Master de Sciences et Technologies
Mention Biologie Moléculaire et Cellulaire
Spécialité Immunologie

Module optionnel X - Génomique/Protéomique (6 ECTS)

Bilan des Conférences proposées

réalisé par les étudiants d'IT2004 ayant suivi dans ce module

Lieu de l'enseignement :
Centre des Enseignements Institut Pasteur, 25-28 rue du Docteur Roux, 75015 Paris

La génomique se définit comme l'étude exhaustive et multidisciplinaire des génomes, tant en ce qui concerne le nombre et la disposition précise des gènes sur les chromosomes, que leur séquence et leur fonction. La discipline est en passe de modifier de façon très profonde les approches scientifiques visant à étudier les fonctions biologiques des êtres vivants, en nous livrant les clés du déterminisme des grands processus cellulaires et de leur intégration au niveau des organismes entiers. Les projets de génomiques devraient également aboutir à éclairer notre compréhension des principes qui sous-tendent les mécanismes de développement, d'adaptation au milieu et de filiation évolutive par l'intermédiaire notamment de l'établissement de nouveaux modèles d'études plus aboutis. En dehors de la vision purement fondamentale, la génomique, et en particulier les études génomiques à grande échelle, devrait être à l'origine de progrès décisifs (pour certains déjà d'actualité) en médecine, dans l'univers agro-alimentaire ou encore dans l'industrie.

En quoi consiste la génomique?

En premier lieu, la discipline vise à séquencer les génomes ou fragments de génomes d'espèces d'intérêt et tente ensuite de les cartographier, d'identifier des mécanismes ou des gènes candidats, d'élaborer des modèles d'études adaptés à la problématique, de déterminer la structure et la fonction des gènes d'intérêts (activité, domaines, partenaires), d'établir les profils d'expression des gènes …

Pour cela, la génomique s'appuie en fonction de la thématique de l'étude sur un ensemble de disciplines et de compétences très variées : génétique moléculaire et humaine, protéomique, biologie moléculaire et cellulaire, biochimie, biophysique, bioinformatique, biologie et physiologie des organismes, physiologies végétale et animale, immunologie … Elle s'appuie également sur de nombreux procédés : méthodes de séquençages, PCR, utilisation de modèles informatiques, spectrométrie de masse

Comment se fait une étude de génomique ?

Un projet de génomique ne peut être mené si la séquence du génome de l'organisme d'intérêt n'est pas connue. Le séquençage peut donc en constituer la première étape. Sa durée dépend alors essentiellement de la taille du génome et de sa complexité. Dans un deuxième temps, un énorme travail bioinformatique est réalisé afin d'annoter le génome, de le caractériser, de prédire la position des séquences d'intérêt pour au final disposer d'une base de données fiable. Ces deux premières étapes peuvent être considérablement longues si bien que de tels projets entiers durent généralement de 2 à 8 ans, surtout si la version préliminaire ne permet pas une analyse suffisamment fine pour l'étude.

Une fois la base de données au point, les approches les plus variées sont possibles pour mettre en évidence un mécanisme ou pour découvrir des gènes candidats dans un contexte précis. Il s'agit à ce moment d'un travail relativement rapide (beaucoup plus que les approches classiques de génétiques comme le clonage positionnel par exemple) qui est basé le plus souvent sur des études comparatives inter-espèces et un gros effort de screening par analogie de séquence ou de fonction.

Si l'approche du sujet s'est avérée pertinente, il est possible de dégager un petit nombre d'hypothèses issues des analyses *in silico* ; hypothèses qui doivent par la suite être confirmées ou infirmées par des méthodes plus classiques de biologie moléculaire, de génétique, de biologie cellulaire, … L'approche génomique s'avère donc très utile dans la mise en évidence de pistes de recherches et d'hypothèses de travail, le tout en un minimum de temps de part la puissance de l'outil informatique. Cependant, la réussite d'une telle étude repose grandement sur la pertinence de la question scientifique posée initialement et qui permet de définir les bornes de recherche *in silico*.

L'établissement d'un ou plusieurs modèles est généralement l'objectif final, celui-ci visant à regrouper l'ensemble des informations pertinentes apportées par l'étude et l'ensemble des connaissances préalables sur le sujet.

Quel peut être l'apport de la génomique à l'immunologie et aux immunotechnologies ?

La seule limite véritable de l'approche génomique est la dépendance vis-à-vis de la séquence des génomes d'intérêts. Ce problème mis à part, il est possible de mener des études dans des domaines très variées et certains domaines de l'immunologie s'y prêtent tout particulièrement :

- Immunité anti-tumorale
- Dérèglements du système immunitaire
- Immunité anti-bactérienne, anti-virale, anti-infectieuse
- Etude de la recombinaison des gènes des immunoglobulines
- Identification de nouvelles cibles thérapeutiques …

Mais la génomique, n'est pas utilisée uniquement pour les maladies génétiques, toute pathologie implique à un moment ou à un autre les gènes et leur expression :
- Prédisposition transmise (maladies héréditaires comme la dystrophie musculaire, certains cancers), MAIS AUSSI:
- Problèmes de réparation / mutation somatique (les modifications permettant le développement et le maintien des cancers ont leur origine dans l'ADN).
- Problèmes d'expression (tous les dysfonctionnements se traduisent tôt ou tard par une modification des profils d'expression des gènes.).
Même dans le cas des maladies infectieuses, la génomique peut nous apporter beaucoup:
- Gènes microbiens impliqués dans la pathogénicité (par ex. gouvernant la spécificité d'infection)
- Gènes impliqués dans les résistances aux antibiotiques.

