UNIVERSITE MONTPELLIER II
SCIENCES ET TECHNIQUES DU LANGUEDOC

Document de synthèse pour obtenir le diplôme d'

HABILITATION A DIRIGER DES RECHERCHES

Ecole Doctorale: Biologie des Systèmes Intégrés, Agronomie, Environnement

Bioinformatique pour la Génomique Evolutive

présenté par

Nicolas GALTIER

Jury pressenti:

Christian GAUTIER, Professeur, Lyon	Rapporteur
Laurent EXCOFFIER, Professeur, Berne	Rapporteur
Vincent LAUDET, Professeur, Lyon	Rapporteur
Olivier GASCUEL, Directeur de Recherches, Montpellier	
François BONHOMME, Directeur de Recherches, Montpellier	
Emmanuel DOUZERY, Professeur, Montpellier	

Bioinformatique pour la génomique évolutive

1. (Introduction) L'évolution moléculaire hier et aujourd'hui.

Mon activité de recherches relève de l'évolution moléculaire, une discipline apparue à la fin du XXe siècle, et actuellement en phase de maturité. L'exercice veut que je débute ce document en situant le contexte, c'est-à-dire en donnant mon point de vue sur l'histoire de l'évolution moléculaire, ses origines, sa justification, et les grandes questions qui l'animent; c'est l'objectif de ce chapitre introductif.

1.1. Des néo-Darwiniens aux génomes complets

C'est dans les années 1960 et 70 que l'on voit apparaître les premières publications relevant de l'évolution moléculaire: premières phylogénies moléculaires (Fitch & Margoliash 1967), premières réflexions sur l'horloge moléculaire (Zuckerkandl & Pauling 1965), premières controverses entre neutralistes et sélectionnistes (revue dans Kimura 1983, Gillespie 1991). La biologie évolutive est alors à une époque charnière. La pensée néo-Darwinienne des Fisher, Wright, Dobzhansky, Haldane, Muller, Malécot et consort avait, au cours des décennies précédentes, fait la synthèse de la théorie de l'évolution et de la génétique Mendélienne. En modélisant les mécanismes de l'évolution des gènes dans les espèces et les populations, en formalisant les notions de mutation, sélection, dérive, structuration, migration, etc.… , ces fondateurs ont posé les bases théoriques de la **génétique des populations**, sur lesquelles nous nous appuyons encore à l'heure actuelle. Parallèlement, les systématiciens découvrent et s'approprient (dans la douleur) la cladistique, une philosophie reliant la classification des espèces à leur histoire évolutive, c'est-à-dire leur **phylogénie**. Il est donc remarquable de constater que l'accès aux **données moléculaires** se produit postérieurement aux efforts théoriques permettant leur exploitation. Ces données vont néanmoins assez radicalement bouleverser le paysage.

Pour les systématiciens, elles sont progressivement apparues comme un marqueur phylogénétique à la puissance incomparable. On ne fait guère plus aujourd'hui de phylogénies que via l'outil moléculaire, et les programmes actuels d'inventaire de la biodiversité font appel au séquençage en batterie d'échantillons naturels, frais ou de musée. Pour les généticiens des populations, les données moléculaires ont permis la mise à l'épreuve des modèles précédemment développés, autour notamment de la controverse entre neutralistes et sélectionnistes: quelle part de la variabilité moléculaire observée est expliquée par la sélection naturelle? L'examen du contenu des génomes conduira par ailleurs à l'émergence de concepts nouveaux, tels que l'ADN égoïste et les niveaux de sélection, les gènes mutateurs, l'usage des codons et l'évolution des systèmes génétiques, par exemple. L'inexorable accumulation de jeux de données de plus en plus volumineux a placé la modélisation et l'inférence statistiques au centre de la discipline: on reconstruit l'histoire, on estime des paramètres, on teste des hypothèses. Enfin, les données moléculaires ont clairement contribué au rapprochement entre macro-et micro-évolutionnistes, les données étant pertinentes pour les deux niveaux temporels.

Les années 2000 constitueront à n'en pas douter un deuxième tournant dans l'histoire de l'évolution moléculaire: l'industrialisation des techniques de séquençage fait entrer la biologie dans l'ère de la **génomique**. Les génomes complets, procaryotes d'abord, eucaryotes ensuite, permettent de traiter des questions d'évolution moléculaire jusqu'alors hors de portée. On peut confronter le contenu en gènes des espèces à leur écologie (Tringe et al. 2005, Eichinger et al. 2005), comparer cartes génétiques et données de séquences (Kong et al. 2002), décrire en détail les processus de substitution nucléotidique passés et actuels (Arndt et al. 2003, Meunier & Duret 2004), analyser la dynamique des éléments transposables (Biémont et al. 2003, Hedges et al. 2004), examiner les spécificités des chromosomes sexuels (Lahn & Page 2003, Charlesworth et al 2005),

reconstruire l'histoire des chromosomes, des duplications, des fusions (Jaillon et al 2004), analyser la totalité des gènes impliqués dans un processus métabolique donnée (Richard et al. 2005). La taille des jeux de données disponibles est décuplée: on fait la phylogénies de dizaines de marqueurs, la génétique de centaines de locus, la comparaison de milliers de gènes, l'analyse de millions de nucléotides. On résout en bonne partie la phylogénie des grands groupes d'eucaryotes (Baldauf et al. 2000), des plantes (Davies et al. 2004), des Mammifères (Murphy et al. 2004), et leurs principales dates de divergence (Douzery et al 2004). On découvre l'importance des transferts horizontaux chez les procaryotes (Bapteste et al. 2004, Daubin et al. 2003). On dissèque les patrons de polymorphisme et de divergence chez l'homme (Nielsen et al. 2005), la drosophile (Glinka et al. 2003), l'arabette (Nordborg et al. 2005). On touche du doigt le déterminisme des systèmes de reproduction chez les plantes (Charlesworth et al. 2005). On évalue la distribution des coefficients de sélection des mutations (Piganeau & Eyre-Walker 2003), le taux de mutation délétère (Eyre-Walker & Keightley 1999), de substitutions avantageuses (Bierne & Eyre-Walker 2004). On décrit les processus généraux de l'évolution des protéines, et on les relie à leurs propriétés biochimiques (Yang et al. 2003, King-Jordan et al. 2005). On reconstitue les circonstances génétiques de la domestication du maïs (Tenaillon et al. 2004), de l'acquisition de l'autofécondation chez *Arabidopsis thaliana* (Shimizu et al. 2004), de l'augmentation du volume crânien chez l'homme (Stedman et al. 2004). On comprend la mise en place des voies métaboliques (Canback et al. 2002), des systèmes génétiques complexes (Bertrand et al. 2004, Danchin et al. 2004). Enfin, les progrès de la génétique moléculaire et cellulaire ouvrent la possibilité de mettre en regard évolution moléculaire et construction des phénotypes au travers de l'évo-dévo, chez les angiospermes (Irish & Litt 2005) et les vertébrés (Holland et al. 2004, Garcia-Fernandez 2005) notamment. L'évolution moléculaire se fait désormais appeler "**génomique évolutive**", et devient intimement liée à une autre discipline apparue avec le génome, la bioinformatique.

1.2. La nécessaire bioinformatique

Le génome a achevé de convaincre les biologistes dans leur ensemble qu'ils avaient besoin de l'informatique. C'est du domaine de l'évidence pour ce qui est des ressources **matérielles**: il faut de gros disques durs pour stocker l'information, et des processeurs rapides pour la traiter. Il apparaît tout aussi clairement que nous avons besoin de compétences dans le domaine de la **gestion des données** et de leur **analyse**. De nombreuses interactions entre biologistes et informaticiens se sont donc mises en place autour du génome, entraînant la création d'une discipline d'interface appelée **bioinformatique**.

L'évolution moléculaire me semble avoir significativement contribué à, et bénéficié de, cette montée en puissance de la bioinformatique, et ceci pour diverses raisons. Tout d'abord, étudier l'évolution implique de faire appel à la **modélisation**, le passé n'étant par définition pas observable (aux fossiles près). De ce fait, la communauté des évolutionnistes est marquée par une forte tradition de **biologie théorique**, qui remonte au début du siècle dernier. Les abondantes données de la biodiversité ont par ailleurs suscité le développement de la **biométrie** et des **biostatistiques**, en écologie notamment. Enfin, l'évolution biologique est typiquement représentée par des **arbres** (phylogénies, généalogies), qui s'avèrent être un des objets formels favoris des algorithmiciens, lesquels ont donc naturellement offert leurs compétences pour enrichir le champ de la phylogénie moléculaire et de la théorie de la coalescence.

La bioinformatique est donc omniprésente en génomique évolutive, et fait partie du quotidien des chercheurs du domaine, développeurs ou utilisateurs. Bases de données spécialisées, logiciels d'analyses et de visualisation, développements algorithmiques et

statistiques, services web, environnements de développement: la liste des outils disponibles s'allonge de semaines en semaines, avec, il faut bien le dire, un taux de redondance élevé. La standardisation fait défaut dans le domaine, chacun réinventant la roue dans son coin, et de manière peu réutilisable. La littérature grossit également de publications souvent très descriptives, dans lesquelles le logiciel X est appliqué au jeu de données Y, sans question spécifique ou résultat digne d'intérêt. La profusion des données rend en effet les tests extrêmement puissants, et les *p*-values facilement significatives (Semon et al. 2005), ce qui facilite la publication de résultats même peu intéressants.

Ces excès ne doivent pas faire oublier les véritables avancées théoriques, algorithmiques et logicielles que la génomique a suscitées. Dans certains domaines, tels que l'algorithmique des arbres, les modèles markoviens ou l'inférence bayésienne, informatique, statistique et biologie ont avancé de concert, la course de l'offre et de la demande jouant à plein. Le deuxième chapitre de ce document donnera un aperçu des développements récents des modèles et des méthodes dans certains de ces domaines. Il faut aussi très chaudement encourager les efforts (bénévoles ou presque) tels que ceux de Julien Dutheil, qui a pris sur lui de mettre en place au cours de sa thèse, et en collaboration avec d'autres membres de l'équipe, une librairie C++ dédiée à l'évolution moléculaire (**Bio++**, Dutheil et al., soumis), dont on espère qu'elle pourra contribuer à structurer et mutualiser les développements bioinformatiques dans la discipline, à l'échelle nationale et internationale.

1.3. Les "grandes" questions

La force de la théorie de l'évolution est sa **généralité**, et cette propriété s'étend à l'évolution moléculaire. On y traite des questions qui relèvent de l'organisation et du fonctionnement du monde vivant dans son ensemble, dans le passé comme dans le présent. Parmi les questions qui me semblent aujourd'hui dignes d'intérêt, je citerai par exemple, dans le désordre et sans aucun souci d'exhaustivité:

- les premières étapes de la vie biochimique et cellulaire
- l'origine de la cellule et du génome eucaryote
- l'origine et l'évolution des introns, des transposons
- l'évolution des génomes cytoplasmiques
- la caractérisation des processus mutationnels
- la coévolution, l'épistasie à l'échelle moléculaire
- l'horloge moléculaire, la datation moléculaire
- le déterminisme génétique de la spéciation
- le rôle des duplications dans l'adaptation
- l'impact du sexe et de la recombinaison sur la structure des génomes
- la phylogénie des animaux et la mise en place des plans d'organisation
- le déterminisme écologique du polymorphisme génétique, des taux de substitution
- l'évolution des voies métaboliques
- l'évolution des chromosomes sexuels

Il est intéressant de noter que ces questions (et celles qui manquent), qui pourraient être vues comme délimitant les contours d'une discipline unitaire, sont en réalité traitées par une communauté de chercheurs d'origines diverses, aux centres d'intérêt distincts. Nous avons déjà évoqué les systématiciens et les généticiens des populations, qui, ayant fait de concert appel à l'outil moléculaire, sont amenés à traiter des questions d'évolution moléculaire – quoique le plus souvent à des échelles de temps différentes. On a également cité les statisticiens et informaticiens, qui, à la recherche de champs d'applications pour leur méthodes et algorithmes, apportent une contribution non-négligeable aux progrès de la discipline. S'y rajoutent, enfin, les biologistes moléculaires, qui ont bien compris combien l'approche comparative était utile à la compréhension du fonctionnement des génomes – on comprend mieux comment fonctionne un génome si on sait comment il a été "fabriqué". Cette **multiplicité des cultures** enrichit incontestablement la discipline, anime les congrès, et égaye le quotidien du chercheur lambda, qui trouve tous les jours matière à apprendre de (enseigner à) ses collègues des disciplines voisines.

Malgré leurs disparités, la grande majorité des questions traitées par l'évolution moléculaire ont en commun de s'intéresser au mode d'action de la **sélection naturelle**. Les données moléculaires ne se prêtent que très mal au paradigme Darwinien de l'adaptation comme moteur de l'évolution. De nombreuses régions des génomes sont non-fonctionnelles, et donc ne sont pas impliquées dans l'adaptation. Par ailleurs, les mutations qui se produisent dans les régions fonctionnelles sont le plus souvent défavorables aux individus qui les reçoivent. La majorité des différences observables entre deux génomes correspondent donc en réalité à des changement **neutres**. Kimura a été le premier à réaliser ces propriétés spécifiques des données moléculaires, et les a "figées" dans sa fameuse théorie neutraliste (Kimura 1983). Du débat houleux qui a suivi (Gillespie 1991) il reste aujourd'hui que (i) la compréhension des forces sélectives agissant à l'échelle moléculaire, et notamment la détection d'événements adaptatifs, est la question principale de notre discipline, sous-jacente à nombre d'applications agronomiques ou médicales, et (ii) le modèle neutre reste l'hypothèse nulle de l'évolution moléculaire, qu'il faut rejeter avant de pouvoir discuter en termes adaptatifs les patrons de variations observés. Le troisième chapitre de cette thèse, qui résume certains des résultats biologiques que j'ai contribué à obtenir au cours de ces dernières années, se veut une illustration – partielle – de cette déclaration d'intention.

2. (Outils) Modèles et méthodes en génomique évolutive.

L'objectif de ce chapitre est d'introduire au non-initié les principales méthodes utilisées dans le domaine de la génomique évolutive. L'accent est mis sur les idées, plus que sur les équations. Deux grands types de méthodes sont présentées: les tests de détection de la sélection basées sur des échantillons populationnels, et les algorithmes de la phylogénie moléculaire, utilisant des données inter-spécifiques. Cette revue est précédée d'une brève introduction aux forces évolutives à l'origine des patrons de variation moléculaire entre individus et entre espèces.

2.1. Les mécanismes de l'évolution moléculaire

(population, mutation)

La variation génétique est générée par la **mutation**: à chaque génération, dans une **population**, un individu peut présenter, à une position donnée de son génome, un état différent de celui de ses parents; on dit alors qu'un nouvel **allèle** (état) apparaît, créant un polymorphisme au **locus** (position) correspondant. Le devenir de cet allèle est incertain. La

fréquence de l'allèle dans la population, initialement très faible, varie au cours du temps au grès du succès reproducteur de ceux qui le portent. Si le mutant initial disparaît sans se reproduire, la mutation est perdue. Ceci peut aussi se produire à la génération suivante, ou encore quelques générations plus tard. La mutation peut également échapper à la perte, et finalement envahir la population; on parle alors de **fixation** (fin de l'état polymorphe transitoire), et de **substitution** de l'allèle précédent par le nouvel allèle.

Les mutations que l'on rencontre fréquemment quand on compare des génomes sont des changements ponctuels de nucléotides A, C, G ou T, des insertions ou délétions, des élongations/raccourcissement de motifs répétés, des duplications de gènes ou de segments génomiques, des transpositions, des inversions, etc.… Ce manuscrit se concentre sur les mutations ponctuelles, qui représentent la plus grosse partie de la littérature de la génomique comparative. Parmi celles-ci, on distinguera notamment les mutations **non-codantes**, qui se produisent à l'extérieur des gènes, les mutations **synonymes**, qui modifient la séquence codante mais pas la protéine codée, et les mutations **non-synonymes**, qui modifient les protéines.

(dérive)

Les variations aléatoires de fréquence allélique liées au hasard de la reproduction sont appelées **dérive génétique**. L'intensité de la dérive est dépendante du nombre de reproducteurs dans la population, ou **taille efficace** (*Ne*). Plus la taille efficace est faible, plus les variations de fréquences alléliques sont fortes à chaque génération – on comprend qu'un allèle en fréquence 10% peut très bien être perdu par hasard à la génération suivante dans le cas d'une population de taille 10, mais pas pour une population de taille 10^6. La durée de vie moyenne d'une mutation neutre (avant fixation ou perte) est proportionnelle à la taille efficace de la population. Ceci explique pourquoi l'espérance du **taux de polymorphisme** (neutre) observable dans une population à l'équilibre est proportionnelle au produit du taux de mutation par la taille efficace: les grandes populations sont aussi les plus diverses. La taille efficace, en revanche, ne joue pas sur les **taux de substitution**, dans le cas de mutations neutres. La probabilité de fixation d'une mutation neutre est égale à sa fréquence, qui vaut initialement $1/Ne$ (dans le cas haploïde). Comme les mutations se produisent dans la population au taux $Ne.\mu$, où μ est le taux de mutation par individu et par génération, le taux de substitution neutre vaut $Ne.\mu /Ne$, c'est-à-dire qu'il est égal au taux de mutation quel que soit *Ne*. En grande population, les mutations (neutres) sont plus nombreuses mais se fixent moins fréquemment, et les deux effets se compensent.

(sélection)

La dérive génétique est (en première approximation) la seule force qui s'applique aux locus **neutres**, c'est-à-dire dont les variations n'ont aucune influence sur le succès reproducteur des individus. Si une mutation confère un avantage ou un désavantage à ses porteurs, son avenir devient alors également sous l'influence de la **sélection naturelle**. Contrairement à la dérive, la sélection est une force directionnelle qui à tendance à "tirer" la fréquence allélique vers zéro (mutations défavorables) ou un (mutations favorables, figure 1). La probabilité de fixation est bien-sûr affectée par la sélection: elle est plus forte [faible] pour les mutations favorables [défavorables] que pour les neutres. Ceci explique que la sélection influence le **taux de substitution**, et donc la divergence entre espèces: un locus sous sélection purificatrice (auquel la sélection élimine des mutations défavorables) évolue moins vite qu'un locus neutre, alors qu'un régime adaptatif (fixation récurrente de mutations avantageuses) augmente le taux de substitution (figure 1). La variabilité des régimes sélectifs entre gènes, entre sites, et dans le temps fait de la modélisation des taux d'évolution un des problèmes principaux de la phylogénie moléculaire.

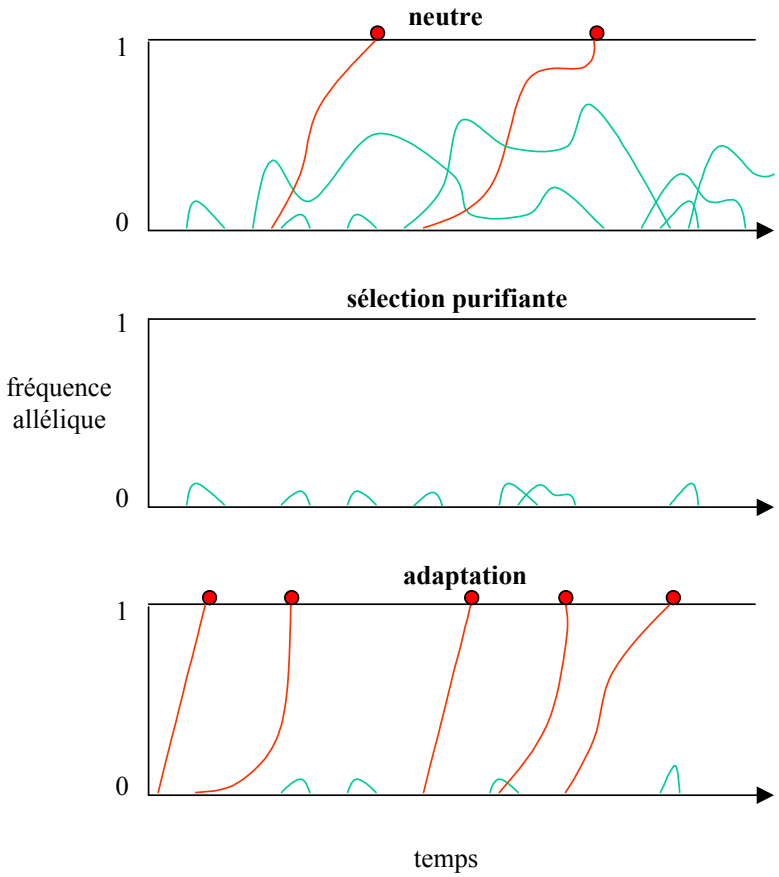


Figure 1. Devenir des mutations sous sélection directionnelle.
Chaque trajectoire correspond à une nouvelle mutation, qui apparaît en fréquence très faible 1/2*Ne*, et qui évolue au cours du temps. Les mutations peuvent être perdues (trajectoires vertes) plus ou moins rapidement, ou atteindre la fixation et réaliser ainsi une substitution (trajectoires rouges). Par rapport au cas neutre, la sélection négative = purifiante diminue le taux de substitution et favorise les fréquences alléliques faibles. La sélection positive = adaptation augmente au contraire le taux de substitution, et augmente la densité des fréquences fortes.

Par ailleurs, la sélection a également un effet sur les patrons de **polymorphisme** intra-spécifique. En favorisant la fixation/perte des variants avantageux/désavantageux, la sélection directionnelle tend à diminuer la durée de vie des mutations, et donc le taux de polymorphisme. La distribution attendue des fréquences des allèles (sachant qu'ils ségrègent) est également modifiée, par rapport au cas neutre: les allèles défavorables sont, en moyenne, en fréquence plus faible, et les favorables en fréquence plus forte. D'une manière générale, la sélection directionnelle augmente la probabilité des fréquences alléliques extrêmes. D'autres formes de sélection, telle que la sélection fréquence-dépendante ou la super-dominance, favorisent au contraire le polymorphisme, en retardant la fixation/perte d'allèle. Nous n'évoquerons que très peu ce type de sélection, qui semble spécifique à un petit nombre de locus dans les génomes.

