look at homology. We might consider that at one extreme there are tribes of sequences that are mostly isolated "closed" tribes (fig. 1, Family A) but that

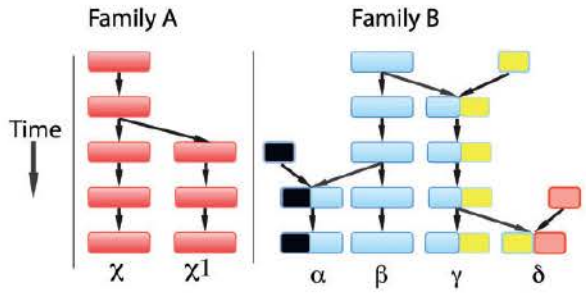


FIG. 1. Two extremes of family evolution. Family A is a closed family shown to evolve according to a strict tree-like process, Family B is an open family that evolve by horizontal and vertical evolutionary processes. Its members display family resemblances, as they can be connected through intermediates and relationships of GT homology (see main text).

there are also tribes that are more "open" in terms of tribal mergers and divisions (fig. 1, Family B; Boucher and Bapteste 2009). In the case of Family B, it would be standard practice to split the family into four tribes to carry out phylogenetic analyses, thereby missing out the context in which the entire family has evolved. These open tribes are not readily analyzed using current phylogenetic methods, because the components of some of the sequences have separate origins and separate roots (in our toy example, the black, blue, yellow, and red gene parts all have separate roots). In other words, evolution has frequently occurred through introgression (Bapteste et al. 2012) with genes and parts of genes acting as goods (McInerney et al. 2011) that can be shared, such that a homology concept that only accommodates STT/PNT is likely to be incomplete as a basis for categorizing and describing the evolutionary histories (Bapteste et al. 2012). To demonstrate this, we explore the assumptions and expectations of STT/PNT.

Building on the historical role of morphology in the study of homology, STT/PNT considers either a complete organ or a significant part of an organ. This perspective has some consequences for the breadth and depth of analyses that can be carried out. The first consequence is that the organ should be clearly defined as a 1:1 correspondence. In contrast, most new genes are constructed from existing parts; fusions of genes, promoters, introns, exons, and motifs are common (Levitt 2009; Bapteste et al. 2012). This means that different parts of proteins can be expected to have different evolutionary histories. The different parts of a protein-coding gene might themselves be homologs of one another and may have arisen by tandem duplication or introgression of previously spatially separated DNA sequences (Bapteste et al. 2012). Even within morphology, it has been recognized that partial homologies offer a much broader view of evolution (Sattler 1984).

The second consequence of STT/PNT-based explanations of homology is that the notion of homology being indivisible is easy to understand—two organs/genes are either homologs or they are not. The problem we have with molecular sequence data is that we now know that a great number of molecular sequences are related to a great many other molecular sequences with varying amounts of structural (e.g.,

domain content) similarity (Adai et al. 2004; Halary et al. 2010; McInerney et al. 2011; Bapteste et al. 2012; Alvarez-Ponce et al. 2013). Consider the thought experiment where we have four proteins (see table 1), each protein has two domains and we have four domains in total. Gene1 has domains A and B, Gene2 has domains B and C, Gene3 has domains C and D, and Gene4 has domains A and D. All four proteins have particular kinds of relationships to the others that cannot be described by an "all or nothing" model. This problem affects both the homology concept and the homology definition. We will refer to this thought experiment when dealing with real data in "case 4" later.

Current STT/PNT thinking does not address most of the issues we have just raised, because, being founded on an assumption of tree-like evolution, it produces a bias against the detection of introgressive processes. Relied upon exclusively, it prevents us from investigating those non-tree-like evolutionary events and relationships that could be revealed through a more pluralistic view of homology. In the following three examples, we use a standard set of analytical tools to demonstrate how our views of what constitutes a homologous family are influenced by the use of such heuristic approaches. We use BlastP (Altschul et al. 1997) and then pass the data through the MCL software (Enright et al. 2002) using default parameters.

Table 1. An Illustration of Four Hypothetical Genes That Manifest a History of Introgressive Processes.

Gene	*Domain Structure*			
Gene1	A	B		
Gene2		B	C	
Gene3			C	D
Gene4	A			D

NOTE.—Each gene consists of two domains, the colors are the same for homologous domains.

Case 1: A Ten-Gene Data Set from Four Enteric Bacterial Genomes

In this case, we analyze data from four enteric bacterial genomes—one *Escherichia coli*, one *Salmonella*, one *Yersinia*, and one *Shigella* genome (data available as supplementary information S1, Supplementary Material online, Case1.aln). Homologous proteins with a helix-turn-helix motif are found ten times in these four genomes using a standard similarity search algorithm (Altschul et al. 1997). However, these genes are short and quite variable. Short gene length reduces the possibility that Blast can detect significant sequence similarity. Figure 2 depicts the gene similarity network that can be constructed from this gene family when an all-versus-all Blast analysis is carried out with a cutoff e-value of 10^{-6}. As can be seen, not all genes show significant sequence similarity with all other genes according to this analysis. However, using Clustal Omega (Sievers et al. 2011), the alignment shown in figure 2 can be produced, and using FastTree with the default parameters (Price et al. 2010), the tree shown in figure 2 can be produced from that alignment. The Blast network also shows an analysis of what happens if the MCL software (Enright et al. 2002) is used to identify homologs with the default inflation value set at 2.0. MCL cuts this graph into three tribes. The color coding of the sequences on the Blast graph, the alignment, and the phylogenetic tree reflects how MCL would carve up the data. The STT/PNT proposition is that a gene family would be characterized by all members of the gene family recognizing all other members in a similarity analysis. This does not happen, so MCL divides up the gene family into three tribes. In these tribes, all members recognize all other members.

One of the features of note in this alignment is that the proteins are quite variable in length, and indeed, this is likely to be part of the reason why Blast does not produce a completely connected component where all sequences show significant similarity to all other sequences. The four sequences shaded in brown contain a conserved 18-amino acid stretch that has either been gained by these sequences or lost in the

Box 1.

Term	Meaning
Homologs	Having a relationship through descent from at least one common ancestor
Family resemblance	Having an evolutionary relationship through intermediate sequences and common descent
Clique	A subgraph in a network where every member of the subgraph is connected to all other members
STT	Strong tree thinking: A perspective that sees homology statements as valid when the homologs have evolved down the branches of a bifurcating phylogenetic tree
PNT	Phylogenetic network thinking: A perspective that sees homology statements as valid when the homologs have evolved through tree-like processes, but allowing for some homologous recombination, thereby making a phylogenetic network.
GT	Goods thinking: A perspective that sees homology relationships encompass illegitimate recombination, fusion, and fission of evolving entities in addition to vertical descent. Gene evolution is expected at times to be very complex and involve merging of evolving entities.
N-rooted fusion networks	A new kind of network that depicts rooted networks with at least one fusion node and at least two roots.
TRIBES	Homologs that have a 1:1 correspondence in terms of being homologous for most or all their length.
TribeMCL	One of the most successful approaches to finding communities in networks of gene similarity.

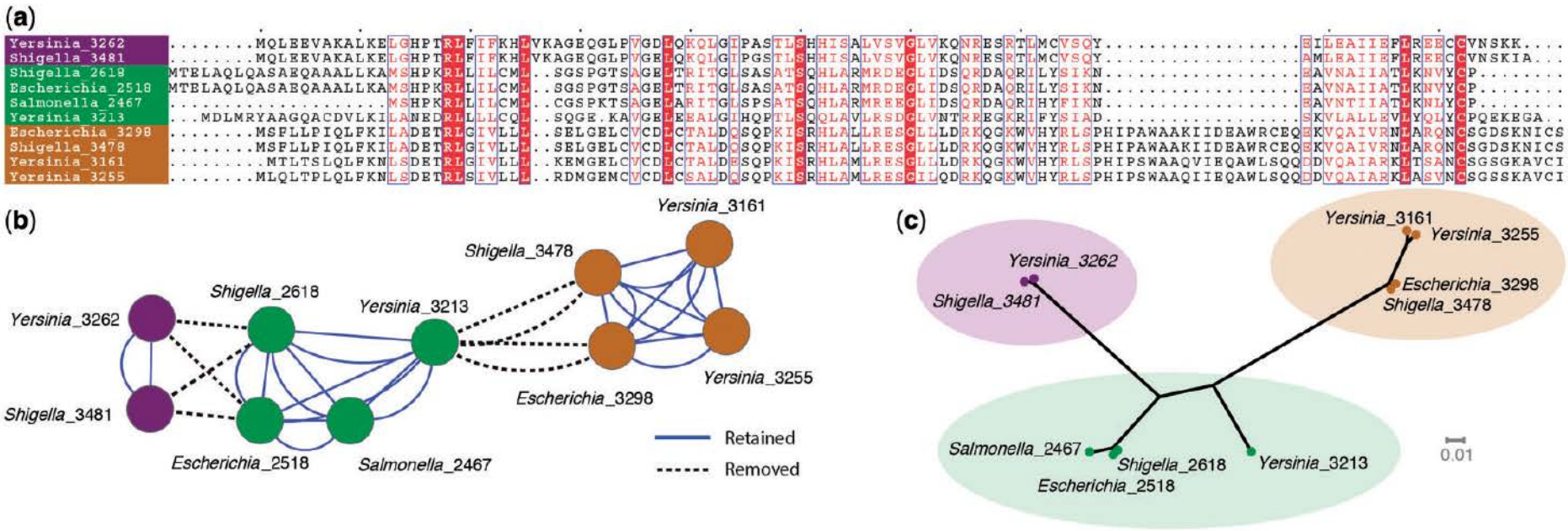


FIG. 2. A ten-gene data set of a family of short proteins with considerable variation, including segmental length variation and possibly a chimeric history. (*a*) Multiple sequence alignment with completely conserved positions in the alignment are indicated by columns with a red background and white typeface, strongly conserved positions are in red typeface and surrounded by a blue rectangle, and the most variable positions are in black typeface. Positions where there is no homologous residue are represented by a dot. (*b*) Results of an all-versus-all Blast search with all proteins represented by a node and all significant hits represented by an arc drawn between nodes. Arcs drawn as dashed lines reflect those edges that are removed in a standard TribeMCL (Enright et al. 2002) analysis. (*c*) A phylogenetic tree inferred from the alignment. See main text for details of analysis.

others. In addition, there is considerable length variation at the N- and C-termini of the sequences.

In the analysis of this alignment of sequences, an STT/PNT perspective will be faced with a conundrum. The four sequences at the bottom of the alignment, identified by the brown taxon labels on the alignment, brown nodes on the network, and brown tip labels in the tree, have an 18-amino acid stretch that is clearly homologous among these four sequences and is absent in the other six sequences. Although the sequences clearly manifest homologous relationships, should this process of insertion or deletion occur frequently, then descendants of these sequences might not contain any amino acid residues homologous to the residues that exist today. As a thought experiment, imagine if we discovered a series of proteins that differed from each other by the presence or absence of small domains, eventually leading to a collection of sequences where at the two extremes of this series we have proteins that do not share any domains (as in table 1). Then, both the STT and PNT perspectives would say that these proteins at the extremes do not have a relationship through common ancestry, whereas a GT perspective would say that they do. A GT model for homology that we could designate as the "open tribes" or family resemblance model would better accommodate this kind of situation, which we show later in this manuscript to be a very common situation.

Case 2: A Set of Orthologs from Closely Related Yeast Species

The Candida Gene Order Browser (CGOB) database (http://cgob.ucd.ie, last accessed December 11, 2013) is a carefully curated data set of 13 yeast (mostly *Candida*) genomes that have been aligned so that any particular gene can act as a "focus point," and all its orthologs (if present) are presented to the viewer as pillars and their neighbors are also visible (see Fitzpatrick et al. 2010 for details). Figure 3*a* shows an example from this database. Open reading frame 19.2962 from the genome of *Candida albicans* is the focus gene and its orthologs are displayed underneath it. On the left and the right of this gene are two genes that are strongly conserved in all species. Orthology is easily recognized in these neighbors using standard analyses of similarity. In the pillar that is in focus (the orf19.2962 pillar) are 11 orthologs of this gene, with the ortholog being absent in the genome of *C. lusitaniae*. Figure 3*b* shows a network representation of the all-versus-all database search for this set of orthologs. The nodes in green produce a significant Blast hit when compared with orf19.2962. As can be seen, only three genes produce a significant result. The other orthologs are included in the network only as a consequence of the full analysis of Blast searches. Applying MCL to this data set results in six Blast hits (statements of homology) being discarded and consequently splits the data into two communities. In standard phylogenomic analyses, this set of orthologs that are weakly conserved in sequence but strongly conserved in genomic location might be analyzed as though they are two separate families, when in fact by any reasonable criterion, they should be analyzed as a single, albeit quite variable, family.

We used the CGOB database to explore how often the standard Blast approach to detecting orthology would fail to detect de facto orthologs. The CGOB contains 6,548 orthology pillars that obviously contain two or more orthologs. Of these, 707 contain at least one ortholog that would be missed in an all-against-all Blast search of the database. They have been manually included in orthology pillars based on synteny and weak Blast hits. This constitutes ~10.8% of CGOBs orthology pillars, where on the basis of Blast alone, the orthologs

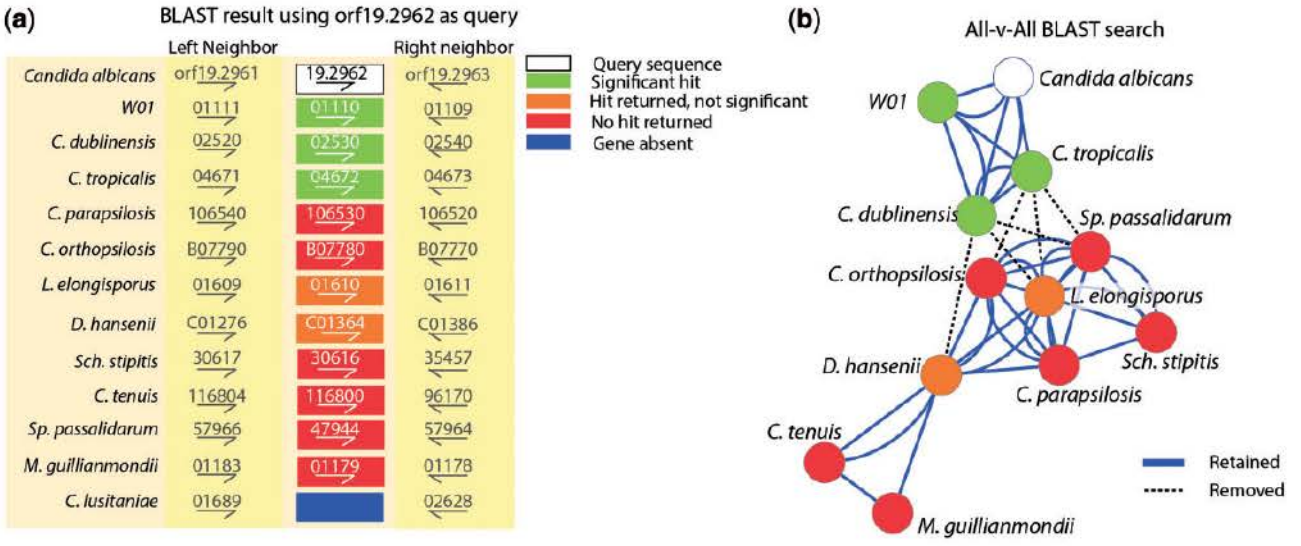


FIG. 3. Analysis of a family of divergent orthologs in *Candida* and close relatives. (*a*) A view of three pillars from the CGOB database showing orthologous genomic locations for 13 organisms, with gene names as per the CGOB database. The focus gene (orf19.2962) is colored in white, the three orthologs that are identified in a standard Blast search are colored in green, and the other orthologs are colored in orange if they are identified in the Blast search as a "hit" but not a significant hit an the gene is colored red if no hit was returned. (*b*) A sequence similarity network constructed using Cytoscape (Shannon et al. 2003) based on the pattern of significant Blast hits from an all-versus-all search. Dashed lines indicate where the MCL algorithm splits the data into two partitions.

would be split into more than one family. Naturally, we anticipate that this figure will increase substantially as we move toward examining organisms that are more distantly related than a group of closely related Ascomycota. Many or most "unknown" proteins could or should be placed into existing families, were it not for the limitations on computational tools and—very specifically—approaches.

Case 3: *N*-Rooted Networks

Figure 4 presents the results of two analyses of proteins that have likely experienced a gene fusion. This gene fusion is clearly seen in figure 4*a*, which is a sequence similarity network based on Blast searches. There are in fact two maximal cliques (completely connected subgraphs that do not exist exclusively within the vertex set of a larger clique) in this network. The collection of three genes on the left of the network and the three genes in the middle of the network collectively form a clique. In addition, the three genes in the center of the network and the seven genes on the right also form a clique. The three genes on the left and the seven genes on the right are not directly connected to each other. This kind of graph topology strongly suggests a gene fusion or fission event. In this example, we are going to assume that the three genes in the middle clique are derived fusion genes and not ancestral (note that the following will be true for any genuine product of gene fusion even if this specific network is not).

One set of proteins (the three genes on the left in fig. 4*a*) are members of the COG1123 family as defined in the COG database (Tatusov et al. 2000) and they function as ATP-binding proteins. The second family of seven genes on the right belong to COG0842 and function as ABC-2 type transporters. The fusion genes are bifunctional ATP-binding and transport proteins. For this analysis, we aligned the fusion proteins (a total of three proteins) separately with the COG1123 proteins (a total of three proteins in this family, resulting in a six-sequence alignment), and separately, we aligned the three fusion proteins with the seven members of COG0842. These two alignments were merged into a

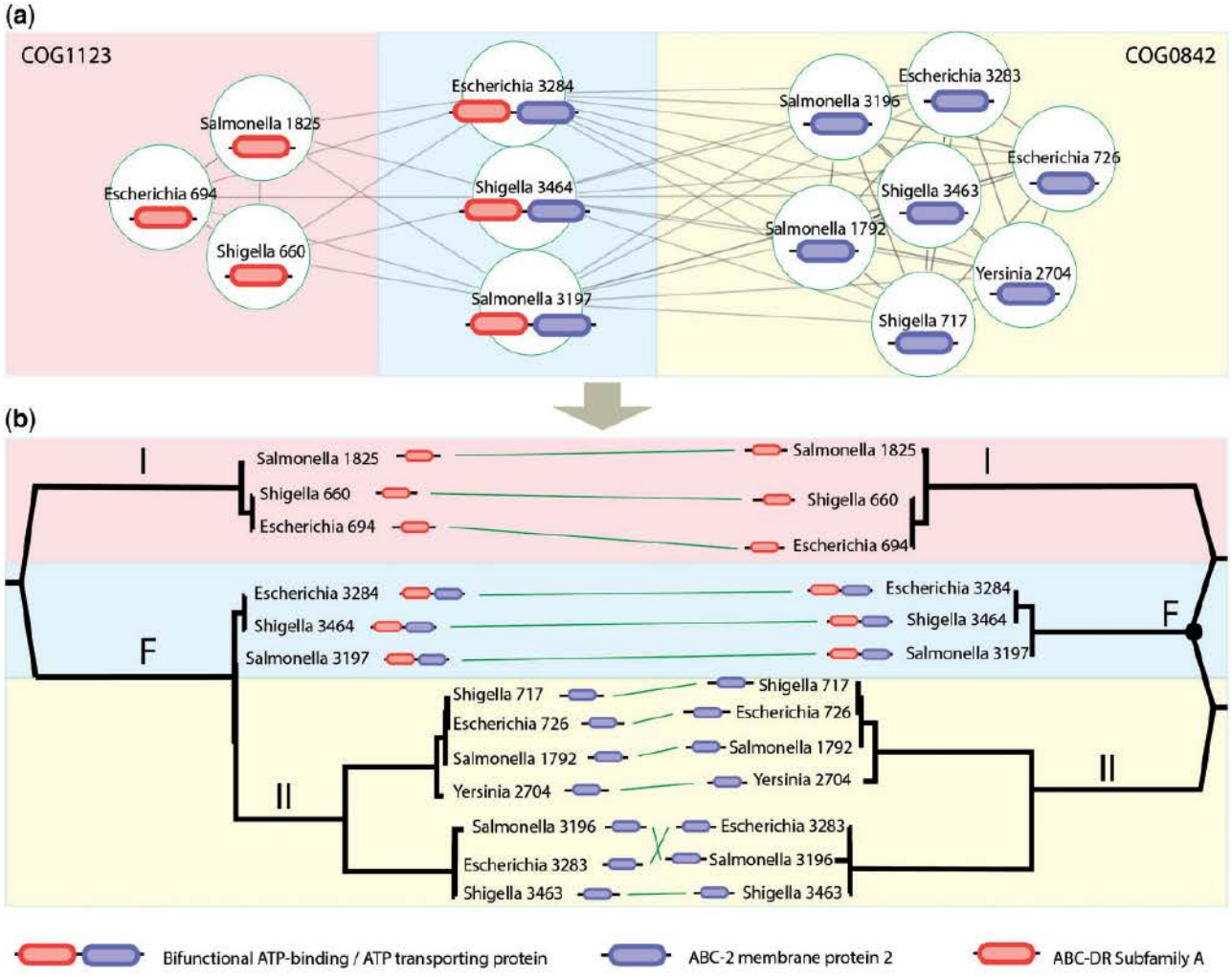


FIG. 4. An example of a data set that cannot fit onto a conventional phylogenetic tree diagram. The sequence similarity network displays the significant similarity results from a Blast search of the collection of proteins against one another. The tree on the left is the tree recovered from a concatenated data analysis and rooted arbitrarily on the internal branch separating the COG1123 proteins from the rest. The network on the right is what we call an *N*-rooted network (in this case $N = 2$, so it is a two-rooted network).

single alignment (available in supplementary information S1, Supplementary Material online) and two analyses were carried out.