Les applications de la génomique sont donc diverses. L'utilisation la plus courante pour l'industrie est la possibilité d'identifier :
- Les gènes impliqués dans les pathologies: cibles pharmaceutiques ou marqueurs diagnostiques
- De nouveaux gènes permettant de synthétiser des molécules d'intérêt
- Des gènes responsables de résistances (microbiologie ou agronomie)
- Des gènes permettant de mieux comprendre des mécanismes-clés: cancer, vieillissement, etc.

La génomique comparative à grande échelle est un dérivé de la génomique classique ce qui permet la comparaison de génomes entièrement séquencés et donc :
- Aide l'annotation en identifiant les régions fonctionnelles •Identifier le jeu de gènes de chaque organisme
- Comprendre les solutions trouvées par des organismes différents pour une même fonction
- Etudier des gènes/fonctions particuliers par comparaison de séquence (voir cours de bioinformatique)
- Autres questions spécifiques: adaptation, résistance, pathogénicité, etc.

Enfin, un autre aspect est étudié par la génomique à savoir: les variations du génome au sein d'une même population. Ceci est très important médicalement parlant pour tout ce qui est Pharmacogénomique: comment chaque patient répond aux drogues et pour définir les Marqueurs de susceptibilité aux maladies, et ceci du fait du Polymorphismes dans le génome humain (Insertions, délétions, duplications, réarrangements, …).

Protéomique d'expression : méthodes, performances et limites

Thierry RABILLOUD

La protéomique a pour objectifs d'identifier et de quantifier l'ensemble des protéines intervenant dans un contexte cellulaire donné. La discipline tente également de mettre en évidence les interactions protéiques clés impliquées dans le phénomène étudié. Si l'on peut considérer le génome comme la « mémoire morte » d'un organisme aux mutations somatiques près, le protéome, tout comme le transcriptome, varie énormément d'un type cellulaire à l'autre selon la situation physio-pathologique et le moment considéré. Ainsi, un même individu ne possède pas un protéome, mais des protéomes dont le nombre égale celui des contextes biologiques rencontrés par chaque cellule. L'approche protéomique vise donc à caractériser des mécanismes moléculaires en fonction de ces différents contextes, et ce, en employant principalement deux types de méthodes :

- L' Electrophorèse bidimensionnelle : séparation des protéines

Cette méthode permet la séparation sur gel des protéines issues d'extraits cellulaires complexes en se basant sur les différences de masse et de charge de ces protéines, puis leur quantification par analyse informatique des spots obtenus. C'est actuellement la seule méthode fiable pour estimer l'abondance et les modifications post-traductionnelles d'un grand nombre de protéines en parallèle. Ces dernières années, la résolution et la reproductibilité ont été considérablement améliorées et des logiciels permettant l'analyse numérique ont été développés permettant ainsi la séparation de plusieurs centaines de protéines en une seule opération.

- La Spectrométrie de masse : analyse des protéines

Il en existe deux variantes principales : MALDI-TOF et ESI/MS/MS. Ces méthodes assurent l'identification de protéines provenant de spots d'intérêt en mesurant la masse des peptides issus de la digestion enzymatique de ces protéines. L'identification se fait par comparaison du spectre peptidique obtenu à des spectres-référence stockés dans des bases de données. Ces bases de données permettent de prédire la masse théorique des fragments de protéolyse d'une protéine. Cette approche, bien que très puissante, reste cependant assez lourde, chère de part l'appareillage requis, et dépendante des banques de données donc de la bonne caractérisation des génomes références.

Les méthodes actuelles d'analyse des protéines n'ont pas la capacité de haut débit, ni la reproductibilité et l'automatisation nécessaires pour suivre le rythme imposé par la génomique à grande échelle. Pour combler ces lacunes, il sera indispensable d'optimiser et d'améliorer l'aspect analytique pour gagner en temps, en finesse d'analyse et pour résoudre certains biais physicochimiques observés dans certains cas (protéines membranaires …).

Il est possible d'envisager certaines applications de la protéomique à l'immunologie. En effet, l'approche protéomique par profils d'expression devrait permettre de constituer des modèles d'intérêt très utiles pour l'immunologie. Ainsi, l'analyse de ces modèles et de leurs altérations dans différents contextes cellulaires pourrait permettre d'accélérer l'identification de nouveaux mécanismes d'interactions moléculaires dans des domaines aussi variés que la cancérogenèse, l'infectiologie ou l'immunologie fondamentale. A titre d'exemple, il devrait être envisageable d'analyser les modifications du protéome chez un patient avant et après traitement afin d'identifier l'ensemble des protéines impliquées dans le phénomène pathologique ou dans la restauration du phénotype et de suivre leur évolution dans le temps.