Il faut noter que les effets de la sélection sur les patrons de variation dépendent de la **force** de l'avantage (désavantage) sélectif associé aux mutations, le point de référence étant, là encore, lié à la taille des populations. Si la valeur absolue du coefficient de sélection (s, où $1+s$ est le ratio du succès reproducteur mutant/sauvage) est très supérieure à $1/Ne$, alors la sélection l'emporte sur la dérive, et l'avenir de la mutation ne dépend plus de la taille efficace: elle sera perdue à coup sûr (pour une délétère), ou se fixera avec probabilité $\sim 2s$ (pour une avantageuse, Haldane 1927). Si s est très inférieur à $1/Ne$, alors la dérive l'emporte et tout se passe comme pour un locus neutre. Pour des coefficients de sélection de l'ordre de $1/Ne$, le devenir des allèles est influencé à la fois par la sélection et la dérive – on invoque typiquement ce type de sélection faible dans le cas de l'optimisation de l'usage des codons, par exemple. Les conséquences d'un effet sélectif d'intensité donnée varient donc, en théorie, selon la taille efficace de la population (de l'espèce) considérée.

(auto-stop, recombinaison)

La sélection naturelle influence directement, comme décrit ci-dessus, le devenir des mutations qui modifient la fitness des individus. Les effets visibles de la sélection, toutefois, ne s'arrêtent pas là. Les variations de fréquences d'allèles se propagent en effet mécaniquement aux locus avoisinants, en **déséquilibre de liaison** avec le locus sous sélection. Ce phénomène, couramment appelé "**auto-stop génétique"** (genetic hitch-hiking), est bien illustré par l'exemple d'une substitution après mutation avantageuse, aussi appelée **balayage sélectif** (Maynard-Smith & Haig 1974, figure 2). Un balayage élimine toute la variation pré-existante au locus considéré, mais également aux locus neutres en liaison. Les événements de **recombinaison** qui se produisent au cours du balayage limitent spatialement sa portée (Barton 2000). Ceci implique qu'un balayage sélectif se produisant sur un chromosome non-recombinant, tel que le chromosome Y, entraîne une diminution du polymorphisme sur toute sa longueur. La sélection purificatrice a également des effets qui se propagent par liaison génétique, un phénomène appelé **sélection d'arrière plan** ("background selection", Charlesworth et al. 1993). La sélection d'arrière plan, comme les balayages sélectifs, conduit à une diminution locale du taux de polymorphisme. Cette relation liaison/sélection est invoquée pour interpréter la corrélation entre taux de recombinaison et taux de polymorphisme: les régions d'un génome recombinant peu sont, en moyenne, moins polymorphes que celles qui recombinent beaucoup (Begun & Aquadro 1992). La liaison génétique joue aussi sur l'interférence entre plusieurs locus sélectionnés: la sélection naturelle multi-locus est moins efficace quand les locus sont liés (Barton 1995), un effet connu sous le nom d'**interférences Hill-Robertson**.

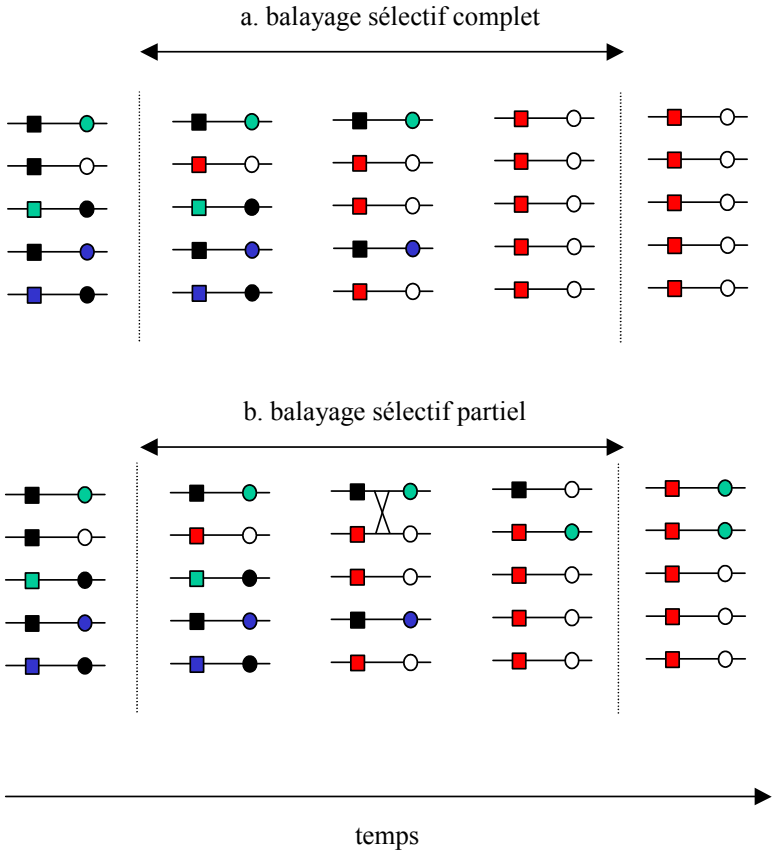


Figure 2. Balayage sélectif et auto-stop génétique.
Une mutation avantageuse (allèle rouge) apparaît au locus carré. Cet allèle augmente en fréquence jusqu'à fixation dans la population, éliminant tout le polymorphisme pré-existant: on parle de balayage sélectif. Ce balayage entraine par auto-stop génétique la fixation de l'allèle blanc ("chanceux") au locus neutre rond, en liaison avec le locus sélectionné (a). Si des événements de recombinaison se produisent au cours du balayage, d'autres allèles du locus rond (allèle vert) peuvent se retrouver associés à l'allèle avantageux et "bénéficier" du balayage (b). Dans les deux cas, le balayage a entrainé une baisse du niveau de polymorphisme au locus sélectionné et au locus neutre en liaison, l'intensité de cette baisse dépendant du rapport entre coefficient de sélection et taux de recombinaison.

(démographie, structure)

Nous avons considéré à ce stade une population panmictique caractérisée par une taille efficace constante dans le temps. En réalité, les populations sont fréquemment structurées (écart à la panmixie), et ont des effectifs variables au cours du temps. Ces facteurs **démographiques** influencent les patrons de variations moléculaire à l'échelle intra-spécifique. Un goulot d'étranglement, par exemple, a des conséquences comparables à celui d'un balayage sélectif (Galtier et al. 2000), c'est-à-dire une réduction brutale de la variabilité. La différence est que les effets démographiques s'appliquent à l'ensemble du génome, tandis que les effets sélectifs sont localisés (voir ci-dessus). Une approche multi-locus est donc souvent nécessaire pour discriminer les deux sources d'écart au modèle nul neutre-panmictique-Ne constant.

2.2. Tests de neutralité et théorie de la coalescence

Muni de ce bagage théorique minimal, le génomicien des populations essaye d'extraire des données de polymorphisme de séquence l'information concernant l'histoire des populations, mais aussi (et surtout) les effets de la sélection naturelle. Les données typiques sont donc un **échantillon** de séquences obtenues à un ou plusieurs locus pour un certain nombre d'individus prélevés dans les populations naturelles. La question posée est typiquement: mon locus se comporte-t-il comme un locus neutre, ou doit-on invoquer des effets sélectifs pour expliquer les patrons de variation? Si oui, de quel type de sélection s'agit-il: adaptation, ou sélection purificatrice (ou autre)? Pour permettre de répondre à cette question, un arsenal de méthodes basées sur les prédictions des modèles de génétique des populations a été développé.

(polymorphisme/divergence)

Une première idée assez naturelle consiste à mesurer le taux de polymorphisme de notre (nos) locus d'intérêt. La sélection directionnelle, en effet, a pour principal effet une diminution de la diversité, par rapport à un locus neutre, ce qui fait qu'un locus peu polymorphe est candidat à être sous sélection. D'autres facteurs que la sélection, ceci dit, influencent le taux de polymorphisme. Des variations de taux de mutation, en particulier, peuvent expliquer des variations de taux de polymorphisme sous un modèle neutre. Le **test HKA** (Hudson et al. 1987) a été imaginé pour contrôler ces facteurs confondants (figure 3). Ce test implique de disposer d'au moins deux locus, avec pour chacun d'eux un représentant d'une espèce distincte de l'espèce focale, appelée groupe externe. Sous l'hypothèse de neutralité, la divergence avec le groupe externe mesure le taux de mutation de chaque locus. Cette information est utilisée pour apprécier la significativité des différences de taux de polymorphisme entre loci (Hudson et al. 1987). Une approche comparable a été développée par Schlotterer (Schlotterer 2002) pour des données de micro-satellites, en utilisant cette fois-ci le taux de polymorphisme dans plusieurs populations apparentées, plutôt que la divergence avec un groupe externe, comme témoin du taux de mutation locus-spécifique.

Une autre approche, réservée aux séquences codantes, consiste à mesurer le taux de mutation au travers des changements synonymes, supposés neutres, pour comprendre les effets sélectifs associés aux changements non-synonymes. Le test **McDonald-Kreitman** (McDonald & Kreitman 1991) compile le nombre de polymorphismes synonymes et non-synonymes (intra-spécifiques), et le nombre de divergences synonymes et non-synonymes (inter-spécifiques), puis applique le principe que la sélection purificatrice devrait diminuer la proportion de non-synonymes qui se fixent, tandis que l'adaptation devrait augmenter cette proportion (figure 4).

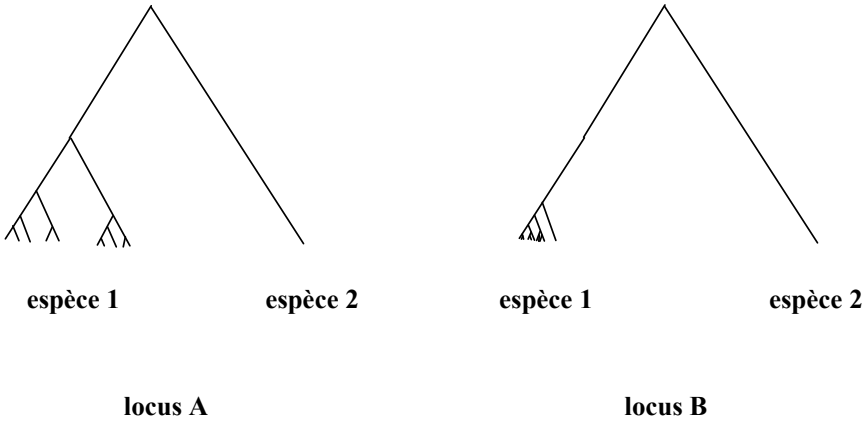
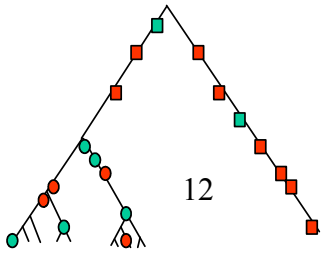


Figure 3. Le test HKA.
Le locus B est moins polymorphe que le locus A dans l'espèce 1, et l'examen des patrons de divergence nous indique que c'est la sélection naturelle qui est à l'origine de cette différence. En effet, une différence de taux de mutation entre locus (sous le modèle neutre) devrait se traduire par un niveau de divergence entre l'espèce 1 et l'espèce 2 plus élevé au locus A, ce qui n'est pas le cas. La formalisation mathématique du test est disponible dans Hudson et al (1987).



espèce focale groupe externe

Cette approche a été étendue par A. Eyre-Walker et ses collaborateurs (Piganeau & Eyre-Walker 2003) pour estimer les taux de substitution de mutations délétères et avantageuses chez divers organismes, ces auteurs rapportant notamment un taux remarquablement élevé de substitutions avantageuses chez la Drosophile (Bierne & Eyre-Walker 2004).

Les méthodes ci-dessus ont pour principe de **confronter patrons de polymorphisme et patrons de divergence,** celui-ci étant censé refléter celui-là sous le modèle neutre où toutes les mutations ont la même probabilité de fixation. Ce point de vue est toutefois limité par la nécessité de disposer de séquences d'un groupe externe suffisamment proche pour que la comparaison soit valide. Le patron de divergence peut s'éloigner du patron de polymorphisme pour tout un tas de raisons liées à l'histoire "ancienne" des espèces, compliquant l'interprétation en cas de rejet de l'hypothèse nulle. Pour éviter cet écueil, de nombreuses méthodes de détection de la sélection ne faisant pas référence à un groupe externe ont été développées.

(coalescence)

Ces tests se basent sur le fait que les patrons attendus de polymorphisme de séquence sous l'hypothèse nulle de neutralité sont bien connus, et décrits de manière synthétique par la **théorie de la coalescence** (encadré 1). Cette théorie prend le partie d'envisager les données de génétique des populations du point de vue **généalogique**, dans le sens inverse du temps. Pour un fragment de génome non-recombinant (=locus) donné, deux "individus" de notre échantillon (on parlera plutôt de "gène", pour désigner l'information portée par un chromosome donné d'un individu donné à un locus donné) ont nécessairement un ancêtre commun dont ils ont dérivé par transmission du gène générations après génération. On peut donc typiquement demander quelle est la probabilité pour que cet ancêtre ait existé à la génération précédente, ou il y a k générations, ou encore la probabilité pour que deux gènes parmi les n de notre échantillon aient un ancêtre commun à la génération précédente, ou il y a k générations. Le fait de posséder un ancêtre commun signifie que deux lignées de la généalogie (prise à rebours) fusionnent en une seule; on appelle cela un événement de **coalescence**. La coalescence est un processus markovien dont les propriétés ont été analysées en détail par divers auteurs (revue dans Rosenberg & Nordborg 2002), fournissant des prédictions sur la **forme** que devraient avoir les généalogies de gènes sous tel ou tel modèle, prédictions auxquelles on peut ensuite confronter les données réelles. Une force de la théorie de la coalescence est qu'elle permet de **simuler** très rapidement des données de polymorphisme: la complexité des algorithmes est typiquement en o(n) (taille de l'échantillon), alors que des simulations dans le sens direct du temps sont en o(Ne^2) (carré de la taille de la population).

(D de Tajima et consort)

Le **coalescent standard** est la distribution attendue des généalogies (topologie, temps de coalescence) sous l'hypothèse de neutralité, panmixie (pas de structure des populations) et taille efficace constante. Les propriétés du coalescent standard sont décrites figure 5. Lorsque la sélection agit, les généalogies attendues sont distordues, et présentent des formes plus ou moins spécifiques de tel ou tel type de sélection, et plus ou moins distinctes du coalescent standard. Pour détecter de tels écarts, qu'on interprétera éventuellement comme révélant des effets sélectifs, diverses statistiques-résumés ont été proposées. Ces statistiques décrivent (sommairement) un aspect des données, et leur valeur observée est comparée à la distribution attendue sous le coalescent standard, typiquement obtenue par simulation.

Encadré 1: les bases de la théorie de la coalescence.

Considérons une population diploïde **panmictique** de **taille efficace** (**constante** dans le temps) $2Ne$ dans laquelle nous avons échantillonné n gènes à un locus **neutre** (par exemple en séquençant les deux allèles de $n/2$ individus à ce locus). La théorie de la coalescence s'intéresse aux propriétés de la **généalogie** des gènes de l'échantillon, et ceci en considérant le processus dans le sens inverse du temps. Partant des n gènes actuels et remontant dans le temps, la généalogie se décrit comme la succession des événements de **coalescence**, c'est-à-dire de réunion de deux lignées en une lignée unique, lorsqu'un **ancêtre commun** est rencontré.
Formellement, considérons tout d'abord le cas où n=2 gènes, et demandons-nous quelle est la probabilité que les deux gènes de notre échantillon aient un ancêtre commun à la génération précédente (c'est-à-dire soient des "frères"). Les parents étant en nombre $2Ne$ et équiprobables (neutralité), cette probabilité vaut:

$$\text{Proba\_coal}(2, 2Ne) = 1/2Ne$$

Passons maintenant à un échantillon de taille n quelconque. La probabilité que deux gènes (au moins) parmi les n coalescent à la génération précédente vaut:

$$\text{Proba\_coal}(n, 2Ne) = 1- (1-1/2Ne)^{n(n-1)/2}$$

Cette expression s'interprète de la façon suivante: $n(n-1)/2$ est le nombre de paires de gènes distinctes dans l'échantillon, $1-1/2Ne$ est la probabilité, pour une paire, de ne pas coalescer, donc $(1-1/2Ne)^{n(n-1)/2}$ est la probabilité qu'aucune paire ne coalesce, et son complémentaire à 1 la probabilité qu'une paire au moins coalesce. Lorsque $1/Ne \ll 1$, cette expression s'approxime par:

$$\text{Proba\_coal}(n, 2Ne) \approx n(n-1)/4Ne$$

Lorsque $n \ll Ne$, la proabilité de coalescences multiples à la même génération est négligeable. Le processus de coalescence peut alors être décrit comme suivant, à chaque génération, une loi de Bernoulli de probabilité Proba_coal(k, $2Ne$), où k est le nombre courant de lignées (d'ancêtres) non-encore coalescées. En passant en temps continu, ce processus devient un processus de **Poisson** de taux $k(k-1)/4Ne$, ce qui signifie que les événements de coalescence sont de plus en plus rares au fur et à mesure que l'on remonte dans le temps (et que k diminue).

Ce processus poissonien implique que le temps d'attente avant la prochaine coalescence suit une distribution **exponentielle** d'espérance $4Ne/k(k-1)$, et la nature markovienne du processus assure **l'indépendance** des temps de coalescence. Le temps moyen passé à k=2 lignées vaut donc $2Ne$, et le temps de coalescence total de l'échantillon, $4Ne\ \Sigma_{2,n}\ 1/\ k(k-1)$, temps vers $4Ne$ quand n devient grand (figure 5).

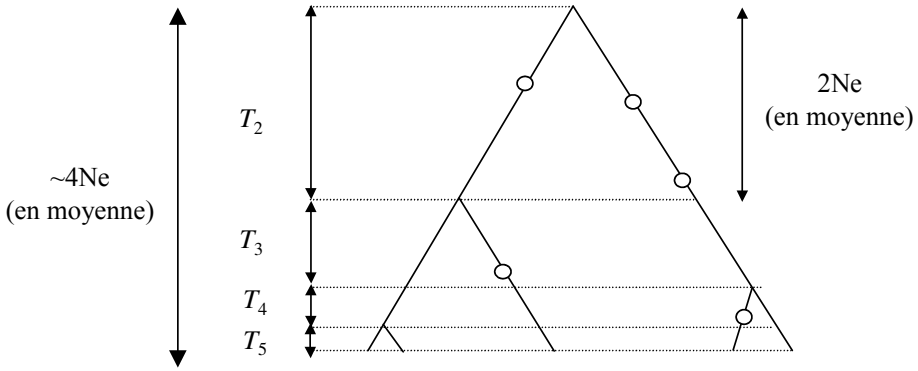


Figure 5: le coalescent standard (pour n = 5 gènes).
La distribution attendue de la généalogie des gènes échantillonnés dans le cas neutre, panmictique et à taille efficace constante obéit a une loi simple:
- toutes les topologies étiquetées sont équiprobables
- les temps de coalescence successifs T_i suivent des lois exponentielles indépendantes d'espérances $2Ne/i.(i-1)$

Une conséquence est que le temps moyen passé à deux lignées (dernière coalescence) est égal à $2Ne$, soit environ la moitié du temps de coalescence total moyen (voir encadré 1).

Parmi celles-ci, citons le célèbre ***D* de Tajima** (Tajima 1989), qui est défini comme la différence standardisée de deux estimateurs du taux de polymorphisme censés être égaux sous l'hypothèse nulle, le ***F* de Fu et Li** (Fu & Li 1993), qui détecte un excès de singletons (sites polymorphes en fréquence absolue 1), ou encore le ***H* de Depaulis et Veuille** (Depaulis & Veuille 1998) et le ***H* de Fay et Wu** (Fay & Wu 2000), développés pour détecter les balayages sélectifs.

Ces tests basés sur des statistiques-résumés atteignent rapidement leurs limites. Tout d'abord, leur interprétation n'est pas immédiate. L'hypothèse nulle est en effet composite (neutralité, mais aussi panmixie et équilibre mutation/dérive), ce qui fait qu'un écart peut être expliqué par de nombreux phénomènes distincts (sélection, mais aussi structure des populations ou histoire démographique). Par ailleurs, cet tests ne sont pas adaptés aux données multi-locus, pour lesquelles on se trouve confronté au délicat problème des tests multiples. A l'heure de la génomique des populations, on utilise d'ailleurs volontiers le *D* de Tajima, par exemple, à titre exploratoire, comme une statistique descriptive (par exemple Nordorg et al. 2005).

(vraisemblance, bayésien, échelle génomique)

L'alternative à ce type de tests est la **modélisation** des processus de l'évolution génétique, suivie de l'estimation des paramètres par l'ajustement du modèle aux données. L'ajustement au maximum de vraisemblance (Griffith & Tavaré 1995, Kuhner et al. 2000) ou bayésien (Beaumont & Rannala 2004) de modèles de génétique des populations est un objectif techniquement difficile, qui fait actuellement l'objet de développements nombreux. Modéliser la sélection n'est malheureusement pas facile: une propriété fondamentale du coalescent standard, l'indépendance entre probabilité de coalescence et état allélique, n'est valide que sous le modèle neutre. Sans cette hypothèse, raisonner en termes généalogiques relève du tour de force (Neuhauser & Krone 1997), et calculer des vraisemblances est un objectif hors de portée. Pour contourner cette difficulté, il est possible (i) d'approximer les effets sélectifs par des effets démographiques qui auraient des conséquences semblables sur les patrons de variation, et (ii) de faire l'hypothèse que les données correspondent à des variations neutres se produisant à des locus en liaison avec les locus sous sélection (Coop & Griffiths 2004) – une hypothèse probablement réaliste pour de nombreux jeux de données. J'ai proposé une telle approche pour détecter les balayages sélectifs en liaison, en utilisant la quasi-correspondance entre balayages et goulots d'étranglement, qui ont des effets très similaires sur la forme des généalogies à un locus donné (Galtier et al. 2000).