The first analysis is seen in figure 4*b*, left tree. This tree was constructed from a complete alignment of the data, with missing parts padded out in the alignment using gap characters. The resulting tree is manifestly incorrect from two perspectives. First, COG1123 and COG0842 should have two different roots because they have two different origins, yet this diagram depicts a single origin of the entire tree. Second, there is no rooting of this tree that can depict the fusion event properly. This is because this representation—a tree—is not how the data have arisen. A fusion event is accurately represented by a node with an in-degree of two, and standard phylogenetic trees do not contain such nodes. The network on the right of figure 4*b* is an accurate representation of how the data have arisen. In this case, the N-terminal end of the fusion proteins were aligned to the COG1123 sequences (resulting in a six-sequence alignment) and the C-terminus portions of the fusion proteins were aligned to COG0842. The FastTree software (Price et al. 2010) was used to construct two maximum likelihood trees from the data, and then these trees were midpoint rooted and merged manually using the Adobe Illustrator software (naturally, there is more than one way to generate such a graph, but for illustration purposes, we chose this method). The resulting network, which we call an *N*-rooted fusion network, is a more accurate representation of the evolutionary history of these sequences. The two roots of the network are indicated, and the approximate location of the fusion event is indicated using the black dot. We note that this is an ad hoc placement of the fusion event—future work can focus on methods for accurately investigating the location of a fusion node. We cannot rule out the possibility or indeed likelihood that the genes described here are in fact related through some ancient undetectable community of descent. This would mean that, for the two-rooted network in figure 4*b*, we would simply be leaving out the edges of the network that would unite the two root edges further back in time, turning this two-rooted network into a more classic phylogenetic network, as expected in PNT. Of course, it is also possible that these two roots would join other kinds of families that would join other kinds of families and so forth, consistent with GT. Thus, although this simple example has two root nodes (it is a two-rooted fusion network), large multidomain proteins probably need to have their evolutionary history represented by 3-, 4-, or *N*-rooted networks, as indicated by our next example.

Is it possible that COG1123 and COG0842 are indeed homologous in the PNT sense, but this homology cannot be detected? As we have said earlier and as seen in Cases 1 and 2, there is a severe technical limitation that means that many homologies are not detected. This affects our homology definitions more than our homology concepts. Even if there is deep, undetected homology of the PNT variety between these two groups of genes, *N*-rooted networks are useful for providing a more complete picture of evolutionary relationships.

Case 4: Composite Genes in the Genomes of 15 Eukaryotes

The single connected component shown in figure 5 illustrates the value of GT (McInerney et al. 2011) in the study of homology. To generate this figure, we have used sequences from a total of 15 eukaryotic genomes (see supplementary information S1, Supplementary Material online). The total number of genes was 199,592. A similarity network was constructed from this data set using the BlastP program (Altschul et al. 1997) with the cutoff set to an e-value threshold of 1e−10. We searched through this network for composite genes, using the program FusedTriplets.py (Jachiet et al. 2013) and a verification test at 1e−20 threshold. Thus, a gene C is identified as composite if there are two component genes A and B such that: A and C are similar, with an e-value less than 1e−20; B and C are similar with an e-value less than 1e−20. In addition, A and B Blast matches on C do not overlap, and A and B are not similar, with an e-value greater than 1e−10. Next, we looked for multicomposite genes, which is the name we give to composite genes whose component genes (A and B) are themselves composites.

The similarity network has a giant connected component (GCC). This GCC contains 41.4% of the nodes (82,702) and more than 90% of the edges (8,826,323). It is very dense, with a mean degree of 200. This makes it impossible to visualize with Cytoscape (Shannon et al. 2003) or Gephi (Bastian et al. 2009).

Interestingly, we have a situation for this relatively small data set of just 15 genomes, where we can find a chain of significant sequence similarity between any two pairs of genes for almost half of the genes in the network. Under the conventional homology concept, the distant homology between any pair of dissimilar sequences is only retrieved by a chain of homologous intermediates with entire length similarities. An alternative GT-based explanation is that sequences with different ancestors recombine to create intermediate sequences that share partial homology with both of their ancestral sequences. Figure 5 illustrates that this alternative explains most of this pattern in the data. This is the situation that is outlined in table 1. In this case, we do not suggest that we alter the meaning of homology so extensively that sequences that have no ancestor–descendent relationship to one another are still considered homologous. Instead, homologous relationships are those where descent from at least one common ancestor has occurred and family resemblance relationships (Wittgenstein 2009) are those where a path of significant similarity can be found through a graph like we see in figure 5 that links the two sequences.

Composite sequences as identified by FusedTriplets (Jachiet et al. 2013) uncover this kind of nontransitive relationship that may result from nonhomologous recombination, domain shuffling, gene fusion, or indeed fission events. Most of the represented communities—and almost all of the largest and central communities—contain at least a small proportion of such composite sequences. A total of 24% of the sequences in the GCC contain a composite signature (which explains the yellowish look of the result), to be compared with the 6% proportion of composite sequences for the

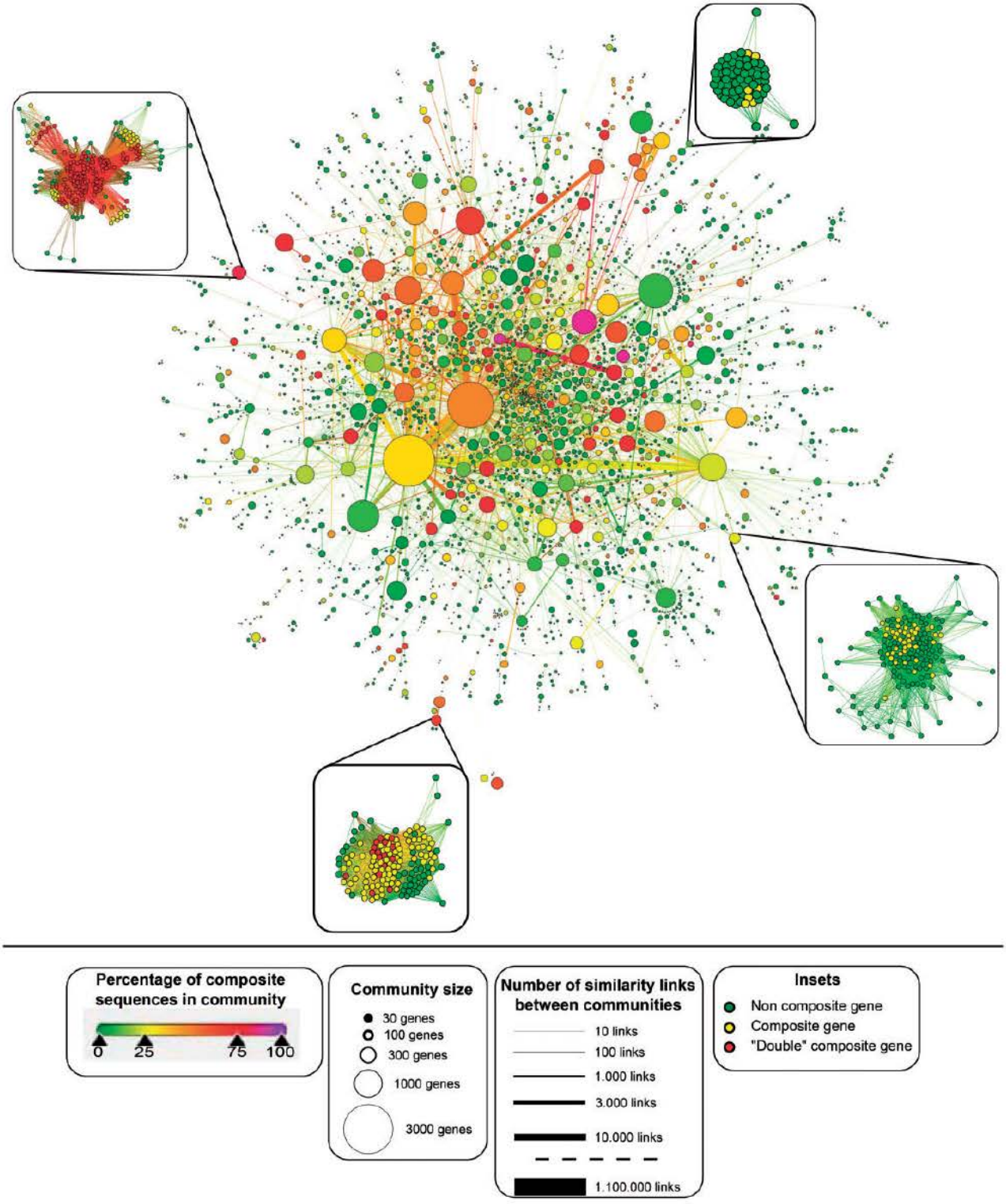


FIG. 5. GCC from all-against-all BlastP search of 15 eukaryotic genomes. Nodes represent communities as identified using a single pass of the Louvain algorithm. Node area representing size of community and edge thickness is the square root of the number of edges connecting two nodes, with the exception of the largest edge that has its size represented by a thickness five times smaller (corresponding to 220,000 edges instead of the actual 1,100,00). Nodes on the left diagram are colored according to the proportion of composite genes in the community (from green = 0% to purple = 100%). Subnetworks of four communities are displayed around the figure. These communities have been chosen along the range of composite proportion (from light green to light red) to illustrate the variety of community structures. Nodes from these insets are colored in green for noncomposite sequences, yellow for composite sequences, and red for multicomposite sequences, that is, composites sequences whose component genes are themselves composites. See supplementary figure S1 (Supplementary Material online) for a pie chart representation of the proportion of noncomposite, composite, and multicomposite genes in each community.

rest of the network (outside the GCC). Furthermore, some composite sequences also tend to recombine, with 10% of sequences identified as multicomposite sequences in the GCC. The structure of the GCC (and of some communities) exhibits large cycles without chords (holes), which also provides evidence of multiple introgressive events in the history of these proteins. This demonstrates the extent to which we can see non-tree-like evolution in many places in this data set.

Phylogenetic software tools or methods have not tackled the evolution of composite molecular sequences, despite the pervasiveness of introgression. The complex, yet real relationships between remodeled genes remains a blind spot for most analyses, because most analyses are performed at a much more local scale after the clustering steps. It is not clear why this perspective is the one usually adopted, because there are several databases of multidomain proteins (Majumdar et al. 2009), and a high level of interest in how domains combine (Sonnhammer and Kahn 1994; Park and Teichmann 1998; Enright et al. 1999; Marcotte et al. 1999; Apic, Gough, Teichmann 2001b; Wuchty 2001; Enright et al. 2003; Portugaly et al. 2006; Song et al. 2008). However, the dominant concept of STT homology, the focus on tree thinking as the prism through which we should view evolutionary histories, has undoubtedly played a role.

A Pluralistic Account of Homology

The concept of homology is defined as "descent from a common ancestor." However, unless we include situations where the number of ancestors is greater than one, then it is necessary to ignore many real relationships—at the moment, this is a very common situation. The standard classifications of homologs place them into the category of ortholog (originating as a consequence of speciation), paralog (created by gene duplication), xenolog (created by horizontal gene transfer of an entire sequence), or ohnolog (created by whole genome duplication), all of whom are divergent events that are expected to appear under the standard concept of homology and are adequately analyzed using phylogenetic trees or phylogenetic networks. In contrast, the merger of two evolving entities (Bapteste et al. 2012) is not expected under a tree-thinking perspective and the standard concept of homology. Very little software has been developed to take account of this kind of process, and indeed, where software has been developed to analyze introgressive events, the resulting homologs have been described as not being homologs at all (Song et al. 2008).

Evolutionary biologists might wish to know about the evolution of more complex gene families, for example, the origins of entire connected components in a gene network and not just members of the same tribe. Alternatively, it might not be interesting to carry out such a broad-scale analysis and instead a narrower focus on a closed family or a subset of members of an open tribe is desired. If the latter, is it possible to clearly articulate why this subset of evolutionary events are the only ones to be studied? We do not say that this is an invalid thing to do—far from it, but it is necessary to be clear a priori why this is the only kind of evolutionary event that is to be studied when a more pluralistic account of evolutionary processes is possible. Gene evolutionary analyses and phylogenetic analyses are not the same thing (Bapteste et al. 2009). Complete reliance upon only full-length homologs in phylogenetic analysis has the potential to censor our understanding of nature (see Dagan [2011] for instance). The pervasive contribution of introgression is a strong incentive to develop tools to handle data created by such events.

It would be absurd to suggest that all the genes in figure 5 are homologs of one another (in the traditional sense); however, it is clear that there are relationships that can be explored that are outside what is conventionally expected of homologs. Going back to our earlier thought experiments with four genes and four domains, with each gene having two domains (see table 1), these genes will form a ring structure in a network analysis (a situation we see repeatedly in the empirical data used to construct fig. 5). We can clearly see that Gene1 has partially homologous relationships with Gene2 and Gene4. Likewise, Gene2 has partially homologous relationships with Gene3 and Gene1. Gene3 has partially homologous relationships with Gene4 and Gene2. Gene4 has partially homologous relationships with Gene1 and Gene3. We can also say that Gene1 and Gene3 have a family resemblance relationship that is only evident because of the presence of intermediates. Gene2 and Gene4 also have a family resemblance relationship. This is to say that they are not related through common ancestry but through intermediate gene sequences that show a line of common ancestry. In the vernacular form, it might be said that they are related through marriage (a union of their relatives). In terms of network analyses, two nodes that are directly connected to one another on a network are homologs (shortest path length of 1), while two nodes that are connected with a shortest path length that is greater than 1 can be considered to have a family resemblance (whose origin can be explored: do they display STT/PNT homology that is no longer detectable by Blast? Are they made of components that are shared within an open tribe?, etc.). Thus, molecular data are complex, with pairs of genes that have only one last common ancestor and other pairs that have more than one last common ancestor.

By stating a pluralistic concept of homology, emphasizing the possibility of both partial homology and linkages that lead to family resemblances between pairs of sequences in the absence of any direct homology (partial or complete), we wish to offer some ways to deal more inclusively with a greater range of homologies and similarities in sequences. For the most part, such pluralistic homology relationships have been depicted using connected components in sequence similarity networks (Dagan et al. 2008; Dagan and Martin 2009; Dagan 2011; Kloesges et al. 2011; Bapteste et al. 2012; Jachiet et al. 2013; also, see figs. 2–5 in this manuscript). However, in this article, we have also introduced the idea of using *N*-rooted fusion networks as an additional means of analyzing such data. Thus, a combination of gene similarity networks and *N*-rooted fusion networks could provide a more inclusive analysis and visualization approach with the ability to deal with multiple (>1) multiple sequence alignments, generating multiple phylogenetic trees or networks

that can be fused together to reflect evolutionary histories more realistically.

The timing of fusions could be estimated using, for instance, maximum likelihood or Bayesian approaches, by reference to a fossil record or some such external timing. Relative or absolute timescales for fusion events can place them in the context of environmental change, for instance. Currently, estimating historical dates is restricted to ramifications on bi- or multifurcating phylogenetic trees (e.g., Tamura et al. 2012). However, the amount of introgression we see in figure 5 suggests the presence of large-scale introgressive events whose timing and context are poorly understood.

Enzymatic properties and how they change can be mapped onto these new structures, and the frequency of "emergent" properties (Fani et al. 2007) or shifts in selective pressures on individual amino acids can be estimated with respect to *N*-rooted fusion networks. Currently, tracing functional evolution is most often carried out by mapping traits onto phylogenetic trees of full-length homologs (e.g., see Feuda et al. 2012 and also Adai et al. 2004). The hotly debated "ortholog conjecture" states that orthologs are more similar in function despite being in different species, compared with paralogs that are to be found in the same species (Nehrt et al. 2011; Altenhoff et al. 2012; Chen and Zhang 2012). Sequence similarity network and *N*-rooted fusion networks offer the possibility of tracing functional evolution in a much more inclusive manner. We can ask whether functional variation and family resemblance are strongly or weakly linked and whether there are patterns that can emerge from such an analysis. Because there are many constraints on the kinds of genetic goods that can be joined together (see e.g., the content of the fusionDB database that clearly shows patterns of fusions are not random), a "family resemblance conjecture," for instance, would suggest that nonhomologous sequences that have a closer family resemblance relationship are more similar in function than sequences that lack or have a more distant family resemblance relationship.

Adjusting our models to the data may well demonstrate whether there are as-yet unknown barriers to introgression, whether gene fusion occurs at different rates at different times and in different contexts and whether there are preferred routes for introgression and preferred partners. Although it is well known that homology relationships strongly suggest functional similarities, analysis of networks could reveal additional functional connections through the analysis of extended family resemblances (Bapteste et al. 2012). It has already been shown that additional evolutionary information can be obtained by the analysis of extended gene similarity networks (Alvarez-Ponce et al. 2013; Jachiet et al. 2013); however, there are further analyses that can be carried out.

In figures 2–4, we show that a rush to "atomize" evolutionary relationships and to only use a conservative perspective when analyzing homologies can completely blind us to interesting evolutionary events. Similarity network analyses can be used not only to understand recombination and fusion but also to find if there are transitive homology statements that can be made (Alvarez-Ponce et al. 2013). Distant homologies may be recognized through intermediate sequences, so if GeneX and GeneY manifest homology along a particular region and GeneY and GeneZ manifest homology along the same region, then even if a tool such as Blast cannot directly detect the homology between GeneX and GeneZ, we can use the network information to assign homology, even though our standard software tools might not see this homology.

Concluding Remarks

At this stage, we know much more about evolutionary relationships than we did 26 years ago when Reeck et al. (1987) felt the need to clarify the terminology. It is now much clearer that fusion and fission (Snel et al. 2000; Kummerfeld and Teichmann 2005; Pasek et al. 2006; Durrens et al. 2008; Jachiet et al. 2013) of (parts of) molecules is a frequent process and a significant source of genetic and genomic novelty. The consequent muddying of gene-level relationships affect sequence relationships to a point that justifies proposition of an extended notion of evolutionary relationships and of what constitutes a gene family. Overlooking of introgressive processes is causing considerably fewer evolutionary events to be appraised than would be the case if family relationships were defined more broadly. For this reason, future notions about homology should be explicit about the kind of homologous relationship that is observed—the model must be informed by the data and not just assumed at the cost of excluding massive amounts of data. Recognizing different homology and family resemblance concepts (STT, PNT, and GT) is useful and important. In other words, any operational definition of homology must be pragmatically oriented. Under that condition, reconsidering how we define relationships between genes may open the door to a new biology.

To conclude, adopting a more pluralistic view of homology entails that a number of methodological issues need to be resolved. Proteins or genes must be allowed to be a member of more than one family. Sequence similarity that is due to extensive remodeling (e.g., Epaktology [Nagy, Bányai, et al. 2011; Nagy, Szláma, et al. 2011]) must be distinguished from similarity that is not due to remodeling. Methods for assessing the importance of family resemblance relationships need to be developed—whether such family resemblances are relevant for function, for instance, or whether they are not. Statistically robust approaches for constructing *N*-rooted networks need to be developed in addition to methods for timing introgressive events on these structures. The analysis of connected component topological features must be developed so that we can understand the relationship between topology, protein function, and evolutionary history. Embedding phylogenetic trees or networks into networks of gene sharing can allow a far greater level of detail in assessing evolutionary histories.

Supplementary Material

Supplementary information S1 and figure S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors acknowledge the generous funding of Science Foundation Ireland to L.H., J.McI. (RFP EOB2510), D.P. (RFP EOB3106), and M.J.O.C. (RFP EOB2673). M.W. was funded by the Biotechnology and Biological Sciences Research Council grant BB/K007440/1. The authors also thank Slim Karkar and two anonymous reviewers for helpful discussions.

References

Abel O. 1910. Kritische Untersuchungen über die palaogenen Rhinocerotiden Europas. *Abhandlungen Kaiserlich-Koenigliche Geologische Reichsanstal.* 20:1–22.

Adai AT, Date SV, Wieland S, Marcotte EM. 2004. LGL: creating a map of protein function with an algorithm for visualizing very large biological networks. *J Mol Biol.* 340:179–190.

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. 2012. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput Biol.* 8:e1002514.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.

Alvarez-Ponce D, Bapteste E, Lopez P, McInerney JO. 2013. Gene similarity networks provide new tools for understanding eukaryote origins and evolution. *Proc Natl Acad Sci.* 110(17):E1594–1603.

Apic G, Gough J, Teichmann SA. 2001a. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol.* 310: 311–325.

Apic G, Gough J, Teichmann SA. 2001b. An insight into domain combinations. *Bioinformatics* 17(Suppl 1), S83–89.

Apic G, Russell RB. 2010. Domain recombination: a workhorse for evolutionary innovation. *Sci Signal.* 3:pe30.

Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. 2009. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PloS One* 4:e4345.

Bapteste E, Lopez P, Bouchard F, Baquero F, McInerney JO, Burian RM. 2012. Evolutionary analyses of non-genealogical bonds produced by introgressive descent. *Proc Natl Acad Sci U S A.* 109:18266–18272.

Bapteste E, O'Malley MA, Beiko RG, Ereshefsky M, Gogarten JP, Franklin-Hall L, Lapointe FJ, Dupré J, Dagan T, Boucher Y, et al. 2009. Prokaryotic evolution and the tree of life are two different things. *Biol Direct.* 4:34.

Bastian M, Heymann S, Jacomy M. 2009. Gephi: an open source software for exploring and manipulating networks. In International AAAI Conference on Weblogs and Social Media.

Boucher Y, Bapteste E. 2009. Revisiting the concept of lineage in prokaryotes: a phylogenetic perspective. *Bioessays* 31:526–536.

Brigandt I. 2003. Homology in comparative, molecular, and evolutionary developmental biology: the radiation of a concept. *J Exp Zool B Mol Dev Evol.* 299:9–17.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973.

Chen X, Zhang J. 2012. The ortholog conjecture is untestable by the current gene ontology but is supported by RNA sequencing data. *PLoS Comput Biol.* 8:e1002784.

Corpet F, Servant F, Gouzy J, Kahn D. 2000. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.* 28:267–269.

Dagan T. 2011. Phylogenomic networks. *Trends Microbiol.* 19(10):483–491.

Dagan T, Artzy-Randrup Y, Martin W. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci U S A.* 105:10039–10044.

Dagan T, Martin W. 2009. Getting a better picture of microbial evolution en route to a network of genomes. *Philos Trans R Soc Lond B Biol Sci.* 364:2187–2196.

Dessimoz C, Gabaldon T, Roos DS, Sonnhammer EL, Herrero J. 2012. Toward community standards in the quest for orthologs. *Bioinformatics* 28:900–904.

Dobzhansky T. 1955. A review of some fundamental concepts and problems of population genetics. *Cold Spring Harb Symp Quant Biol.* 20:1–15.