Génomique structurale et haut débit

Arnaud DUCRUIX

La génomique structurale est une discipline récente visant à attribuer une structure tridimensionnelle aux protéines solubles identifiées lors des programmes de séquençage systématique. Les méthodes utilisées sont principalement des méthodes de biologie structurale (microscopie haute résolution …) et de biophysique (diffraction aux rayons X, RMN, …) mais le recours à la biologie moléculaire ou à la biochimie est courant.

La génomique structurale consiste donc en une approche pluridisciplinaire ce qui se traduit sur le terrain par la constitution de plateformes techniques (Institut Pasteur, Gif sur Yvette, …) regroupant des spécialistes d'horizons divers : microbiologistes, biologistes cellulaires, biologistes moléculaires, biochimistes, cristallographes, RMNistes et bio-informaticiens.

Cependant, ce domaine d'étude se heurte à certains problèmes et notamment à des problèmes techniques : analyse RMN limitée par la taille de l'échantillon, analyse aux rayons X nécessitant des structures figées … Une autre limite majeur de cette approche structurale la quasi-impossibilité d'étudier les protéines membranaires. Or, plusieurs études récentes ont indiqué qu'elles représentent près de 1/3 des protéines totales. De plus, de nombreuses molécules doivent être étudiés dans des conditions très différentes des conditions cellulaires optimales, provoquant parfois des altérations profondes de leur structure tridimensionnelle. Si dans certaines situations la structure oriente sur la fonction d'une protéine donnée, ceci est loin d'être vérifié en pratique de part ces altérations conformationnelles. Dans ce cas, l'approche structurale s'avère très discutable et des approches moins directes de biologie classique doivent être envisagées (étude de protéines mutantes chez la souris …). Il est donc réellement nécessaire de regrouper les plateformes technologiques comme celles de biologie structurale au sein d'équipes ou d'instituts multidisciplinaires.

Les applications directes de biologie structurale restent néanmoins nombreuses puisque à l'heure actuelle, et malgré l'importance et la diversité de l'ensemble des méthodes et des paramètres cliniques, biologiques ou immunologiques, il est courant que des patients qui souffrant d'une même pathologie, présentent des évolutions cliniques et des réponses thérapeutiques complètement différentes. Une part non négligeable dans ce phénomène serait associé à des altérations tridimensionnelles (*misfolding*) indétectables au sein de séquences nucléiques ou protéiques.

Parmi les nombreux travaux actuels de biologie structurale, une approche pourrait intéresser les immunologistes et virologistes. Cette approche consiste en l'identification de la structure de toutes les protéines codées par certains virus impliqués en pathologie humaine, et notamment le VIH, avec une attention particulière pour les protéines impliquées dans la réplication virale. De tels travaux pourraient à terme permettre de dégager de nouvelles pistes pour le développement de nouvelles molécules thérapeutiques hautement spécifiques.

The evolution of mycobacterial pathogenicity: clues from comparative genomics

Stewart COLE

La recherche en microbiologie et immuno-pathologie rencontre actuellement de nombreuses difficultés pour résoudre les problèmes de l'augmentation incessante du nombre de souches multi-résistantes aux antibiotiques et de l'absence de vaccin efficace contre les mycobactéries. C'est dans ce contexte difficile que de plus en plus de spécialistes de la génomique viennent apporter leur aide dans cette branche de la recherche. Chez les mycobactéries, il existe environ 80 espèces réparties en deux classes selon leur temps de génération. La 1ère classe regroupe les espèces dites de « génération rapide » et la 2ème les espèces dites de « génération lente ». Afin de mieux comprendre la complexité de ces organismes et donc de mieux appréhender la question du développement d'alternatives aux problèmes actuels, de nombreuses études de génomiques ont été menées.

Les équipes de l'Institut Pasteur se sont intéressées à trois mycobactéries à génération lente pathogènes chez l'homme en adoptant une démarche de génomique :

1) La lèpre, dont l'agent responsable est *mycobacterium leprae* (intracellulaire)
2) La tuberculose, dont l'agent responsable est *mycobacterium tuberculosis* (intracellulaire)
3) L'ulcère de Buruli, dont l'agent responsable est *mycobacterium ulcerans* (extracellulaire)

Le terme *mycobacterium tuberculosis* ne désigne pas une unique espèce mais définit en fait un complexe auquel appartiennent *M. tuberculosis, M. africanum, M. canetti, M. microti, M. bovis* et *M. bovis BCG*. Les génomes *tuberculosis* et de *bovis*, récemment séquencés, comptent 4,4 Mb et 4000 gènes dont (40% orphelins). Au sein de ce complexe *Mycobacterium tuberculosis,* où les single-nucleotide polymorphisms (SNPs) sont rares, les évènements d'insertions et de délétions (InDels) et les transpositions sont les principales sources de la plasticité du génome. L'approche génomique comparative de ces souches intra et intercomplexe que la diversité touchait essentiellement les gènes codant les polyketides synthase (PKS), enzymes qui permettent la formation des protéines de la paroi. Elle a également permis de corriger la phylogénie du complexe, de souligner la prédominance des gènes du métabolisme lipidique, ou encore de montrer que les gènes codant la paroi étaient les plus susceptibles d'évolution : ce sont là autant de pistes utiles pour la compréhension du pathogène.