Les développements méthodologiques évoqués ci-dessus sont actuellement mis sur la sellette par la **dimension génomique** que prend la génétique des populations. Pour un nombre encore petit, mais croissant, d'espèces, le nombre de locus analysés se compte en centaines, interdisant une analyse "classique" locus après locus. La modélisation multi-locus étant, comme indiqué plus haut, encore balbutiante et très coûteuse en temps de calcul, les publications récentes ont essentiellement adapté les méthodes standard pour extraire au mieux l'information pertinente des données. Il me semble qu'il y a ici la place pour des améliorations méthodologiques d'importance dans les années à venir, les données ayant, peut-être pour la première fois dans l'histoire de la génétique des populations, apparemment dépassé les méthodes. L'objectif principal d'une analyse de génomique des populations devrait être d'extraire la composante gène-spécifique, qui reflète les effets sélectifs, du patron de variation "collectif" à l'ensemble des gènes, qui traduit l'histoire démographique et migratoire des populations – et ceci en respectant des temps de calcul raisonnables malgré la taille des jeux de données. Lorsque cet objectif sera atteint, l'écueil suivant sera peut-être celui de la liaison génétique entre les locus échantillonnés: la densité des marqueurs s'intensifiant, on ne pourra plus faire l'hypothèse que leurs généalogies sont indépendantes, et la vie sera alors bien compliquée pour les statisticiens de la génétique des populations.

2.3. Phylogénie moléculaire et modèles markoviens

(phylogénèse)

Les considérations ci-dessus concernent les variations entre individus d'une même espèces, apparues relativement récemment, et pas encore fixées dans la population. L'autre échelle de temps pertinente pour comprendre l'évolution moléculaire est celle des **divergences entre espèces**. Les différentes espèces que le monde vivant a connues sont apparues par **spéciations** successives, et ont pour la plupart, mais pas toutes, disparu par **extinction**. Les espèces actuelles peuvent donc conceptuellement être reliées par un **arbre** dit **phylogénétique**, qui récapitule leurs liens de parentés, ou encore leur histoire de spéciation, comme décrit par Darwin il y a deux siècles. Les molécules que portent ces espèces ont elles-aussi évolué (en première approximation) le long de cette phylogénie, accumulant des substitutions qui se traduisent aujourd'hui par des différences observables entre génomes. L'examen de ces différences permet d'espérer remplir un double objectif: celui de **reconstruire la phylogénie** du vivant, qui nous est *a priori* inconnue, et celui de **caractériser les processus** de l'évolution moléculaire, et notamment le rôle joué par la sélection naturelle.

(l'approche statistique)

La bonne approche pour atteindre ces objectifs, c'est la **modélisation** de l'évolution des séquences. Une vingtaine d'années de débats parfois vifs (Sober 2004) ont permis de convaincre la communauté dans son ensemble ou presque, et malgré quelques soubresauts (Kolaczkowski & Thornton 2004), de la supériorité de l'approche statistique (maximum de vraisemblance, approche bayésienne) sur des méthodes plus intuitives (maximum de parcimonie) pour reconstruire des phylogénies moléculaires (voir par exemple Galtier 1997). Quant à l'objectif de caractériser les processus, il implique de lui-même l'usage de ces méthodes. En effet, le vrai débat de fond – ou du moins, le seul qui me semble digne d'intérêt – entre cladistes (qui soutiennent la parcimonie maximale) et tenants d'une approche statistique concerne l'existence d'un processus évolutif, c'est-à-dire d'une relative prédictibilité des changements évolutifs: connaissant les patrons de variations aux sites 1 à 99 d'une molécule, ai-je gagné de l'information à propos de l'évolution du 100^{e} site? Pour ce qui est des données moléculaires, la réponse à cette question me semble être très généralement affirmative: les quantités d'évolution, mesurées en nombre de substitution ou **longueurs de branches**, sont évidemment dépendantes du temps écoulé, et ceci de manière commune à l'ensemble des sites; de plus, les **processus mutationnels** sont essentiellement partagés par tous les sites d'une molécule, ce qui se traduit typiquement par des compositions en bases (pourcentage de A, C, G et T) gène- ou génome-spécifiques. Ces caractéristiques évolutives "globales" ont suffisamment d'impact sur les patrons de variations entre génomes pour mériter le nom de processus, et la modélisation formelle. Il faut noter que la sélection naturelle, en revanche, influence l'évolution des molécules de manière bien souvent imprédictible, créant des spécificités de vitesse/patron évolutif à l'échelle du site, de la lignée, ou de l'interaction site/lignée. Comprendre les processus sélectifs par la phylogénie moléculaire, ce qui est notre objectif principal, implique donc l'usage de modèles complexes (ratio des taux non-synonymes sur synonymes variable entre site, covarions, coévolution), où l'on atteint parfois les limites de l'approche statistique en raison de l'inflation du nombre de paramètres (par exemple Yang & Nielsen 2002).

Encadré 2. Modèles markoviens pour l'évolution des séquences.

On modélise l'évolution d'une séquence d'ADN ou de potéine en faisant l'hypothèse que les différents sites évoluent tous indépendamment selon un processus de **Markov** qui leur est commun. Le temps est supposé continu, et le processus prend ses états dans un espace discret qui est typiquement celui des nucléotides {A, C, G, T}, des amino-acides ou des codons.

La dynamique **instantanée** du processus est décrite par les équations différentielles telles que (pour des nucléotides):

$A(t+dt)=A(t) - A(t)\ (m_{AC}+m_{AG}+m_{AT})dt + C(t)m_{CA}dt + G(t)\ m_{GA}dt + T(t)\ m_{TA}dt$

Dans cette expression, $X(t)$ est la probabilité d'être dans l'état X au temps t, et m_{XY} est le taux instantané de changement de X vers Y. L'équation ci-dessus peut également être écrite pour les trois autres états (C, G et T). En appelant $\mathbf{F}(t)$ le vecteur vertical $(A(t), C(t), G(t), T(t))t$ et $\mathbf{M}$ la matrice des m_{XY} (les termes diagonaux étant tels que les colonnes somment à zéro), ce système d'équations différentielles s'écrit économiquement comme:

$\mathbf{F}(t+dt)=\mathbf{F}(t)+\mathbf{M}.\mathbf{F}(t).dt$

dont la résolution (classique) donne:

$\mathbf{F}(t)=e^{\mathbf{M}.t}.\mathbf{F}(0)$

Cette dernière expression introduit l'exponentielle d'une matrice, qui se définit par le développement limité $e^{\mathbf{M}}=\mathbf{I}+\mathbf{M}+\mathbf{M}^2/2!+...$, et qui se calcule facilement pour des matrices diagonalisables, ce qui est le cas des matrices de transition utilisées ici.

La matrice $\mathbf{P}(t)=e\mathbf{M}t$ fournit donc les probabilités $p_{XY}(t)$ d'atteindre l'état Y après évolution selon **M** pendant un temps t à partir de l'état X. Ces probabilités de transition de long terme permettent la **simulation** de données, mais aussi le calcul de **vraisemblance** en phylogénie moléculaire (cf. encadré 3).

Le générateur **M** est typiquement paramétré via des hypothèses concernant les valeurs des 12 m_{XY} ($X \neq Y$). Le modèle le plus simple fait l'hypothèse que tous ces taux sont égaux. Le modèle le plus général autorise les 12 taux à différer. Les modèles les plus populaires présentent un nombre de degrés de libertés intermédiaires; c'est le cas du modèle de Tamura & Nei (1992, repris par Galtier & Gouy 1998), qui contrôle le pourcentage de G+C d'équilibre et le rapport entre taux de transition et taux de transversion.

(modèles markoviens)

Voyons maintenant comment ces modèles markoviens sont utilisés en pratique en phylogénie moléculaire. Le jeu de données typique, on l'aura compris, est constitué d'un ensemble de *n* séquences alignées sans gap de longueur *p*, qui définissent donc *p* sites. Ces séquences sont supposées avoir évolué à partir d'une séquence ancestrale par l'accumulation de changements qui se sont produits le long des branches d'un arbre phylogénétique. Les **paramètres** du modèle de base seront donc (i) la topologie de l'arbre (ordre de branchement), (ii) les longueurs de branche, qui représentent l'aspect quantitatif de l'évolution, et (iii) une matrice de transition, de dimension 4x4 pour l'ADN, explicitant les taux relatifs des différents types de changements (figure 6). Cette matrice est appelée le **générateur** du processus markovien. Les différents sites de la molécule sont supposés évoluer indépendamment et selon le même processus.

Les mathématiques associés à un tel processus sont connus de longues dates, et impliquent simplement d'écrire les équations différentielles décrivant la dynamique instantanée des probabilités des quatre états, comme indiqué dans l'encadré 2. Ces équations se résolvent de manière formelle pour aboutir à la dynamique à long terme, c'est-à-dire la probabilité d'atteindre l'état *j* après une durée d'évolution *t* sous le processus **M** sachant l'état initial *i*. Ces probabilités permettent, par exemple, de **simuler** l'évolution de séquences fictives le long d'un arbre donné, avec longueurs de branches et processus donnés.

(vraisemblance)

Mais l'objectif principal est bien-sûr celui de **l'estimation** des paramètres inconnus du modèle par **ajustement** aux données réelles. Ceci implique le calcul de la **vraisemblance**, qui est définie comme étant la probabilité du jeu de données sachant les paramètres du modèle. La vraisemblance est une quantité relativement peu intuitive, qui mesure la chance (infime) que l'on a eue d'obtenir les données dont on dispose, et non pas un autre jeu de données. La vraisemblance est ensuite typiquement maximisée sur l'espace des paramètres: on cherche à déterminer le jeu de paramètres qui rendent les données le plus plausible.

Le calcul de vraisemblance en phylogénie moléculaire a été introduit dès 1981 par Joe Felsenstein (Felsenstein 1981), qui a exhibé un **algorithme récursif** impliquant un seul passage sur l'arbre, et donc de complexité o(*n*) par site, malgré le nombre exponentiel de combinaisons d'états ancestraux possibles, qui sont toutes formellement prises en compte (cf. encadré 3). L'hypothèse d'indépendance est ici critique en ce qu'elle permet un traitement séparé des différents sites, dont les probabilités sont ensuite multipliées. La probabilité d'un site est obtenue en sommant (intelligemment, donc) les contributions de tous les scénarios possibles, un scénario étant défini comme un ensemble d'états ancestraux aux nœuds internes de l'arbre.

Les éléments ci-dessus permettent de calculer la vraisemblance d'un jeu de paramètres donné. L'estimation au maximum de vraisemblance implique, comme son nom l'indique, de maximiser cette fonction sur l'espace des paramètres. Il s'agit d'un problème d'**optimisation** sur un espace mixte, certaines dimensions étant discrètes (topologies) et d'autres continues (longueurs de branches et autres paramètres numériques). C'est l'exploration des topologies qui pose des problèmes algorithmiques difficiles, sur lesquels des avancées récentes ont permis d'accélérer considérablement les programmes de maximum de vraisemblance en phylogénie (Guindon & Gascuel 2003, Stamatakis et al. 2005).

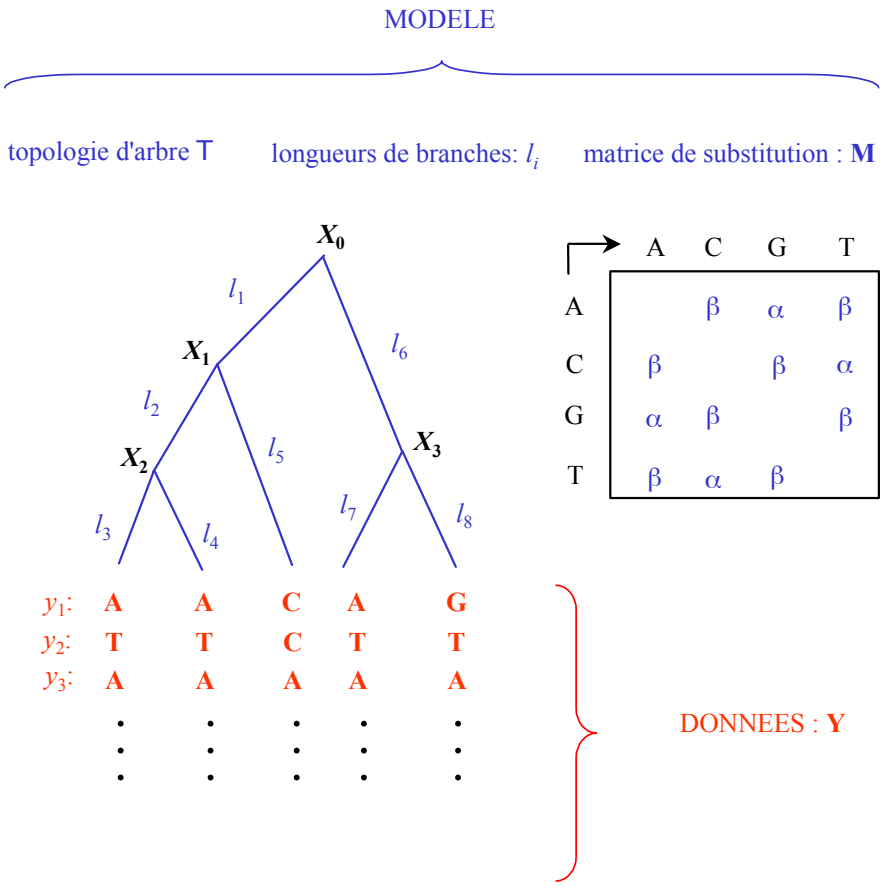


Figure 6: modélisation markovienne en phylogénie moléculaire.
Les séquences (données) sont présentées en colonnes. Chaque site (y_i) est supposé avoir évolué à partir d'un état ancestral X_0 le long des branches de l'arbre T selon le processus décrit par le générateur **M** (ici celui proposé par Kimura 1980). Les X_i sont des variables aléatoires correspondant aux états ancestraux inconnus.

Encadré 3. Calcul de vraisemblance en phylogénie moléculaire.

Pour un modèle et un jeu de données D donnés, la **vraisemblance** d'un jeu de paramètres θ est définie comme la probabilité des données sachant les paramètres:

$L(\theta) = Pr(D/\ \theta)$

Cette fonction est à la base de nombreuses procédures d'estimation statistique, en particulier les approches au maximum de vraisemblance et bayésienne.

En phylogénie moléculaire, les données sont les sites $\mathbf{Y}=\{y_k\}$ supposés indépendants, et les paramètres du modèle sont la topologie de l'arbre T
, les longueurs de branches l_i et les paramètres de la matrice de transition **M** du processus sous-jacent (figure 6). La vraisemblance se calcule en utilisant tout d'abord l'hypothèse d'indépendance des sites:

$$L(\mathsf{T},\{l_i\},\mathbf{M}) = Pr(Y / \mathsf{T},\{l_i\},\mathbf{M}) = \prod_k Pr(y_k / \mathsf{T},\{l_i\},\mathbf{M})$$

Le problème se réduit donc au calcul de la vraisemblance pour un site. Considérons le site y_1 de la figure 6. On a:

$$Pr(y_1 / \mathsf{T},\{l_i\},\mathbf{M}) = \sum x_0 \sum x_1 \sum x_2 \sum x_3\ Pr(X_0=x_0)\ p_{x0x1}(l_1)\ p_{x1x2}(l_2)\ p_{x2A}(l_3)\ p_{x2A}(l_4)\ p_{x1C}(l_5)\ p_{x0x3}(l_6)\ p_{x3A}(l_7)\ p_{x3G}(l_8)$$

Cette équation somme les probabilités conférées par chacun des scénarios, un scénario correspondant à un ensemble d'états ancestraux possibles $\{x_0, x_1, x_2, x_3\}$. Les termes de la forme $p_{xy}(t)$ sont calculés à partir de la matrice de transition **M** comme expliqué dans l'encadré 2.

Felsenstein (1981) a montré que cette expression se simplifie grâce à la factorisation de nombreux termes:

$$Pr(y_1 / \mathsf{T},\{l_i\},\mathbf{M}) = \sum x_0\ \ Pr(X_0=x_0)\ \sum x_1\, p_{x0x1}(l_1)\ p_{x1C}(l_5)\ \sum x_2\, p_{x1x2}(l_2)\ p_{x2A}(l_3)\ p_{x2A}(l_4)\ \sum x_3\, p_{x0x3}(l_6)\ p_{x3A}(l_7)\ p_{x3G}(l_8)$$

ce qui permet de calculer la vraisemblance **récursivement** par un seul parcours de l'arbre, en un temps proportionnel au nombre de séquences comparées.

La fonction de vraisemblance à le mérite de rassembler toute **l'information** disponible dans les données concernant le modèle choisi pour les représenter. La méthode d'estimation au maximum de vraisemblance retourne fort logiquement le jeu de paramètres qui rendent les données les plus probables. On peut toutefois reprocher à cette approche de fournir avant tout des estimateurs **ponctuels**. Mesurer l'incertitude sur ces estimations n'est pas immédiat dans le contexte du maximum de vraisemblance (le fait que tel paramètre confère aux données une probabilité 10 fois inférieure à tel autre suffit-il à l'exclure?), même si des pratiques empiriques telles que le ré-échantillonnage par bootstrap (Felsenstein 1985) peuvent fournir à l'utilisateur une mesure raisonnable du niveau de confiance.

(bayésien)

C'est, me semble-t-il, entre autre pour traiter au mieux la question de l'incertitude que la philosophie **bayésienne** est apparue et s'est développée. Grâce au théorème de Bayes, et au prix d'**hypothèses *a priori*** sur la distribution des paramètres (indépendamment des données), le point de vue bayésien retourne la situation en définissant la probabilité (dite postérieure) des paramètres sachant les données, ce qui permet donc de définir aisément les intervalles **(zones) de confiance** autour des estimations. Au rang des avantages, cette approche permet également de gérer des modèles extrêmement complexes, incluant de très nombreux paramètres dont seulement certains nous intéressent. Les paramètres dits de nuisance, présents par construction mais dont nous ne souhaitons pas connaître la valeur, peuvent être facilement écartés par intégration sur leur dimension; on se focalise donc aisément sur la distribution postérieure **marginale** des paramètres d'intérêt.

Ces bénéfices ne vont pas sans un certain nombre de coûts, bien-entendu. Il y a d'abord le problème d'avoir à spécifier des hypothèses *a priori* sur la distribution des paramètres. C'est un exercice parfois naturel, parfois plus périlleux, et qui peut modifier significativement le résultat de l'analyse, ce qui est alors inquiétant. Le deuxième point noir est l'implémentation informatique. L'intégration bayésienne ne pouvant quasiment jamais être réalisée analytiquement, une procédure numérique introduite par Metropolis (Metropolis et al. 1953) et Hastings (1970), connue sous le nom de Chaînes de Markov Monte-Carlo (**MCMC**), est mise en œuvre. Il s'agit de parcourir l'espace des paramètres par essais-erreurs (certaines erreurs étant tolérées, à la manière du recuit-simulé) de sorte que les valeurs de paramètres rencontrées échantillonnent cet espace conformément à la distribution postérieure. Bien que fascinantes par la simplicité de leur principe, les MCMC demandent souvent en pratique beaucoup d'ingéniosité (et de travail) de la part du programmeur pour fournir un résultat convaincant, le choix des "mouvements" autorisés sur l'espace des paramètres étant crucial.

L'approche bayésienne a récemment envahi la logithèque internationale en phylogénie (par exemple Yang & Rannala 1997, Huelsenbeck & Ronquist 2001) et en génétique des populations (par exemple Pritchard et al 2000, Dawson & Belkhir 2001, Excoffier et al. 2005). Il faut toutefois bien avouer que le succès du programme MrBayse en phylogénie moléculaire, par exemple, a été plus vraisemblablement dû à sa (relative) rapidité d'exécution, par rapport aux programmes de maximum de vraisemblance disponibles à l'époque, qu'aux propriétés statistiques de la philosophie bayésienne. La mesure bayésienne de l'incertitude phylogénétique, qui pourrait être un point fort de l'approche, est par exemple très controversée (Douady et al. 2003). Les empiristes semblent d'ailleurs se retourner volontiers vers des méthodes au maximum de vraisemblance depuis que des progrès algorithmiques ont significativement accéléré les programmes implémentant ces dernières. L'approche bayésienne, malgré ses possibles lourdeurs, est toutefois quasi-irremplaçable (et fournit des résultats remarquables) lorsque les

modèles considérés sont très complexes, au point de rendre parfois le calcul de vraisemblance analytiquement impossible (par exemple Lartillot & Philippe 2004).

(diversité des modèles)

La méthode du maximum de vraisemblance, comme l'approche bayésienne, a été initialement développée – avec succès, comme discuté ci-dessus – comme une méthode de reconstruction d'arbres. Le modèle évolutif sous-jacent était choisi avec l'idée de représenter au mieux les données, mais l'objet de l'estimation était avant tout la topologie; cela reste d'ailleurs l'usage le plus fréquent des méthodes statistiques en phylogénie moléculaire. A partir de la fin des années 90, la littérature a toutefois également grossi d'analyses phylogénétiques au maximum de vraisemblance (ou bayésiennes) destinées à comprendre les processus de l'évolution moléculaire. On s'est donc penché plus avant sur le choix et la pertinence des modèles markoviens utilisés, avec l'idée de représenter/tester des hypothèses biologiques spécifiques. Ces développement sont par exemple passés en revue dans Galtier et al. 2005.

Le modèle standard présenté dans l'encadré 2 est défini pour les séquences d'ADN et implique de nombreuses hypothèses contraignantes. L'alphabet ADN à 4 lettres peut être étendu à 20 (protéines) ou 61 (codons), et les hypothèses de base peuvent (parfois) être relâchées, et donc testées sur des jeux de données spécifiques, sous la forme de "le gain de ce degré de liberté améliore-t-il suffisamment l'adéquation modèle/données pour mériter notre considération?" – c'est le principe sous-jacent aux tests de rapport de vraisemblance.