Doherty A, Alvarez-Ponce D, McInerney JO. 2012. Increased genome sampling reveals a dynamic relationship between gene duplicability and the structure of the primate protein-protein interaction network. *Mol Biol Evol.* 29:3563–3573.

Duret L, Mouchiroud D, Gouy M. 1994. HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res.* 22:2360–2365.

Durrens P, Nikolski M, Sherman D. 2008. Fusion and fission of genes define a metric between fungal genomes. *PLoS Comput Biol.* 4: e1000200.

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. 1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402:86–90.

Enright AJ, Kunin V, Ouzounis CA. 2003. Protein families and TRIBES in genome sequence space. *Nucleic Acids Res.* 31:4632–4638.

Enright AJ, Ouzounis CA. 2000. GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics* 16: 451–457.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30: 1575–1584.

Epp CD. 1997. Definition of a gene. *Nature* 389:537.

Ereshefsky M. 2007. Psychological categories as homologies: lessons from ethology. *Biol Philos.* 22:659–674.

Fani R, Brilli M, Fondi M, Lió P. 2007. The role of gene fusions in the evolution of metabolic pathways: the histidine biosynthesis case. *BMC Evol Biol.* 7:S4.

Felsenstein J. 2004. Inferring phylogenies. Sunderland (MA): Sinauer Associates.

Feuda R, Hamilton SC, McInerney JO, Pisani D. 2012. Metazoan opsin evolution reveals a simple route to animal vision. *Proc Natl Acad Sci U S A.* 109:18868–18872.

Fitch WM. 2000. Homology a personal view on some of the problems. *Trends Genet.* 16:227–231.

Fitzpatrick DA, O'Gaora P, Byrne KP, Butler G. 2010. Analysis of gene evolution and metabolic pathways using the *Candida* Gene Order Browser. *BMC Genomics* 11:290.

Greider CW, Blackburn EH. 1985. Identification of a specific telomere terminal transferase activity in *Tetrahymena* extracts. *Cell* 43: 405–413.

Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E. 2010. Network analyses structure genetic diversity in independent genetic worlds. *Proc Natl Acad Sci U S A.* 107:127–132.

Heger A, Holm L. 2003. Exhaustive enumeration of protein domain families. *J Mol Biol.* 328:749–767.

Hillis DM. 1994. Homology in molecular biology. In: Hall B, editor. Homology, the hierarchical basis of comparative biology. San Diego (CA): Academic Press. p. 483.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol.* 23:254–267.

Huson DH, Scornavacca C. 2011. A survey of combinatorial methods for phylogenetic networks. *Genome Biol Evol.* 3:23.

Ingolfsson H, Yona G. 2008. Protein domain prediction. *Methods Mol Biol.* 426:117–143.

Jachiet PA, Pogorelcnik R, Berry A, Lopez P, Bapteste E. 2013. MosaicFinder: identification of fused gene families in sequence similarity networks. *Bioinformatics* 29(7):837–844.

Kloesges T, Popa O, Martin W, Dagan T. 2011. Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. *Mol Biol Evol.* 28:1057–1074.

Koonin EV, Wolf YI, Karev GP. 2002. The structure of the protein universe and genome evolution. *Nature* 420:218–223.
Kummerfeld SK, Teichmann SA. 2005. Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet.* 21:25–30.
Levitt M. 2009. Nature of the protein universe. *Proc Natl Acad Sci U S A.* 106:11079–11084.
Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13:2178–2189.
Majumdar I, Kinch LN, Grishin NV. 2009. A database of domain definitions for proteins with complex interdomain geometry. *PloS One* 4: e5084.
Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. 1999. A combined algorithm for genome-wide prediction of protein function. *Nature* 402:83–86.
McInerney JO, Pisani D, Bapteste E, O'Connell MJ. 2011. The public goods hypothesis for the evolution of life on Earth. *Biol Direct.* 6:41.
Miele V, Penel S, Daubin V, Picard F, Kahn D, Duret L. 2012. High-quality sequence clustering guided by network topology and multiple alignment likelihood. *Bioinformatics* 28:1078–1085.
Mindell DP, Meyer A. 2001. Homology evolving. *Trends Ecol Evol.* 16: 434–440.
Nagy A, Bányai L, Patthy L. 2011. Reassessing domain architecture evolution of metazoan proteins: major impact of errors caused by confusing paralogs and epaktologs. *Genes* 2:516–561.
Nagy A, Szláma G, Szarka E, Trexler M, Bányai L, Patthy L. 2011. Reassessing domain architecture evolution of metazoan proteins: major impact of gene prediction errors. *Genes* 2:449–501.
Natale DA, Galperin MY, Tatusov RL, Koonin EV. 2000. Using the COG database to improve gene recognition in complete genomes. *Genetica* 108:9–17.
Nehrt NL, Clark WT, Radivojac P, Hahn MW. 2011. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput Biol.* 7:e1002073.
Noble WS, Kuang R, Leslie C, Weston J. 2005. Identifying remote protein homologs by network propagation. *FEBS J.* 272:5119–5128.
O'Hara RJ. 1997. Population thinking and tree thinking in systematics. *Zoologica Scripta* 26:323–329.
Owen R. 1868. On the archetype and homologies of the vertebrate skeleton. London: Richard and John E. Taylor.
Park J, Teichmann SA. 1998. DIVCLUS: an automatic method in the GEANFAMMER package that finds homologous domains in single- and multi-domain proteins. *Bioinformatics* 14:144–150.
Park J, Teichmann SA, Hubbard T, Chothia C. 1997. Intermediate sequences increase the detection of homology between sequences. *J Mol Biol.* 273:349–354.
Pasek S, Risler JL, Brozellec P. 2006. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics* 22:1418–1423.
Perriere G, Duret L, Gouy M. 2000. HOBACGEN: database system for comparative genomics in bacteria. *Genome Res.* 10:379–385.
Portugaly E, Harel A, Linial N, Linial M. 2006. EVEREST: automatic identification and classification of protein domains in all protein sequences. *BMC Bioinformatics* 7:277.
Price MN, Dehal PS, Arkin AP. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS One* 5: e9490.
Reeck GR, de Haën C, Teller DC, Doolittle RF, Fitch WM, Dickerson RE, Chambon P, McLachlan AD, Margoliash E, Jukes TH, et al. 1987. "homology" in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell* 50:667.
Roth VL. 1988. The biological basis of homology. In: Humphries CJ, editor. Ontogeny and systematics. New York: Columbia University Press. p. 236.
Sapp J. 2009. The new foundations of evolution. On the tree of life. New York: Oxford University Press. p. 425.
Sasson O, Vaaknin A, Fleischer H, Portugaly E, Bilu Y, Linial N, Linial M. 2003. ProtoNet: hierarchical classification of the protein space. *Nucleic Acids Res.* 31:348–352.
Sattler R. 1984. Homology-a continuing challenge. *Syst Bot.* 9: 382–394.
Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13:2498–2504.
Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol.* 7:539.
Snel B, Bork P, Huynen M. 2000. Genome evolution-gene fusion versus gene fission. *Trends Genet.* 16:9–11.
Song N, Joseph JM, Davis GB, Durand D. 2008. Sequence similarity network reveals common ancestry of multidomain proteins. *PLoS Comput Biol.* 4:e1000063.
Sonnhammer EL, Kahn D. 1994. Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.* 3:482–492.
Tamura K, Battistuzzi FU, Billing-Ross P, Murillo O, Filipski A, Kumar S. 2012. Estimating divergence times in large molecular phylogenies. *Proc Natl Acad Sci U S A.* 109:19333–19338.
Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28:33–36.
Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. *Science* 278:631–637.
Tillyard RJ. 1919. The panorpoid complex. Part 3: the wing venation. *Proc Linn Soc N S W.* 44:533–717.
Weston J, Elisseeff A, Zhou D, Leslie CS, Noble WS. 2004. Protein ranking: from local to global structure in the protein similarity network. *Proc Natl Acad Sci U S A.* 101:6559–6563.
Wittgenstein L. 2009. Philosophical investigations. Wiley-Blackwell.
Wong S, Ragan MA. 2008. MACHOS: Markov clusters of homologous subsequences. *Bioinformatics* 24:i77–i85.
Wuchty S. 2001. Scale-free behavior in protein domain networks. *Mol Biol Evol.* 18:1694–1702.
Yona G, Linial N, Linial M. 2000. ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res.* 28:49–55.

Chapitre 4 - Application de la notion d'air de famille à des jeux de données de virus

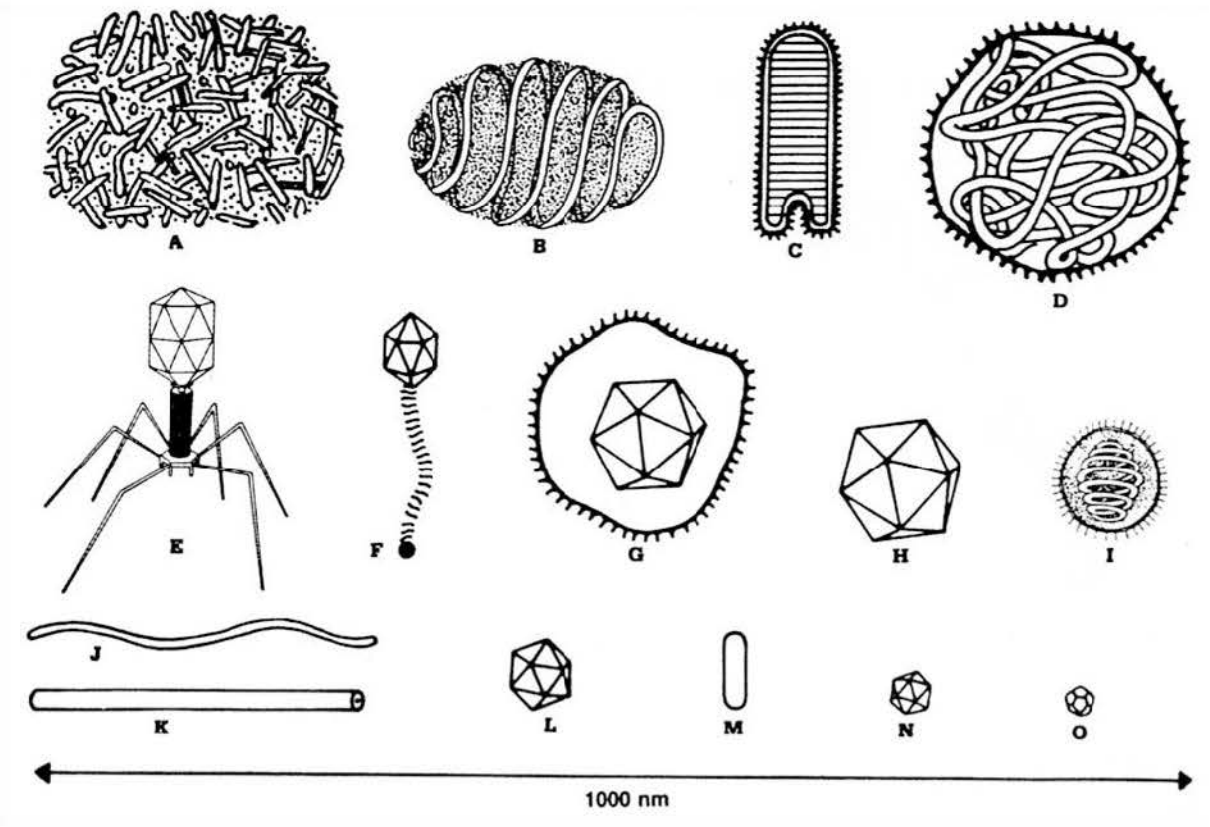


Figure 4-1 : Diversité morphologique des virus

4.1. La diversité des virus est immense et très peu connue 80
4.2. Différentes classifications des virus 80
4.3. Evolution des virus 83
4.4. Etude systématique des phénomènes combinatoires chez les virus 84
4.5. Enjeux nouveaux abordés lors de cette étude 85

Ce troisième chapitre présente une application des méthodes et réflexions présentées dans les deux premiers à l'étude de la diversité génétique des virus.

4.1. La diversité des virus est immense et très peu connue

Les virus sont les entités biologiques les plus abondantes, les plus diverses et les moins bien connues. Le nombre de particules virales sur Terre est estimé à 10^{31}, soit 10 virus pour une bactérie. Les virus présentent une très grande variété de tailles, de morphologies, d'hôtes, de génomes ou de cycles de vie [Breitbart, Rohwer, 2005 ; Suttle, 2005]. Nous connaissons encore très mal cette gigantesque nano-biodiversité, au point que notre conception des virus change régulièrement. Les virus ont longtemps été considérés comme des entités simples et de petite taille, jusqu'à ce que l'on découvre récemment que certains parasites d'amibes, que l'on prenait pour des bactéries, étaient en fait des virus géants [La Scola et al., 2003]. Le pandoravirus par exemple, découvert en 2013 dans des sédiments au Chili et en Australie, contient plus de 2000 gènes, c'est-à-dire autant que certains eucaryotes parasites. Parmi ces gènes, certains sont impliqués dans des fonctions que l'on pensait réservées aux organismes cellulaires, telles que des protéines chaperonnes. Ainsi, la frontière séparant les virus des organismes cellulaires parasites s'atténue. En outre, il devient difficile de distinguer nettement les virus des autres éléments génétiques mobiles, tels que les plasmides et transposons. On découvre en effet des agents viraux très simples, qui ne codent pas pour leur propre réplication (virus satellites, viroïdes)[Francki, 1985 ; Chen et al., 1986] ou qui ne subissent pas de phase extracellulaire (rétrovirus endogènes)[Löwer et al., 1996]. Par ailleurs, la perception des rôles écologiques et évolutifs des virus s'est considérablement élargie au-delà de celle de simples parasites pathogènes, impliqués dans une course aux armements avec leur hôte. On connaît aujourd'hui de nombreux exemples de virus en relation symbiotique avec leur hôte, et d'adaptations génétiques clés apportées par les virus [Roossinck, 2011]. Enfin, la diversité génétique des virus est considérable. Lors du séquençage de génomes ou métagénomes viraux, on découvre systématiquement une grande proportion de gènes entièrement nouveaux, sans homologues dans les bases de donnés existantes (ORFans) [Yin, Fischer, 2008]. En outre, parmi les gènes viraux qui trouvent des homologues, très peu sont annotés fonctionnellement à partir d'études expérimentales [Kristensen et al., 2011], si bien que la biologie des virus nous est largement inconnue. Ces quelques remarques montrent que l'exploration et la compréhension de la diversité des virus n'en sont qu'à leurs débuts.

4.2. Différentes classifications des virus

La diversité foisonnante des virus est difficile à organiser. Il semble impossible de les classer par une approche phylogénétique. Contrairement aux organismes cellulaires, qui possèdent des gènes universels sur la base desquels les comparer (par exemple la petite sous-unité du ribosome), les virus ne partagent aucun caractère commun [Koonin et al.,

2006]. Les génomes viraux sont petits et évoluent rapidement, de sorte que peu d'information de similarité est conservée. De plus, un arbre phylogénétique n'est sans doute pas un modèle théorique adapté pour résumer l'évolution des virus, dans laquelle les processus combinatoires jouent un rôle important. Il est en effet fréquent que des virus assemblent du matériel génétique de plusieurs origines lors de la réplication. L'histoire ancienne des virus est par ailleurs très incertaine [Forterre, 2006 ; Moreira, López-García, 2009]; peut être ne sont-ils pas tous apparentés et les principales lignées de virus seraient apparues plusieurs fois indépendamment [Koonin et al., 2006].

De nombreux systèmes de classification des virus ont donc été proposés, qui diffèrent beaucoup les uns des autres en fonction des caractères étudiés. Les premières classifications, basées sur les hôtes des virus ou sur leur morphologie, ne coïncident pas du tout. Aujourd'hui, trois principales classifications font référence.

Premièrement, David Baltimore a proposé en 1971 [Baltimore, 1971] d'organiser les virus en fonction de la nature de leur acide nucléique (ADN ou ARN, simple ou double brin), des étapes de synthèse des ARN messagers viraux et des étapes de réplication de leur génome (Figure 4-2). Cette classification repose donc, non pas sur l'information génétique elle-même, mais sur son support physique. Elle comporte sept catégories de virus : à ADN double-brin (I), à ADN simple-brin (II), à ARN double-brin (III), à ARN simple-brin à polarité positive (IV), à ARN simple-brin à polarité négative (V), à ARN simple-brin à polarité positive intégré au génome ADN de l'hôte après rétrotranscription (VI), à ADN double-brin répliqué par un intermédiaire ARN et une rétrotranscription (VII).

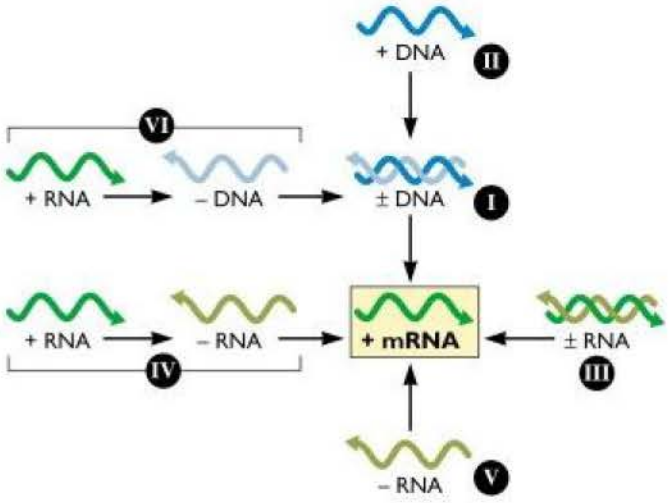


Figure 4-2 : Classification des virus proposée par David Baltimore
Cette classification repose sur la nature de l'acide nucléique et sur les étapes de synthèse des ARN messagers viraux (source). La classe VII n'est pas représentée dans ce schéma.

Le Comité International sur la Taxonomie des Virus (ICTV) propose et tient à jour une seconde classification hiérarchique des virus, inspirée du système de nomenclature proposé par Linné, avec les niveaux suivants : Ordre, Famille, Sous-famille, Genre, Espèce. Cette classification n'est pas phylogénétique : les taxons ne sont pas définis par descendance depuis un ancêtre commun. Elle est « logique » pour les taxons supérieurs au genre, qui sont

définis par une liste de propriétés nécessaires et suffisantes. Les espèces virales sont en revanche polythétiques, c'est-à-dire que les virus d'une espèce partagent un certain nombre de propriétés, sans qu'aucune prise isolément ne soit discriminante[1]. L'ICTV procède en fait à une classification experte ; il organise la diversité des virus pour mieux la résumer et l'appréhender[2]. Cette classification est d'ailleurs incomplète puisque de nombreuses espèces virales sont les seules de leur genre, voire de leur famille, et que de nombreuses familles ne sont associées à aucun ordre. Plutôt que d'un arbre, il s'agit plutôt d'une forêt, composé de 7 grands arbres au niveau de l'ordre et de nombreux petits sous-arbres et branches isolées. Ce système de taxonomie est utile, et sa méthode de constitution éclaire la nature de notre compréhension de la diversité virale. Nous n'avons cependant pas employé la classification de l'ICTV dans notre étude, car les taxons supérieurs sont logiques et n'ont pas de sens biologique directement interprétables.

Enfin, une troisième classification de référence a été proposée en 2005 par Eugene Koonin et ses collaborateurs. Cette classification divise les virus en cinq groupes monophylétiques, c'est-à-dire en cinq lignées évolutives qui descendent d'origine distinctes. Cette classification repose sur un scénario d'émergence des virus qui remonte à l'apparition des cellules (Figure 4-3). Les lignées proposées sont très diverses, elles regroupent des virus qui ont des stratégies de réplications différentes. Elle est cependant soutenue par des gènes conservés entre virus du même groupe. Nous emploierons cette classification, non pas parce qu'elle regroupe des virus présumés apparentés, mais parce qu'elle délimite des frontières entre lesquelles aucune trace de partage d'histoire évolutive n'est attendue. De nombreux virus ne rentrent pas dans cette classification.

---

[1] Il est intéressant de faire un parallèle entre cette procédure et la notion d'air de famille, proposée au chapitre 3 pour étendre la notion de famille homologue. L'ICTV défini les espèces virales comme des « classes polythétiques qui constituent une lignée de réplication et occupe une niche écologique particulière ». Les classes polythétiques correspondent donc à des clusters de virus dans le réseau de partage de caractères, définis de façon empirique par expertise humaine plutôt que par un algorithme. L'unité de ces groupes est justifiée par une relation évolutive (lignée de réplication) et fonctionnelle (niche écologique), sans que ces relations soient définies précisément. Il s'agit en fait plutôt d'une hypothèse explicative de l'air de famille partagé par les virus d'une même espèce que d'un critère de délimitation. Cette explication est d'ailleurs relativisée par l'affirmation que les taxons sont des « abstractions conceptuelles », de nature différente des virus qui ont une existence physique.

[2] Il est également très intéressant de faire un parallèle entre cette procédure, le groupement de gènes par leur *air de famille*, et les considérations de M. Mirbel en 1810 sur la classification naturelle des plantes [Mirbel, 1810]. M. Mirbel refuse une justification logique, aristotélicienne, de la classification. Il explique que cette justification ne fait pas de sens car elle nécessite de subordonner certains caractères à d'autres (par exemples le système végétatif par rapport au système reproductif), et qu'elle ne reflète souvent pas la réalité des regroupements effectuées par les experts *[« Mais si l'on s'est trompé dans la théorie, il faut convenir que l'on ne s'est pas trompé dans la pratique »]*. Il explique que les familles naturelles sont délimités par l'*air de famille* qu'elles entretiennent ; soit que tous les membres de cette famille se ressemblent deux à deux pour tous les caractères (clique), soit que les membres d'une famille sont reliés par des chaînes de similarité pour certains caractères (communauté moins dense) [*« On peut passer par des nuances insensibles de la première espèce jusqu'à la dernière ; mais on ne peut rapprocher subitement les extrêmes en écartant les espèces intermédiaires, attendu que les différences qui les séparent sont trop grandes et trop multipliées »*].]