Une même approche génomique a été utilisée pour étudier *Mycobacterium leprae*. Son génome compte 3,3 Mb, comprend 1605 gènes, est « génomiquement proche » de *M. tuberculosis* (1433 gènes conservés) et aurait vraisemblablement subi une évolution réductrice. L'analyse de la plasticité génétique a ici montré que les SNPs étaient encore moins fréquent. D'autres informations ont été apportées avec notamment l'identification de nouveaux gènes codant la paroi et l'identification d'un gène codant une glycoprotéine impliquée dans le neurotropisme du pathogène. D'autres analyses sont en cours pour tenter d'identifier les gènes spécifiques du bacille de la lèpre. Les spécificités génétiques de *M. leprae* expliquant son faible taux de croissance pourraient probablement être dévoilées. Cela pourrait permettre de lever un frein majeur à l'étude du bacille de la lèpre : l'impossibilité, aujourd'hui, de cultiver cette bactérie au laboratoire.

Enfin, l'étude du génome de *Mycobacterium ulcerans*, responsable d'une maladie émergente : l'ulcère de Buruli, a permis d'apprendre beaucoup sur ce pathogène jusque-là peu étudié. Le génome de *M. ulcerans* présente la particularité d'être constitué d'un chromosome circulaire et également d'un plasmide ; l'ensemble représentant 5,8Mb. Le pathogène est la seule mycobactérie à sécréter une toxine : la mycolactone, synthétisée par une PKS présente sur le plasmide. C'est la première fois qu'un plasmide est impliqué dans la pathogénicité. Même si *M. ulcerans* reste encore partiellement méconnu, les informations apportées par l'analyse de son génome ouvre la voie à de nouvelles pistes de développement de thérapeutiques.

Ainsi, il paraît clair que la génomique peut apporter énormément à l'infectiologie et l'immuno-pathologie en permettant de mieux connaître la biologie de pathogènes humains et en dégageant des pistes pour le développement de nouvelles thérapeutiques, le tout en un minimum de temps par rapport aux méthodes de génétiques classiques.

Organisation structurelle des génomes et étude de la dynamique de réplication : une approche sur la molécule individuelle de l'ADN

Aaron BENSIMON

Quand on étudie les remaniements chromosomiques ou géniques importants, l'analyse cytologique et caryotypique est souvent suffisante. En revanche, si les lésions sont plus petites, la coloration artificielle des chromosomes par des sondes spécifiques marquées (sensibilité 1 Mb) et les méthodes de PCR (sensibilité 1 pb) sont nécessaires. Il existe tout de même un manque de moyen d'étude pour les analyses de l'ordre de 0,1 à 1 kb. La méthode du peignage moléculaire qui se base sur l'analyse de molécules individuelles d'ADN a pour objectif de combler ce déficit.

La méthode du peignage moléculaire se fait en plusieurs étapes :
→ Prélèvement de l'échantillon et extraction de l'ADN
→ Déprotéinisation de l'ADN par électrophorèse en champs pulsé
→ Plaquage des molécules nues sur une lame de verre (d'autres supports sont envisageables) dans un tampon adapté. Certaines molécules adhèrent par leurs deux extrémités, d'autres par une seule, les molécules restantes n'adhèrent pas
→ Lorsque l'on retire la lame de verre, le ménisque exerce une force qui va plaquer et linéariser de façon irréversible les molécules ayant adhéré par une seule extrémité. En subissant cette linéarisation, les molécules se fragmentent en éléments d'environ 700 pb
Une variante appelée DMC ou « Dynamic Molecular Combing », permet de fixer l'ADN à un support vertical, et ainsi, de coller environ 100 génomes par plaque.

Si l'on a travaillé avec une lame de verre, il est alors possible de réaliser une hybridation *in situ* par FISH et la taille des différents fragments ainsi que leurs positions relatives pourront être analysées. Le grand intérêt de la méthode réside dans le fait que l'on dispose de nombreuses copies du génome donc de nombreux loci d'intérêt pour l'étude des remaniements ou altérations géniques, mêmes uniques, de l'ordre du kb ou de la 100n pb. Ceci est d'autant plus intéressant que l'approche est quantitative de par le marquage spécifique. De plus, l'aspect linéaire de l'ADN favorise considérablement les hybridations ce qui confère à la méthode une grande sensibilité.

Le peignage moléculaire a été utilisé pour caractériser les régions d'ADNr humaines et il est ressorti de l'analyse que celles-ci s'organisent en profils (canonique, non-canonique de type I ou II, …). L'aspect quantitatif a notamment permis de mettre en évidence une sur-représentation de l'ADNr en configuration non canonique de type II chez les patients souffrant de la maladie de Werner par rapport à des sujets sains. A l'opposé, il a pu être observé que l'ADNr non canonique était sous-représenté dans certains cancers : ces séquences non canoniques favoriseraient-elles l'apoptose physiologiquement ?