Parmi les exemples, citons les modèles **non-stationnaires** (Galtier & Gouy 1998), qui permettent de prendre en compte les variations de composition en bases entre espèces et d'estimer les compositions en bases ancestrales (Galtier et al. 1999), les modèles **codons** (Goldman & Yang 1994), qui permettent de détecter des gènes/sites/lignées sous un régime de sélection positive, caractérisés par un taux de substitution non-synonyme plus élevé que le taux synonyme, les modèles relâchant l'hypothèse **d'indépendance** des sites, par chaîne de Markov cachée (Felsenstein & Churchill 1996, Robinson et al. 2003) pour les sites voisins, ou entre paires de sites arbitraires (Pollock et al. 1999) pour détecter la **coévolution**, les tentatives de modélisation de l'écart à l'**horloge moléculaire** (différences de vitesse entre lignées, Thorne & Kishino 2002, Huelsenbeck et al. 2000), utilisées pour estimer des dates absolues de divergence à confronter aux données fossiles (Douzery et al. 2004), ou encore les modèles prenant en compte les variations de vitesse site-spécifiques (ou covarion ou hétérotachie, Tuffley & Steel 1997, Galtier 2001, Galtier & Jean-Marie 2004), qui permettent potentiellement de détecter des épisodes adaptatifs au cours de l'histoire d'une molécule. Le paragraphe 3.2 illustrera ce type d'approche, et son apport potentiel en évolution moléculaire.

3. (Applications) Etre ou ne pas être sous sélection…

Voyons maintenant un certain nombre d'applications, auxquelles je me suis consacré ces dernières années, des méthodes présentées au chapitre 2 aux questions évoquées au chapitre 1. Ces applications sont multiples et parfois déconnectées les unes des autres, reflétant les étapes de mon parcours scientifique, et les collaborations diverses que j'ai pu établir. Cela aura l'avantage, du point de vue pédagogique, d'illustrer diverses facettes de la recherche en évolution moléculaire. Le premier aspect, celui de l'évolution des isochores, relève typiquement de la génomique comparative: ayant séquencé son génome favori, le biologiste moléculaire cherche à en comprendre le sens au travers de la dimension évolutive. Le deuxième paragraphe est plus méthodologique: il approfondit et met à l'épreuve les modèles

et méthodes présentées au chapitre 2. Le troisième point, enfin, présente mon projet de recherche principal du moment, celui de l'évolution des organites cellulaires.

3.1. Isochores et biais de conversion génique

La **composition en bases** (pourcentage de A, C, G et T) des gènes et des génomes est un centre d'intérêt récurrent en évolution moléculaire. Une raison est que cette composition reflète le processus de substitution nucléotidique sous-jacent, lui-même déterminé par les forces évolutives qui régissent la variation moléculaire (mutation, réparation, recombinaison, sélection, dérive), que l'on cherche à caractériser.

La composition en bases des génomes de mammifères présente une structuration spatiale appelée "**isochores**": de grandes (10^5-10^6 nucléotides) régions **riches** et **pauvres** en GC alternent le long des chromosomes (Bernardi et al. 1995, IHGSC 2001). La question de l'origine évolutive et du rôle potentiel de cette structuration se pose avec acuité. En effet, la structure en isochores est corrélée avec d'autres caractéristiques importantes de l'organisation spatiale de ces génomes, comme la distribution des dinucléotides, des éléments répétés, et, remarquablement, la **densité en gènes**. On se demande donc si cette structure est adaptative, c'est-à-dire "utile" au fonctionnement du génome, ou résulte simplement de biais mutationnels variables spatialement, indépendamment de toute notion d'adaptation. Une théorie prometteuse proposait que les isochores constituaient une adaptation à l'homéothermie (Bernardi 1993), mais ceci a été réfuté par l'observation d'isochores chez plusieurs espèces de vertébrés à sang-froid (Hughes et al. 1999). Les données de polymorphisme de séquence, cependant, montrent un patron de ségrégation incompatible avec la neutralité, suggérant un avantage (mais lequel?) des allèles C ou G par rapport à A ou T dans les isochores riches (Eyre-Walker 1999, Smith & Eyre-Walker 2001, Lercher et al. 2002).

Cette revue bibliographique rapide révèle un paradoxe : les données de polymorphisme suggèrent l'existence d'une pression de sélection favorable aux allèles G et C, mais aucun scénario sélectif crédible n'a pu être mis en évidence. Laurent Duret et moi-même avons amorcé une réflexion sur cette apparente contradiction, réflexion qui nous a amenés à proposer que la **conversion génique biaisée** soit le (un des) mécanisme(s) moléculaire(s) à l'origine de la structure en isochores chez les mammifères. Il s'agit d'un biais dans la fixation des mutations lié à la recombinaison: l'hypothèse est qu'un hétérozygote A/G produira une proportion plus élevée de gamètes G que de gamètes A en raison d'événements de conversion de A par G au cours de la méiose (voir Marais 2003 pour revue). Ce mécanisme est **neutre** au sens où la fitness des individus est indépendante du contenu en GC de leur génome, mais prédit des patrons de polymorphisme présentent un **écart apparent à la neutralité**, les allèles GC ayant une probabilité de transmission (donc, ultimement, de fixation) supérieure aux allèles AT, conformément au message des données de polymorphisme humain. Le biais de conversion génique étant associé à la recombinaison, il induirait un taux de GC plus élevé dans les régions à fort taux de recombinaison que dans les régions à faible taux de recombinaison, d'où la structure en isochores.

Nos arguments en faveur de cette hypothèse résultent essentiellement d'analyses des données de la génomique accessibles via les bases de données publiques. Ces arguments sont l'existence d'un biais de réparation vers GC démontré expérimentalement, l'existence d'une corrélation entre taux de recombinaison et taux de GC chez les mammifères, et le taux de GC élevé des gènes connus pour avoir un taux de conversion génique élevé (Galtier et al. 2001). Une analyse fine d'une de ces familles de gènes, les histones, montre que les séquences présentes en copies nombreuses et presque identiques dans le génome (subissant fréquemment des

événements de conversion génique) ont un taux de GC significativement plus élevé que les copies plus ou moins uniques (qui ne font que peu ou pas de conversion), suggérant très fortement que le mécanisme de conversion génique est biaisé vers GC (Galtier 2003). L'évolution des chromosomes sexuels des mammifères reflète également l'existence de ce biais. Les chromosomes X et Y se sont différenciés à partir d'une paire autosomale ancestrale par épisodes successifs d'arrêt de la recombinaison (Lahn & Page 1999). Nous montrons par l'analyse du locus clé *amel* que cet arrêt de la recombinaison s'est accompagné d'une diminution du taux de GC des régions concernées, conformément à l'hypothèse d'un biais de conversion génique associé à la recombinaison (Marais & Galtier 2003). Nous montrons enfin que si ce biais est encore effectif, il est probablement moins fort que ce qu'il fut par le passé, puisque la structure en isochores est **en voie d'érosion** : le taux de G+C des régions riches en GC diminue chez les rongeurs, les primates et les artiodactyles, comme l'analyse de séquences codantes nous en a convaincus (Duret et al. 2002, Belle et al. 2004).

Nous proposons donc que la sélection naturelle n'est pas le déterminant de la structure en isochores des génomes de mammifères, ou du moins pas directement. Si sélection il y a, c'est sur les mécanismes de réparation de l'ADN qui contrôlent les biais de conversion. Reste la question de la corrélation entre la structure en isochores et la densité en gènes : pourquoi les régions riches en GC sont-elles beaucoup plus denses en gènes que les autres ?

Pour aborder cette question, nous avons mis en place, en collaboration avec P. Boursot et J. Montoya, une **approche expérimentale** exploitant l'histoire particulière du gène *fxy* chez la souris. Chez la souris domestique (*Mus musculus*), le gène *fxy* a été récemment transloqué d'une région du génome quelconque vers un point chaud de recombinaison, en l'occurrence la **région pseudoautosomale**. Suite à ce changement de localisation, le taux de GC du gène *fxy* a augmenté spectaculairement (Perry & Ashworth 1999), en accord avec l'hypothèse d'un biais vers GC associé à la recombinaison. Nous avons donc séquencé ce gène (introns et exons) chez plusieurs espèces de souris de manière à caractériser son mode d'évolution.

Après avoir clarifié l'histoire évolutive du gène, nous montrons que cette translocation a généré une réduction significative de la taille des introns, donc une augmentation de la densité en gènes (figure 7). D'autres caractéristiques des isochores riches en GC sont également apparues chez *fxy* suite à sa translocation, et notamment une augmentation de la différence entre le taux de GC des exons et des introns, et l'apparition de nombreux locus minisatellites. Il apparaît donc que le fragment d'ADN transloqué dans une région fortement recombinante a très rapidement mimé un isochore riche, suggérant fortement que la recombinaison, via le biais de conversion génique, est le déterminant premier de la structure en isochores (Montoya-Burgos et al 2003). La comparaison avec les données équivalentes chez l'homme (Yi et al 2004), chez qui l'histoire de la région pseudoautosomale est bien différente, est en accord avec ce scénario (Galtier 2004a).

Ces travaux soulèvent la question du rôle de la conversion génique biaisée en évolution moléculaire. Ce distorteur de ségrégation généralisé, négligé jusqu'alors, s'avère avoir des conséquences importantes sur la structure des génomes de mammifères. La question de sa **distribution taxonomique** et son rôle dans la genèse des patrons de variation moléculaire dans d'autres espèces se pose donc. Il faut rappeler que la conversion génique induit un biais de fixation des allèles, tout comme la sélection naturelle, suggérant que de nombreux patterns de variation interprétés comme de la sélection pourraient en fait "n'être que" de la conversion biaisée. Une étude récente en collaboration avec Nicolas Bierne suggère par exemple que la conversion biaisée vers GC influe sur la composition en bases du génome de la Drosophile, et interfère avec la sélection pour l'usage des codons (Galtier et al. 2006). L'autre question naturelle,

dépendante de la précédente, est celle de l'origine évolutive du biais de conversion génique: cette distorsion vers GC est-elle adaptative, et qu'est-ce qui détermine ses variations d'intensité entre espèces?

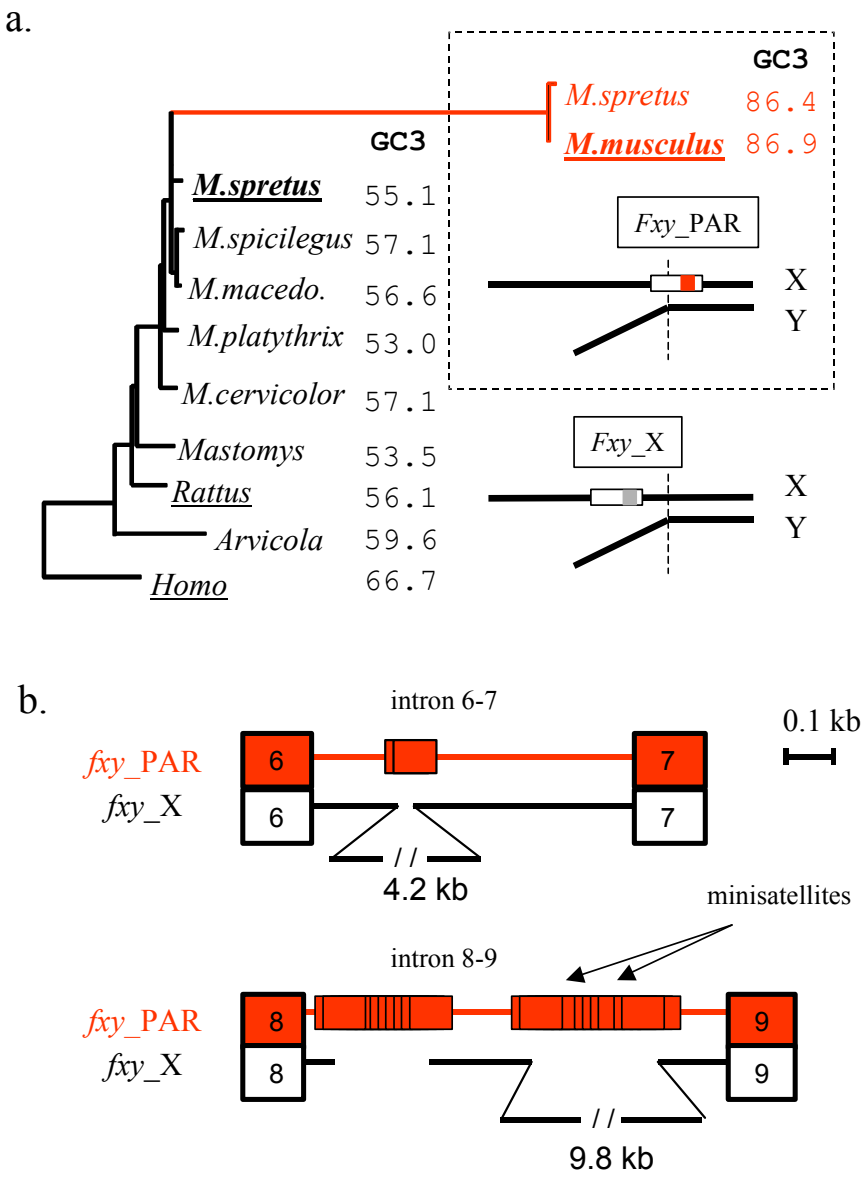


Figure 7. Evolution du gène Fxy chez la souris.
Après son transfert (partiel) dans la région pseudoautosomale (PAR) chez la souris, le gène Fxy a subi une forte augmentation de sa vitesse d'évolution, de son taux de GC (a), une réduction de la longueur de ses introns, et l'apparition de locus minisatellites (b), mimant ainsi un isochore GC-riche.

Plus généralement, la leçon que je tirerais de ce travail est que l'hypothèse nulle, en évolution moléculaire, est plus que jamais celle de la neutralité sélective. L'augmentation de la taille des jeux de données fournit de plus en plus de puissance pour rejeter cette hypothèse, ce qui se explique sans doute le penchant sélectionniste relativement marqué de la littérature actuelle. Profitons donc de cet exemple pour rappeler que le rejet d'une hypothèse nulle (ici, la neutralité) ne valide pas immédiatement l'hypothèse alternative que l'on avait en tête (ici, la sélection): d'autres facteurs non pris en compte (ici, la conversion génique biaisée) peuvent expliquer les écarts détectés.

3.2. Détecter l'adaptation moléculaire: une approche phylogénétique

La plupart des gènes du génome évoluent sous un régime de **sélection purificatrice**. Les mutations sont majoritairement neutres ou défavorables et, dans ce dernier cas, elles peuvent être éliminées par sélection naturelle. Cette description, conforme à la théorie neutraliste de Kimura, n'explique pourtant pas l'adaptation des espèces et des populations à leur environnement. Des **mutations avantageuses** se produisent nécessairement dans les génomes, même si c'est en proportion faible. Les détecter et les caractériser est, à mon sens, un défi important de l'évolution moléculaire. Plusieurs approches sont disponibles pour atteindre cet objectif. Dans ce paragraphe, j'aborde le point de vue **phylogénétique** de l'adaptation moléculaire: il s'agit de rechercher des traces de l'adaptation dans les patrons de variation de séquences codantes échantillonnées dans diverses espèces. Ces travaux utilisent la théorie des modèles markoviens en évolution moléculaire brièvement présentée ci-dessus, et font l'objet d'une collaboration de longue durée avec plusieurs collègues informaticiens (Olivier Gascuel, Alain Jean-Marie, Nicolas Lartillot).

3.2.1. Taux d'évolution synonyme/non synonyme, duplication de gènes.

Une approche standard consiste à contraster les taux de substitution **synonyme** (*dS*) et **non-synonyme** (*dN*) d'un gène. Par substitutions synonymes on désigne les changements de nucléotide qui ne modifient pas l'acide aminé codé, en raison de la dégénérescence du code génétique. Ces changements sont supposés neutres en première approximation. Ils s'accumulent donc à une vitesse de base correspondant au taux de mutation. Les mutations non-synonymes modifient la séquence protéique, et donc potentiellement la fitness des individus qui les reçoivent. Pour la plupart des gènes, ces mutations sont majoritairement délétères et donc éliminées par sélection purificatrice, générant un déficit de changements (observables) non-synonymes par rapport aux synonymes ($dN/dS<1$). Dans un petit nombre de cas, on observe toutefois un excès de substitutions non-synonymes ($dN/dS>1$). Ceci est interprété comme la preuve d'une dynamique d'adaptation : le gène en question a reçu un grand nombre de mutations favorables, qui se fixent avec une probabilité plus forte que les mutations neutres, et s'accumulent donc à une vitesse élevée. Ce principe est exploité au moyen d'analyses phylogénétiques au maximum de vraisemblance faisant intervenir des modèles de Markov sur les codons (Goldman & Yang 1994, Yang 1998, Yang et al. 2000).

Ces techniques ont été utilisées par **Emilie Guldner** au cours de sa thèse (coencadrée avec B. Godelle) pour l'analyse de l'évolution de **l'hémoglobine des plantes**, gène pour lequel une dynamique d'adaptation récurrente, en liaison avec la **symbiose azotée**, est pressentie. Ce

gène est dupliqué (c'est-à-dire présent en deux exemplaires) chez la plupart des plantes, une des deux copies ayant un rôle constitutif, et l'autre ayant un rôle régulateur dans la symbiose avec la bactérie *Rhizobium* chez les Légumineuses, et avec la bactérie *Frankia* chez les plantes actinorhiziennes. Nous nous proposions d'essayer de mettre en évidence un possible phénomène de Reine Rouge (adaptations récurrentes de chaque partenaire de l'interaction aux changements de son interlocuteur) à l'échelle moléculaire par l'examen des patrons d'évolution de ce gène candidat.

Emilie Guldner a tout d'abord cloné et séquencé le gène d'hémoglobine chez le **nénuphar** *Euryale ferox* (Nymphaeaceae), une espèce branchée à la base de l'arbre des angiospermes, de manière à pouvoir enraciner l'arbre phylogénétique de l'hémoglobine. Ce travail a permis de mettre en évidence la présence de deux gènes chez le nénuphar, indiquant une **duplication ancienne** du gène, positionnant avec certitude la racine de l'arbre, et permettant une interprétation plus aisée des processus d'évolution de ce gène (Guldner et al 2004a). Par la suite, E. Guldner a analysé les variations de pression de sélection (mesurée par le rapport *dN*/*dS*) reçues par les différents sites de la molécule dans les différentes lignées phylogénétiques, grâce à une batterie de modèles, dans le contexte de la méthode du maximum de vraisemblance. Ces analyses ont permis de montrer que la symbiose s'était accompagnée d'un changement significatif de mode évolutif de la protéine (accélération du taux de substitution non-synonyme), et a révélé un site aminoacide particulièrement rapide, candidat à un rôle dans l'adaptation récurrente de cette molécule. Ces travaux ont également permis de discuter des limites de la méthodologie employée, et ont souligné le rôle possible dans l'adaptation de duplications récentes du gène de l'hémoglobine chez les espèces symbiotiques (Guldner et al 2004b).

3.2.2. *Episodes adaptatifs et variation de vitesse site-spécifique.*

Une limitation importante de la comparaison synonyme/non synonyme est que le signal est moyenné entre codons et/ou dans le temps : on estime, selon les modèles, un taux de substitution synonyme/non-synonyme commun à l'ensemble des codons du gène considéré, ou à la totalité de l'histoire évolutive des séquences analysées. Or, l'adaptation à l'échelle moléculaire se produit sans doute fréquemment sous forme d'épisodes localisés : un gène subit pendant un temps court un certain nombre de substitutions avantageuses qui concernent un nombre limité de codons. Puis, sa nouvelle fonction acquise, le gène en question reprend un régime de sélection purificatrice, préservant en quelque sorte ce nouvel état.

Pour détecter de tels **épisodes adaptatifs**, j'ai proposé une approche basée sur la détection de **variations de vitesse d'évolution site-spécifique**. L'idée est la suivante: lorsqu'une une protéine acquiert une nouvelle fonction, les **contraintes fonctionnelles** s'exerçant sur les divers acides aminés vont changer. Certains acides aminés (sites) étaient cruciaux pour la fonction de la protéine dans la structure initiale, mais n'ont plus de rôle spécifique dans la nouvelle structure; d'autres, au contraire, étaient précédemment sans fonction spécifique, mais deviennent désormais importants. Une conséquence est que la vitesse d'évolution de ces sites devrait varier. Un niveau de contrainte fonctionnelle fort se traduit en effet par une vitesse d'évolution faible (la plupart des mutations étant éliminées par sélection purificatrice), tandis que les sites neutres évoluent librement, et plus rapidement. L'idée est donc de contraster les vitesses d'évolution site-spécifiques mesurées avant et après l'épisode pressenti, des variations importantes traduisant probablement un changement fonctionnel pour la protéine considérée. Les processus impliquant un changement site-spécifique de vitesse d'évolution sont improprement nommés "**covarion**" (Fitch 1971), ou plus correctement "hétérotaches" (Lopez et al. 2002).