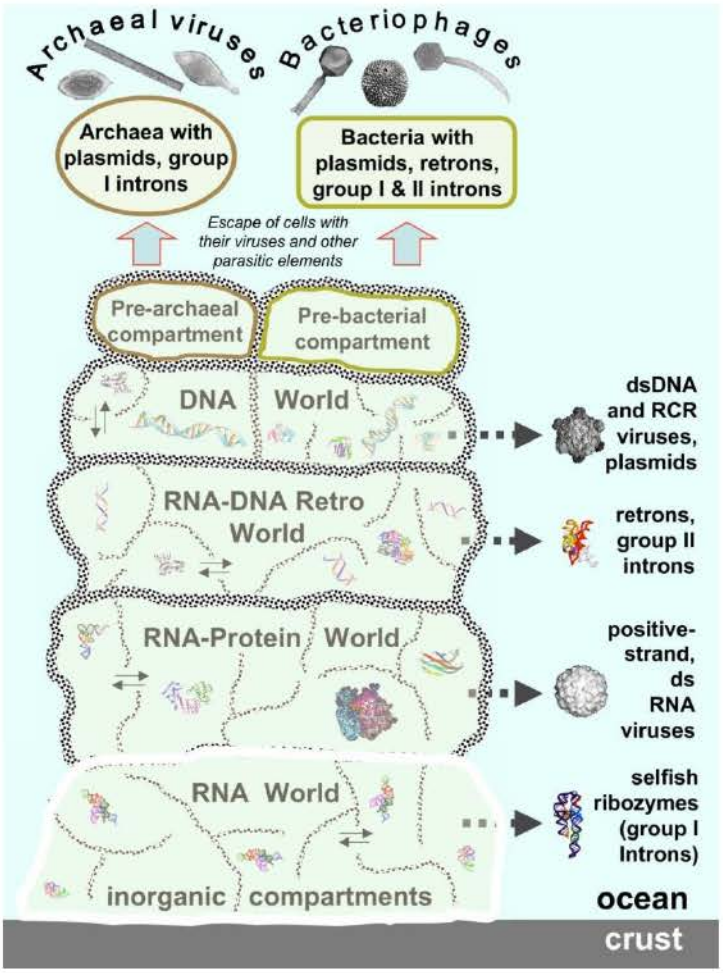


Figure 4-3 : Scénario d'apparition des lignées virales
Scénario d'apparition des lignées virales à partir du pool de gène primordial, proposé par Koonin, Eugene V., Tatiana G. Senkevich, and Valerian V. Dolja. "The Ancient Virus World and Evolution of Cells." Biology Direct, 2006

4.3. Evolution des virus

Certains virus constituent un excellent modèle d'évolution expérimentale, parce qu'ils sont faciles à cultiver, ont des temps de génération très courts, et leurs petits génomes sont faciles à séquencer. Certains aspects de la microévolution des virus sont donc bien connus.

Une première caractéristique des génomes viraux est leur très grande compacité. Les gènes sont très peu espacés, voire peuvent se chevaucher dans des cadres de lectures différents [Krakauer, 2000]. Cette tendance à la compacité est le fruit de deux forces évolutives. D'une part, la petite dimension des capsides virales limite physiquement la taille des génomes. D'autre part, un génome plus petit se réplique plus rapidement. Puisque les virus ont de très grandes tailles de populations, des mutations très légèrement avantageuses peuvent être fixées, comme la perte de petites régions inter-géniques inutiles.

Une autre caractéristique bien étudiée des virus est leur taux de mutation ponctuelle, qui est l'un des facteurs déterminant leur vitesse d'évolution. Ce taux de mutation varie beaucoup entre les virus, selon le support de l'information génétique. Les virus à ARN ont

ainsi un taux de mutation plus élevé que les virus à ADN, parce que leur polymérase n'a pas de fonction correctrice d'erreur. Les virus à ARN évoluent donc plus vite que les virus à ADN, mais ne peuvent pas avoir de grands génomes qui seraient trop instables [Barr, Fearns, 2010].

Les virus évoluent aussi beaucoup par combinaison en raison de la façon dont ils se répliquent. Contrairement aux cellules, il ne s'agit pas d'une réplication isolée dans des compartiments suivie par une division binaire. Les virus s'assemblent dans un milieu cellulaire (noyau ou cytoplasme) où il y a beaucoup de matériel génétique, viral ou cellulaire. Plus particulièrement, certains virus tels que celui de la grippe ont une évolution fortement combinatoire. Cela est du à leur structure de génome en segments. Lorsque plusieurs virus de ce type infectent une même cellule, ils peuvent donc échanger facilement des segments de matériel génétique pour former un nouveau génome chimérique [Lei, Shi, 2011].

Mais d'une manière générale, les connaissances sur l'évolution des virus concernent principalement les génomes. L'évolution des familles de gènes viraux est moins étudiée et encore peu connue. Elle est habituellement étudiée à partir des modèles développés sur les organismes cellulaires (évolution par mutation ponctuelle suivant les mêmes matrices de transition et les mêmes coûts d'ouverture des gaps). Si les familles de gènes évoluent selon d'autres processus, plus réticulé, une étude par des réseaux de similarité pourrait avantageusement compléter ces approches.

4.4. Etude systématique des phénomènes combinatoires chez les virus

Dans l'article présenté ici, nous nous sommes intéressés à l'existence de phénomènes combinatoires chez les virus au niveau des gènes, plus précisément à la présence et aux fonctions des gènes composites dans les différentes classes de virus. Cette question n'avait pas encore été traitée de façon systématique. Puisque les phénomènes combinatoires sont importants au niveau des génomes viraux, on pourrait penser qu'ils le sont également pour les gènes. Mais les gènes viraux n'ont pas l'organisation en intron-exon des gènes eucaryotes, qui leur permet de recombiner facilement des modules. On pourrait donc aussi penser que les phénomènes combinatoires sont rares dans l'évolution des gènes viraux. Répondre à cette question n'est pas anecdotique : la majorité des gènes sur terre est portée par des virus. La compréhension de leur évolution est donc importante pour comprendre la diversité génétique en général ; d'autant plus qu'elle n'est pas isolée mais sert de réservoir pour les génomes cellulaires, qui acquièrent régulièrement des gènes viraux.

L'article « Extensive gene remodeling on Earth: functionally biased gene composition in all viral classes contributes to saltatory evolution in the mobilome network » est le fruit d'une collaboration avec le virologue Philippe Colson. Nous avons constitué un jeu de données comprenant tous les génomes viraux complètement séquencés disponibles sur GenBank en novembre 2012, soit un total de 3008 virus. J'ai construit un grand réseau de similarité de séquences avec les 122 392 gènes de ces virus.

J'ai étudié le partage de gènes entre virus appartenant à des classes virales différentes, au sens de Baltimore ou de Koonin. Nous avons vérifié qu'en général les classes virales constituent des « mondes génétiques » isolés. En effet l'indice d'assortativité des classes virales, mesure de leur isolement dans le réseau, est très élevé. Nous nous sommes alors concentrés sur la recherche des gènes voisins (et donc semblables) appartenant à des virus de différentes classes (et donc normalement très différents). Ces gènes sont intéressants car ils montrent le partage d'information génétique entre des classes virales fondamentalement distinctes. Nous avons trouvé une quarantaine de ces cas où la distribution des gènes transgresse ainsi les frontières des familles virales. Cette analyse du réseau nous a d'ailleurs permis de facilement corriger l'annotation semi-automatique de certains virus, qui étaient mal classés puisque leurs gènes étaient tous voisins de gènes d'une autre classe.

Nous avons recherché de manière systématique les gènes composites dans ce réseau. Nous avons détecté environ 10 % de gènes viraux composites. Ils sont présents dans toutes les classes virales et pour toutes les fonctions biologiques, mais de façon non-uniforme. De façon remarquable, les fonctions enrichies en gènes composites chez les virus sont des fonctions importantes pour la perpétuation du cycle de vie des virus.

4.5. Enjeux nouveaux abordés lors de cette étude

Cette étude sur le réseau de gène de virus a permis d'aborder trois enjeux. Le premier est qu'en raison du grand nombre d'événements de composition, les relations entre gènes viraux sont complexes. Un gène peut être le produit de plusieurs événements de composition successifs. Cela n'a alors pas de sens de le classer dans une famille composite. Le principe de MosaicFinder repose sur la notion de famille composite. Il n'est donc pas adapté pour étudier des gènes qui ont connu des événements composition successifs. Pour mesurer le fait qu'un gène a été le produit de plusieurs événements de composition, nous avons introduit la notion de gène multi-composite, défini comme étant un gène composite, dont les composants sont eux même des éléments de gènes composites. Ce patron du réseau de gène ne correspond pas à une histoire évolutive unique. Lorsque l'on détecte de tels gènes, il n'est pas possible de déterminer leur histoire précise à partir de la topologie du réseau, mais l'on sait en revanche qu'elle a été complexe. La présence de nombreux multi-composites explique la difficulté de classer en familles nettement séparées un large ensemble de gène viraux.

Un deuxième enjeu a été de quantifier le partage de gènes entre les différentes classes virales. Nous avons pour cela employé la mesure de mélange assortatif proposée par Newman en 2003. Cette mesure vaut 1 lorsque les classes n'ont aucun mélange (c-à-d lorsque les séquences d'une classe virale ne sont connectées que entre elles), 0 lorsque le mélange est aléatoire. Les séquences des classes virales ont toutes eu des assortativités supérieures à 0.9, sauf les classes VI et VII qui partagent des gènes entre elles qui ont une assortativité de 0.7. L'assortativité de ces classes augmente à 0.9 lorsqu'elles sont regroupées en une, ce qui montre qu'elles sont mélangées entre elles mais pas avec les

autres. Ces classes sont composées de virus qui ont des acides nucléiques différents, mais qui emploient toutes deux la *reverse transcriptase.*

Un troisième enjeu a été de mesurer la présence de grands cycles dans les réseaux de similarité de séquences. Nous nous étions intéressés à ces grands cycles parce qu'ils posent des problèmes pour utiliser MosaicFinder. Nous nous sommes interrogés sur leur signification biologique. Contrairement aux triplets intransitifs qui peuvent être dus à de l'homologie distante, ces patrons ne peuvent pas apparaître suite à une évolution purement arborescente. Nous avons mesuré la présence de ces cycles de manière approchée en recherchant et coloriant les cycles dans le réseau de communautés de Louvain des séquences de virus. Suite à un échange sur la question avec le théoricien des graphes Michel Habib, ce dernier a proposé un sujet de stage d'informatique sur ce sujet.

En résumé, notre analyse illustre la complexité et l'intrication génétique affectant l'évolution des virus et propose d'étudier les adaptations moléculaires au moyen de réseaux de similarité de séquences.

Article, Discoveries

Extensive gene remodeling in the viral world: new evidence for non-gradual evolution in the mobilome network

PA Jachiet[1], P Colson[2, 3], P Lopez[1], E Bapteste[1]*

[1]UMR CNRS 7138 Evolution Paris Seine, Université Pierre et Marie Curie. 7 quai Saint-Bernard, 75005 Paris, France.

[2]URMITE UMR CNRS 6236 IRD 198, Facultés de Médecine et de Pharmacie, Université de la Méditerranée. 27 boulevard Jean Moulin, 13385 Marseille cedex 5, France.

[3]Pôle des Maladies Infectieuses et Tropicales Clinique et Biologique, Fédération de Bactériologie-Hygiène-Virologie, Centre Hospitalo-Universitaire Timone. 264 rue Saint-Pierre, 13385 Marseille cedex 5, France.

*To whom correspondence should be addressed

Eric Bapteste

Email: eric.bapteste@snv.jussieu.fr

Phone: +33 1 44 27 34 70

Abstract

Complex non-gradual evolutionary processes such as gene remodeling are difficult to model, to visualize and to investigate systematically. Despite these challenges, the creation of composite (or mosaic) genes by combination of genetic segments from unrelated gene families was established as an important adaptive phenomena in eukaryotic genomes. By contrast, almost no general studies have been conducted to quantify composite genes in viruses. While viral genome mosaicism has been well-described, the extent of gene mosaicism and its rules of emergence remain largely unexplored. Applying methods from graph theory to inclusive similarity networks, and using data from more than 3000 complete viral genomes, we provide the first demonstration that composite genes in viruses are (i) functionally biased, (ii) involved in key aspects of the arm race between cells and viruses, and (iii) can be classified into two distinct types of composite genes in all viral classes. Beyond the quantification of the widespread recombination of genes among different viruses of the same class, we also report a striking sharing of genetic information between viruses of different classes and with different nucleic acid types. This latter discovery provides novel evidence for the existence of a large and complex mobilome network, which appears partly binded by the sharing of genetic information and by the formation of composite genes between mobile entities with different genetic material. Considering that there are around 10E31 viruses on the planet, gene remodeling appears as a hugely significant way of generating and moving novel sequences between different kinds of organisms on Earth.

Introduction

The assembly of genetic material from different gene families, producing composite genes (Enright et al. 1999; Jachiet et al. 2013), has been mostly described in eukaryotic genomes. Individual studies have shown that the combination of domains (Wang and Caetano-Anollés 2009) and the fusion of genes account for important aspects of biological complexity, from the evolution of distinct signaling systems to possible key evolutionary transitions such as animal multicellularity (Koonin et al. 2002). Genetic fragments common to all cellular beings are combined in specific ways in each domain of life, affecting as many as two thirds of the proteins in unicellular organisms to over 80% in metazoa (Apic et al. 2001). However, the extent to which composite gene genesis is observed across th viral world is unquantified.

If one considers the mechanisms by which genomes of these major numerous evolutionary players evolve, it can immediately be noted that viruses exploit a vast pool of genes and that viral genomes are structurally and evolutionary highly constrained. Most viral genes are under purifying selection (Holmes 2003; Koonin and Wolf 2010) and intragenomic gene duplication is rare (Liu et al. 2006; Simon-Loriere and Holmes 2013) (with the exception of large and giant DNA viruses (Shackelton and Holmes 2004; Filée 2009)). Frequent mutations, insertion/deletions and hyperplastic regions allow viruses to go through their life cycle by

escaping their hosts immune systems and defense mechanisms (Arias et al. 2009; Sanjuán et al. 2010). Moreover, many mechanisms could be, in principle at least, involved in the making of composite genes. More precisely, many viral genomes, such as double-stranded (ds) DNA bacteriophages (Casjens 2008; Hatfull 2008) and RNA viruses (Lai 1992; Barr and Fearns 2010; Jackwood et al. 2012), are highly recombinogenic (Lima-Mendez et al. 2008). Viral gene repertoire is thus commonly expanded by strand-switching, the use of incompletely replicated genetic material as a primer for another strain, by crossing-over of non-homologous segments (Liu et al. 2006; Arias et al. 2009; Savolainen-Kopra and Blomqvist 2010), by genetic reassortment of fragments of genomes (Lei and Shi 2011), by the use of specific proteins enhancing recombination (Martinsohn et al. 2008), by transposition and illegitimate recombination joining pieces of DNA with limited homology (Crawford-Miksza and Schnurr 1996), and by the shuffling of groups of genes (modules) between genomes (Botstein 1980).

These processes may not only generate composite genomes but also composite genes in viruses. However, processes producing composite genes have not been systematically analyzed for these taxa, although an estimated 6-8% of viral genes have been reported to be multi-domain (Hatfull 2008; Kristensen, Cai, et al. 2011; Kristensen et al. 2013), as well as few occasional cases of gene fusions between viruses of the same class (involving some tail fiber and replication genes (Highton et al. 1990), and two cyanophage photosynthetic genes (Sharon et al. 2009)). We seek to perform such a systematic analysis of composite genes in viral genomes, and in the process address three complementary questions. First, we tested whether composite genes link genetic material from different kinds of sequences in all viral classes based on 3 major classifications. Second, we tested whether these composite genes fulfilled central or less essential functions for the completion of the viral life cycle. Third, we investigated whether composite genes may be constituted from genetic segments from viruses belonging to different viral classes (e.g. DNA and RNA viruses), hence from distantly related or even unrelated viruses.

Systematic studies of composite genes are best formulated within the framework of sequence similarity networks (Adai et al. 2004). In these graphs, each node is an individual sequence, connected to others when they share some alignable regions with significant similarity (Atkinson et al. 2009). Composite genes act as detectable bridges that connect sequences harboring segments from unrelated gene families (Adai et al. 2004; Jachiet et al. 2013). Constant progress in sequencing technologies, computing power and memory capacities, network display (Shannon et al. 2003; Bastian et al. 2009) and analyses (Song et al. 2008; Berry et al. 2010; Jachiet et al. 2013) now permit the analysis of the structure of these graphs for datasets of thousands of viral genomes. Here, we mined the genes of 3,008 viral genomes and detected 8-15% composite sequences. These composite genes were found in all viral classes (according to three classifications), mostly encoding important functions for the viral life cycle. The emergence of composite genes operated beyond the frontiers of both viral classes and gene families, meaning that numerous viral adaptations are best understood from a global

perspective, because boundaries or viral classes are not forbidding sharing of gene segments from different gene families.

Results

Extensive gene remodeling in viruses

We compared 122,392 sequences from 3,008 viruses in a BLAST all versus all analysis, searching for sequences with significant similar genetic fragments, called hits. Sequences were included as nodes in sequence similarity networks. Two sequences were connected when at least one of the pairwise BLAST comparisons returned a hit with an E-value < 10E-5. At this stringency threshold, false positive hits between non-homologous sequences are not expected (Medini et al. 2006; Fokkens et al. 2010), although genuine homology between very divergent sequences can be missed. Using simple linkage, we partitioned the graph into 24,092 singletons and 12,506 clusters or connected components of 2 sequences or more. Homologous genes that have not diverged beyond recognition by BLAST typically produce such clusters. Composite sequences indirectly bridge several different homologous families in the graph, when distinct regions of composite sequences present similarity with distinct families. Thus composite sequences produce larger connected components, uniting sequences from different gene families (Enright and Ouzounis 2000; Kristensen, Wolf, et al. 2011; Jachiet et al. 2013). The largest connected component present in the network comprised 18,033 sequences (15% of the data set), demonstrating that composite genes involved genetic segments from numerous and diverse homologous families.

The topology of this network was explored to find candidate composite genes, using FusedTriplets (Jachiet et al. 2013). Composite genes fulfill three conditions: (i) they fall at the center of a non-transitive triplet of nodes; (ii) the hits between a candidate composite sequence and each of its two direct neighbors in such triplets must not overlap by more than 20 amino acids. (These short windows of potential overlap account for BLAST tendency to slightly extend a hit between 2 similar regions over non homologous regions by a few amino acids; this overlap criterion did not affect our results, since they were virtually unchanged when removing it – identifying 9,177 composite genes instead of 9,872, and 2,959 multi-composite genes instead of 3,351, see below). (iii) Along a non-transitive triplet, the edges between sequences with component fragments and the candidate composite sequence must present a similarity above the twilight zone (Rost 1999) (an E-value of < 10E-10 instead of the E-value of < 10E-5 used for network building), so no similarity, however weak, is found between component sequences. This latter condition ensures that non-transitive triplets do not comprise homologous divergent sequences, aligned over distinct regions. There were 423 million triplets to investigate, out of which 123 million were non-transitive (i), 85 million also fulfilled condition (ii) and 53 million fulfilled all three conditions. Within these latter, we counted 9,872 composite genes (8% of the

data set, 10% of the sequences present in the network when singletons are excluded from the data set). Without enforcing the stringency condition (iii), 12% of the sequences (15% of the sequences present in the network without the singletons) were diagnosed as composite.

Of course, such composite genes may be the outcomes of two distinct types of processes occurring in viral genomes, or in their cellular hosts: fusion events (when components of composite genes originate from different gene families) and fission events (when components of composite genes terminate in different gene families). Here, we did not attempt to distinguish between these two processes. Rather we focused on another observation: all viral classes contained at least one composite gene (Table 1).

a. Baltimore classes	Dataset	Composite	Multi-composite
1 : dsDNA	109324	7488 (6,8%)	2372 (2,2%)
2 : ssDNA	3071	732 (23,8%)	12 (0,4%)
3 : dsRNA	819	35 (4,3%)	4 (0,5%)
4 : +ssRNA	6000	1218 (20,3%)	763 (12,7%)
5 : -ssRNA	983	94 (9,6%)	31 (3,2%)
6 : +ssRNA DNA intermediate	394	148 (37,6%)	80 (20,3%)
7 : dsDNA RNA intermediate	283	43 (15,2%)	41 (14,5%)
Unknown	1518	114 (7,5%)	48 (3,2%)
b. Nucleic acid	**Dataset**	**Composite**	**Multi-composite**
DNA	112941	8270 (7,3%)	2431 (2,2%)
RNA	8212	1495 (18,2%)	878 (10,7%)
Unknown	1239	107 (8,6%)	42 (3,4%)
c. Monophyletic groups	**Dataset**	**Composite**	**Multi-composite**
1	6587	1212 (18,4%)	702 (10,7%)
2	675	188 (27,9%)	118 (17,5%)
3	3241	731 (22,6%)	10 (0,3%)
4	59937	3683 (6,1%)	1330 (2,2%)
5	23765	2458 (10,3%)	652 (2,7%)
NA	28187	1600 (5,7%)	539 (1,9%)
	Dataset	**Composite**	**Multi-composite**
d. Total	122392	9872 (8,1%)	3351 (2,7%)

Table 1: Composite and multi-composite genes in viral classes

Number and percentages of composite and multi-composite genes in Baltimore and major monophyletic viral classes and by type of nucleic acid.

Furthermore, we detected an additional class of composite genes, called multi-composite genes. These multi-composite genes exploit sets of genetic segments found in sequences that were themselves identified as composite by the above protocol. For instance, patterns indicating multi-composite genes occur as a result of two successive steps when genetic fragments from distinct composite genes are subsequently assembled into a new sequence. Moreover, sets of multi-composite genes will also be observed when sequences diagnosed as composite are directly connected in the network, since these sequences evolved

from different yet overlapping combinations of a common pool of genetic fragments (Fig. 1). We detected these multi-composite genes by applying the three search conditions described above to a subset of the network, retaining only the sequences already identified as composite. We found 3,351 multi-composite viral sequences (3% of the dataset, 4% of the sequences in the network without singletons). This is the first report of this class of composite sequences in viral genomes. Again, all viral classes contained at least one multi-composite gene (Table 1).