D'autres applications sont possibles en cancérologie et en virologie : analyse de l'intégration du HPV 18 dans le génome et incidence sur le déclenchement du cancer du sein, analyse du cycle de réplication et de la cinétique de réplication des cellules à fort renouvellement, … En immunologie, cette technique pourrait permettre de visualiser l'état polyclonal ou monoclonal d'une population de lymphocytes B, afin de pouvoir détecter un lymphome par la présence d'une majorité de clones monoclonaux. Les recombinaisons VDJ étant visualisées par FISH, après fixation de l'ADN sur un support, grâce à des sondes spécifiques des régions Vβ.

Dynamique cellulaire et approche génomique, Analyse du transcriptome chez* S. cerevisae *»

Claude JACQ

S'il est possible aujourd'hui d'analyser génome et transcriptome chez une espèce donnée afin d'identifier des lots de gènes impliqués dans une fonction d'intérêt, les méthodes à grande échelle et l'apport du séquençage des génomes devraient permettre dans les années à venir de comparer ces lots de gènes entre de nombreuses espèces pour élucider les mécanismes sous-jacents.

Dans cette optique, les outils d'analyse du transcriptome, puces à ADN et bioinformatique au premier plan, offrent à présent la possibilité de travailler sur l'identification de gènes impliqués dans un phénomène donné à l'échelle du génome mais également de suivre l'évolution de l'activité de ces gènes au cours du temps en adoptant une démarche cinétique.

L'équipe du Pr JACQ a proposé une approche en 2 temps pour identifier les gènes impliqués dans la réaction au stress chimique chez *S. cerevisae* et tacher de suivre l'évolution de leur état d'activation :

1) **Approche cinétique** : Observation précoce des variations au cours du temps (de 30s à 40 min) du transcriptome de *S. cerevisae* suite à l'ajout d'un anti-mitotique dans le milieu. Le protocole, basé sur une adaptation des puces à ADN, a permis de mettre en évidence un groupe de gènes responsables de la réponse précoce et transitoire au stress chimique. Parmi ces gènes, le gène codant pour le facteur de transcription YAP-1 a été suivi et son rôle dans la réponse a été confirmé grâce à l'utilisation de mutants de ce facteur chez qui le profil transcriptionnel en réponse à la drogue est très altéré.

2) **Approche statique** : Localisation de YAP-1 dans la cellule via la méthode d'immunoprécipitation de la chromatine. Cette technique permet de déterminer quels sont les promoteurs avec lesquels YAP-1 interagit pour lancer la 2è vague de réponse au stress chimique. Celui-ci interagit avec de nombreux promoteurs mais seuls quelques uns seront activés sur le plan de la transcription. En répétant ce processus à d'autre facteurs, il est possible d'établir des réseaux d'interaction et ainsi de proposer des modèles théoriques de réponse au stress chimique.

La même approche a été conduite pour identifier 2 lots de gènes nucléaires à haut ou faible degré d'association avec les gènes mitochondriaux. Leur étude a par la suite permis de déterminer l'importance des protéines correspondantes au sein des complexes membranaires mitochondriaux (sous-unités indispensables et structurales respectivement).

Cette approche se heurte néanmoins au problème du **choix judicieux du champ à analyser et du protocole en conséquence**. En effet, la quantité de données à analyser peut rapidement s'avérer ingérable et la complexité d'analyse insurmontable si la question initiale de l'étude est mal formulée.

Archae et évolution des génomes

Patrick FORTERRE

Le concept d'archaebactérie (ou archéobactérie) est relativement récent dans l'histoire de la biologie puisqu'il date de la fin des années 1970. On découvre à cette époque des bactéries présentant diverses caractéristiques originales : de l'ARN ribosomal, des lipides particuliers (liaisons éthers entre isoprénol et glycérol), une architecture membranaire atypique ou encore de l'ADN dont l'organisation s'avère différente de celle rencontrée chez la majorité des procaryotes et eucaryotes. Elles furent reconnues à ce titre comme un nouveau groupe d'êtres vivants dont un grand nombre d'espèces vivent dans des milieux extrêmes bouleversant la classification acceptée jusqu'alors.

L'intérêt des archaebactéries dans le cadre de l'évolution de la vie résulte de l'hypothèse universellement acceptée d'une origine unique des trois grands groupes d'êtres vivants, eubactéries, archæbactéries et eucaryotes, étayée par les travaux de phylogénie moléculaire. Les archaebactéries sont divisées en deux groupes : les euryarchaeota de phénotype divers et les crenarchaeota le plus souvent hyperthermophiles. Elles peuvent survivre à des températures allant jusqu'à 113°C grâce à leur monocouche lipidique composée de molécules d'isopropanol pontées garantissant l'imperméabilité aux H+ même à haute température.

Depuis 1996 et le premier génome d'archae séquencé, de nombreux projets de séquençage ont été menés et les chercheurs disposent de 25 génomes complets d'archae à l'heure actuelle. A partir de cette base d'étude très riche, une approche génomique a pu être réalisée pour tenter de mieux comprendre l'évolution des archaebactéries. Grâce à une analyse comparative de génomes de bactéries et d'archae par BLAST, un transfert de gène entre ces deux groupes a été mis en évidence. Ainsi, il semble que certaines archae vivant en milieu chaud aient transféré des gènes à certaines bactéries, les transformant en souches hyperthermophiles. Le phénomène de transfert de gène joue donc un rôle majeur dans l'évolution des bactéries.