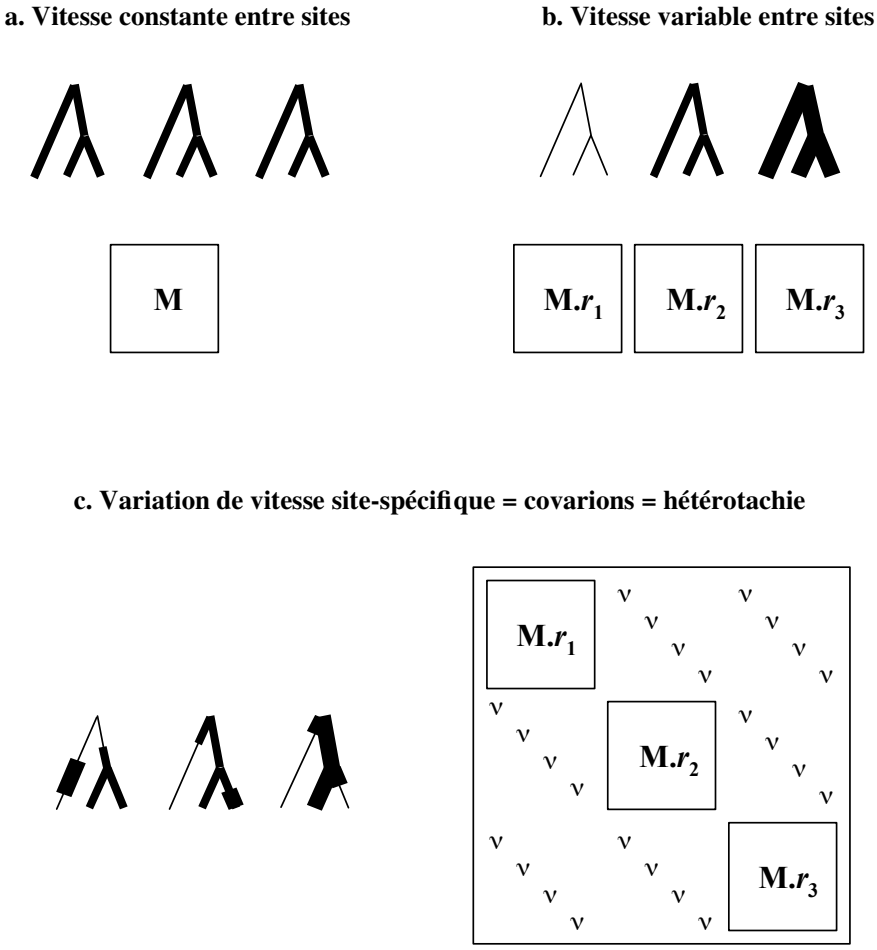


Figure 8. **Trois modèles pour la distribution des vitesses d'évolution entre sites**.
a. Tous les sites évoluent à la même vitesse. Le calcul de vraisemblance se fait comme indiqué dans l'encadré 3.
b. Certains sites évoluent lentement (trait fin), d'autres rapidement (trait épais). Le calcul de vraisemblance (Yang 1994) implique de sommer sur l'ensemble des vitesses possibles.
c. La vitesse d'évolution d'un site peut changer au cours du temps. Le calcul de vraisemblance implique l'usage de la matrice composite dans laquelle un site peut changer d'état (selon le processus **M** et avec sa vitesse courante *ri*) mais aussi de vitesse, et ceci au taux ν (Galtier 2001).

J'ai développé au cours de ces dernières années plusieurs travaux théoriques formalisant et exploitant cette notion. Dans le premier (Galtier 2001), je propose un **modèle markovien** d'évolution des séquences d'ADN sous un processus de type covarion, ainsi qu'une méthode d'analyse phylogénétique au maximum de vraisemblance sous ce modèle. Cette approche permet donc de prendre en compte le fait que les divers sites d'une séquence n'ont pas une vitesse d'évolution constante dans le temps, et de quantifier cet effet par l'estimation d'un paramètre mesurant le taux de changement de vitesse site-spécifique (figure 8). Une collaboration avec Alain Jean-Marie a permis de généraliser ce modèle, et de proposer un algorithme efficace de traitement (Galtier & Jean-Marie 2004) qui accélère les calculs de vraisemblance, notamment pour les données de type aminoacides (20 états possibles) et codons (61 états possibles).

Une deuxième approche, entreprise en collaboration avec Tal Pupko, un collègue israélien, est plus directement reliée à l'objectif de détection d'effets sélectifs (Pupko & Galtier 2002). Nous proposons un **test de rapport de vraisemblance** permettant de détecter des variations de vitesse site-spécifiques entre deux groupes prédéterminés de taxons. Appliqué aux protéines mitochondriales de mammifères, ce test suggère que pour un nombre significatif de sites, les contraintes fonctionnelles s'exerçant sur les acides aminés de ces protéines chez les **primates** sont distinctes de celles qui s'exercent chez les **autres mammifères** : certains acides-aminés sont très conservés chez les primates mais fortement variables chez les autres mammifères, et réciproquement (figure 9). Ceci indique que la structure tridimensionnelle des protéines mitochondriales a changé chez les primates, un résultat assez inattendu vue la fonction de ces protéines, que nous discutons (cf. chapitre 3).

Cette approche me parait avoir un avenir prometteur, mais la méthodologie actuelle, y compris notre contribution, n'est pas pleinement satisfaisante. Une limite du test que nous avons proposé est que l'utilisateur doit spécifier *a priori* la branche de l'arbre phylogénétique où l'épisode adaptatif est pressenti. Il apparaît souhaitable de mettre au point une méthode plus exploratoire, qui mesurerait le taux de covarion associé à chaque branche, et détecterait les branches de l'arbre où ce taux est supérieur au bruit de fond, signe possible de l'adaptation. C'est là un des projets que **Julien Dutheil** et moi-même souhaitons traiter au cours des années à venir.

3.2.3. *Coévolution et non-indépendance des sites*.

La plupart des modèles représentant l'évolution des séquences font l'hypothèse **d'indépendance des sites**. Cette hypothèse est clairement violée par les données réelles, pour lesquelles des contraintes structurales et fonctionnelles s'exercent sur des groupes de sites (par exemple les paires Watson-Crick dans les tiges des ARN ribosomiques et de transfert), qui co-évoluent. En oubliant ces contraintes communes, on simplifie les modèles et leurs prédictions, mais d'une part on **sur-estime le niveau de confiance** des reconstructions phylogénétiques tel que mesuré par la technique du bootstrap (Galtier 2004b), et d'autre part on s'interdit de détecter/prendre en compte ces effets sélectifs pourtant majeurs. L'épistasie intra-génique, qui fait que le coefficient de sélection associée à une mutation donnée dépend très largement de l'état allélique aux autres sites de la molécule, semble en effet être un facteur primordial de l'évolution des séquences fonctionnelles, comme le montrent les travaux récents d'A. Kondrashov (Kern & Kondrashov 2004).

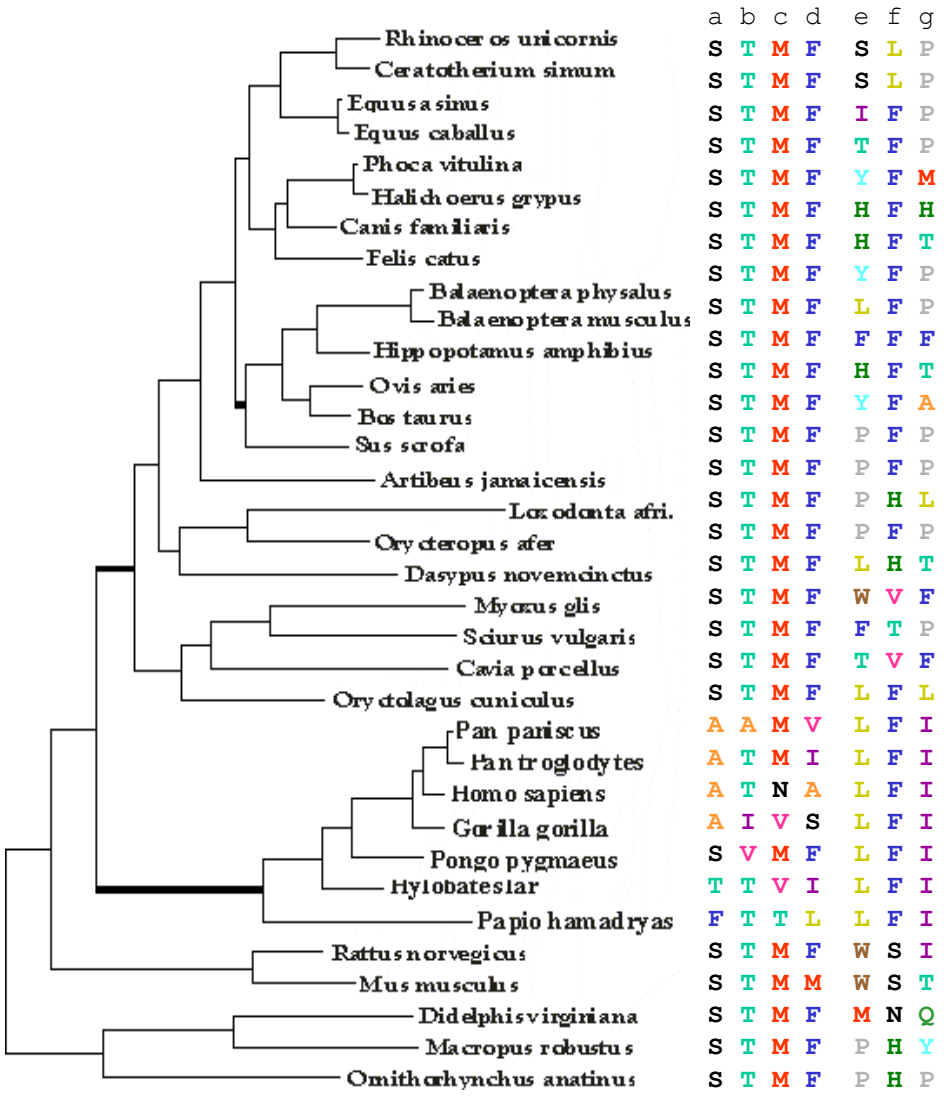


Figure 9: Phylogénie mitochondriale des Mammifères et covarions.
Les variations entre espèces de Mammifères à 7 sites amino-acides de protéines mitochondriales sont représentées en contrastant Primates (encadrés) et non-Primates. Certains sites (a-d) sont très conservés chez la plupart des Mammifères mais variables chez les Primates. D'autres (e-g) montrent le patron inverse (Pupko & Galtier 2002).
Ces variations de vitesse d'évolution reflètent probablement des variations dans le temps du niveau de contrainte fonctionnelle s'exerçant sur ces sites.

Pour traiter cette question, Julien Dutheil a mis en place une méthode d'analyse phylogénétique basée sur la cartographie des événements de substitution pour chaque site sur un arbre, permettant la mise en évidence de **coévolution entre sites d'une molécule**. Un calibrage de la méthode sur données d'ARN ribosomiques s'est avérée prometteur (Dutheil et al. 2005, figure 10), et son extension aux données protéiques est maintenant réalisée, ce qui n'était pas un mince exploit au vu du pessimisme de la littérature à ce sujet (Tufféry & Darlu 2000) – il a fallu notamment prendre en compte les propriétés biochimique des aminoacides, et, via une méthode de classification hiérarchique des sites, le fait que des groupes de plus de deux résidus peuvent co-évoluer. Ce projet pourrait avoir des implications importantes en biologie structurale. Nous cherchons notamment à identifier d'éventuelles **règles générales** de coévolution dans les motifs de structure secondaire, voire tertiaire, des protéines.

Les perspectives dans le domaine de la modélisation markovienne pour l'évolution moléculaire me semblent nombreuses. Tout d'abord, je ne pense pas que tous les modèles pertinents, loin de là, aient été décrits et mis en œuvre. On peut progresser, me semble-t-il, dans la représentation des sources de variations des vitesses d'évolution moléculaire (échelle du gène, du site, variations dans le temps, et leurs interactions). Les phénomènes divers qualifiés de "variation de vitesse entre sites", "écart à l'horloge moléculaire" et "covarion", par exemple, pourraient être réunis en un modèle généralisé, de manière à quantifier leur importance relative. Le phénomène de transfert horizontal de gènes mériterait également une formalisation mathématique. Mais le point faible majeur que j'identifierais dans la littérature actuelle est que les applications de ces modèles et méthodes sophistiqués relèvent presque toujours de l'étude de cas: on décortique l'évolution moléculaire d'un ou deux gènes pour illustrer l'approche, ou parce que c'est notre gène favori sur le plan fonctionnel. Il manque, à mon sens, un effort de synthèse, une tentative de description des processus de l'évolution moléculaire en général (pour autant que cette notion ait un sens; mais si elle n'en a pas, démontrons-le, et si elle en a un, exhibons-le), qui pourrait ainsi être comparées aux efforts similaires menés en ce sens en génétique des populations.

3.3. L'ADN mitochondrial, marqueur neutre et clonal?

Les chloroplastes et les mitochondries sont d'anciennes bactéries, qui ont intégré la cellule eucaryote par **endosymbiose**. Les mitochondries sont le lieu de la **respiration** cellulaire (production d'ATP par le catabolisme des glucides). Elles dérivent d'une α-protéobactérie ancestrale, et sont présentes chez tous les eucaryotes, sauf cas de perte secondaire (Gray et al. 2001). Les chloroplastes réalisent la **photosynthèse** (réception de l'énergie lumineuse et synthèse de composés organiques). Ils dérivent de cyanobactéries ancestrales. Il semble que les chloroplastes de la lignée verte (algues vertes et plantes) et ceux des algues rouges et brunes aient des **origines distinctes** – la photosynthèse aurait donc été acquise plusieurs fois indépendamment chez les eucaryotes (Delwiche 1999). Les rôles principaux de ces organites (respiration, photosynthèse) sont des fonctions métaboliques de base, communes à l'ensemble des cellules/organismes qui les portent, et qui ont donc été fortement **conservées** au cours de l'évolution.

Les mitochondries sont **transmises maternellement** chez les animaux (à l'exception exclusive, à ma connaissance, du cas de la moule chez qui il y a bi-uniparentalité, c'est-à-dire une lignée mitochondriale mâle et une lignée femelle). Cette uniparentalité implique une évolution **clonale**, sans échange génétique entre lignées. Chez les plantes, la règle est aussi

celle de la transmission maternelle des organites, mais les exceptions sont assez nombreuses (Reboud & Zeil 1994).

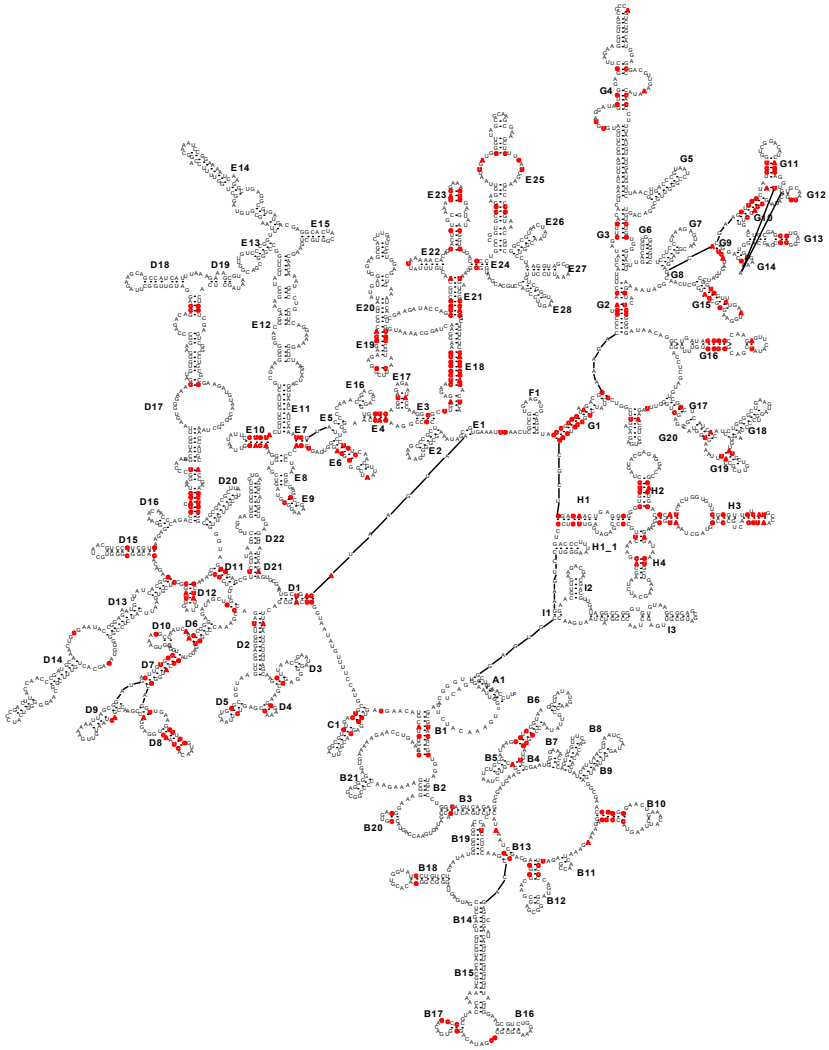


Figure 10. Cévolution dans l'ARN ribosomique.
Une approche basée sur la cartographie phylogénétique des événements de Substitution et la comparaison des cartes entre sites a permis de détecter 182 Paires de sites coévoluant dans l'ARNr 16S bactérien. Parmi ces paires candidates, 162 Sont des paires structurales documentées dans les tiges (en rouge), et 17 correspondent à des interactions confirmées par la visualisation de la structure 3D. 85% des paires Watson-Crick détectables ont été détectées (Dutheil et al. 2005)

Les génomes des organites ont été fortement **réduits** au cours de leur évolution depuis leur ancêtre bactérien. Alors qu'une bactérie typique porte plusieurs milliers de gènes, le plus complet des génomes mitochondriaux connus, celui du protiste *Reclinomonas americana*, en contient moins de 100. On en trouve une petite quarantaine chez les animaux, et jusqu'à moins d'une dizaine chez *Plasmodium falciparum*. Sur le millier de protéines qui composent le protéome mitochondrial des animaux, seulement 13 sont codées par le génome mitochondrial. Les autres sont codées par le génome nucléaire, et importées depuis le cytoplasme. La **taille** du génome des organites a également été fortement réduite, jusqu'à 16 kilobases (kb) pour les mitochondries animales, et quelques dizaines/centaines de kb pour les chloroplastes et mitochondries des plantes (un génome bactérien typique fait 2-5 mégabases). La **structure** de ces génomes, enfin, est très variable à l'échelle du monde eucaryote, allant d'un unique chromosome circulaire à une ou plusieurs molécules linéaires, qui recombinent parfois entre elles (Burger et al. 2003).

L'évolution de ces génomes a fait couler beaucoup d'encre, tout d'abord parce que l'ADN mitochondrial est un des **marqueurs privilégiés de la génétique des populations**, et qu'à ce titre son mode d'évolution se devait d'être caractérisé, et ensuite parce que la dynamique des ces génomes symbiotiques, apparemment "domestiqués" par le génome nucléaire, nous fascine à plus d'un titre, et a probablement joué un rôle important lors des **premiers stades de l'évolution des eucaryotes**. De nombreuses questions demeurent malgré tout. Concernant l'évolution récente, il semble que les paradigmes qui ont été avancés pour justifier l'usage massif de l'ADN mitochondrial – clonalité, quasi-neutralité – soient remis en question par les données disponibles aujourd'hui. Il serait bon de clarifier ces aspects, vue l'importance de ce marqueur génétique. Les mitochondries sont-elles vraiment clonales, et uniquement sous sélection purificatrice ? Concernant les tendances à long terme, les questions critiques ont été posées à plusieurs reprises, mais pas toujours résolues de manière satisfaisante: pourquoi ces pertes de gènes, à un rythme erratique, et pourquoi jamais la perte de tous les gènes? Cette dynamique est-elle neutre ou sous sélection, et quels niveaux de sélection (organite, cellule, organisme) sont ceux qui agissent/ont agi? J'ai choisi d'aborder ces questions au cours des années à venir, et je résume ci-dessous les premiers résultats que nous avons obtenus, et les voies vers lesquelles nous pensons nous diriger pour éclaircir les processus de l'évolution mitochondriale.

3.3.1. Mitochondries et recombinaison

L'évolution des mitochondries animale est traditionnellement supposée **clonale**: du fait de la transmission maternelle, les diverses lignées mitochondriales ne peuvent pas se rencontrer dans un zygote pour recombiner. C'est d'ailleurs un des arguments mis en avant par les généticiens des populations pour justifier le choix de ce marqueur. Cette hypothèse de travail a toutefois été remise en cause en 1999, lorsque trois articles plutôt provocateurs ont détecté des évidences de **recombinaison mitochondriale** chez l'homme. Le premier (Eyre-Walker et al. 1999), exploitant une vingtaine de génomes mitochondriaux humains complets, rapportait une quantité d'**homoplasie** (incompatibilité phylogénétique entre les sites) remarquablement élevée, qu'il semblait difficile d'expliquer par des phénomènes mutationnels. Le deuxième (Awadalla et al. 1999) montrait avec ces mêmes données que le niveau d'incompatibilité entre sites s'accroît avec la distance physique, comme attendu si des événements de recombinaison viennent casser les déséquilibres de liaison. Le troisième de ces articles (Hagelberg et al. 1999), enfin, mettait en avant une convergence surprenante dans la région de contrôle mitochondriale: les deux haplotypes présents sur une île de l'archipel du Vanuatu,

généalogiquement fort éloignés, présentaient une mutation en forte fréquence qui leur était commune, et qui n'apparaissait dans (quasiment) aucune des milliers d'autres régions de contrôles séquencées chez l'homme, suggérant fortement que c'est par échange génétique et recombinaison sur cette île que ladite mutation était passé d'une lignée à l'autre.

Ces trois publications ont provoqué, comme on pouvait s'y attendre, un débat intense et des vérifications acharnées. Le recul nous enseigne qu'aucune d'entre elles n'y a résisté. La relation entre déséquilibre de liaison et distance (Awadalla et al. 1999) n'est pas robuste: elle disparaît lorsque la taille du jeu de données augmente (Ingman et al. 2000, Herrnstadt et al. 2002), ou lorsque les méthodes d'analyse changent (Innan & Nordborg 2002). La convergence dans le Vanuatu, quant à elle, provenait simplement d'une erreur d'alignement (Hagelberg 2003). Restait toutefois le taux anormalement élevé d'homoplasie intra-spécifique (Eyre-Walker et al. 1999), qu'il était difficile d'expliquer sans recombinaison. Grâce à une analyse des patrons de polymorphisme mitochondriaux dans des groupes d'espèces proches chez les Mammifères, nous montrons que ce niveau d'homoplasie élevé peut s'expliquer en réalité par la présence de **points chauds mutationnels** nombreux, et **changeants** au cours du temps (Galtier et al. 2006): les espèces d'un même genre ont tendance à montrer du polymorphisme aux même sites, révélant des points chauds mutationnels, mais cette relation diminue rapidement (et disparaît) lorsque l'on compare des espèces plus éloignées. Ces points chauds expliquent l'homoplasie intra-spécifique (sans nécessité d'invoquer la recombinaison), et leur courte durée de vie explique qu'Eyre-Walker et ses collègues n'aient pas pu en retrouver la trace chez les autres espèces de primates qu'ils ont analysées, trop éloignées de l'homme. Le troisième mousquetaire de la fronde anti-clonalité mitochondriale tombe donc également.