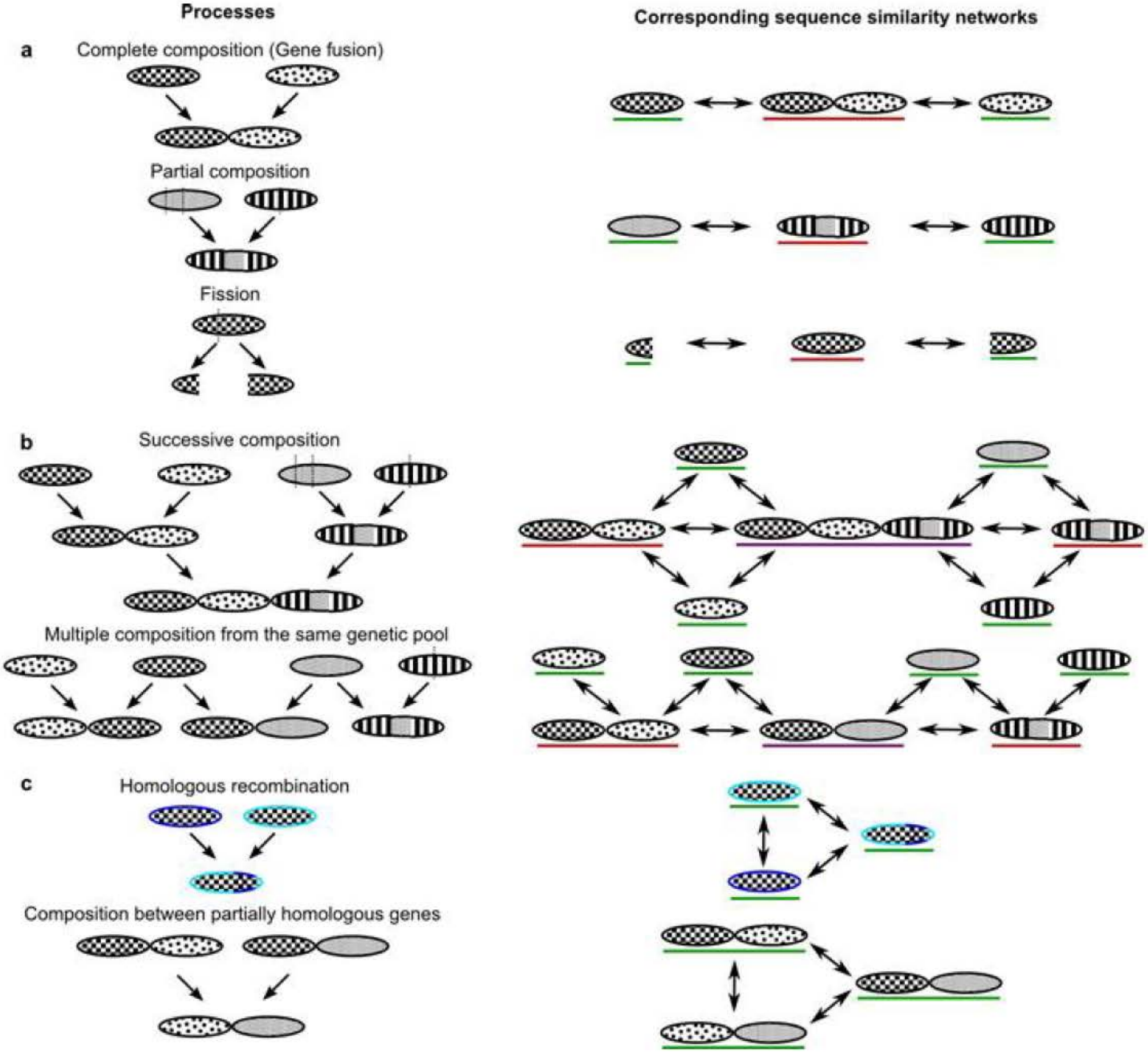


Figure 1. Processes producing composite genes and characteristic similarity patterns

Composite genes result from processes (left) that produce typical similarity networks (right). Shared inner motifs (e.g. wavelets) between genes indicate common ancestry. Underlined in color are genes detected as composite (red), as multi-composite (violet) or not detected as such (green). (**A**) Fusions and fissions lead to the detection of composite genes. Fissioned genes are composite because they combine fragments that exist as independent genes. (**B**) Multiple compositions lead to the detection of multi-composite genes. (**C**) Composition between homologous genes produce transitive similarity relationships and are not detected by this protocol.

These proportions of composite sequences indicate that the fixation of composite genes is a general phenomenon in virus evolution. The number of composite genes is likely an underestimate, since some leave undetectable traces in sequence similarity networks (Fig. 1). This minimal estimate of composite genes is consistent, yet provides new information with respect to former analyses of multidomain genes by Kristensen et al. (Kristensen et al. 2013), because composite genes can be built from segments outside the boundaries of protein domains, and since estimates of composite genes for each viral class and functional categories have not been considered previously (see below).

Remodeling of genes essential to the viral life cycle

Composite genes were found in all functional categories in different proportions (Fig. 2), confirming that they broadly contribute to the range of genetic diversity in viruses. Due to the strong selective pressures acting on viral genomes, one could argue that most of these composite genes are likely adaptative, as viruses have large population sizes theses composite genes would be eliminated. One could argue that some neutral ratchet-like mechanism (a form of constructive neutral evolution) is responsible for the fixation of composite genes in viral genomes. One argument in favor of the adaptive interpretation of this extended distribution of composite genes is provided by the fact that these genes are over-represented in specific functional categories, i.e. the fixation is non-random. More precisely, we defined functional classes as important for viruses using a larger comparative dataset including cellular organisms from all branches of Life for a total of 740,842 sequences. The comparison with this dataset showed functional categories enriched in viruses with respect to cellular organisms. Such categories include replication, recombination and repair (DNA modifications), transcription, RNA processing and modification, chromatin structure and dynamics, post-translational modification, protein turnover and chaperones, nucleotide transport and metabolism, cytoskeleton, cell wall/membrane/envelope biogenesis, extracellular structures, defence mechanisms, and unknown or precisely unknown functions. Remarkably, most of the categories that are functionally important for viruses were also enriched in viral composite genes (with the exception of post-translational modification, protein turnover and chaperones, RNA processing

and modification, defence mechanisms and cytoskeleton). This trend of enrichment in viral composite sequences in functional categories important for viruses was most significant (p-value 0,05) for chromatin structure and dynamics, nucleotide transport and metabolism, cell wall/membrane/envelope biogenesis, and extracellular structures. Therefore, the fixation of composite genes in viruses is biased with respect to functional categories, and composite genes for the most part belong to functions that are essential for the completion of the viral cycle. Noteworthy, several functions particularly enriched in composite genes (e.g. ribonucleotide reductase and thymidylate synthase) are encoded by genomes from large and giant DNA viruses (Boyer et al. 2010). Other composite genes of note encode ankyrin repeat containing-proteins that are known to mimic or manipulate various host functions (Al-Khodor et al. 2010).

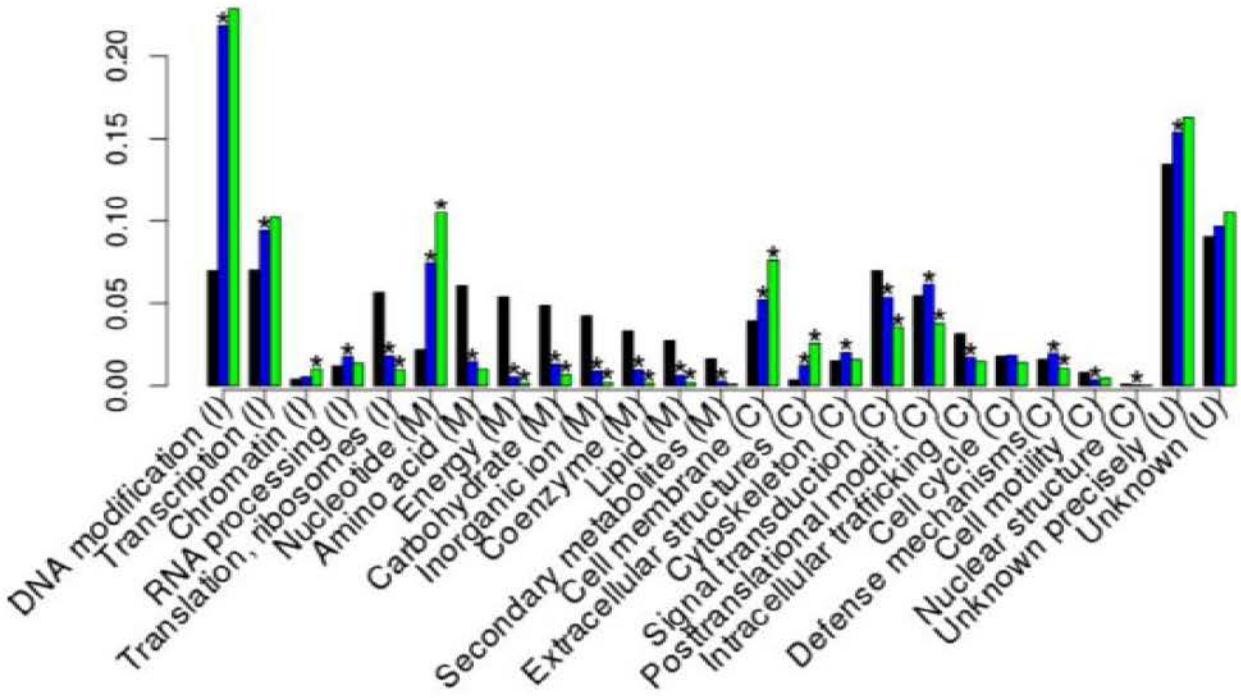


Figure 2. Functional distribution of cellular, viral and viral composite genes

The proportion of genes in each functional category was plotted for the reference "cellular" dataset (black), the viral dataset (blue) and the viral composite subset (green). Genes assigned to multiple categories were redistributed evenly into each of the specified categories. Unannotated genes were not considered. Stars highlight functional categories significantly depleted or enriched in the viral dataset with respect to the cellular dataset, and in viral composite subset with respect to viral dataset (Fisher test, overall significance level of 0.05). Final letter indicates broad functional categories, (I): Information storage and processing, (C): Cellular processes and signalling, (M): Metabolism, (U): Poorly characterized.

Indeed, there is a non-gradual process of molecular evolution at the origin of such composite genes, because both genetic fission and genetic fusion differ from punctual mutations, and may be responsible of larger, potentially more damaging changes in the sequences. Remarkably, functionally important viral categories presented composite genes, even though changes in such key genes may be generally deleterious for their viral hosts. However, in large viral populations, such changes may be highly adaptive and therefore are relatively frequently observed in extant genomes as shown in our analysis. If composite genes within these functional categories are of benefit at least to some members of the population, for example by enhancing their potential to interact with their cellular hosts, to escape their immune systems and defense mechanisms, then that arms race between cells and viruses, composite genes are important players. These composite genes can be formed through a combinatorial process mixing gene lineages that sustains viral life cycles in all viral classes through (lucky) adaptive changes in key viral genes. If this adaptative interpretation is correct, this result proposes a novel instance of the red queen process in evolution, where intimate genetic transformations involving material beyond the boundaries of the gene family allow for the persistence of a lineage.

The quantitative measures of composite genes proposed for each viral class and functional category depends on the quality of sequence annotation of viral genomes, and thus may vary as the annotation improves. We assessed the impact of quality of genomes on our conclusions, by restricting our analyses of the taxonomical and functional distribution of composite genes to a very stringently defined 'safest' subset of 6,144 composite genes, using three additional conditions. We removed from our analysis all genes from non-transitive triplets in which components were found embedded in a composite from the same (this was to circumvent the issue of overlapping genes). We also removed composite genes found exclusively in one non-transitive triplet where the two component genes came from a single genome (to remove false positives due to genes artefactually split during the annotation process of that genome would do). Finally, we additionally removed all composite genes that were only found in one host genome, without homologs in any other genome (to reduce the possibility of including genes artefactually 'fused' during the annotation process of that genome). The 'safest' composite genes are found in all viral classes (following Baltimore classes, major monophyletic classes, or nucleic acid types). We recovered the exact same trends as previously described concerning functional categories (Fig. S1). In addition, 1,920 'safest' multi-composite genes were identified. Consequently, we do not suspect major biases in the trends detected here (although we cannot insulate against overall noise in the data from poorly sequenced genomes or misannotated genes).

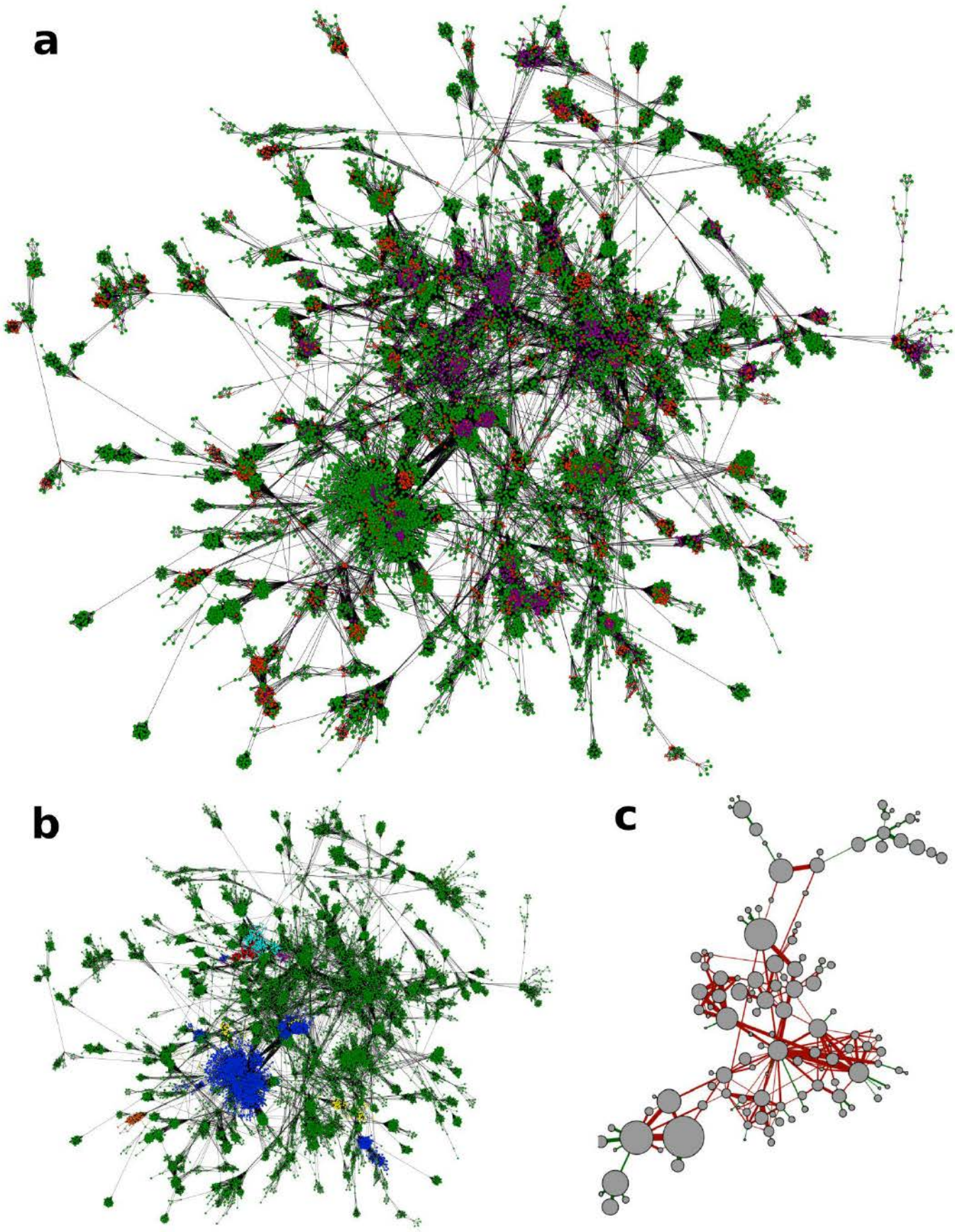

a
b
c

Figure 3. Giant connected component of the viral gene similarity network

This graph contains 15 % of the sequences, held together by composite genes. (**A**) Nodes are individual sequences, edges connect similar sequences (BLAST E-value < 10E-5). Composite are in red, multi-composite in violet, and other genes in green. (**B**) Same graph with colors corresponding to Baltimore classes (dsDNA: green, ssDNA: orange, dsRNA: yellow, +ssRNA: dark blue, -ssRNA: purple, +ssRNA with DNA intermediate: light blue, dsDNA with RNA intermediate: red). (**C**) Simplification of the graph by pooling together densely connected groups of sequences. Super node area is proportional to community size. Edge width is proportional to 1 + log(number of inter-community edges). Edges participating to cycles are colored in red

An informative network view of molecular changes in viruses

The emergence of composite genes operates on a scale that is broader than gene families. Its study requires a more global perspective. The sequence similarity network, describing the viral sequence space, provides a suitable framework. We analyzed the topological properties of our graph to confirm that the detection of composite genes by means of intransitive triplets had successfully identified composite sequences acting as bridges between unrelated protein families. Indeed, composite sequences have a 17 times higher average betweenness than non-composite sequences (2.7e-5 versus 1.6e-6).

We showed that these composite sequences bridge many densely connected regions (called graph communities, identified by Louvain community detection algorithm) into a giant connected component (Fig. 3). . Moreover, composite genes introduce cycles between these graph communities. Such cycles indicate that sequences in this giant connected component have not simply diverged from a last common ancestor. Indeed, while sequence divergence lowers the density of connections between homologous sequences in a sequence similarity network but it does not produce cycles. Homologous sequences presenting little conservation (i.e. a lesser sequence similarity across them than the threshold at which the network is constructed) will eventually produce chains of sequences. Instead, we demonstrated that similarities across sequences found in viruses presented cycles, that we visualized by pooling densely connected groups of nodes together in a super node in the graph (Fig. 3c). These cycles constitute a unique network pattern to diagnose extensive gene remodeling (and non-gradual evolutionary processes).

Importantly, other informative patterns of connections between viral sequences are also observed in the graph. A first major observation from the graph is that genes have a high tendency to be similar to genes from the same viral class, as measured by their assortativity (1 means perfect assortativity). The overall assortativity score for the Baltimore classes is 0.992. Thus, Baltimore classes overall assortativity is 0.992 (class I: 0.994, class II: 0.998, class III: 0.910, class IV: 0.997, class V: 0.987, class IV: 0.706, class VII: 0.705). Regarding classes VI and VII

composed of viruses with different types of nucleic acid but all encoding a reverse transcriptase, their assortativity rises to 0.9 when aggregated. In addition, major monophyletic classes overall assortativity is 0.941 (class 1: 0.954, class 2: 0.929, class 3: 1.000, class 4: 0.903, class 5: 0.916). This preferential connection of like with like, e.g. genes from the same viral class linked with one another, means that full or partial homologs are not usually readily detected in genomes across the viral classes considered here. The sharing or mixing of genetic material is not the rule for viruses from such distinct groups (which should not be confused with lower level classes such as ICTV families, for example, for which some sharing can be observed).

Although generally viruses from different groups have different genes, composite genes are not limited to associations of genetic material within a given viral class. Indeed, some viruses from different classes harbor sequences that are sufficiently similar to connect together in our graph. Consequently, densely connected sets of sequences from different viral classes or exploiting different nucleic acids, fall into the same connected component. Despite the major structural and phylogenetic differences between their members, groups of sequences from viruses from all Baltimore and monophyletic classes (Fig. 3b, fig S2) are indirectly aggregated into the giant connected component, and in some other connected components. This complex pattern is expected when composite genes associate with genetic fragments from different gene families of distinct viral origins into a single composite sequence, or when fragments of a composite sequence are inherited by different gene families from different viral classes. In either case, genetic information present in a given viral class can be effectively remodeled to work into another class of viruses. Figure 3b illustrates such cross-combination of genetic material from RNA and DNA viruses.

These results expand our view on the remarkable plasticity of viral genomes: here illustrated by the combinations of information encoded in genetic material of different types and in unrelated entities (rather than by the more standard acquisition of stand-alone genes from viruses of the same class). Consistently, this holistic network reveals 40 instances of similar sequences distributed across Baltimore viral classes, 20 of them across RNA and DNA viruses, which represents further evidence that information (in particular coding the manipulation of DNA molecules) can be used by multiple members in the viral world, irrespective of biological support (e.g. RNA or DNA) (figs S3 and S4). Some large scale gene sharing between very different mobile entities (i.e. viruses and capsid-less mobile elements) has recently been described elsewhere, giving rise to the concept of a mobilome network. Typically, virophages, polintons, some transposable elements, transpovirons, adenoviruses, and some bacteriophages were reported to form a network of evolutionary relationships, held together by overlapping sets of shared genes (Desnues et al. 2012; Yutin et al. 2013). Our findings on composite genes originating from different viral host lineages provides a fundamentally novel line of evidence for

the recognition of the broad scope of the mobilome network, and for the true genetic intricacy and fluidity within it.