Des expériences de génomique à petite échelle ont ensuite permis de découvrir les mécanismes à l'origine de l'évolution de ces phyla. Avant 1996, une approche génétique par clonage des origines de réplication permit l'identification d'une première ORI spécifique des archaebactéries et impliqua les gènes de la réplication ce qui fut un grand apport de la génomique à l'époque. Les chercheurs se sont également intéressés aux protéines se fixant sur ces régions et grâce aux puces à ADN ont montrés que ces protéines se fixaient préférentiellement sur ces ORI. La génomique comparée entre espèce proche a ainsi permis de mettre en évidence un mécanisme essentiel dans l'évolution des génomes. En effet, des réplications, réparations et transcriptions induisent des recombinaisons non homologues et intra génomiques induisant des perturbations dans le biais des séquences.

Par des approches de génomique comparées à grande échelle, les chercheurs ont tentés de construire des arbres génomiques en étudiant l'ensemble du génome. Ces « super-arbres » ont mis en évidence l'existence des trois domaines connus du monde vivant (eucaryotes, procaryotes et archaebactéries) mais différents au sein de chaque groupe par rapport aux analyses antérieures comparant les ARN16S. La phylogénie basée sur les protéines informationnelles (traduction, transcription) permet donc de construire des phylogénies fiables en diminuant l'impact des transferts de gène tout en construisant des phylogénies individuelles.

Le grand nombre de génomes séquencés et les informations apportées par la génomique autant à petite qu'a grande échelle en comparant les génomes d'archaebactéries permirent de déboucher sur des perspectives générales de l'évolution des génomes.

ADN répété et contrôle épigénétique de l'activité du génome : le modèle Arabidopsis thaliana

Vincent COLOT

L'épigénétique se définit comme étant l'étude des changements d'expression, transmissibles au travers des mitoses, et des méioses dans certains cas, qui ne peuvent s'expliquer par des altérations de séquence. Les modifications épigénétiques sont souvent associées à un phénotype altéré.

Il existe différents phénomènes associés au contrôle épigénétique :

→ Méthylation de l'ADN au niveau des ilôts CpG
→ Acétylation des histones
→ Phosphorylation des histones
→ Ubiquitinylation des histones

Définissent un code pouvant être lu par une machinerie nucléaire et corrigé au besoin

Ces modification véhiculent des informations fondamentales : détermination du brin matrice par rapport au brin néo-synthétisé non méthylé, signature euchromatine ou hétérochromatine assurée par les modifications des histones avec une notion de mémoire épigénétique ... L'inactivation du chromosome X chez les mammifères, et le phénomène d'empreinte parentale, sont également des processus épigénétiques, de même que la transmission clonale de l'état actif ou réprimé des gènes homéotiques chez la drosophile, contrôlée par les protéines du groupe polycomb et thritorax.

Chez les plantes, il existe aussi des phénomènes d'empreinte parentale, et ces processus épigénétiques sont souvent associés à des variants génétiques naturels. L'apparition d'un nouveau variant peut être dû à l'extinction d'un gène par méthylation, via l'insertion d'un retro élément, ou éléments transposables. C'est une « épimutation ». Ce phénomène a été également retrouvé chez la souris, où l'extinction du gène agouti a conduit à l'apparition de souris jaunes. Mais y a t'il corrélation entre la variation génétique et l'origine de l'épimutation ?

Une étude a été entreprise chez *Arabidopsis thaliana* chez qui l'hétérochromatine (constituée essentiellement d'ADN répété et comptant de nombreux transposons et séquences satellites) et l'euchromatine ont été analysées et comparées. Il a été mis en évidence que l'hétérochromatine dérive en fait de l'euchromatine, par des phénomènes de duplication segmentaire et d'insertions réitérées d'éléments transposables. Des techniques d'analyses de puces à ADN, d'immunoprécipitation de la chromatine ou « CHIP », et d'analyse des fragments de restrictions, digérés par une enzyme dont l'activité est sensible à l'état de méthylation de son site de coupure, ont pu mettre en évidence une corrélation entre hétérochromatine et méthylation. Enfin des analyses de RT-PCR et de génomiques ont montré que les régions « éteintes », par l'insertion d'un transposon, souvent méthylées, restent potentiellement transcriptibles. Il a également été prouvé qu'une mutation, dans un gène codant pour un facteur de remodelage de la chromatine pouvait « réactiver » l'élément transposable et permettre la transcription du gène dans lequel l'élément était inséré.

Les modifications épigénétiques sont associées à des changements de l'état de la chromatine, où l'hétérochromatinisation d'une région entraîne son extinction, dans la plupart des cas. L'altération d'expression épigénétique est très souvent associée à un gène ayant subi une insertion d'éléments transposables. L'analyse quantitative de l'expression des gènes est en projet et vise à mettre en évidence l'association d'un caractère quantitatif (QTL) avec un phénotype, dans le but de déterminer si l'importance de l'impact phénotypique est associé avec une épimutation.

Les récents progrès dans la recherche des mécanismes moléculaires associés aux surdités héréditaires

Marie-Christine SIMMLER

L'audition, comme le toucher ou l'équilibre, est un sens qui met un jeu une transduction très rapide de signaux, donnée qui complexifie l'étude des mécanismes moléculaires de surdités héréditaires. L'absence de messagers secondaires et d'informations biochimiques relatives au phénomène d'audition rend difficile l'étude du mécanisme de l'audition.