Il est bon de noter que malgré les erreurs et les limites de ces articles finalement infirmés, le pavé dans la marre de 1999 a eu des conséquences très positives. Le dogme de la clonalité une fois remis en cause, de nombreux auteurs se sont engagés dans la brèche et ont essayé de mettre en évidence par des méthodes indirectes des phénomènes de recombinaison mitochondriale chez diverses espèces animales. Ces tentatives semblent avoir été couronnées de succès, le cas le plus probant me semblant être celui des scorpions du genre *Buthus* et *Mesobuthus* (Piganeau et al. 2004, Gantebain et al. 2005). Mais le résultat le plus remarquable concerne encore une fois l'espèce humaine, puisque Schwartz & Vissing (2002) ont découvert, chez un patient atteint d'une myopathie, des traces d'ADN mitochondrial paternel, et même des molécules recombinées entre les lignées paternelle et maternelle (Kraytsberg et al. 2004). La recombinaison mitochondriale existe donc bien chez l'homme, et sans doute chez d'autres animaux, et si nous savons cela aujourd'hui, c'est grâce aux trois publications erronées de 1999. Reste à quantifier l'importance de ce phénomène à l'échelle évolutive, et sur ce plan, il semble bien que pour l'instant on ne puisse accorder à ce mécanisme qu'un statut anecdotique. Ceci dit, il a fallu, pour le démontrer, déployer un arsenal de données et de méthodes, décortiquer les patrons évolutifs mitochondriaux, et cette recherche profitera aux gens sérieux qui utilisent l'ADN mitochondrial comme outil de gestion de la biodiversité ou à des fins biomédicales.

3.3.2. Mitochondries et adaptation

L'autre argument maintes fois avancé pour justifier un usage sans modération du marqueur mitochondrial en génétique des populations est sa **quasi-neutralité**. Les gènes mitochondriaux codent en effet pour des ARN ribosomiques et de transferts, et pour 13 protéines de la chaîne respiratoire. Il s'agit donc de fonctions du métabolisme de base de la cellule, qui existe depuis des milliards d'années, et qui semble intuitivement peu à même de s'adapter à des changement environnementaux. On est loin des gènes de la perception

sensorielle ou de l'immunité, les champions du *dN*/*dS* chez l'homme (Nielsen et al. 2005), impliqués dans les interactions durables entre espèces et donc poussés par la Reine Rouge vers une évolution rapide. Les gènes mitochondriaux sont bien, pense-t-on, sous sélection purificatrice, mais cela ne gène pas réellement l'analyse phylogéographique: certaines mutations sont éliminées et ne comptent donc plus, et puis voilà.

Je dois à la vérité de reconnaître que le paragraphe ci-dessus est volontairement naïf, et que si, dans les faits, les analyses des données mitochondriales de génétique des populations se font effectivement le plus souvent sous l'hypothèse de quasi-neutralité, leurs auteurs sont en revanche bien conscients que ce n'est là qu'une hypothèse. Et ceux-ci admettent d'ailleurs que, depuis un certain nombre d'années, les arguments soutenant l'existence de processus adaptatifs dans l'ADN mitochondrial s'accumulent. Des preuves directes de sélection ont été décrites chez la drosophile (de Stordeur 1997, James & Ballard 2003), par exemple. Le groupe de L.I. Grossman a pris pour objet d'étude les protéines de la chaîne respiratoire (incluant les 13 protéines mitochondriales) des primates, et a détecté plusieurs gènes/lignées pour lesquels le ratio des vitesses non-synonyme/synonyme était élevé, suggérant une histoire jalonnée d'événements de fixation de mutations favorables (revue dans Grossman et al 2004). Nos travaux sur la détection de changements de vitesse d'évolution site-spécifique (Pupko & Galtier 2002) nous ont également amenés à suggérer l'existence d'événements adaptatifs chez les primates, et ceci avec une approche distincte. Chez l'homme, une hypothèse d'adaptation de l'ADN mitochondrial aux conditions climatiques de température (Ruiz-Pesini et al. 2004) a fait un scoop récent (bien que le soutien m'apparaisse peu clair), et une situation analogue pourrait s'être produite chez des salmonidés (Doiron et al 2002) et chez les lièvres européens (Alves et al 2003).

Le nombre croissant d'études de cas suggérant des événements adaptatifs au cours l'évolution mitochondriale souligne la nécessité d'une approche plus globale visant à quantifier l'importance de ce phénomène: est-ce une propriété générale du génome mitochondrial, ou simplement un petit nombre d'anecdotes exagérément mises en avant? Nous avons tenté d'apporter une réponse à cette question en traitant l'ensemble des données de polymorphisme intra-spécifique mitochondrial disponibles chez les animaux.

Avant d'exposer ces résultats, je me dois de présenter l'outil qui nous a permis de rassembler ce jeu de données exhaustif. Il s'agit de **Polymorphix**, une base de données dédiée au polymorphisme de séquence, qu'a construite Eric Bazin (en collaboration avec Laurent Duret et le Pôle Bioinformatique Lyonnais, Bazin et al. 2005) au cours de sa thèse. Grâce à une procédure complexe et astucieuse faite de comparaisons de séquences, filtres bibliographiques et critères de similarité, cette base réalise la prouesse d'extraire de Genbank/EMBL, la base de données généraliste fourre-tout de la génomique, les quelques milliers de jeux de données de polymorphisme de séquence qui y sont inclus, pour ne pas dire noyés. Cette base permet, via une dernière étape de "nettoyage" fin, la mise sur pied rapide de jeux de données prêts à l'analyse. Polymorphix nous a permis d'initier en peu de temps un bon nombre de méta-analyses de génétique des populations et évolution moléculaire (Galtier et al. 2006a, Galtier et al. 2006b, Glémin et al., en préparation, Nabholz et al., en préparation), alors que chacune de ces études aurait réclamé de longues semaines d'un travail fastidieux si les données avaient dû être réunies manuellement depuis la littérature.

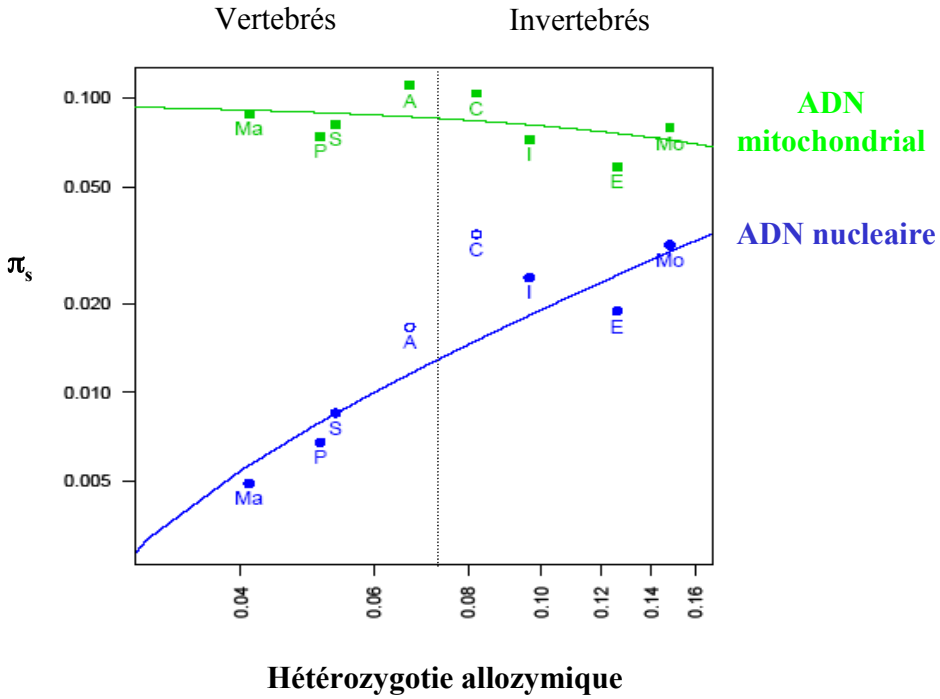


Figure 10. Distribution taxonomique de la diversité nucléotidique synonyme moyenne nucléaire et mitochondriale chez les animaux.
L'hétérozygotie allozymique moyenne est prise comme point de comparaison. La divversité mitochondriale ne reflète pas la différence de taille efficace moyenne entre Vertébrés en Invertébrés, sans doute en raison de balayages sélectifs récurrents.
Ma: Mammifères (allozymes: 184 espèces, ADN nuc: 25, ADN mt: 311); P: Poissons (Teleostei + Chondrichtyes; 183, 11, 248). S: Sauropsidés (Reptiles + Oiseaux; 116, 18, 348); A: Amphibiens (61, 4, 80); C: Crustacés (122, 2, 58); I: Insectes (122, 69, 451); E: Echinodermes (15, 22, 26); Mo: Mollusques (46, 11, 107).

Eric Bazin a donc pu mesurer, via Polymorphix, la distribution taxonomique du polymorphisme intra-spécifique nucléaire et mitochondrial environ 1700 espèces animales, données qui ont été comparées à la méta-analyse de polymorphisme allozymique publiée par Nevo et ses collaborateurs (Nevo et al. 1984). Les résultats concernant les données nucléaires et allozymiques furent conformes aux attendus intuitifs. Rappelons que sous le modèle neutre (ou quasi-neutre), la diversité d'une espèce donnée est contrôlée avant tout par sa taille efficace (cf. paragraphe 2.1). Conformément à cette prédiction, notre analyse a révélé un polymorphisme nucléaire plus élevé chez les invertébrés (par rapport aux vertébrés), chez les espèces marines (par rapport aux terrestres/eau douce), et chez les organismes de petite taille (par rapport aux organismes de grande taille), et ceci à l'échelle de l'ensemble des animaux ou via des comparaisons intra-phylum. Les données mitochondriales, en revanche, ne reflètent en aucune manière ces différences de taille efficace entre groupes d'organismes (figure 11), et ceci malgré la quantité de données disponible (1629 espèces animales analysées). Ce résultat très surprenant ne peut guère être interprété que par l'adaptation: nous proposons que des **balayages sélectifs fréquents** réduisent de manière récurrente le taux de polymorphisme mitochondrial des espèces à grande taille efficace (Bazin 2005, Bazin et al. soumis), un modèle appelé le "**genetic draft**" par J. Gillespie (2001) – cette hypothèse est d'ailleurs corroborée par l'analyse des patrons de substitutions synonymes/non-synonymes (Bazin et al. soumis). Le génome mitochondrial, non-recombinant et très dense en gènes, est particulièrement sensible aux effets de la sélection en liaison (l'auto-stop génétique), ce qui explique les particularités de son patron de variation.

Outre son intérêt en évolution moléculaire en général, avec une remise en cause sérieuse du modèle neutre dans le cas de l'ADN mitochondrial, ce résultat a des implications importantes dans le monde de la recherche appliquée. Il montre en effet, et c'est confirmé par une analyse spécifique, menée par **Benoît Nabholz**, du déterminisme du polymorphisme mitochondrial chez les Mammifères, que l'ADN mitochondrial n'atteint sans doute que très rarement l'équilibre mutation-dérive, au point que la diversité mitochondriale ne renseigne pas le moins du monde sur la taille efficace des espèces et des populations. Or une bonne partie de la **génétique de la conservation**, cette discipline visant à s'aider de la génétique pour prendre des décisions concernant la gestion et la sauvegarde de la biodiversité, s'appuie majoritairement sur les données mitochondriales, et raisonne souvent en reliant diversité à abondance – on préservera en priorité les populations les plus polymorphes. Nos résultats suggèrent, en accord avec Hurst & Jiggins 2005, que le marqueur mitochondrial était malheureusement peut-être un des plus mauvais choix pour ce type d'analyse.

3.3.3. Pourquoi l'hyper-mutation mitochondriale?

On réalise, au travers des aspects abordés ci-dessus combien notre compréhension des mécanismes de l'évolution mitochondriale peut encore progresser, et combien cette recherche fondamentale peut influer sur les pratiques appliquées. Outre ces aspects qui concernent les mécanismes actuels de l'évolution mitochondriale, on ne peut que difficilement s'empêcher de s'interroger sur les origines, les causes ultimes des spécificités de ce génome. Les deux questions qui me viennent immédiatement à l'esprit concernant la mitochondrie animale sont: "pourquoi conserver un génome?" et "pourquoi un taux de mutation si élevé?".

La première de ces deux questions a déjà été abordée à plusieurs reprises. Des hypothèses "mécanistes" ont été proposées pour expliquer le maintien d'un système génétique mitochondrial. On a notamment invoqué la difficulté à transférer du noyau vers les mitochondries les protéines trop hydrophobes, telles que le cytochrome *b* ou les sous-unités 1

et 2 de la cytochrome oxydase (Popot & de Vitry 1990, Daley et al. 2002). D'autres auteurs ont proposé que ces protéines pourraient être toxiques dans le cytoplasme (Martin & Herrmann 1998). D'autres enfin ont suggéré que leur expression devait être régulée par le statut oxydo-réducteur de la matrice, témoin du niveau d'activité des mitochondries (Race et al. 1999), ce qui implique une localisation dans cet organite. Enfin, plusieurs auteurs ont proposé que les codes génétiques non-standard de beaucoup de mitochondries (notamment animales) soient un frein fort au transfert (Adams & Palmer 2003). Ces facteurs ont pu jouer un rôle sur les modalités de transfert de gènes. Ce n'est clairement pas un hasard si les gènes "restants" sont notamment ceux de la chaînes respiratoires, et si les protistes qui perdent la respiration peuvent alors (et ils sont les seuls) perdre la totalité de leur génome mitochondrial.

Les hypothèses ci-dessus invoquent diverses **"impossibilités"** structurales ou fonctionnelles à "terminer" le processus de perte de génome cytoplasmique. Ce point de vue, quoique très raisonnable, me semble sous-estimer la capacité des organismes et des génomes à "parvenir à leurs fins". J'ai le sentiment, certes spéculatif, que s'il était avantageux pour un animal, par exemple, de perdre son génome mitochondrial, une lignée au moins aurait trouvé une "solution" aux "impossibilités" mentionnés ci-dessus. Je propose donc de considérer l'hypothèse que les génomes mitochondriaux des animaux sont conservés parce qu'ils sont "utiles", et non pas par défaut de perte. Mon idée est la suivante: demandons-nous ce qui se passerait si un zygote parvenait à transférer l'ensemble de son génome mitochondrial dans le noyau. Une conséquence immédiate est que cet individu ne pourrait plus contrôler le développement et la **sénescence** de sa lignée somatique.

Il est en effet clairement établi que les mitochondries jouent un rôle primordial dans le vieillissement cellulaire, au moins chez les Mammifères (par exemple Tanaka et al. 1998). La **théorie mitochondriale du vieillissement**, défendue par A.M. Harman, propose que les radicaux libres oxydants produits par un (léger) disfonctionnement des mitochondries soient mutagènes, et que les mutations induites conduisent à une plus forte production de radicaux libres (Harman 1956, Harman 1992), ce cercle vicieux conduisant à la mort cellulaire et au vieillissement. De nombreuses évidences expérimentales soutiennent aujourd'hui cette théorie (revues dans Alexeyev et al. 2004). Il apparaît donc plausible d'imaginer, et c'est l'hypothèse que je me propose de considérer, que **l'apparition de mutations dans le génome mitochondrial joue un rôle de signal dans le cycle cellulaire somatique et le cycle de vie** des animaux. La mitochondrie serait alors un des effecteurs de la pléiotropie antagoniste (compromis reproduction/longévité) menant au vieillissement.

La question du maintien d'un génome mitochondrial pourrait donc rejoindre celle de l'hypermutation. Cette dernière caractéristique est habituellement considérée comme un fardeau indésiré, mais ne peut-on pas y voir une nécessité au bon fonctionnement du "chronomètre mutationnel mitochondrial"? A ce titre, il faut noter que chez les plantes, qui ont clairement une relation bien différente au cycle de vie et au vieillissement (il n'y a pas de vraie lignée somatique chez les plantes), le taux de mutation mitochondrial n'est pas plus fort que celui du noyau (Wolfe et al. 1987). Il est également remarquable que le seul transfert de gènes **vers** un génome mitochondrial connu à ce jour est celui d'un gène de réparation de l'ADN (*mutS*), découvert dans la mitochondrie d'un corail (Pont-Kingdon et al. 1998). On peut penser que le plan de développement de ce Cnidaire, peut-être plus proche de celui des plantes, implique des dynamiques de sénescence bien différentes de celles de la plupart des animaux, et donc peut-être un taux de mutation moins élevé.

Les éléments de réflexion ci-dessus ne constituent pas un projet de recherche concret et immédiat, mais ils me paraissent porter les prémisses de développements qui pourraient, si l'hypothèse sous-jacente se confirme, contribuer à expliquer diverses caractéristiques encore mal comprises (ou mal interprétées) des génomes mitochondriaux, telles qu'un taux de

mutation élevé, une compaction forte, et le maintien obligatoire d'un génome. Quoique parfois mentionnées (par exemple Ballard & Whitlock 2004), le rôle de la mitochondrie dans le vieillissement n'a jamais été considéré, à ma connaissance, comme une force évolutive potentielle. Les pistes de recherche que j'envisage pour explorer cette idée incluent :

(i) la modélisation de l'évolution du taux de mutation et de la taille des génomes dans un contexte de sélection multi-niveau (mitochondrie, cellule, organisme), en prenant en compte explicitement l'existence ou non d'une lignée somatique.

(ii) des tests empiriques d'un lien entre taux de mutation mitochondrial et longévité des espèces ou plan de développement. Le taux de mutation mitochondrial peut être mesuré indirectement par la comparaison de séquences. Les mammifères longévifs montrent moins de dommages oxydatifs mitochondriaux que les espèces à durée de vie plus courte (Barja & Herrero 2000). Est-ce dû à un taux de mutation plus faible, comme le prédit la théorie mitochondriale du vieillissement?

(iii) l'analyse de l'évolution moléculaire de l'ADN polymérase γ, en relation, là encore, avec la longévité. Voit-on se produire au cours de l'évolution des substitutions d'aminoacides susceptibles de modifier la fidélité, et peut-on les relier à des variations de longévité, chez les Mammifères, par exemple? Dans le même ordre d'idée, les gènes régulant l'activité oxydative des radicaux libres (superoxyde dismutase, peroxydases, enzymes de biosynthèse des hormones du vieillissement) pourraient également faire l'objet d'une étude évolutive et corrélative.

4. (Perspectives) Pourquoi on fera toujours de l'évolution moléculaire.

J'en ai donc fini avec la présentation de mes travaux récents et de mes projets à moyen terme, et le moment d'un bilan introspectif est dans doute venu. Au final, je dois convenir que cette revue plutôt rapide me laisse un double et paradoxal sentiment d'inachevé. La première raison de ce sentiment provient de la nature disparate et désordonnée, manquant, pour tout dire, de logique interne, des différents aspects considérés dans ce manuscrit. Pourquoi sauter sur les mitochondries alors que l'origine des isochores n'est pas encore élucidée, et que l'importance de la conversion génique biaisée vient d'être découverte? Pourquoi s'intéresser à la coévolution dans les protéines si la question biologique d'intérêt est celle de l'hypermutation mitochondriale? Ces questions sont d'autant plus pertinentes que, et c'est la 2^{e} raison du sentiment évoqué ci-dessus, l'ensemble de ces travaux, malgré leur diversité, sont une goutte d'eau dans l'océan de la littérature en évolution moléculaire, qui peut elle-même apparaître, quoique j'ai pu en dire en introduction, comme une discipline finalement quelque peu annexe de la biologie actuelle.

Chassons bien vite ces doutes inutiles. La dimension mondiale de la recherche moderne veut qu'un individu considéré isolément ne peut qu'apporter sa pierre à un édifice qui le dépasse, aussi frustrant que cela puisse être. Ces réflexions nous conduisent alors naturellement à déplacer notre questionnement, de manière moins égocentrée, sur la pertinence dudit édifice: l'évolution moléculaire mérite-t-elle, au final, qu'on y consacre le temps, l'argent et l'énergie que nous lui consacrons? Pourquoi faire de l'évolution moléculaire, et en ferons nous toujours? Après réflexion, je décide de répondre par l'affirmative à cette série de questions (ce qui ne me coûte rien), et je m'en explique ci-après.

4.1. L'ampleur de la tâche

La première raison qui fait que, selon moi, la génomique évolutive a de beaux jours devant elle, c'est le niveau de complexité des systèmes que nous étudions. On voit paraître dans les journaux spécialisés de nombreuses publications concernant l'évolution moléculaire d'un gène (d'intérêt) dans un taxon (d'intérêt). Or les génomes sont gigantesques, la biodiversité est insondable, de sorte qu'il me paraît extravagant d'imaginer qu'on aura un jour "épuisé" les questions de la génomique évolutive. L'objectif d'annoter les génomes (qui nécessite une approche comparative) et celui de compléter la systématique des êtres vivants (qui nécessite l'outil moléculaire), pour ne citer qu'eux, m'apparaissent en réalité infinis en raison de leur nature quasi-fractale: on peut toujours demander une meilleure résolution, génétique ou taxonomique. Les prochaines étapes semblent être le séquençage complet d'une quinzaine de Mammifères pour annoter le génome humain *via* la recherche de segments conservés, et les projets de code-barre moléculaire pour l'identification des plantes et des animaux. Ce sont des projets sans doute remarquables, mais qui n'ont, me semble-t-il, rien de définitif: d'autres suivront. Pour résumer, tant qu'il y aura de la génétique et/ou de la systématique, il y aura de la génomique évolutive.

Il faut également noter que cette boulimie moléculaire renouvelle sans-cesse les besoins informatiques, et que les méthodes, une fois développées, ne demandent qu'à être appliquées. La bioinformatique est devenue indissociable de la génomique, et je ressens un certain auto-entretien de ce système, qui semble parfois, à l'extrême, pouvoir avancer par inertie même en l'absence de question biologique intéressante.