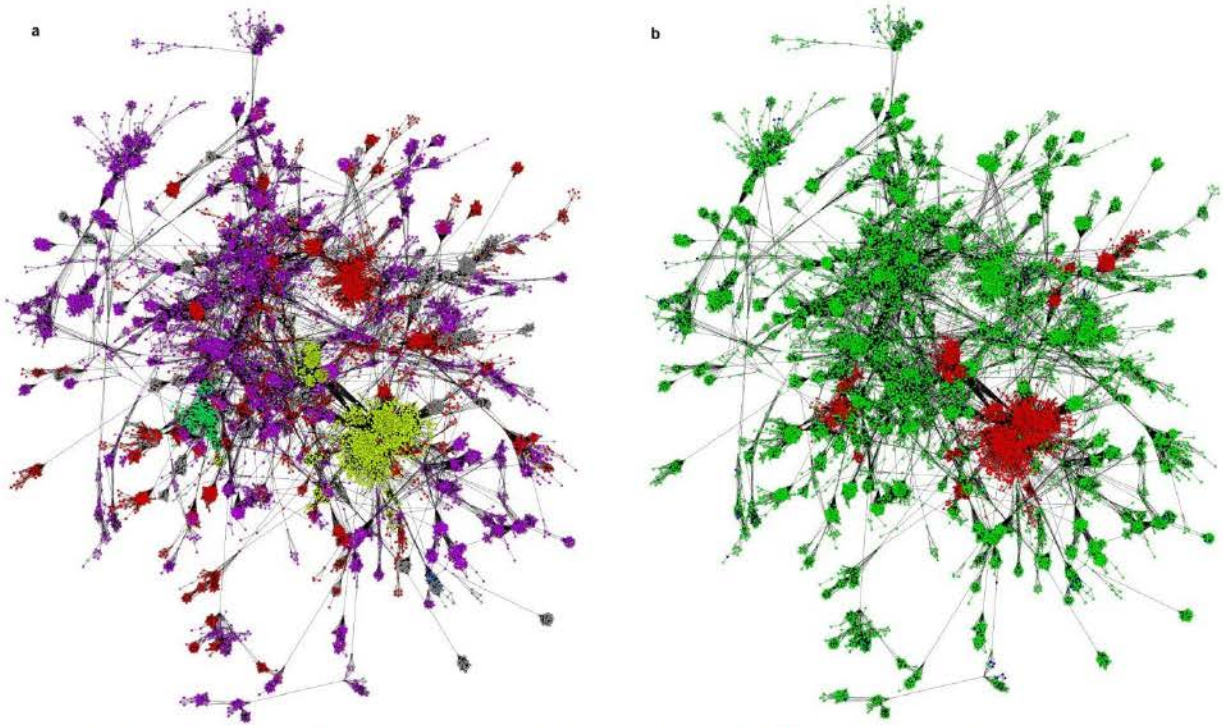


Figure. S2: Giant connected component of the viral gene similarity network

Nodes are individual sequences, edges represent similarity of BLAST E-value < 1e-5. **(A)** Node colors correspond to major monophyletic classes of viruses (1: yellow, 2: green, 3: blue, 4: purple, 5: red). **(B)** Node colors correspond to types of nucleic acids (DNA: green, RNA: red).

Discussion

Our systematic large scale analysis of composite sequences in viral genomes suggests that the fixation of composite genes is a general fundamental phenomenon in virus evolution. Since composite genes were mostly found in functionally important gene categories (this suggests that they play a key role in persistence), in all viral classes. We report the existence of two classes of composite genes, involving genetic components from sequences belonging to distinct, eventually already composite, gene families. These results are relatively unexpected because unlike eukaryotic genomes, viral genomes are not characterized by the presence of intron-exon structure, or junk nucleic acids, that may ease the process of emergence of composite genes. Furthermore, we report composite genes involving information encoded on distant, and even unrelated viral classes, such as RNA viruses and DNA viruses. Viral genomes thus benefit from molecular evolution having occurred in distant lineages, possibly because this information, irrespective of which substrate it is encoded on, allows effective interactions with the machinery of cellular hosts, or alternatively because the functions encoded by some genetic

fragments and compatible with any genome type may trace back to a profound connection between RNA and DNA viruses. Noteworthy, chimeras between RNA and ssDNA viruses were recently proposed to have resulted from recombination (Diemer and Stedman 2012; Roux et al. 2013).

We propose that the emergence of composite genes, relying on the combination of genetic material from different gene families, and occasionally from dramatically different classes of viruses, may be seen as a non-gradual instance of the red queen process. Viral lineages benefit from introgressive combinations of genetic fragments that transform their functionally life cycle important genes, allowing these lineages to survive in the cells-viruses arm-races. Overall, the recognition of composite genes evolving from the association of genetic material beyond the scale of individual viral gene families and from distinct viral lineages provides further evidence that genome mosaicism is a general feature of viruses (Georgiades and Raoult 2012). This finding encourages the development of increasingly combinatorial models and network based analyses of viral evolution. Future finer-grained analyses of the rules of combination of domains in viral genes is definitely one such option. Already, considering that there are around 10E31 viruses on the planet, our results indicate that gene remodeling is a hugely significant way of moving novel sequences between different kinds of organisms.

Materials and methods

Datasets

The viral dataset contains 122,392 protein sequences from 3,008 completely sequenced viral genomes, including all of those available at NCBI on November 2012, and additional genomes from members of the proposed order Megavirales (viruses in Supplementary Information Table 1). The larger comparative dataset, used to define important functional classes for viruses, includes protein sequences from completely sequenced plasmids (all available at NCBI) and a phylogenetically balanced selection of cellular organisms from all of life, resulting in a total of 740,842 sequences. Repartitioning of sequences into genetic vectors is summarized in Table S1, and Supplementary Information Table 1 details all included genomes. Taxonomical annotation was based on (i) classification of viruses into families by the International Commitee on Taxonomy of Viruses (ICTV) (http://talk.ictvonline.org/files/ictv_documents/m/msl/4440.aspx), (ii) Baltimore classification that classified viruses according to the nature of their genome and their replicative strategy (Baltimore 1971), and (iii) classification into five monophyletic classes of viruses and selfish genetic elements as demonstrated by Koonin et al. (Koonin et al. 2006).

Functional annotations

Sequences were functionally annotated by the category (Tatusov et al. 1997) of their best RPSBLAST match (if E-value < 10E-5) against COG (baCteria) and KOG (euKaryota) orthologous

groups (Tatusov et al. 2003). Sequences with no such significant hit were not considered in functional analyses (74% of viral dataset and 50% of larger comparative dataset). We did not used POGs (*24*) orthologous groups, built on viral genomes, since those have not been grouped into higher functional classes.

Statistical test

To determine if a functional category was significantly enriched in one gene set with respect to another, we performed a two-sided Fisher exact test of this category against the combination of every other category. To account for multiple testing on 25 functional categories, we used the conservative Bonferroni correction and considered significant only those categories for which p-value < 0.02 = 0.05/25.

Sequence similarity networks construction and analyses

We used the result of an all-against-all BLAST+ (Camacho et al. 2009) (softmasking with segmasker) comparison to build a sequence similarity network for this dataset, joining pairs of sequences with an E-value < 10E-5. We symmetrized the network and removed multiple edges by keeping the best E-value hit between each pairs of sequences. We mined this network to detect composite genes using FusedTriplets (Jachiet et al. 2013), with a stringency E-value of 10E-10. We searched for multi-composite genes by using the same protocol on the subnetwork of previously identified composite genes. We clustered nodes into densely packed groups as determined by the first pass of Louvain community detection algorithm (Blondel et al. 2008). We used NetworkX (Hagberg et al. 2008) Python library to compute several networks metrics: assortativity (Newman 2003) of viral classes in the network, an approximate betweenness (Brandes and Pich 2007) of nodes using k=5000 random pivots, and a cycle basis of Louvain community network (to find edges participating in cycles). We produced the displays of sequence similarity networks using Cytoscape 2.0 (Shannon et al. 2003) with Force Directed Layout, and the display of Louvain community network Gephi (Bastian et al. 2009) with ForceAtlas2 Layout.

Acknowledgments

We thank David M. Kristensen (NCBI) for helpful discussion about Phage Orthologous Groups (POGs), and Prs. Didier Raoult, Mary J. O'Connell, François Lapointe, and Hervé Le Guyader for critical reading of the manuscript.

References

Adai AT, Date SV, Wieland S, Marcotte EM. 2004. LGL: creating a map of protein function with an algorithm for visualizing very large biological networks. J. Mol. Biol. 340:179–190.

Apic G, Gough J, Teichmann SA. 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J. Mol. Biol. 310:311–325.

Arias CF, Escalera-Zamudio M, de los Dolores Soto-Del Río M, Georgina Cobián-Güemes A, Isa P, López S. 2009. Molecular Anatomy of 2009 Influenza Virus A (H1N1). Arch. Med. Res. 40:643–654.

Atkinson HJ, Morris JH, Ferrin TE, Babbitt PC. 2009. Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. PLoS ONE 4:e4345.

Baltimore D. 1971. Expression of animal virus genomes. Bacteriol. Rev. 35:235–241.

Barr JN, Fearns R. 2010. How RNA viruses maintain their genome integrity. J. Gen. Virol. 91:1373–1387.

Bastian M, Heymann S, Jacomy M. 2009. Gephi: An open source software for exploring and manipulating networks.

Berry A, Pogorelcnik R, Simonet G. 2010. An Introduction to Clique Minimal Separator Decomposition. Algorithms 3:197–215.

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. 2008. Fast unfolding of communities in large networks. arXiv:0803.0476 [Internet].

Botstein D. 1980. A THEORY OF MODULAR EVOLUTION FOR BACTERIOPHAGES*. Ann. N. Y. Acad. Sci. 354:484–491.

Boyer M, Madoui M-A, Gimenez G, La Scola B, Raoult D. 2010. Phylogenetic and Phyletic Studies of Informational Genes in Genomes Highlight Existence of a 4th Domain of Life Including Giant Viruses. PLoS ONE 5:e15530.

Brandes U, Pich C. 2007. Centrality estimation in large networks. Int. J. Bifurc. Chaos 17:2303–2318.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Casjens SR. 2008. Diversity among the tailed-bacteriophages that infect the Enterobacteriaceae. Res. Microbiol. 159:340–348.

Crawford-Miksza LK, Schnurr DP. 1996. Adenovirus Serotype Evolution Is Driven by Illegitimate Recombination in the Hypervariable Regions of the Hexon Protein. Virology 224:357–367.

Desnues C, La Scola B, Yutin N, et al. 2012. Provirophages and transpovirons as the diverse mobilome of giant viruses. Proc. Natl. Acad. Sci. U. S. A. 109:18078–18083.

Diemer GS, Stedman KM. 2012. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. Biol. Direct 7:13.

Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA. 1999. Protein interaction maps for complete genomes based on gene fusion events. Nature 402:86–90.

Enright AJ, Ouzounis CA. 2000. GeneRAGE: A Robust Algorithm for Sequence Clustering and Domain Detection. Bioinformatics 16:451–457.

Filée J. 2009. Lateral gene transfer, lineage-specific gene expansion and the evolution of Nucleo Cytoplasmic Large DNA viruses. J. Invertebr. Pathol. 101:169–171.

Fokkens L, Botelho S, Boekhorst J, Snel B. 2010. Enrichment of homologs in insignificant BLAST hits by co-complex network alignment. BMC Bioinformatics 11:86.

Georgiades K, Raoult D. 2012. How microbiology helps define the rhizome of life. Front. Cell. Infect. Microbiol. 2:60.

Hagberg A, Swart P, S Chult D. 2008. Exploring network structure, dynamics, and function using NetworkX. Los Alamos National Laboratory (LANL)

Hatfull GF. 2008. Bacteriophage genomics. Curr. Opin. Microbiol. 11:447–453.

Highton P j., Chang Y, Myers R j. 1990. Evidence for the exchange of segments between genomes during the evolution of lambdoid bacteriophages. Mol. Microbiol. 4:1329–1340.

Holmes EC. 2003. Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. J. Virol. 77:11296–11298.

Jachiet P-A, Pogorelcnik R, Berry A, Lopez P, Bapteste E. 2013. MosaicFinder: Identification of fused gene families in sequence similarity networks. Bioinformatics [Internet].

Jackwood MW, Hall D, Handel A. 2012. Molecular evolution and emergence of avian gammacoronaviruses. Infect. Genet. Evol. 12:1305–1311.

Al-Khodor S, Price CT, Kalia A, Abu Kwaik Y. 2010. Functional diversity of ankyrin repeats in microbial proteins. Trends Microbiol. 18:132–139.

Koonin EV, Senkevich TG, Dolja VV. 2006. The ancient Virus World and evolution of cells. Biol. Direct 1:29.

Koonin EV, Wolf YI. 2010. Constraints and plasticity in genome and molecular-phenome evolution. Nat. Rev. Genet. 11:487–498.

Koonin EV, Wolf YI, Karev GP. 2002. The structure of the protein universe and genome evolution. Nature 420:218–223.

Kristensen DM, Cai X, Mushegian A. 2011. Evolutionarily Conserved Orthologous Families in Phages Are Relatively Rare in Their Prokaryotic Hosts▽ . J. Bacteriol. 193:1806–1814.

Kristensen DM, Waller AS, Yamada T, Bork P, Mushegian AR, Koonin EV. 2013. Orthologous Gene Clusters and Taxon Signature Genes for Viruses of Prokaryotes. J. Bacteriol. 195:941–950.

Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. 2011. Computational methods for Gene Orthology inference. Brief. Bioinform. 12:379–391.

Lai MM. 1992. RNA recombination in animal and plant viruses. Microbiol. Rev. 56:61–79.

Lei F, Shi W. 2011. Prospective of Genomics in Revealing Transmission, Reassortment and Evolution of Wildlife-Borne Avian Influenza A (H5N1) Viruses. Curr. Genomics 12:466–474.

Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. 2008. Reticulate Representation of Evolutionary and Functional Relationships between Phage Genomes. Mol. Biol. Evol. 25:762 –777.

Liu J, Glazko G, Mushegian A. 2006. Protein repertoire of double-stranded DNA bacteriophages. Virus Res. 117:68–80.

Martinsohn JT, Radman M, Petit M-A. 2008. The λ Red Proteins Promote Efficient Recombination between Diverged Sequences: Implications for Bacteriophage Genome Mosaicism. PLoS Genet 4:e1000065.

Medini D, Covacci A, Donati C. 2006. Protein Homology Network Families Reveal Step-Wise Diversification of Type III and Type IV Secretion Systems. PLoS Comput Biol 2:e173.

Newman ME. 2003. Mixing patterns in networks. Phys. Rev. E 67:026126.

Rost B. 1999. Twilight zone of protein sequence alignments. Protein Eng. 12:85–94.

Roux S, Enault F, Bronner G, Vaulot D, Forterre P, Krupovic M. 2013. Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses. Nat. Commun. [Internet] 4.

Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. 2010. Viral Mutation Rates. J. Virol. 84:9733–9748.

Savolainen-Kopra C, Blomqvist S. 2010. Mechanisms of genetic variation in polioviruses. Rev. Med. Virol. 20:358–371.

Shackelton LA, Holmes EC. 2004. The evolution of large DNA viruses: combining genomic information of viruses and their hosts. Trends Microbiol. 12:458–465.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 13:2498–2504.

Sharon I, Alperovitch A, Rohwer F, et al. 2009. Photosystem I gene cassettes are present in marine virus genomes. Nature 461:258–262.

Simon-Loriere E, Holmes EC. 2013. Gene duplication is infrequent in the recent evolutionary history of RNA viruses. Mol. Biol. Evol. 30:1263–1269.

Song N, Joseph JM, Davis GB, Durand D. 2008. Sequence Similarity Network Reveals Common Ancestry of Multidomain Proteins.Vogel C, editor. PLoS Comput. Biol. 4:e1000063.

Tatusov RL, Fedorova ND, Jackson JD, et al. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science 278:631–637.

Wang M, Caetano-Anollés G. 2009. The Evolutionary Mechanics of Domain Organization in Proteomes and the Rise of Modularity in the Protein World. Structure 17:66–78.

Yutin N, Raoult D, Koonin EV. 2013. Virophages, polintons, and transpovirons: a complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. Virol. J. 10:158.

Chapitre 5 - Conclusion

5.1. Analyse des réseaux de similarité de séquences 105
5.1.1. Détection des phénomènes d'évolution combinatoire 105
5.1.2. Etude de grands réseaux de similarité 106
5.1.3. Vers une bibliothèque d'outils pour manipuler les réseaux 109
5.2. Étude de l'évolution combinatoire des gènes 110
5.2.1. Etude systématique dans de grands jeux de données biologique 111
5.2.2. Modélisation de l'évolution des gènes 112
5.3. Pour une démarche pluraliste en évolution 114

Au cours de cette thèse, j'ai développé des méthodes d'analyse de réseaux de similarité de séquences pour étudier l'évolution combinatoire des gènes. La première composante de ce travail est donc bioinformatique, marquée par la manipulation de gros jeux de données. J'ai employé de nombreux programmes existants et en ai écrit de nouveaux. La deuxième composante de ce travail est l'étude de l'évolution combinatoire des gènes. J'ai étudié systématiquement la distribution de ces phénomènes dans le vivant, et participé à une proposition d'extension de la notion d'homologie qui les prenne en compte. Dans cette conclusion, je reprendrai et discuterai séparément ma contribution selon ces deux aspects bioinformatique et biologique.

5.1. Analyse des réseaux de similarité de séquences

5.1.1. Détection des phénomènes d'évolution combinatoire

En premier lieu, mon travail a consisté au développement d'algorithmes pour détecter des phénomènes d'évolution combinatoire, à partir des réseaux de similarité de séquences. Mes premières approches étaient basées sur des mesures de centralité et des algorithmes de clustering. J'ai finalement proposé d'identifier les familles de gènes composites comme des cliques séparatrices minimales du réseau de similarité de séquence. En collaboration avec un informaticien spécialiste des cliques séparatrices minimales, j'ai développé le programme MosaicFinder qui détecte automatiquement des familles composites selon ce patron. J'ai testé MosaicFinder en créant un simulateur d'évolution de gènes composites. Je l'ai également comparé aux méthodes antérieures, qui détectaient des gènes composites isolément et non en familles. J'ai implémenté une version générique de ces méthodes dans le programme FusedTriplets, facile d'emploi et fonctionnant à partir de la donnée standard d'un résultat BLAST [Jachiet et al., 2013].

La comparaison de FusedTriplets et de MosaicFinder sur des données simulées a montré que MosaicFinder produit très peu de faux positifs, et qu'il est donc plus adapté pour des analyses automatiques. Le patron employé par MosaicFinder est plus fiable parce qu'il porte sur un grand nombre de relations de similarité, tandis que FusedTriplets n'en fait intervenir que trois.

MosaicFinder a en revanche une moins bonne sensibilité que FusedTriplets, car il ne détecte que les familles composites qui sont des cliques, c'est-à-dire dont les séquences n'ont pas trop divergées depuis l'événement de composition. Une solution à ce problème serait de définir un patron approché, de détection de quasi-cliques[1] séparatrices minimales. Il n'existe cependant pas encore de fondement théorique à ce problème en théorie des graphes, et les expérimentations à partir d'une modification de MosaicFinder n'ont pas été concluantes.

Une autre difficulté apparaît sur les jeux de données réels, liée à la présence de grands cycles dans le réseau de similarité. Ces grands cycles relient indirectement les composants de certaines familles composites, qui ne sont alors plus séparatrices et ne sont plus détectées par MosaicFinder. Ces cycles posent également un problème à l'algorithme de détection de cliques minimales séparatrices, qui passe beaucoup de temps à les traiter. Alors que MosaicFinder est bien plus rapide que FusedTriplets sur des petits jeux de données qui n'ont pas ces cycles, il devient très lent sur la composante géante de grands réseaux de similarité de séquences. Une solution à ce problème serait de rechercher des séparateurs locaux, qui séparent leur voisinage à une petite distance donnée. Cependant, de même que pour les quasi-cliques séparatrices minimales, il n'existe pas de théorie décrivant ces patrons et permettant de les énumérer de façon efficace. L'étude des *quasi-cliques séparatrices minimales locales* présente ainsi une piste de recherche intéressante pour détecter des gènes composites.

5.1.2. Etude de grands réseaux de similarité

Un des enjeux du laboratoire est de travailler sur des jeux de données toujours plus importants. Au cours de ma thèse, j'ai donc été amené à construire et étudier de très grands réseaux de similarité pour étudier ensemble un maximum de séquences génétiques. Cependant, l'étude de grands réseaux pose des problèmes informatiques, de complexité en mémoire pour les stocker et de complexité en temps pour les traiter. Une part importante de mon travail a donc consisté à développer des méthodes pour traiter ces grands réseaux, en écrivant des programmes dédiés et en les combinant à des programmes ou librairies existantes. On peut distinguer deux phases qui posent des problèmes spécifiques, la construction du réseau d'une part, son analyse et sa visualisation d'autre part.

5.1.2.1. Construction des grands réseaux

[1] Une quasi-clique est un ensemble de sommets qui est *presque* une clique dans le réseau, avec une proportion d'arêtes manquante inférieure à un seuil donné

La construction d'un réseau de similarité de séquence débute par la comparaison avec BLAST de chaque séquence à toutes les autres du jeu de données. Puisque le nombre de ces comparaisons augmente comme le carré du nombre de séquences, les calculs à effectuer deviennent rapidement trop longs pour être traités en un temps raisonnable sur un seul ordinateur. S'il faut une semaine pour traiter un jeu de données, il faut un mois pour un jeu de données seulement deux fois plus gros ; or la tendance est plutôt à une explosion de la taille des jeux de données étudiés.

Une solution pour effectuer ces comparaisons est de paralléliser les calculs, c'est-à-dire de les diviser en sous-calculs puis de les distribuer sur plusieurs ordinateurs, ou sur les nombreux processeurs d'un cluster de calcul. Nous ne disposons pas au laboratoire de cluster de calcul, et la distribution manuelle de calculs entre la dizaine d'ordinateurs personnels du laboratoire est fastidieuse et source d'erreur. J'ai donc mis en place le système de calcul distribué Hadoop au laboratoire. Ce logiciel libre est développé par et pour des grandes sociétés comme Amazon ou Facebook. La distribution des calculs fonctionne sur le principe du *Map-Reduce*, c'est-à-dire la division d'une tâche en sous-tâches (map) et l'assemblage des résultats des sous-tâches (reduce). Ce modèle est général et s'adapte à de nombreux problèmes, possibilités qu'a explorées Slim Karkar, post-doctorant au laboratoire. Pour ma part, j'ai programmé des scripts pour gérer automatiquement la comparaison de gros fichiers de séquences avec BLAST. En associant les 60 processeurs en moyenne non utilisés au laboratoire, nous pouvons désormais traiter un fichier de 2,5 millions de séquences en une semaine, ce qui aurait pris 2 mois sur un ordinateur personnel tournant à plein régime. Les fichiers de résultat BLAST de la comparaison de jeux de données aussi massif mesurent plusieurs centaines de Giga octets. Il devient impossible de charger en mémoire vive une telle quantité de données pour construire le réseau de similarité de séquence. J'ai donc écrit un programme spécifique qui travaille directement à partir de la mémoire du disque dur.

Le problème de la construction de grands réseaux de similarité n'est cependant résolu que temporairement puisque la taille des jeux de données continue d'augmenter rapidement. Une perspective pour accélérer les comparaisons serait d'utiliser une méthode plus rapide que BLAST, comme celle des k-mers. Cette comparaison ne repose pas sur un alignement des séquences, mais uniquement sur des statistiques de partage de petits mots entre séquences. Une autre perspective serait de rassembler au préalable les séquences très similaires dans des groupes, qui ne seraient représentés chacun que par une séquence (noyau) lors des comparaisons et dans le réseau. Le groupement des séquences peut se faire par un algorithme glouton comme celui de CD-Hit et UCLUST, en comparant successivement les séquences aux noyaux des groupes déjà existant, et en créant un nouveau groupe avec cette séquence comme noyau si elle ne peut être rattachée à aucun noyau existant.