L'audition, comme tout autre modalité sensorielle, possède une voie et des centres primaires dédiés totalement à cette fonction, et des voies dites « non primaires » ou convergent l'ensemble des modalités. Schématiquement le son arrive sous forme de vibration de l'air au niveau du tympan. Derrière le tympan, on définie l'oreille moyenne qui comporte des osselets dont l'étrier qui amplifie le son. Celui-ci arrive ensuite au niveau de l'oreille interne et de la cochlée, organe constitué de 3 tubes enroulés en spirale et remplis de liquide très riche en ions K+ (ce qui permet de générer une force électrique). La cochlée comporte également un organe sensori-nerveux : l'organe de Corti, qui est composé de cellules sensorielles très solidaires (cellules ciliées) à l'origine d'un influx nerveux. On comprend alors l'implication possible des cellules sensorielles dans certaines surdités…

Les surdités héréditaires représentent un problème de Santé Publique (1 enfant sur 1000 atteint de surdité profonde) et il n'existe pas de traitements curatifs pour les surdités sévère (les traitements orthophoniques et prothétiques restent les seuls recours). L'étude des pathologies liées à l'audition se heurte à deux obstacles en particulier :

- Les bases moléculaires de l'audition restent mal connues (faible connaissance des molécules impliquées dans la fonction auditive, pauvreté des banques de cDNA …)
- Difficultés dans l'analyse génétique : hétérogénéité génétique, cause « environnementale » et génétique dans une même famille, fréquence d'union entre malentendants élevée, absence de critères cliniques…

Trois approches doivent être réalisées pour une meilleure compréhension des mécanismes moléculaires associés aux surdités :

1) Clonage positionnel : recherche de gènes impliqués dans la pathologie
2) Mutagenèse a grande échelle chez la souris : induction de diverses mutations chez la souris (utilisation d'un agent chimique) et recherche de gènes associés au phénotype surdité.
3) Constitution d'une banque cDNA « oreille interne » voire banque cDNA « cellules sensorielles », par la technique de soustraction de banques cDNA
On peut alors combiner ces approches à des techniques de microarrays, microchips afin de mieux connaître les gènes surexprimés ou induits de l'embryon de souris (P0) à l'adulte (P30)… L'utilisation de la puce Affymetrix a ainsi permis d'isoler des gènes candidats associés à des syndromes de surdité (syndrome de Usher…)

On peut envisager comme perspective de faire un catalogue complet des gènes responsables des déficits primaires et des voies physiologiques mises en jeu, avec des approches multidisciplinaires (biophysique, électrophysiologie, vectorologie….).

*Le Projet Génome d'*Anopheles gambiae

Paul BREY – Karin EGLMEIER

Chaque année, le paludisme est responsable de la mort de plus d'un million d'individus, dont une majorité d'enfants des régions sub-sahariennes, et de près de 500 millions de cas cliniques à travers le monde. Le principal vecteur africain de ***plasmodium falciparum*** est ***anopheles gambiae***. Depuis une vingtaine d'années, la lutte contre le paludisme est devenue excessivement difficile de part le développement de résistances aux insecticides et l'accroissement du nombre de souches du plasmodium résistantes aux médicaments conventionnels. Dans ce contexte très pessimiste, les **interactions entre le parasite et l'insecte** sont de plus en plus étudiées, en vue notamment de trouver chez le moustique des voies pour inhiber le développement du parasite. Disposant à présent du génome du plasmodium, du vecteur et de l'homme, les travaux de génomique se multiplient dans l'espoir de trouver enfin une parade à la propagation du fléau.

Anopheles gambiae : Approches Génomiques

En 1998, un consortium international (regroupant entre autres Celera, Ensembl, le Génoscope et l'Institut Pasteur) est constitué afin de séquencer le génome *d'Anopheles gambiae* en employant la méthode « whole genome shotgun » mise au point par Celera. La version préliminaire de ce génome est accessible en 2002 mais plusieurs difficultés sont apparues lors de ce travail : manque de séquences associées au chromosome Y, lacunes parfois difficiles à combler, obtention au finale de **91 % de la séquence** … L'effort d'annotation - la délimitation et la caractérisation des gènes – a ainsi été considérablement perturbé, notamment par la complexité extraordinaire du génome de l'anophèle. En outre, le polymorphisme de la souche utilisée pour cette étude – issue d'un croisement entre moustiques collectés en Afrique et d'une souche de laboratoire porteuse d'une mutation - a entraîné des difficultés et des incertitudes lors de la phase d'assemblage. Au final, la **version préliminaire** du génome d'*Anopheles gambiae* est apparue encore **largement perfectible** et partiellement incomplète.