4.2. L'appel des disciplines appliquées

L'autre raison qui me fait prédire un avenir florissant à la génomique évolutive, c'est qu'elle est le ferment fondamental de nombreuses recherches appliquées. Les agronomes et les médecins ont besoin de comprendre la biodiversité pour mieux appréhender le lien génotype-phénotype. Ainsi génotype-t-on des dizaines d'espèces de blé ou de maïs, des centaines d'hommes et de femmes, mais aussi des multitudes de souches de parasites ou de pathogènes, et l'interprétation de ces données fait bien-sûr appel aux concepts de la génomique des populations. Les politiques de gestion et de conservation de la biodiversité s'appuient elles-aussi en partie sur les données génétiques, que ce soit pour gérer les ressources naturelles exploitées ou pour préserver les écosystèmes sauvages menacés. Enfin, bien que l'on puisse difficilement les qualifier d'appliquées, il faut noter que la génomique évolutive est également au service de la paléontologie et de l'archéologie, ces disciplines visant à retracer notre histoire, qui ont un impact majeur sur notre perception de la place de l'homme dans la nature.

4.3. Les vraies raisons

Tout cela, ce sont les raisons officielles, celles que l'on invoque dans les projets de recherche pour obtenir des financements. Je m'étonne d'ailleurs de n'avoir pas mentionné, et je répare immédiatement cet oubli, que la génomique évolutive devra s'adapter, à l'avenir, aux progrès technologiques inévitables, puces à ADN, transcriptomique, protéomique et autres mots en "ique" – voilà qui est dit.

Par bonheur, il me semble que les vraies raisons qui poussent certains d'entre nous vers l'évolution moléculaire sont souvent autres. Je pense qu'il y a le plus souvent une vraie fascination pour l'évolution biologique, dont les aspects moléculaires ne sont qu'une facette. On ne se lasse pas de disséquer les mécanismes de l'adaptation des espèces, leurs inter-relations, de construire des modèles permettant d'expliquer l'existence ou les variations de tel ou tel trait. Il y a un aspect jouissif à replacer dans un contexte évolutif un système biologique

quelconque (et les génomes ne font pas exception), à deviner les "raisons" de sa structure ou de son fonctionnement, et ceci indépendamment de toute velléité d'application à court ou long terme. Nous qui avons été si longtemps fixistes et religieux n'avons pas encore, me semble-t-il, complètement digéré le Darwinisme, et sommes toujours dans la phase d'exaltation – peut-être simplement parce que les enfants grandissent encore majoritairement avec ce point de vue fixiste si naturel, et ont donc la chance d'être un jour émerveillés par la découverte des concepts de l'évolution.

4.4. Est-ce bien raisonnable?

Il y a, dans le paragraphe précédent, l'aveu implicite que nous menons ces recherches pour "nous amuser", pour satisfaire nos questionnements personnels, et ceci aux frais du contribuable, ce qui paraît difficilement acceptable. Je m'empresse de nuancer cette interprétation en rappelant que (i) il n'y a pas de recherche innovante sans passion, et (ii) il n'y a pas de recherche appliquée sans recherche fondamentale, même si les liens peuvent sembler lointains – mieux comprendre comment le monde fonctionne est un objectif en soi, et tout progrès en ce sens a de bonnes chances de devenir utile un jour ou l'autre.

Je pense néanmoins qu'il y a dans notre discipline des risques de dérive, contre lesquels il n'est pas inutile d'essayer de lutter. On peut, par exemple, se laisser facilement enfermer dans un réductionnisme excessif. Ayant découvert un système biologique d'intérêt que l'on érige alors au rang de "système modèle", la facilité est de décortiquer à n'en plus finir ses propriétés, voire ses spécificités, qui peuvent, pour certaines, n'avoir qu'un intérêt très limité en dehors du système lui-même. Je pense par exemple, au risque de fâcher, qu'on est peut-être allé trop loin dans la génomique évolutive de *Drosophila melanogaster;* je préfèrerais disposer d'une image moins nette, mais dans plus d'espèces, des processus gouvernant l'usage des codons, ou de la relation polymorphisme/recombinaison. Un autre écueil potentiel me semble être l'élitisme théoricien. Compliquer encore et toujours un modèle pour prendre en compte le maximum d'effets, et ceci au prix de prouesses mathématiques ou informatiques qui font l'admiration des collègues, n'est pas toujours une bonne idée. Faisant cela, on spécialise en réalité son modèle, sans forcément augmenter sa pertinence vis-à-vis du monde réel. La difficulté technique ou théorique d'un projet ne mesure pas son intérêt biologique.

Une manière d'éviter de tels écueils est peut-être de se demander quelles informations seraient utiles aux collègues non-spécialistes pour qu'ils traitent au mieux leurs systèmes biologiques, interprètent au mieux leurs données. Cette remarque m'amène, et ce sera ma conclusion, à insister sur la mission pédagogique de la biologie évolutive. Peut-être serions-nous plus utiles, plutôt que d'approfondir encore notre compréhension des mécanismes les plus complexes et spécifiques, à enseigner les concepts de base de l'évolution à l'ensemble des biologistes. La biologie évolutive est traditionnellement associée à l'écologie, probablement parce que les écologistes ont rapidement eu besoin de la théorie de l'évolution pour comprendre la formation/prédire l'avenir des écosystèmes. Mais l'ensemble de la biologie gagnerait à incorporer les concepts de l'évolution, et la coupure (au moins en France) entre biologie moléculaire et cellulaire, d'une part, et biodiversité, d'autre part, me paraît regrettable et contre-productive. J'espère donc que les années à venir permettront de rapprocher ces points de vue, encouragé en cela par le récent redécoupage des sections du Comité National de la Recherche Scientifique, où le mot "évolution" apparaît dans le descriptif de trois sections différentes ("biodiversité", mais aussi "génomique" et "développement").

Références.

Adams K.L., Palmer J.D. 2003. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Mol. Phylogenet. Evol. 29:380-395.

Alexeyev M.F., Ledoux S.P., Wilson G.L. 2004. Mitochondrial DNA and aging. Clin. Sci. 107:355-364.

Alves P.C., Ferrand N., Suchentrunk F., Harris D.J. 2003. Ancient introgression of *Lepus timidus* mtDNA into *L. granatensis* and *L. europaeus* in the Iberian Peninsula. Mol. Phylogenet. Evol. 27:70-80.

Arndt P.F., Petrov D.A., Hwa T. 2003. Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation. Mol. Biol. Evol. 20:1887-1896.

Awadalla P., Eyre-Walker A., Maynard-Smith J.1999. Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science 286:2524-2525.

Baldauf S.L., Roger A.J., Wenk-Siefert I., Doolittle W.F. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science. 290:972-977.

Ballard J.W., Whitlock M.C.. 2004. The incomplete natural history of mitochondria.Mol. Ecol. 13:729-744.

Bapteste E., Boucher Y., Leigh J., Doolittle W.F. 2004. Phylogenetic reconstruction and lateral gene transfer. Trends Microbiol. 12:406-411.

Barja G., Herrero A. 2000. Oxidative damage to mitochondrial DNA is inversely related to maximum life span in the heart and brain of mammals. FASEB J. 14:312-318.

Barton N.H. 1995. Linkage and the limits to natural selection. Genetics 140:821-841.

Barton N.H. 2000. Genetic hitchhiking. Philos. Trans. R. Soc. Lond. B Biol. Sci. 355:1553-1562.

Bazin E. 2005. Etude du déterminisme de la diversité génétique des métazoaires par une approche bioinformatique. Thèse de doctorat. Université Montpellier 2.

Bazin E., Duret L., Penel S., Galtier N. 2005. Polymorphix, a polymorphism sequence data base. Nucleic Acids Res. 33:D481-D484.

Bazin E., Glémin S., Galtier N. 2006. Population size does not influence mitochondrial genetic diversity in animals. (soumis)

Beaumont M.A., Rannala B. 2004. The Bayesian revolution in genetics. Nat. Rev. Genet. 5:251-261.

Begun D.J., Aquadro C.F. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature. 356:519-20.

Belle E., Galtier N., Duret L., Eyre-Walker A. 2004. The decline of isochores in mammals: an assessment of the GC-content variation along the mammalian phylogeny. J. Mol. Evol. 58: 653-660.

Bernardi G. 1993. The vertebrate genome: isochores and evolution. Mol. Biol. Evol. 10:186-204.

Bernardi G., Olofsson B., Filipski J., Zerial M., Salinas J., Cuny G., Meunier-Rotival M., Rodier F. 1985. The mosaic genome of warm-blooded vertebrates. Science 228:953-958.

Bertrand S., Brunet F.G., Escriva H., Parmentier G., Laudet V., Robinson-Rechavi M. 2004. Evolutionary genomics of nuclear receptors: from twenty-five ancestral genes to derived endocrine systems. Mol. Biol. Evol. 21:1923-1937.

Biémont C., Nardon C., Decelière G., Lepetit D., Loevenbruck C., Vieira C. 2003. Worldwide distribution of transposable element copy number in natural populations of *Drosophila simulans*. Evolution 57:159-167.

Bierne N., Eyre-Walker A.2004. The genomic rate of adaptive amino acid substitution in *Drosophila*. Mol. Biol. Evol. 21:1350-1360.

Burger G., Gray M.W., Lang B.F. 2003. Mitochondrial genomes: anything goes. Trends Genet. 19:709-716.

Canback B., Andersson S.G., Kurland C.G. 2002. The global phylogeny of glycolytic enzymes. Proc. Natl. Acad. Sci. USA. 99:6097-6102.

Charlesworth B., Morgan M.T., Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289-1303.

Charlesworth D., Charlesworth B., Marais G. 2005. Steps in the evolution of heteromorphic sex chromosomes. Heredity 95:118-128.

Charlesworth D., Vekemans X., Castric V., Glemin S. 2005. Plant self-incompatibility systems: a molecular evolutionary perspective. New Phytol. 168:61-69.

Coop G., Griffiths R.C. 2004. Ancestral inference on gene trees under selection. Theor. Popul. Biol. 66:219-232.

Daley D.O., Clifton R., Whelan J. 2002. Intracellular gene transfer: reduced hydrophobicity facilitates gene transfer for subunit 2 of cytochrome c oxidase. Proc. Natl. Acad. Sci. USA. 99:10510-10515.

Danchin E., Vitiello V., Vienne A., Richard O., Gouret P., McDermott M.F., Pontarotti P. 2004. The major histocompatibility complex origin. Immunol. Rev. 198:216-232.

Daubin V., Lerat E., Perriere G. 2003. The source of laterally transferred genes in bacterial genomes. Genome Biol. 4:R57.

Davies T.J., Barraclough T.G., Chase M.W., Soltis P.S., Soltis D.E., Savolainen V. 2004. Darwin's abominable mystery: Insights from a supertree of the angiosperms. Proc. Natl. Acad. Sci. USA.101:1904-1909.

Dawson K.J., Belkhir K. 2001. A Bayesian approach to the identification of panmictic populations and the assignment of individuals. Genet Res. 78:59-77.

Delwiche C.F. 1999. Tracing the thread of plastid diversity through the tapestry of life. Am Nat. 154:164-177.

Depaulis F, Veuille M. 1995. Neutrality tests based on the distribution of haplotypes under an infinite-site model. Mol. Biol. Evol. 15:1788-1790.

De Stordeur E. 1997. Nonrandom partition of mitochondria in heteroplasmic *Drosophila*. Heredity 79:615-623.

Doiron S., Bernatchez L., Blier P. 2002. A comparative mitogenomic analysis of the potential adaptive value of Arctic charr mtDNA introgression in brook charr populations (*Salvelinus fontinalis* Mitchill). Mol. Biol. Evol. 19:1902-1909.

Douady C.J., Delsuc F., Boucher Y., Doolittle W.F., Douzery E.J. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol. Biol. Evol. 20:248-254.

Douzery E.J., Snell E.A., Bapteste E., Delsuc F., Philippe H. 2004. The timing of eukaryotic evolution: does a relaxed molecular clock reconcile proteins and fossils? Proc. Natl. Acad. Sci. USA 101:15386-15391.

Duret L., Semon M., Piganeau G., Mouchiroud D., Galtier N. 2002. Vanishing GC-rich isochores in mammalian genomes. Genetics 162: 1837-1847.

Dutheil J., Pupko T., Jean-Marie A., Galtier N. 2005. A model-based approach for detecting coevolving positions in a molecule. Mol. Biol. Evol. 22:1919-1928.

Dutheil J., Gaillard S., Bazin E., Glémin S., Ranwez V., Galtier N., Belkhir K. 2006. Bio++: a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. (soumis)

Eichinger L. et al. (97 auteurs) 2005. The genome of the social amoeba *Dictyostelium discoideum*. Nature 435:43-57.

Excoffier L., Estoup A., Cornuet J.M. 2005. Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. Genetics 169:1727-38.

Eyre-Walker A. 1999. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. Genetics 152:675-683.

Eyre-Walker A., Keightley P.D. 1999. High genomic deleterious mutation rates in hominids. Nature. 397:344-347.

Eyre-Walker A., Smith N.H., Maynard-Smith J. 1999. How clonal are human mitochondria? Proc. Roy. Soc. Sci. London B 266: 477-483.

Fay J.C., Wu C.I. 2000. Hitchhiking under positive Darwinian selection. Genetics 155:1405-1413.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368-376.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783-791.

Felsenstein J., Churchill G.A. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. Mol. Biol. Evol. 13:93-104.

Fitch W.M. 1971. Rate of change of concomitantly variable codons. J. Mol. Evol. 1:84-96.

Fitch W.M., Margoliash E. 1967. Construction of phylogenetic trees. Science 155:279-284.

Fu Y.X., Li W.H. 1993. Statistical tests of neutrality of mutations. Genetics 133:693-709.

Galtier N. 1997. L'approche statistique en phylogénie moléculaire: influence des compositions en bases variables. Thèse de doctorat, Université C. Bernard Lyon 1.

Galtier N. 2001. Maximum likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. 18:866-873.

Galtier N. 2003. Gene conversion drives GC-content evolution in mammalian histones. Trends Genet. 19: 65-68.

Galtier N. 2004a. Recombination, GC-content, and the human pseudoautosomal boundary paradox. Trends Genet. 20: 347-349.

Galtier N. 2004b. Sampling properties of the bootstrap support in molecular phylogeny: influence of non-independence between sites. Syst. Biol. 53: 38-46.

Galtier N., Gouy M. 1998. Inferring pattern and process : maximum likelihood implementation of a non-homogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. 15: 871-879.

Galtier N., Jean-Marie A. 2004. Markov-modulated Markov chains and the covarion process. J. Comp. Biol. 11: 727-733.

Galtier N., Tourasse N.J., Gouy M. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science 283: 220-221;

Galtier N., Depaulis F., Barton N.H. 2000. Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics 155:981-987.

Galtier N., Piganeau G., Mouchiroud D., Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159:907-911.

Galtier N., Gascuel O., Jean-Marie A. 2005. An introduction to Markov models in molecular evolution. in "Statistical Methods in Molecular Evolution", R. Nielsen Ed., Springer.

Galtier N., Bazin E., Bierne N. 2006a. GC-biased segregation of non-coding polymorphisms on Drosophila. Genetics (sous presse).

Galtier N., Enard D., Radondy Y., Bazin E., Belkhir K. 2006b. Mutation hot-spots in mammalian mitochondrial DNA. Genome. Res. (sous presse)

Gantenbein B., Fet V., Gantenbein-Ritter I.A., Balloux F. 2005. Evidence for recombination in scorpion mitochondrial DNA (Scorpiones: Buthidae). Proc. Roy. Soc. Sci. London B 272:697-704.

Garcia-Fernandez J. 2005. Hox, ParaHox, ProtoHox: facts and guesses. Heredity 94:145-152.

Gillespie J.H. 1991. The causes of molecular evolution. Oxford University Press.

Gillespie J.H. 2001. Is the population size of a species relevant to its evolution? Evolution 55:2161-2169.

Glinka S., Ometto L., Mousset S., Stephan W., De Lorenzo D. 2003. Demography and natural selection have shaped genetic variation in Drosophila melanogaster: a multi-locus approach. Genetics 165:1269-1278.

Goldman N., Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11:725-736.

Gray M.W., Burger G., Lang B.F. 2001. The origin and early evolution of mitochondria. Genome Biol. 2:REVIEWS1018

Griffiths R.C., Tavare S. 1995. Unrooted genealogical tree probabilities in the infinitely-many-sites model. Math. Biosci. 27:77-98.

Grossman L.I., Wildman D.E., Schmidt T.R., Goodman M. 2004. Accelerated evolution of the electron transport chain in anthropoid primates.Trends Genet. 20:578-585.

Guindon S., Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696-704.

Guldner E., Desmarais E., Galtier N., Godelle B. 2004a. Molecular evolution of plant hemoglobin: two hemoglobin genes in Nymphaeaceae *Euryale ferox*. J. Evol. Biol. 17: 48-54.

Guldner E., Godelle B., Galtier N. 2004b. Molecular adaptation in plant hemoglobin, a duplicated gene involved in plant-bacteria symbiosis. J. Mol. Evol. 59: 416-425.

Hagelberg E. 2003. Recombination or mutation rate heterogeneity? Implications for mitochondrial Eve. Trends Genet. 19:84-90.

Hagelberg E., Goldman N., Lio P., Whelan S., Schiefenhovel W., Clegg J.B., Bowden D.K. 1999. Evidence for mitochondrial DNA recombination in a human population of island Melanesia. Proc. Roy. Soc. Sci. London B 266:485-492.

Haldane J.B.S. 1927. A mathematical theory of natural and artificial selection, part V: selection and mutation. Proc. Camb. Philos. Soc. 28:838-844.

Harman D. 1956. Aging: a theory based on free radical and radiation chemistry. J. Gerontol.11:298-300.

Harman D. 1992. Role of free radicals in aging and disease. Ann. NY Acad. Sci. 673:126-41.

Hastings W.K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97-109.

Hedges D.J., Callinan P.A., Cordaux R., Xing J., Barnes E., Batzer M.A. 2004. Differential alu mobilization and polymorphism among the human and chimpanzee lineages. Genome Res. 14:1068-1075.

Herrnstadt C. et al. (11 auteurs) 2002. Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. Am. J. Hum. Genet. 70:1152-1171.

Holland L.Z., Laudet V., Schubert M. 2004. The chordate amphioxus: an emerging model organism for developmental biology. Cell. Mol. Life Sci. 61:2290-2308.

Hudson R.R., Kreitman M., Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153-159.

Huelsenbeck J.P., Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-755.

Huelsenbeck J.P., Larget B., Swofford D. 2000. A compound poisson process for relaxing the molecular clock. Genetics 154:1879-1892.

Hughes S., Zelus D., Mouchiroud D. 1999. Warm-blooded isochore structure in Nile crocodile and turtle. Mol. Biol. Evol. 16:1521-1527.

Hurst G.D., Jiggins F.M. 2005. Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. Proc. Roy. Soc. Sci. London B 272:1525-1534.

Ingman M., Kaessmann H., Paabo S., Gyllensten U. 2000. Mitochondrial genome variation and the origin of modern humans. Nature. 408:708-713.

Innan H., Nordborg M. 2002. Recombination or mutational hot spots in human mtDNA? Mol. Biol. Evol. 19:1122-1127.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.

Irish V.F., Litt A. 2005. Flower development and evolution: gene duplication, diversification and redeployment. Curr. Opin. Genet. Dev. 15:454-60.

James A.C., Ballard J.W. 2003. Mitochondrial genotype affects fitness in *Drosophila simulans*. Genetics 164:187-194.

Jaillon et al. (61 auteurs) 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431:946-957.

Kern A.D., Kondrashov F.A. 2004. Mechanisms and convergence of compensatory evolution in mammalian mitochondrial tRNAs. Nat. Genet. 36:1207-1212.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotides sequences. J. Mol. Evol. 16:111.120.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge University Press.

King-Jordan I., Kondrashov F.A., Adzhubei I.A., Wolf Y.I., Koonin E.V., Kondrashov A.S., Sunyaev S. 2005. A universal trend of amino acid gain and loss in protein evolution. Nature. 433:633-638.

Kolaczkowski B., Thornton J.W. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980-984

Kong X. et al. (16 auteurs) 2002. A high-resolution recombination map of the human genome. Nat. Genet. 31:241-247.

Kraytsberg Y., Schwartz M., Brown T.A., Ebralidse K., Kunz W.S., Clayton D.A., Vissing J., Khrapko K. 2004. Recombination of human mitochondrial DNA. Science 30:981.

Kuhner M.K., Yamato J., Felsenstein J. 2000. Maximum likelihood estimation of recombination rates from population data. Genetics 156:1393-1401.

Lahn B.T., Page D.C. 1999. Four evolutionary strata on the human X chromosome. Science 286:964-967.

Lartillot N., Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095-1109.

Lercher M.J., Smith N.G., Eyre-Walker A., Hurst L.D. 2002.The evolution of isochores: evidence from SNP frequency distributions. Genetics 162:1805-1810.

Lopez P., Casane D., Philippe H. 2002. Heterotachy, an important process of protein evolution. Mol. Biol. Evol. 19:1-7.

Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. Trends Genet. 19:330-338.

Marais G., Galtier N. 2003. Sex chromosomes: how X-Y recombination stops. Curr. Biol. 13: 641-643.

Martin W., Hermann R.G. 1998. Gene transfer from organelles to the nucleus: how much, what happens and why? Plant Physiol. 118:9-17.

Maynard-Smith J., Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res. 23:23-35.

McDonald J.H., Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351:652-654.

Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., Teller E. 1953. Equations of state calculations by fast computing machines. J. Chem. Phys. 21:1087-1091.

Meunier J., Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. Mol. Biol. Evol. 21:984-990.

Montoya-Burgos J.I., Boursot P., Galtier N. 2003. Recombination explains isochores in mammalian genomes. Trends Genet. 19: 128-130.

Murphy W.J., Pevzner P.A., O'Brien S.J. 2004. Mammalian phylogenomics comes of age. Trends Genet. 20:631-639.

Neuhauser C., Krone S.M. 1997. The genealogy of samples in models with selection. Genetics 145:519-534.