5.1.2.2. Analyses des grands réseaux

Après avoir construit un grand réseau, des difficultés peuvent encore apparaître pour effectuer certaines analyses ; soit que la complexité des calculs les rende trop longs, soit qu'il

soit impossible de charger le réseau en mémoire vive. Les logiciels de visualisation interactifs tels que Gephi ou Cytoscape, peuvent notamment devenir très difficiles à utiliser lorsque le nombre d'objets (nœuds et arêtes) d'un réseau dépasse quelques centaines de milliers. Il est alors beaucoup plus difficile d'employer les approches exploratoires qui font la richesse des méthodes de réseau, pour dégager des patrons dans les données ou pour vérifier la qualité de certains calculs.

Une première solution à ce problème est de découper le réseau en morceaux et de les étudier indépendamment. Cela résout directement les problèmes de mémoire vive, et permet de distribuer les calculs longs sur plusieurs ordinateurs. Lors de ma recherche d'une méthode pour identifier les gènes composites, j'ai étudié le fonctionnement de plusieurs algorithmes de clustering sur les réseaux de similarité de séquences. L'algorithme de Louvain est ressorti comme une méthode à la fois extrêmement rapide et qui produit des clusters qui se comportent bien par rapport à la topologie du réseau et des annotations biologiques que l'on peut en faire. Cet algorithme optimise comme beaucoup d'autres le critère de modularité des communautés, mais a la particularité de produire une classification hiérarchique des nœuds en communautés. Il procède effectivement en optimisant localement la modularité des communautés, puis en améliorant récursivement la modularité par des regroupements dans le réseau des communautés ainsi produites. Les communautés au dernier niveau présentent la meilleure décomposition selon le critère de modularité. En revanche, celles de la première étape, plus petites, ont l'intérêt de ne pas être affectées par le problème de résolution de l'indice de modularité, qui a tendance à agglomérer des petites communautés même si elles sont très faiblement connectées. Je n'ai finalement pas employé l'algorithme de Louvain pour détecter directement des gènes composites. Je m'en suis servi pour décomposer des gros réseaux sur lesquels MosaicFinder était trop lent, en recherchant indépendamment des familles composites dans chaque communauté. Cela permet de distribuer les calculs sur chaque communauté, avec l'intérêt additionnel pour MosaicFinder d'éliminer les grands cycles qui l'handicapent. Le programme est ainsi beaucoup plus rapide même sans parallélisation, et il peut détecter les familles composites qui n'étaient auparavant que des séparateurs *locaux*. En revanche cette décomposition ne permettra pas d'identifier les familles composantes et composites si elles sont divisées dans plusieurs communautés différentes.

La décomposition d'un réseau en parties est une méthode générique pour distribuer les calculs à effectuer. Elle a cependant l'inconvénient d'éliminer les arêtes inter-communautés, et donc de produire un résultat différent de l'analyse du réseau complet. Elle n'est de plus adaptée que lorsque l'on s'intéresse à des propriétés locales. Elle ne serait par exemple pas pertinente si l'on s'intéressait à la communication dans un réseau, car l'on séparerait des agents dans des parties différentes. Dans un réseau de similarité de séquences, deux gènes qui sont séparés par de nombreux intermédiaires (p. ex. 5) n'ont probablement pas de relations évolutives intéressantes à mesurer. Les propriétés pertinentes étudiées sur ces réseaux sont donc généralement locales, et l'on peut appliquer une méthode de décomposition. Pour ne pas perdre d'information, et obtenir des résultats identiques à ceux

obtenus sur le réseau entier, il est possible de décomposer le réseau de façon chevauchante. La difficulté est alors de limiter la redondance des calculs, tout en évitant de produire des parties trop grandes pour être analysées. J'ai implémenté une méthode de "bin-packing" pour tester cette approche sur nos réseaux. Elle s'est montrée prometteuse sur des grands réseaux, et donnera peut être lieu à la publication d'une note d'application dans un journal de bioinformatique.

Il reste cependant souvent difficile de décomposer le « cœur » du réseau, qui est une région très densément connectée de la composante connexe géante. De plus, ce cœur peut lui même devenir trop grand sur les très grands réseaux de similarité de séquences, auquel cas il faut recourir à de nouvelles perspectives. Une première est de diminuer la taille du réseau en groupant les séquences similaires, par exemple celles qui partagent beaucoup de voisins (comme p. ex. l'algorithme de Neighborhood Correlation) et occupent donc des positions équivalentes dans le réseau. Une autre perspective est d'effectuer les calculs sur un système distribué. Il est cependant difficile de distribuer les calculs sur les réseaux de façon générique, et de diviser une tâche en sous-tâches. Le réseau est par nature auto-dépendant, et plus facilement traité dans son ensemble, car chaque nœud peut être connecté à n'importe quel autre. Une solution est d'employer un système de calcul distribué spécialement conçu pour les réseaux, tel que Pregel développé par des ingénieurs de Google [Malewicz et al., 2010]. Ce système adopte une perspective centrée sur les nœuds du réseau. Chaque nœud peut effectuer des calculs à partir d'informations qu'il stocke, et d'informations qu'il reçoit de son voisinage. Il est ainsi possible de distribuer ces calculs, puisque seule la topologie locale autour d'un nœud est nécessaire pour les effectuer. Cela demande cependant de changer la conception des algorithmes, qui doivent faire émerger une solution à partir de solutions locales stockées sur les nœuds.

5.1.3. Vers une bibliothèque d'outils pour manipuler les réseaux

Un argument de cette thèse est qu'il est simple, rapide, et souvent très fructueux d'analyser les informations de similarité entre séquences à l'aide de réseaux. Si l'on souhaite que ces méthodes de réseaux se développent dans la communauté scientifique, il faut que tout chercheur puisse facilement reproduire des analyses existantes et en implémenter de nouvelles. Malgré la simplicité théorique des méthodes de réseaux, elles restent difficiles à mettre en place. Elles nécessitent d'abord apprendre à employer de nombreux outils hétérogènes : la comparaison de séquences (BLAST), les logiciels de calculs sur les réseaux (IGraph, NetworkX), de visualisation (Cytoscape, Gephi), et de statistique (R). Il faut ensuite écrire des scripts spécifiques pour créer le réseau, effectuer des calculs, ou passer d'un format de réseau à l'autre. Il faut enfin assembler toutes ces étapes de façon fluide pour pouvoir les reproduire aisément.

Une solution naturelle est de créer un programme tout-en-un pour analyser les réseaux. Elle ne me semble pas judicieuse car ce programme ne pourrait pas être performant pour toutes les étapes, serait difficile à écrire et difficile à maintenir à niveau dans tous les domaines. De plus, s'il est possible d'employer des programmes généralistes pour traiter des

petits jeux de données, il faut employer des structures de données spécifiques et optimisées à chaque problème pour travailler sur des grands réseaux.

Une solution plus réaliste est d'adopter une approche modulaire, en s'inspirant par exemple des outils développés en phylogénie. La construction d'un arbre phylogénétique demande de nombreuses étapes, qui sont effectuées par des programmes spécifiques. BLAST permet de comparer les séquences, TribeMCL de les grouper en familles, Muscle de créer des alignements multiples, GBLOCKS de nettoyer ces alignements, PhyML de construire des arbres phylogénétiques, Forester/Archéoptérix de visualiser ces arbres. Ces programmes sont faciles à intégrer dans des scripts qui automatisent ces procédures, des « pipelines » d'analyse, car ils fonctionnent en ligne de commande, et communiquent entre eux par des fichiers textes dans des formats standards (FASTA, sortie BLAST, Newick). Cette approche modulaire permet aussi de remplacer facilement un ancien programme par un nouveau plus performant.

Mon idée est donc de créer une bibliothèque d'outils pour créer et manipuler les réseaux, qui fonctionne à partir de fichiers de données simples et standardisés. Ces outils doivent tous fonctionner sur le même modèle, être bien documentés et employer des options claires. Pour réaliser une première version de ces outils, il me semble adapté de s'appuyer sur une librairie de manipulation de réseaux, telle que Igraph, NetworkX, ou graph-tool. NetworkX pourrait être un bon choix, car il est codé en Python, il a une syntaxe claire, et il emploi une structure de graphe très polyvalente, au prix d'une consommation importante de mémoire. Des versions plus efficaces de certains outils, adaptées pour travailler sur des grands réseaux, pourraient être ensuite codées dans des langages et des formats de données plus performants au cas par cas, en respectant la syntaxe de commande des outils précédents.

Il faut de plus associer à cette bibliothèque des tutoriels permettant de reproduire des analyses très courantes, et d'autres pour apprendre à créer des nouvelles analyses à partir des possibilité offertes par les logiciels de calculs sur les réseaux. Un format adapté pour ce type de documentation serait un site internet, statique ou sous forme de blog, qui puisse associer ligne de commande, captures d'écrans et vidéos. Ce travail important pourrait être valorisé par la publication d'articles dans des journaux présentant cette librairie.

5.2. Étude de l'évolution combinatoire des gènes

Ce travail de thèse a également apporté des contributions en biologie évolutive. Il a permis d'améliorer la compréhension des phénomènes combinatoires d'évolution des gènes, de mesurer leur distribution dans les organismes vivants et dans les fonctions biologiques. Il a enfin permis de réfléchir à la nature des relations d'homologie entre séquences, et de proposer une façon alternative de les appréhender directement à partir de réseaux de similarité de séquences.

5.2.1. Etude systématique dans de grands jeux de données biologique

Mon premier apport à ce sujet a été de créer le programme MosaicFinder qui détecte automatiquement des familles de gènes composites à partir d'un réseau de similarité de séquences, là où les méthodes antérieures détectaient des gènes composites isolés. J'ai pour cela formalisé en termes de réseaux les méthodes existantes de détection de gènes fusionnés, qui consistent à identifier des triplets de gènes intransitifs ; c'est-à-dire une chaîne ouverte de trois séquences où la séquence centrale est connectée aux deux autres, sans que celles-ci ne soient connectées. J'ai élargi la signification de ce patron : il détecte des gènes fusionnés mais également des combinaisons partielles entre séquences non-homologues – typiquement l'acquisition d'un domaine accessoire – ou des gènes fissionnés. MosaicFinder présente des intérêts informatiques, présentés plus haut (robustesse, rapidité lorsqu'il n'y a pas de grand cycle), mais il permet surtout d'améliorer les analyses évolutives. Les patrons qu'il identifie correspondent directement à des événements uniques, alors qu'il est difficile d'associer *a posteriori* les gènes composites identifiés séparément à un même événement de composition ancestral. MosaicFinder est actuellement utilisé par le groupe de recherche de Mary J. O'Connell pour identifier des familles de gènes composites dans les génomes de grands singes (communication personnelle). Après avoir analysé les mécanismes moléculaires à l'origine des familles identifiées, ils étudient maintenant le patron d'expression de ces gènes dans plusieurs tissus de différentes espèces.

J'ai étudié de façon systématique la distribution les événements de composition, en les appliquant les programmes MosaicFinder et FusedTriplets à un grand jeu de données comprenant 591.439 séquences provenant d'éléments génétiques mobiles (virus et plasmides) et d'organismes issus des trois domaines du vivant (bactéries, archées, eucaryotes). Cette étude a montré que les événements de composition existent dans tous les domaines du vivant. Les organismes eucaryotes, dont les gènes ont une structure en intron-exon favorisant la recombinaison non-homologue, ont de loin la plus grande proportion moyenne de gènes composites dans leurs génomes. On en trouve plus chez les bactéries que chez les archées, et plus dans les plasmides que dans les virus. Ces derniers écarts sont cependant à relativiser car la proportion de gènes composites identifiés peut varier entre les vecteurs génétiques si l'échantillonnage de leur diversité n'est pas équilibré. La diversité des gènes viraux est par exemple tellement importante que de nombreux gènes ne sont similaires à aucun autre du jeu de données ; un tel gène isolé ne participe à aucun triplet et ne pourra jamais être identifié comme composite. J'ai également étudié la fonction des gènes composites, et montré qu'il en existe dans toutes les catégories fonctionnelles. La composition est donc un phénomène général, qui concerne tous les organismes vivants et toutes leurs fonctions. Dans le futur, une étude systématique des fonctions des composites autour d'une séquence composite devrait permettre d'établir certaines règles de la « composition des gènes ».

5.2.2. Modélisation de l'évolution des gènes

Ce travail m'a amené à réfléchir à la façon dont nous modélisons l'évolution des gènes, et plus précisément à l'apparition de variations dans la diversité des gènes. Nous ne nous intéresserons pas directement aux mécanismes de fluctuation numérique de cette diversité, comme la sélection naturelle ou la fluctuation neutre. Le modèle le plus courant d'évolution des gènes considère qu'ils sont soumis aléatoirement à des petites modifications locales (mutations, insertions, délétions). Ce modèle est employé pour reconstruire des arbres phylogénétiques de gènes, qui retracent les relations de parenté entre gènes à partir du patron des variations ponctuelles dans les séquences. Cependant, les gènes se transforment également par des événements de plus grande amplitude, qui ne sont pas pris en compte par ce modèle standard.

Les recombinaisons homologues créent de nouvelles séquences, qui sont différentes sur des régions entières de leurs séquences parents. Placée dans un arbre phylogénétique, une séquence issue d'une recombinaison homologue contient une information de branchement contradictoire, entre les séquences parents. Ces événements de recombinaisons homologues peuvent être ajoutés au modèle d'évolution des séquences par mutation ponctuelle, et représentés dans un arbre par des jonctions de branches. Il est cependant difficile d'implémenter ce modèle dans une méthode de reconstruction automatique de l'histoire évolutive des séquences comparable aux programmes de phylogénie ; la combinatoire des recombinaisons possibles est gigantesque tandis que l'information évolutive portée par les séquences est limitée. En pratique, il existe plusieurs méthodes pour rechercher indépendamment les recombinaisons homologues, ou le conflit entre les signaux phylogénétiques des séquences peut être représenté avec des réseaux phylogénétiques.

Les événements de recombinaisons non-homologues sont fréquents, ils engendrent une part importante de la diversité génétique, mais ils ne peuvent pas être représentés par un modèle arborescent classique de l'évolution. Il n'est pas possible d'adapter ce modèle, car il suppose une origine évolutive commune et unique aux groupes de séquences étudiées, représentés par la racine de l'arbre ou du réseau phylogénétique. Or les séquences produites par des recombinaisons non-homologues ont plusieurs origines différentes. Cette part de l'histoire évolutive est éliminée lorsque les gènes sont groupés en séquences homologues. Pour la prendre en compte, il faut renouveler notre conception de la notion d'homologie, qui détermine les entités comparables entre organismes.

Habituellement, des séquences sont considérées comme homologues si elles descendent d'*une même et unique* séquence ancestrale. Nous proposons d'étendre la notion d'homologie pour inclure les homologues partiels, et comparer des séquences qui partagent *au moins une* origine évolutive commune. Cette notion d'homologie ne permet pas de justifier un partage strict de la diversité des séquences en familles homologues. Nous argumentons cependant qu'il est possible de réaliser des analyses évolutives pertinentes en groupant des séquences qui partagent un *air de famille*, c'est-à-dire des séquences qui forment des communautés dans un réseau, ici un réseau de similarité de séquence mais

éventuellement un réseau de partage de fonctions, de positions dans les génomes ou autres caractéristiques. Nous montrons que les méthodes de constitution de familles homologues au sens classique fonctionnent de fait selon ce principe.

Nous invitons par ailleurs au développement de nouvelles méthodes d'étude de l'histoire des séquences, tels que les réseaux phylogénétiques à plusieurs racines, ou l'analyse directe des réseaux de similarité de séquences (cf. article chapitre 3). Nos analyses des processus combinatoires à grande échelle, dans tout le vivant ou plus spécifiquement chez les virus, illustrent l'apport de méthodologies complémentaires aux représentations arborescentes. La découverte de séquences composites unissant le matériel génétique de virus utilisant un support héréditaire différent (ARN et ADN), en accord avec la découverte récente de virus hybrides ARN-ADN, démontre le potentiel de notre approche pour découvrir des relations à des échelles sortant du cadre évolutif traditionnel.

Enfin, une perspective intéressante à ce travail serait de combiner l'étude de plusieurs niveaux biologiques pour étendre la portée des analyses évolutives, par exemple en étudiant simultanément l'évolution des gènes et des génomes, ou des domaines protéiques et des gènes. Mon co-directeur de thèse Philippe Lopez explore ainsi l'utilisation de réseaux bipartites pour représenter et étudier les relations d'appartenance des gènes aux génomes. Une autre collaboration en cours de mes directeurs de thèse avec l'informaticien Laurent Viennot a pour but de détecter automatiquement des domaines protéiques à partir de l'analyse d'un réseau de similarité de séquences. De la même façon qu'on étudie la position sociale d'un individu en visualisant les communautés auxquels il appartient dans son réseau social local, on pourrait étudier l'évolution d'un génome (resp. d'un gène) en visualisant sa position dans des communautés définies par le partage de gènes (resp. de domaines).

5.3. Pour une démarche pluraliste en évolution

« Quoiqu'il en soit, une théorie défectueuse, accréditée par les suffrages des maîtres de la science, doit avoir tôt ou tard de graves inconvéniens. L'élève s'égare en suivant cette lumière trompeuse. Accoutumé à ne considérer les objets que sous un point de vue systématique, il finit par ne plus les voir tels qu'ils sont réellement, mais tel qu'il trouve commode de se les représenter. »

MIRBEL, *CONSIDERATIONS SUR LA MANIERE D'ETUDIER L'HISTOIRE NATURELLE DES VEGETAUX*, 1810.

« different theories can successfully describe the same phenomenon through disparate conceptual frameworks. […] Each theory can describe and explain certain properties, and neither theory can be said to be better or more real than the other. »

HAWKING STEPHEN, AND LEONARD MLODINOW, *THE GRAND DESIGN: NEW ANSWERS TO THE ULTIMATE QUESTION OF LIFE*, 2010.

« Given that so much of evolutionary biology is about the unique history of a single instantiation of life known to us and that so much of this history depends on chance and contingency, a concise metanarrative seems to be impossible in principle. The best one may hope for is a tapestry of multiple narratives at different levels of generality and abstraction. »

KOONIN EUGENE V, THE LOGIC OF CHANCE: THE NATURE AND ORIGIN OF BIOLOGICAL EVOLUTION, 2011.

L'étude de l'évolution, comme de tous les phénomènes naturels nécessite le recours à des modèles. Un modèle est un ensemble de concepts et de relations logiques qui les relient, entre eux et aux observations. Un modèle est une représentation du monde qui permet de le rendre intelligible. Lorsque l'on classifie la matière en trois états, solide, liquide et gazeux, on emploie ainsi un modèle, très utile pour mieux la manipuler. Un objectif de la démarche scientifique est d'améliorer les modèles employés, pour qu'ils donnent un résumé plus simple de la réalité, qu'ils soient plus précis dans leurs prédictions, ou encore qu'ils aient une portée plus large. Ces différentes qualités ne sont cependant généralement pas réunies par un même modèle, et plusieurs théories coexistent souvent pour traiter un même phénomène. Selon le pluralisme ontologique, ou sa formulation sous forme de *model-dependent realism* défendue par l'astrophysicien Stephen Hawking [Hawking, Mlodinow, 2010], le monde n'est accessible que par des modèles, il n'a pas de réalité indépendante. La représentation générale que nous avons du monde ne peut donc être autre chose qu'une union de modèles employés pour des aspects particuliers. Si deux modèles s'accordent avec les observations, cela n'a pas de sens de dire qu'un des deux est plus réaliste que l'autre. Il est possible de les choisir à sa convenance, en fonction des questions traitées et des caractéristiques privilégiées.

Lorsque l'on étudie l'évolution des séquences, le modèle le plus courant est la phylogénie. Ce modèle à l'intérêt d'être simple, d'avoir une formulation mathématique

précise et de reconstituer une histoire explicite des événements passés. D'autres modèles ont été proposés qui considèrent des processus évolutifs supplémentaires, quitte à perdre la possibilité de reconstituer une histoire évolutive des séquences dans un schéma précis. Au cours de cette thèse, j'ai ainsi utilisé des réseaux de similarité de séquences pour étudier les phénomènes combinatoires d'évolution des séquences. Cette approche peut sembler incompatible avec la phylogénie moléculaire. On pourrait être tenté d'opposer ces modèles, pour déterminer lequel est le plus proche de la réalité de l'évolution des séquences. Cette opposition risque de ne pas être fructueuse car ces modèles ont des caractéristiques différentes, qui leur permettent de mieux représenter des aspects différents de l'évolution des séquences. Il est cependant peut être possible de les réconcilier en adoptant une posture pluraliste et pragmatique, c'est-à-dire en abandonnant la prétention d'universalité d'un modèle sur les autres, et en choisissant pour chaque question le modèle le plus adapté pour la traiter, tout en ayant conscience de ses limites.

Les arbres hiérarchiques ont longtemps été la façon privilégiée d'organiser la complexité, que ce soit pour établir une clé de détermination taxonomique, pour schématiser un organigramme d'entreprise, ou pour concevoir un système de distribution d'électricité. Les arbres ont de nombreuses propriétés intéressantes, qui les rendent particulièrement intelligibles et faciles à mémoriser pour un être humain. Avec le développement de l'informatique, il devient plus facile et naturel de manipuler les schémas complexes et interconnectés que sont les réseaux. Il n'est plus nécessaire d'organiser les informations de façon hiérarchique si un système de recherche performant permet de les retrouver dans une base de données relationnelle. Les réseaux sont ainsi employés par de nombreux domaines pour schématiser leur organisation : la structure des liens hypertextes d'internet, les systèmes écologiques, ou encore l'organisation sociale. Avec le développement des technologies de l'information, nous prenons l'habitude de penser les systèmes dans leurs relations, de manière globale. L'époque est aux démarches collaboratives, transversales, pluridisciplinaires. Terminons donc cette thèse par une proposition : la biologie évolutive ne pourra que s'enrichir en employant davantage des représentations en réseaux.

Chapitre 6 - Bibliographie

ADAI, A.T., DATE, S.V., WIELAND, S. et MARCOTTE, E.M., 2004. LGL: creating a map of protein function with an algorithm for visualizing very large biological networks. In : *Journal of molecular biology*. 2004. Vol. 340, n° 1, p. 179–190.