Nevo E., Beiles A., Ben-Shlomo R. 1984. The evolutionary significance of genetic diversity: ecological, demographic and life history correlates. Pp 13-213 in Evolutionary dynamics of genetic diversity, G.S. Mani Ed., Springer-Verlag.

Nielsen R. et al. (13 auteurs) 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 3:e170.

Nordborg M. et al. (24 auteurs) 2005 The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol. 3:e196.

Perry J., Ashworth A. 1999. Evolutionary rate of a gene affected by chromosomal position. Curr Biol. 9:987-989.

Piganeau G., Eyre-Walker A. 2003. Estimating the distribution of fitness effects from DNA sequence data: implications for the molecular clock. Proc. Natl. Acad. Sci. U S A. 100:10335-10340.

Piganeau G., Gardner M., Eyre-Walker A. 2004. A broad survey of recombination in animal mitochondria. Mol. Biol. Evol. 21:2319-2325.

Pollock D.D., Taylor W.R., Goldman N. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. J. Mol. Biol. 287:187-198.

Pont-Kingdon G., Okada N.A., Macfarlane J.L., Beagley C.T., Watkins-Sims C.D., Cavalier-Smith T., Clark-Walker G.D., Wolstenholme D.R. 1998. Mitochondrial DNA of the coral *Sarcophyton glaucum* contains a gene for a homologue of bacterial MutS: a possible case of gene transfer from the nucleus to the mitochondrion. J. Mol. Evol. 6:419-431.

Popot J.L., de Vitry C. 1990. On the microassembly of integral membrane proteins. Annu. Rev. Biophys. Biophys. Chem. 19:369-403.

Pritchard J.K., Stephens M., Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945-959.

Pupko T., Galtier N. 2002. A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genome. Proc. Roy. Soc. Sci. London B 269: 1313-1316.

Race H.L., Herrmann R.G., Martin W1999. Why have organelles retained genomes? Trends Genet. 15:364-370.

Reboud X., Zeyl C. 1994. Organelle inheritance in plants. J. Hered. 72:132–140.

Richard G.F., Kerrest A., Lafontaine I., Dujon B. 2005. Comparative genomics of hemiascomycete yeasts: genes involved in DNA replication, repair, and recombination.Mol. Biol. Evol. 22:1011-1023.

Robinson D.M., Jones D.T., Kishino H., Goldman N., Thorne J.L. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol. Biol. Evol. 20(:1692-1704.

Rosenberg N.A., Nordborg M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. 3:380-390.

Ruiz-Pesini E., Mishmar D., Brandon M., Procaccio V., Wallace D.C. 2004. Effects of purifying and adaptive selection on regional variation in human mtDNA. Science 303:223-226.

Schlotterer C. 2002. Towards a molecular characterization of adaptation in local populations. Curr Opin Genet Dev. 12:683-687.

Schwartz M., Vissing J. 2002. Paternal inheritance of mitochondrial DNA. N. Engl. J. Med. 347:576-580.

Semon M., Mouchiroud D., Duret L. 2005. Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance. Hum. Mol. Genet.14:421-427

Shimizu K.K. et al. (11 auteurs) 2004. Darwinian selection on a selfing locus. Science 306: 2081-2084.

Smith N.G., Eyre-Walker A. 2001. Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. Mol. Biol. Evol. 18:982-986.

Sober E. 2004. The contest between parsimony and likelihood. Syst. Biol. 53:644-653.

Stamatakis A., Ludwig T., Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456-463.

Stedman H.H., Kozyak B.W., Nelson A., Thesier D.M., Su L.T., Low D.W., Bridges C.R., Shrager J.B., Minugh-Purvis N., Mitchell M.A. 2004. Myosin gene mutation correlates with anatomical changes in the human lineage. Nature 428:415-418.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 123:585-595.

Tamura K. 1992. Estimating the number of nucleotide substitutions when there are strong transition-transversion and GC-content biases. Mol. Biol. Evol. 9:678-687.

Tanaka M., Gong J.S., Zhang J., Yoneda M., Yagi K. 1998. Mitochondrial genotype associated with longevity. Lancet 351:185-186.

Tenaillon M.I., U'Ren J., Tenaillon O., Gaut B.S. 2004. Selection versus demography: a multilocus investigation of the domestication process in maize. Mol. Biol. Evol. 21:1214-1225.

Thorne J.L., Kishino H. 2002. Divergence time and evolutionary rate estimation with multilocus data. Syst. Biol. 51:689-702.

Tringe S.G. et al. (13 auteurs) 2005. Comparative metagenomics of microbial communities. Science 308:554-557.

Tufféry P., Darlu P. 2000. Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. Mol. Biol. Evol. 17:1753-1759.

Tuffley C., Steel M.A. Modelling the covarion hypothesis of nucleotide substitution, Math. Biosci. 147:63-91.

Wolfe K.H., Li W.H., Sharp P.M. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. USA. 84:9054-9058.

Yang J., Lusk R., Li W.H. 2003. Organismal complexity, protein complexity, and gene duplicability. Proc. Natl. Acad. Sci. USA. 100:15661-15665.

Yang Z. 1994. Maximum-likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306-314.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15: 568-573.

Yang Z., Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19:908-917.

Yang Z., Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Mol. Biol. Evol. 14:717-724.

Yang Z., Nielsen R., Goldman N., Pedersen A.-M. K. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155: 431-449.

Yi S., Summers T.J., Pearson N.M., Li W.H. 2004. Recombination has little effect on the rate of sequence divergence in pseudoautosomal boundary 1 among humans and great apes. Genome Res. 14:37-43.

Zuckerkandl E., Pauling L. 1965. Molecules as documents of evolutionary history. J. Theor. Biol. 8:357-368.

Curriculum vitae.

Né le 10 octobre 1970, Français, marié, deux enfants.

Parcours universitaire

- Thèse, 1997, Biométrie (phylogénie moléculaire), Université Lyon 1
  Directeur: Manolo Gouy
  Jury: F. Bonhomme, O. Gascuel, M. Veuille, C. Gautier, M. Gouy
- Magistère, 1993, Biologie Moléculaire et Cellulaire, ENS Lyon

Parcours professionnel

- 1999 — , Chargé de Recherches CNRS, UMR 5171, Montpellier, France.
- 1998 — 1999, Post-doctorant, ICAPB, Edinburgh (N. Barton & B. Charlesworth)
- 1997 — 1998, Post-doctorant, UMR 5000, Montpellier (F. Bonhomme)

Publications

Articles de revues à comité de lecture

1. **Galtier N.** & Gouy M. 1994. Eubacterial phylogeny: a new multiple-tree analysis method applied to 15 sequence data sets questions the monophyly of Gram-positive bacteria. ***Research in Microbiology*** **145:** 531-541
2. **Galtier N**. & Gouy M. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. ***Proceedings of the National Academy of Science USA*** **92:** 11317-11321
3. **Galtier N**., Gouy M., & Gautier C. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. ***Computer Applications for Biosciences*** **12 :** 543-548
4. **Galtier N.** & Lobry J. 1997. Relationships between genomic G+C content, RNA secondary structures and optimal growth temperature in prokaryotes. ***Journal of Molecular Evolution*** **44:** 632-636
5. Brito R.M., Briolay J., **Galtier N**., Bouvet Y., & Coelho M.M. 1997. Phylogenetic relationships within genus Leuciscus (Pisces:Cyprinidae) in portuguese freshwater based on mitochondrial DNA cytochrome *b* sequences. ***Molecular Phylogenetics and Evolution*** **8:** 435-442
6. **Galtier N.** & Gouy M. 1998. Inferring pattern and process : maximum likelihood implementation of a non-homogeneous model of DNA sequence evolution for phylogenetic analysis. ***Molecular Biology and Evolution*** **15:** 871-879
7. **Galtier N.** & Mouchiroud D. 1998. Evolution of isochores in mammals: a human-like ancestral pattern. ***Genetics*** **150:** 1577-1584
8. Briolay J., **Galtier N**., Brito R.M., & Bouvet Y. 1998. Molecular phylogeny of Cyprinidae inferred from *cytochrome b* DNA sequences. ***Molecular Phylogenetics and Evolution*** **9:** 100-108
9. **Galtier N**., Tourasse N.J. & Gouy M. 1999. A nonhyperthermophilic common ancestor to extant life forms. ***Science*** **283:** 220-221

10. Chenuil A., **Galtier N**., & Berrebi P. 1999. A test of the hypothesis of an autopolyploid *versus* allopolyploid origin for a tetraploid lineage. Application to the genus *Barbus* (Cyprinidae). ***Heredity*** **82:**373-380
11. **Galtier N.** & Boursot P. 2000. A new method to locate nucleotide changes in a tree reveals unequal polymorphism *vs* divergence patterns in mouse mitochondrial DNA. ***Journal of Molecular Evolution*** **50:**224-231
12. **Galtier N.**, Depaulis F., & Barton N.H. 2000. Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. ***Genetics*** **155:**981-987
13. Duret L. & **Galtier N.** 2000. The covariation between TpA deficiency, CpG deficiency and the G+C-content of human isochores is a mathematical artefact. ***Molecular Biology and Evolution*** **17:** 1620-1625
14. **Galtier N.** 2001. Maximum likelihood phylogenetic analysis under a covarion-like model. ***Molecular Biology and Evolution*** **18:**866-873
15. Arnaud S., Monteforte M., **Galtier N.**, Bonhomme F. & Blanc F. 2001. Population structure and genetic variability of pearl oyster *Pinctada mazatlanica* along Pacific coasts from Mexico to Panama. ***Conservation Genetics*** **1:** 299-308
16. **Galtier N.**, Piganeau G., Mouchiroud D. & Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. ***Genetics*** **159:**907-911
17. Dutheil J. & **Galtier N.** 2002. BAOBAB: a java editor for large phylogenetic trees. ***Bioinformatics*** **18:** 892-893
18. Pupko T. & **Galtier N.** 2002. A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genome. ***Proceedings of the Royal Society of Science London B*** **269:** 1313-1316
19. Duret L., Semon M., Piganeau G., Mouchiroud D. & **Galtier N.** 2002. Vanishing GC-rich isochores in mammalian genomes. ***Genetics*** **162:** 1837-1847
20. **Galtier N.** 2003. Gene conversion drives GC-content evolution in mammalian histones. ***Trends in Genetics*** **19:** 65-68.
21. Montoya-Burgos J.I., Boursot P. & **Galtier N.** 2003. Recombination explains isochores in mammalian genomes. ***Trends in Genetics*** **19:** 128-130.
22. Ruault M., Ventura M., **Galtier N.**, Brun M.E., Archidiacono M., Roizès G. & De Sario A. 2003. *BAGE* genes, which were generated by juxtacentromeric reshuffling in the hominidae lineage, are under selective pressure. ***Genomics*** **81:** 391-399.
23. Angers B., Charbonnel N., **Galtier N.** & Jarne P. 2003. The influence of demography, population structure and selection on molecular diversity in the selfish freshwater snail *Biomphalaria pfeifferi*. ***Genetical Research*** **81:** 19-204.
24. Marais G. & **Galtier N**. 2003. Sex chromosomes: how X-Y recombination stops. ***Current Biology*** **13:** 641-643.
25. Guldner E., Desmarais E., **Galtier N.** & Godelle B. 2004. Molecular evolution of plant hemoglobin: two hemoglobin genes in Nymphaeaceae *Euryale ferox*. ***Journal of Evolutionary Biology*** **17:** 48-54.
26. **Galtier N.** 2004. Sampling properties of the bootstrap support in molecular phylogeny: influence of non-independence between sites. ***Systematic Biology*** **53:** 38-46.
27. Belle E., **Galtier N.**, Duret L. & Eyre-Walker A. 2004. The decline of isochores in mammals: an assessment of the GC-content variation along the mammalian phylogeny. ***Journal of Molecular Evolution*** **58:** 653-660.
28. **Galtier N.** & Jean-Marie A. 2004. Markov-modulated Markov chains and the covarion process. ***Journal of Computational Biology*** **11:** 727-733.
29. **Galtier N**., Bonhomme F., Moulia C., Belkhir K., Caminade P., Desmarais E., Duquesne J.J., Orth A., Dod B. & Boursot P. 2004. Mouse biodiversity in the genomic era. ***Cytogenetics and Genome Research*** **105:** 385-394.

30. Guldner E., Godelle B. & **Galtier N.** 2004. Molecular adaptation in plant hemoglobin, a duplicated gene involved in plant-bacteria symbiosis. ***Journal of Molecular Evolution*** 59: 416-425.
31. **Galtier N.** 2004. Recombination, GC-content, and the human pseudoautosomal boundary paradox. ***Trends in Genetics*** **20:** 347-349.
32. Bazin E., Duret L., Penel S. & **Galtier N.** 2005. Polymorphix: a sequence polymorphism data base. ***Nucleic Acids Research*** **33:** 481-484.
33. Dutheil J., Pupko T., Jean-Marie A. & **Galtier N.** 2005. A model-based approach for detecting coevolving positions in a molecule. ***Molecular Biology and Evolution*** **22:** 1919-1928.
34. **Galtier N**., Bazin E. & Bierne N. 2006. GC-biased segregation of non-coding polymorphisms in Drosophila. ***Genetics*** (sous presse)
35. **Galtier N.**, Enard D., Radondy Y., Bazin E. & Belkhir K. 2006. Mutation hot spots in mammalian mtDNA. ***Genome Research*** (sous presse)

Chapitres de livres

1. **Galtier N.** 2002. The statistical approach to molecular phylogeny: evidence for a nonhyperthermophilic common ancestor. Pp 111-121 in "Biological Evolution and Statistical Physics", M. Lässig & A. Valleriani, Eds., Springer
2. **Galtier N.**, Montoya-Burgos J.I., Mouchiroud D., Bonhomme F., Boursot P. and Duret L. 2003. Recombination, gene conversion, and the evolution of isochores in mammalian genomes. Pp 43-54 in "Recent Research Developments in Genetics", Research Signpost Ed., Kerala, India.
3. **Galtier N.**, Gascuel O. & Jean-Marie A. 2005. An introduction to Markov models in molecular evolution. Pp 3-24 in "Statistical Methods in Molecular Evolution", R. Nielsen, Ed., Springer
4. Bryant D., **Galtier N.** & Poursat M.A. 2005. Likelihood calculation in molecular phylogeny. in "Mathematics of Evolution and Phylogeny", O. Gascuel, Ed., Oxford University Press

Conférences sur invitation

Galtier N., Mouchiroud D., Tourasse N.J. & Gouy M. 1998. Modélisation et inférence en phylogénie moléculaire: applications d'un modèle non-homogène. *VIe Rencontres de la Société Francophone de Classification*. Montpellier.
Galtier N. 1999. Non-stationary models of nucleotide substitution and the evolution of base composition. *Colloque de la Société Européenne de Biologie Evolutive*, Barcelone, Espagne.
Galtier N. 1999. Coalescent and the reconstruction of population history: bottlenecks *vs* selective sweeps. *Rencontres "Algorithmique et Biologie"*, Institut Pasteur.
Galtier N., Tourasse N. & Gouy M. 2000. A nonstationary model of DNA sequence evolution: evidence for a non-hyperthermophilic common ancestor. *Conference "Biological Evolution and Statistical Physics"*, Dresde, Allemagne.
Galtier N. 2000. Evolution des génomes dans les populations: l'exemple de la composition en bases. "*Séminaires du Jeudi" des Sciences de la Vie du C.N.R.S.,* Montpellier.
Galtier N. 2000. Molecular evidence for a non-hyperthermophilic common ancestor to extant life forms. *Gordon Conference "Origin of life"*, Plymouth (New Hampshire), USA.
Galtier N. 2001. Evolution moléculaire et coalescence: détection des effets de la sélection naturelle dans les génomes. "*Journée de la Coalescence*", Montpellier

Galtier N. 2001. L'approche statistique en phylogénie moléculaire. *Journée "Applications et méthodes de la phylogénie moléculaire" de l'IFR 41*, Lyon.
Galtier N. 2002. The statistical approach to molecular phylogeny: improved models. *International School of Computational Biology 2002*, Le Havre.
Galtier N. 2003. The statistical approach to molecular phylogeny: improved models. *Conference "Mathematics of Evolution and Phylogeny"*, Institut Henri Poincaré, Paris.
Galtier N. 2004. Biased gene conversion, or the end of a neutral paradigm. *Seminars of the Bioinformatics Research Center*, Aarhus, Danemark.
Dutheil J. & **Galtier N.** 2005. Statistical approaches to the substitution mapping problem and the detection of coevolution between sites. Conference "*Using ancestral sequence reconstructions to understand protein function*", Kristineberg, Suède.
Galtier N. 2005. Population size does not influence mitochondrial genetic diversity in animals. *6th Anton Dohrn Workshop*, Ischia, Italie.
Galtier N. 2005. Population size does not influence mitochondrial genetic diversity in animals. *Seminars of the Department of Biology*, University College London, Royaume-Uni

Distinctions

- 1998, Prix Simon Régnier du jeune chercheur (Société Française de Classification)
- 1999, Prix Maynard-Smith du jeune chercheur (European Society for Evolutionary Biology)
- 2004, médaille de bronze du CNRS

Edition scientifique

- Editeur associé pour *Journal of Molecular Evolution*, 2003 —
- Evaluation de projets de recherche:
  ACI IMPBio, BBSRC, ECOS-Sud, Israel Science Foundation
- Relecteur pour les revues
  Appl. Env. Microbiol., *Am. J. Hum. Genet*., *Bioessays*, *Bioinformatics*, *Biol. J. Linn. Soc.*, *BMC Evol. Biol.*, *BMC Genomics*, *Curr. Biol.*, *Evolution*, *Evol. Bioinform*., *Gene*, *Genet. Sel. Evol*., *Genetica*, *Genetics*, *Heredity*, *J. Evol. Biol.*, *J. Mol. Evol*., *J. Theor. Biol*., *Mol. Biol. Evol*., *Mol. Ecol*., *Mol. Phyl. Evol*., *Parasite*, *Proc. Nat. Acad. Sci. USA*, *Proc. Roy. Soc. Lond. B*., *Syst. Biol*., *Trends Ecol. Evol*., *Trends Genet*.
- Membre de la *FACULTY OF 1000*, un service web de revue de presse scientifique en biologie, 2004 —

Encadrement

Post-docs

- Juan Montoya-Burgos, Financement Ministère de la Recherche, 2000 — 2001
- Cynthia Steiner, Financement ACI, 2003

Etudiants en thèse

- Benoît Nabholz (2004 — , DEA et thèse, 50%) "Déterminisme génétique, écologique et physiologique du taux de mutation mitochondrial chez les Mammifères: une approche phylogénétique"
- Julien Dutheil (2002 — , DEA et thèse, 100%) "Modèles de Markov en phylogénie moléculaire, coévolution, covarion"
- Eric Bazin (2001-2005 , DEA et thèse, 100%) "Etude du déterminisme de la diversité génétique des métazoaires par une approche bioinformatique"
- Emilie Guldner (2000 – 2003, DEA et thèse 50%) "Adaptation moléculaire de l'hémoglobine chez les plantes"

Autres étudiants de DEA/DESS/M2

- Adeline Césaro (2000, DEA, 50%) "Evolution moléculaire du locus *Fxy* chez la souris"
- Alban Guillaumet (2001, DEA, 50%) "Séparation géographique hivernale et coexistence chez les oiseaux migrateurs: le cas des Sylvidés"
- Sylvain Gaillard (2004, DESS, 50%) "PopGenLib, une bibliothèque C++ pour la génétique des populations"
- Philippe Gayral (2005, DEA, 50%) "Evolution des gènes après duplication: approche microévolutive chez la souris"
- Yoan Radondy (2006, M2Pro, 100%) "Co-évolution et motifs structuraux dans les protéines."

Participation à des jurys de thèse

- Gwenaël Piganeau (2001, Lyon, examinateur)
- Gabriel Marais (2002, Lyon, examinateur)
- Vincent Ranwez (2002, Montpellier, examinateur)
- Frédéric Delsuc (2002, Montpellier, examinateur)
- Elodie Gazave (2003, Montpellier, examinateur)
- Mohamad Hassan (2003, Montpellier, examinateur)
- Emilie Guldner (2003, Montpellier, co-directeur)
- Raphaël Leblois (2004, Montpellier, examinateur)
- Julien Meunier (2005, Lyon, examinateur)
- Eric Bazin (2005, Montpellier, (co-)directeur)
- Alexandra Calteau (2005, Lyon, Examinateur)

Enseignement Universitaire

1995-1997: Moniteur à l'Université Lyon 1

Depuis 1999: (horaire réel, année de début de l'année scolaire)
- Maîtrise, Université Montpellier 2, "Phylogénie moléculaire: approche statistique"
  2000-2005: 3h
- Maîtrise/M1, Université Montpellier 2, "Génétique moléculaire des populations"
  2002-2005: 3h

- DESS Bioinformatique: "Modèles de l'évolution des séquences" + TP phylo_win
  2001-2002: 4h, 2003-2004: 2h
- DEA de Parasitologie, Université Montpellier 2, "Evolution moléculaire"
  2000: 2h
- DEA/M2R BEE, Université Montpellier 2, "Evolution moléculaire"
  1999-2001: 2 à 3h, 2002: 6h, 2003-2005: 3h
- DEA RPI, Université Montpellier 2, "Evolution moléculaire"
  2001-2002: 3h, 2003-2004: 6h
- M2R Biométrie, Université Lyon 1, "Théorie de la coalescence"
  2004: 3h
- DEA d'Informatique, Université Montpellier 2, "Méthodes statistiques pour la phylogénie"
  1999: 2h
- Module optionnel "Biométrie" (ENSAM), "Phylogénie moléculaire"
  2002-2005: 3h
- Ecole Doctorale Biologie Intégrative, Montpellier, "Initiation à la programmation"
  2000: 12h d'enseignement, et organisation du module.
- Ecole Doctorale CBS2, Montpellier, "Bioinformatique pour la génomique évolutive"
  2003: 2h, 2005:2h
- Module doctoral , Université de Genève, "Méhodes pour la phylogénie"
  2000: 3h, 2002: 3h