ALTSCHUL, S F et GISH, W, 1996. Local alignment statistics. In : *Methods in enzymology*. 1996. Vol. 266, p. 460‑480.

ALTSCHUL, Stephen F., GISH, Warren, MILLER, Webb, MYERS, Eugene W. et LIPMAN, David J., 1990. Basic local alignment search tool. In : *Journal of Molecular Biology*. 5 octobre 1990. Vol. 215, n° 3, p. 403‑410. DOI 10.1016/S0022-2836(05)80360-2.

ATKINSON, Holly J., MORRIS, John H., FERRIN, Thomas E. et BABBITT, Patricia C., 2009. Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. In : *PLoS ONE*. 3 février 2009. Vol. 4, n° 2, p. e4345. DOI 10.1371/journal.pone.0004345.

BALTIMORE, D., 1971. Expression of animal virus genomes. In : *Bacteriological Reviews*. 1 septembre 1971. Vol. 35, n° 3, p. 235‑241.

BAPTESTE, Eric, 2013. *Les gènes voyageurs*. Paris : BELIN LITTERATURE ET REVUES. ISBN 9782701158853.

BAPTESTE, Eric, LOPEZ, Philippe, BOUCHARD, Frédéric, BAQUERO, Fernando, MCINERNEY, James O et BURIAN, Richard M, 2012. Evolutionary analyses of non-genealogical bonds produced by introgressive descent. In : *Proceedings of the National Academy of Sciences of the United States of America*. 6 novembre 2012. Vol. 109, n° 45, p. 18266‑18272. DOI 10.1073/pnas.1206541109.

BARBER, Alan E, 2nd et BABBITT, Patricia C, 2012. Pythoscape: a framework for generation of large protein similarity networks. In : *Bioinformatics (Oxford, England)*. 1 novembre 2012. Vol. 28, n° 21, p. 2845‑2846. DOI 10.1093/bioinformatics/bts532.

BARR, John N. et FEARNS, Rachel, 2010. How RNA viruses maintain their genome integrity. In : *Journal of General Virology*. 1 juin 2010. Vol. 91, n° 6, p. 1373‑1387. DOI 10.1099/vir.0.020818-0.

BASTIAN, Mathieu, HEYMANN, Sebastien et JACOMY, Mathieu, 2009. Gephi: an open source software for exploring and manipulating networks. In : *ICWSM*. S.l. : s.n. 2009.

BENSON, G, 1999. Tandem repeats finder: a program to analyze DNA sequences. In : *Nucleic Acids Research*. 15 janvier 1999. Vol. 27, n° 2, p. 573‑580.

BORNBERG-BAUER, Erich, HUYLMANS, Ann-Kathrin et SIKOSEK, Tobias, 2010. How do new proteins arise? In : *Current Opinion in Structural Biology*. juin 2010. Vol. 20, n° 3, p. 390 ‑ 396. DOI 10.1016/j.sbi.2010.02.005.

BRANDES, Ulrik, ROBINS, Garry, MCCRANIE, A. N. N. et WASSERMAN, Stanley, 2013. What is network science? In : *Network Science*. 2013. Vol. 1, n° 01, p. 1–15.

BREITBART, Mya et ROHWER, Forest, 2005. Here a virus, there a virus, everywhere the same virus? In : *Trends in microbiology*. juin 2005. Vol. 13, n° 6, p. 278‑284. DOI 10.1016/j.tim.2005.04.003.

CAMACHO, Christiam, COULOURIS, George, AVAGYAN, Vahram, MA, Ning, PAPADOPOULOS, Jason, BEALER, Kevin et MADDEN, Thomas L, 2009. BLAST+: architecture and applications. In : *BMC bioinformatics*. 2009. Vol. 10, p. 421. DOI 10.1186/1471-2105-10-421.

CHEN, Pei-Jer, KALPANA, Ganjam, GOLDBERG, Janet, MASON, William, WERNER, Barbara, GERIN, John et TAYLOR, John, 1986. Structure and replication of the genome of the hepatitis delta virus. In : *Proceedings of the National Academy of Sciences*. 1986. Vol. 83, n° 22, p. 8774–8778.

DONGARRA, Jack et LUSZCZEK, Piotr, 2011. LINPACK Benchmark. In : PADUA, David (éd.), *Encyclopedia of Parallel Computing*. S.l. : Springer US. p. 1033‑1036. ISBN 978-0-387-09765-7, 978-0-387-09766-4.

EDWARDS, Richard J., DAVEY, Norman E. et SHIELDS, Denis C., 2007. SLiMFinder: A Probabilistic Method for Identifying Over-Represented, Convergently Evolved, Short Linear Motifs in Proteins. In : *PLoS ONE*. 3 octobre 2007. Vol. 2, n° 10, p. e967. DOI 10.1371/journal.pone.0000967.

ENRIGHT, A. J, VAN DONGEN, S. et OUZOUNIS, C. A, 2002. An efficient algorithm for large-scale detection of protein families. In : *Nucleic acids research*. 2002. Vol. 30, n° 7, p. 1575.

ENRIGHT, Anton J., ILIOPOULOS, Ioannis, KYRPIDES, Nikos C. et OUZOUNIS, Christos A., 1999. Protein interaction maps for complete genomes based on gene fusion events. In : *Nature*. 4 novembre 1999. Vol. 402, n° 6757, p. 86‑90. DOI 10.1038/47056.

ENRIGHT, Anton J. et OUZOUNIS, Christos A., 2000. GeneRAGE: A Robust Algorithm for Sequence Clustering and Domain Detection. In : *Bioinformatics*. 1 mai 2000. Vol. 16, n° 5, p. 451‑457. DOI 10.1093/bioinformatics/16.5.451.

FORSLUND, Kristoffer et SONNHAMMER, Erik L L, 2012. Evolution of protein domain architectures. In : *Methods in molecular biology (Clifton, N.J.)*. 2012. Vol. 856, p. 187‑216. DOI 10.1007/978-1-61779-585-5_8.

FORTERRE, Patrick, 2006. The origin of viruses and their possible roles in major evolutionary transitions. In : *Virus Research*. avril 2006. Vol. 117, n° 1, p. 5 - 16. DOI 10.1016/j.virusres.2006.01.010.

FORTUNATO, Santo, 2010. Community detection in graphs. In : *Physics Reports*. Février 2010. Vol. 486, n° 3-5, p. 75‑174. DOI 10.1016/j.physrep.2009.11.002.

FRANCKI, R. I. B., 1985. Plant virus satellites. In : *Annual Reviews in Microbiology*. 1985. Vol. 39, n° 1, p. 151–174.

FRITH, Martin C., 2011a. A new repeat-masking method enables specific detection of homologous sequences. In : *Nucleic Acids Research*. 1 mars 2011. Vol. 39, n° 4, p. e23 - e23. DOI 10.1093/nar/gkq1212.

FRITH, Martin C., 2011b. Gentle masking of low-complexity sequences improves homology search. In : *PloS one*. 2011. Vol. 6, n° 12, p. e28819.

FRITH, Martin C., HAMADA, Michiaki et HORTON, Paul, 2010. Parameters for accurate genome alignment. In : *BMC Bioinformatics*. 9 février 2010. Vol. 11, n° 1, p. 80. DOI 10.1186/1471-2105-11-80.

GHERARDINI, Pier Federico, WASS, Mark N., HELMER-CITTERICH, Manuela et STERNBERG, Michael J. E., 2007. Convergent Evolution of Enzyme Active Sites Is not a Rare Phenomenon. In : *Journal of Molecular Biology*. 21 septembre 2007. Vol. 372, n° 3, p. 817 - 845. DOI 10.1016/j.jmb.2007.06.017.

GILBERT, Walter, 1978. Why genes in pieces? In : *Nature*. 1978. Vol. 271, n° 5645, p. 501.

HALARY, S., LEIGH, J. W., CHEAIB, B., LOPEZ, P. et BAPTESTE, E., 2009. Network analyses structure genetic diversity in independent genetic worlds. In : *Proceedings of the National Academy of Sciences*. décembre 2009. Vol. 107, n° 1, p. 127‑132. DOI 10.1073/pnas.0908978107.

HALARY, Sébastien, MCINERNEY, James O., LOPEZ, Philippe et BAPTESTE, Eric, 2013. EGN: a wizard for construction of gene and genome similarity networks. In : *BMC Evolutionary Biology*. 11 juillet 2013. Vol. 13, n° 1, p. 146. DOI 10.1186/1471-2148-13-146.

HAN, Jung-Hoon, BATEY, Sarah, NICKSON, Adrian A., TEICHMANN, Sarah A. et CLARKE, Jane, 2007. The folding and evolution of multidomain proteins. In : *Nature Reviews Molecular Cell Biology* . 14 mars 2007. Vol. 8, n° 4. DOI 10.1038/nrm2144.

HAWKING, Stephen et MLODINOW, Leonard, 2010. *The Grand Design: New Answers to the Ultimate Question of Life*. S.l. : Bantam Books.

JACHIET, Pierre-Alain, POGORELCNIK, Romain, BERRY, Anne, LOPEZ, Philippe et BAPTESTE, Eric, 2013. MosaicFinder: Identification of fused gene families in sequence similarity networks. In : *Bioinformatics*. 30 janvier 2013. DOI 10.1093/bioinformatics/btt049.

JOSEPH, Jacob M. et DURAND, Dannie, 2009. Family classification without domain chaining. In : *Bioinformatics*. juin 2009. Vol. 25, n° 12, p. i45 -i53. DOI 10.1093/bioinformatics/btp207.

KELL, Douglas B et OLIVER, Stephen G, 2004. Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. In : *BioEssays: news and reviews in molecular, cellular and developmental biology*. janvier 2004. Vol. 26, n° 1, p. 99-105. DOI 10.1002/bies.10385.

KOONIN, Eugene V., SENKEVICH, Tatiana G. et DOLJA, Valerian V., 2006. The ancient Virus World and evolution of cells. In : *Biology Direct*. 19 septembre 2006. Vol. 1, n° 1, p. 29. DOI 10.1186/1745-6150-1-29.

KRAKAUER, D C, 2000. Stability and evolution of overlapping genes. In : *Evolution; international journal of organic evolution*. juin 2000. Vol. 54, n° 3, p. 731-739.

KRISTENSEN, David M., CAI, Xixu et MUSHEGIAN, Arcady, 2011. Evolutionarily Conserved Orthologous Families in Phages Are Relatively Rare in Their Prokaryotic Hosts▽. In : *Journal of Bacteriology*. avril 2011. Vol. 193, n° 8, p. 1806-1814. DOI 10.1128/JB.01311-10.

LA SCOLA, Bernard, AUDIC, Stéphane, ROBERT, Catherine, JUNGANG, Liang, DE LAMBALLERIE, Xavier, DRANCOURT, Michel, BIRTLES, Richard, CLAVERIE, Jean-Michel et RAOULT, Didier, 2003. A giant virus in amoebae. In : *Science*. 2003. Vol. 299, n° 5615, p. 2033–2033.

LEI, Fumin et SHI, Weifeng, 2011. Prospective of Genomics in Revealing Transmission, Reassortment and Evolution of Wildlife-Borne Avian Influenza A (H5N1) Viruses. In : *Current Genomics*. novembre 2011. Vol. 12, n° 7, p. 466-474. DOI 10.2174/138920211797904052.

LIU, Jinfeng et ROST, Burkhard, 2004. CHOP proteins into structural domain-like fragments. In : *Proteins*. 15 mai 2004. Vol. 55, n° 3, p. 678-688. DOI 10.1002/prot.20095.

LONG, Manyuan, 2000. A New Function Evolved from Gene Fusion. In : *Genome Research*. 1 novembre 2000. Vol. 10, n° 11, p. 1655-1657. DOI 10.1101/gr.165700.

LÖWER, Roswitha, LÖWER, Johannes et KURTH, Reinherd, 1996. The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. In : *Proceedings of the National Academy of Sciences*. 1996. Vol. 93, n° 11, p. 5177–5184.

MALEWICZ, Grzegorz, AUSTERN, Matthew H., BIK, Aart JC, DEHNERT, James C., HORN, Ilan, LEISER, Naty et CZAJKOWSKI, Grzegorz, 2010. Pregel: a system for large-scale graph processing. In : *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. S.l. : ACM. 2010. p. 135–146.

MARCOTTE, Edward M., PELLEGRINI, Matteo, NG, Ho-Leung, RICE, Danny W., YEATES, Todd O. et EISENBERG, David, 1999. Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. In : *Science*. Juillet 1999. Vol. 285, n° 5428, p. 751 - 753. DOI 10.1126/science.285.5428.751.

MIELE, Vincent, PENEL, Simon, DAUBIN, Vincent, PICARD, Franck, KAHN, Daniel et DURET, Laurent, 2012. High-Quality Sequence Clustering Guided by Network Topology and Multiple Alignment Likelihood. In : *Bioinformatics*. 15 avril 2012. Vol. 28, n° 8, p. 1078 - 1085. DOI 10.1093/bioinformatics/bts098.

MINSHULL, Jeremy et WILLEM STEMMER, P. C., 1999. Protein evolution by molecular breeding. In : *Current Opinion in Chemical Biology*. juin 1999. Vol. 3, n° 3, p. 284‑290. DOI 10.1016/S1367-5931(99)80044-1.

MIRBEL, Ch Fr, 1810. Considérations sur la manière d'étudier l'histoire naturelle des végétaux. S.l. : s.n.

MOREIRA, David et LÓPEZ-GARCÍA, Purificación, 2009. Ten reasons to exclude viruses from the tree of life. In : *Nature Reviews Microbiology*. avril 2009. Vol. 7, n° 4, p. 306 - 311. DOI 10.1038/nrmicro2108.

NEEDLEMAN, Saul B. et WUNSCH, Christian D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. In : *Journal of Molecular Biology*. 28 mars 1970. Vol. 48, n° 3, p. 443‑453. DOI 10.1016/0022-2836(70)90057-4.

NEWMAN, M. E. J., STROGATZ, S. H. et WATTS, D. J., 2001. Random graphs with arbitrary degree distributions and their applications. In : *Physical Review E*. 24 juillet 2001. Vol. 64, n° 2, p. 026118. DOI 10.1103/PhysRevE.64.026118.

O'HARA, ROBERT J., 1997. Population thinking and tree thinking in systematics. In : *Zoologica scripta*. 1997. Vol. 26, n° 4, p. 323–329.

OWEN, Richard, 1848. *On the archetype and homologies of the vertebrate skeleton*. S.l. : John van Voorst, Paternoster Row.

PAREEK, Chandra Shekhar, SMOCZYNSKI, Rafal et TRETYN, Andrzej, 2011. Sequencing technologies and genome sequencing. In : *Journal of Applied Genetics*. 23 juin 2011. Vol. 52, n° 4, p. 413‑435. DOI 10.1007/s13353-011-0057-x.

PASEK, Sophie, RISLER, Jean-Loup et BRÉZELLEC, Pierre, 2006. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. In : *Bioinformatics*. Juin 2006. Vol. 22, n° 12, p. 1418 ‑1423. DOI 10.1093/bioinformatics/btl135.

PATTHY, L, 1999. Genome evolution and the evolution of exon-shuffling--a review. In : *Gene*. 30 septembre 1999. Vol. 238, n° 1, p. 103‑114.

PEARSON, William R., 1990. [5] Rapid and sensitive sequence comparison with FASTP and FASTA. In : *Methods in enzymology*. 1990. Vol. 183, p. 63–98.

ROOSSINCK, Marilyn J., 2011. The good viruses: viral mutualistic symbioses. In : *Nature Reviews Microbiology*. février 2011. Vol. 9, n° 2, p. 99‑108. DOI 10.1038/nrmicro2491.

RUSSELL, Robert B., 1994. Domain insertion. In : *Protein engineering*. 1994. Vol. 7, n° 12, p. 1407–1410.

RUSSELL, Robert B., 1998. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. In : *Journal of molecular biology*. 1998. Vol. 279, n° 5, p. 1211–1227.

SHANNON, Claude E., 1948. A note on the concept of entropy. In : *Bell System Tech. J.* 1948. Vol. 27, p. 379–423.

SHANNON, Paul, MARKIEL, Andrew, OZIER, Owen, BALIGA, Nitin S., WANG, Jonathan T., RAMAGE, Daniel, AMIN, Nada, SCHWIKOWSKI, Benno et IDEKER, Trey, 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. In : *Genome Research*. 1 novembre 2003. Vol. 13, n° 11, p. 2498‑2504. DOI 10.1101/gr.1239303.

SMITH, T. F. et WATERMAN, M. S., 1981. [a]Identification of Common Molecular Subsequences, [o] J. In : *Molecular Biology*. 1981. Vol. 147, p. 195–197.

SONG, Nan, JOSEPH, Jacob M., DAVIS, George B. et DURAND, Dannie, 2008. Sequence Similarity Network Reveals Common Ancestry of Multidomain Proteins. In : VOGEL, Christine (éd.), *PLoS Computational Biology*. mai 2008. Vol. 4, n° 5, p. e1000063. DOI 10.1371/journal.pcbi.1000063.

SOON, Wendy Weijia, HARIHARAN, Manoj et SNYDER, Michael P., 2013. High-throughput sequencing for biology and medicine. In : *Molecular Systems Biology*. 22 janvier 2013. Vol. 9, n° 1. DOI 10.1038/msb.2012.61.

SUTTLE, Curtis A, 2005. Viruses in the sea. In : *Nature*. 15 septembre 2005. Vol. 437, n° 7057, p. 356-361. DOI 10.1038/nature04160.

WITTGENSTEIN, Ludwig, 1968. *Philosophical investigations.* Oxford : Basil Blackwell. ISBN 0631119000 9780631119005.

WONG, Wing-Cheong, MAURER-STROH, Sebastian et EISENHABER, Frank, 2010. More Than 1,001 Problems with Protein Domain Databases: Transmembrane Regions, Signal Peptides and the Issue of Sequence Homology. In : *PLoS Comput Biol*. 29 juillet 2010. Vol. 6, n° 7, p. e1000867. DOI 10.1371/journal.pcbi.1000867.

WOOTTON, John C., 1994. Non-globular domains in protein sequences: Automated segmentation using complexity measures. In : *Computers & Chemistry*. septembre 1994. Vol. 18, n° 3, p. 269-285. DOI 10.1016/0097-8485(94)85023-2.

YIN, Yanbin et FISCHER, Daniel, 2008. Identification and investigation of ORFans in the viral world. In : *BMC Genomics*. 19 janvier 2008. Vol. 9, n° 1, p. 24. DOI 10.1186/1471-2164-9-24.

ZHANG, J. et KUMAR, S., 1997. Detection of convergent and parallel evolution at the amino acid sequence level. In : *Molecular Biology and Evolution*. 1 mai 1997. Vol. 14, n° 5, p. 527-536.

ZHOU, Qi, ZHANG, Guo-jie, ZHANG, Yue, XU, Shi-yu, ZHAO, Ruo-ping, ZHAN, Zubing, LI, Xin, DING, Yun, YANG, Shuang et WANG, Wen, 2008. On the origin of new genes in Drosophila. In : *Genome Research*. 1 janvier 2008. DOI 10.1101/gr.076588.108.

Jachiet Pierre-Alain – Thèse de doctorat – 2014

Résumé

L'accumulation récente de données de séquences génomiques a montré que l'évolution des gènes n'est pas strictement arborescente. De nombreux processus évolutifs, comme l'exon shuffling, la fusion de gènes ou la recombinaison illégitime remodèlent les gènes, créant des structures composites, formées de parties dont les histoires évolutives sont différentes. Le développement de réseaux de similarité de séquences fournit un cadre analytique permettant d'étudier l'impact de ces processus sur l'évolution moléculaire, en structurant les relations de ressemblance entre séquences et en formalisant en termes de graphes la détection de gènes (triplets intransitifs) et de familles de gènes (cliques minimales séparatrices) composites. La taille des jeux de données actuels, de l'ordre de plusieurs millions de séquences, a également requis le développement de nouveaux outils et méthodes : parallélisation des comparaisons de séquences, visualisation de très grands réseaux par simplification en communautés de Louvain et identification de grands cycles. Appliquées à des jeux de données de génomes eucaryotes et viraux, ces méthodes ont démontré la présence de gènes composites dans tout le vivant et les éléments génétiques mobiles. En proportion, les gènes composites sont plus nombreux dans les génomes eucaryotes ; en nombre absolu, ils sont plus nombreux à être portés par des virus. Chez ces derniers, la distribution fonctionnelle des gènes composites est biaisée (enrichissement dans les familles essentielles pour la perpétuation du cycle viral), et les éléments des gènes composites trouvent même parfois leurs origines dans le matériel génétique de classes virales différentes. Plus généralement, l'étendue des processus combinatoires, en révélant des liens évolutionnaires autres que les liens d'homologie au sens fort, justifie une étude pluraliste des relations de similarité entre séquences.

Mots clés : évolution, réseau, bioinformatique, gènes composites, génomique comparative

Using sequence similarity networks to study combinatorial evolution of genes

Abstract

The recent accumulation of genomic sequence data has shown that gene evolution is not strictly tree-like. Many evolutionary processes, like exon shuffling, gene fusion or nonhomologous recombination remodel genes by creating composite structures that are made from parts with different evolutionary histories. The development of sequence similarity networks provides an analytical framework to study the impact of these processes on molecular evolution, by structuring the resemblance relationships between sequences and by formalizing, in terms of graph theory, the detection of composite genes (intransitive triplets) and gene families (clique minimal separators). The size of current data sets, typically several million sequences, has also required the development of new tools and methods: sequence comparison parallelization, large networks visualization with Louvain communities and large cycles identification. When applied to eukaryotic and viral genome data sets, these methods have shown that composite genes are found throughout cellular organisms and mobile genetic elements. Proportionally, composite genes are more numerous in eukaryotic genomes; in absolute number, they are more numerous in viruses. In the latter, composite genes functional distribution is biased (enrichment of genes families that are essential for the perpetuation of the viral cycle), and the various parts of composite genes sometimes even originate from the genetic material of different viral classes. More generally, the extent of combinatorial processes, by unravelling other evolutionary bonds than homology bonds in the strictest sense, legitimates a pluralistic study of similarity relationships between sequences.

Keywords: evolution, networks, bioinformatics, composite genes, comparative genomics