En dépit de ses imperfections, la séquence disponible constitue d'ores et déjà un formidable outil pour les chercheurs travaillant sur le moustique, et plus largement sur le paludisme. En effet, une approche de génomique comparative avec le génome d'un autre diptère bien caractérisé, **la drosophile**, a pu être réalisée. L'analyse a montré que malgré une divergence assez récente entre les deux insectes (250 millions d'années) les génomes ont considérablement divergé malgré quelques régions notablement conservées, preuve d'une évolution rapide et d'une complexification considérable du génome de l'anophèle. Il a néanmoins été possible d'identifier près de 14000 gènes par bioinformatique alors qu'une vingtaine seulement étaient connus avant le séquençage. Les spécialistes doivent à présent vérifier l'existence de ces gènes et les caractériser par les méthodes génétiques classiques mais il apparaît clairement que la connaissance du génome fera grandement avancer nos connaissances de la biologie générale de l'anophèle.

Afin d'obtenir davantage d'informations et tenter d'affiner la version préliminaire, les chercheurs ont entrepris de séquencer les ADNc de l'anophèle femelle adulte sur toute leur longueur : « ADNc full length » obtenus grâce à un marquage d'oligo au niveau du capping de l'ARNm. Ces ADNc offriront un moyen de valider l'annotation et de la corriger par endroit et donneront des informations notamment sur le site d'initiation de la transcription, la région promotrice, l'épissage alternatif… Une étude pilote a permis entre autres de déterminer 543 nouveaux transcrits, de prolonger les régions UTRs de 85% et de préciser que 17% des gènes soit restent encore à découvrir, soit sont incorrects dans les banques de données. Actuellement, les chercheurs travaillent sur la constitution et l'analyse de banques d'ADNc d'anophèle femelle adulte, de larves et d'embryons dans cette même optique d'amélioration et de correction de la version préliminaire du génome de l'insecte.

Parmi les 14 000 gènes potentiels, les chercheurs ont pu identifier certains gènes impliqués dans la transmission du parasite, dans la résistance aux insecticides, dans l'olfaction du moustique, dans son immunité… L'analyse plus poussée de ces gènes et de leur expression (notamment au travers d'études de transcriptomes via des puces à ADN) devrait permettre à moyen et long terme de développer des outils et mettre en place des stratégies plus fiables de contrôle de la transmission du paludisme (répulsifs, meilleures politiques de contrôle des populations de moustiques, nouvelles approches de contrôle vectoriel …)

La connaissance du génome de *Anopheles gambiae* ouvre également la voie à l'analyse des gènes d'autres espèces vectrices comme ***Aedes aegypti***, vecteur de la dengue, dont le séquençage du génome a commencé. Cette tendance de séquençage et d'analyse grande échelle des génomes devrait vraisemblablement s'intensifier dans les années à venir de part le gain de temps considérable dans l'identification des gènes offert par les méthodes *in silico*.

Approche SAGE du transcriptome des glandes salivaires d'*Anopheles gambiae*

Le cycle de transmission du paludisme comprend un certain nombre d'étapes clés dont le blocage pourrait permettre de stopper la transmission de la maladie que l'on agisse chez l'homme, le moustique ou le plasmodium : première piqûre et infection des hématies humaines, libération des mérozoïtes et de gamétocytes dans la circulation, fusion des gamètes chez le moustique, développement de l'ookinète mobile, infiltration de l'ookinète dans la muqueuse intestinale, schizogonies aboutissant à la génération de sporozoïtes qui gagnent les glandes salivaires …

L'épithélium salivaire constitue ainsi une barrière biologique susceptible de réagir à la présence du parasite. Afin d'explorer l'étendue de ces réactions cellulaires et leurs conséquences à la fois sur la composition de la salive et sur la survie du parasite, des chercheurs ont commencé la caractérisation du transcriptome des glandes salivaires d'*Anopheles gambiae* et de ses modifications en réponse à la présence de plasmodium. Pour cela, ils ont utilisé la méthode **SAGE (Serial Analysis of Gene Expression)** particulièrement adaptée à l'analyse qualitative et quantitative du transcriptome dans des conditions où le matériel biologique est peu abondant (limites de travail de l'ordre de quelques µg).

Cette technique permet une analyse de la fréquence d'un ARN messager parmi les milliers produits dans une cellule à un moment donné. Une étiquette (tag) est une séquence de 9 ou 10 nucléotides caractéristique d'un ARNm d'une cellule. Les étiquettes, isolées grâce à l'action des enzymes de restriction, sont liées bout à bout en une seule séquence qui est clonée, amplifiée par PCR et séquencée. Une analyse informatique du taux d'expression des ARNm à partir de cette séquence est ensuite entreprise.

Afin d'étudier la réaction de la cellule du moustique en présence du parasite (la souche *plasmodium berghei* a été utilisée dans ce système expérimental), des banques SAGE des glandes infectées ou non par le parasite ont été comparées ainsi que plusieurs banques correspondant à différentes périodes au cours de l'infection. Les résultats obtenus ont pu montrer qu'effectivement certains transcrits ont une fréquence de transcription modifiée en présence de plasmodium et que, de même, la transcription des gènes est modulée au cours de l'infection. D'autres banques SAGE sont en cours de réalisation, et l'on peut penser que ces transcrits sont les produits de gènes candidats impliqués dans l'interaction parasite-anophèle.

L'étude du transcriptome d'*Anopheles gambiae* pourrait donc permettre de mieux comprendre les interactions parasite-moustique, de caractériser des gènes pouvant intervenir dans la réponse du moustique à l'infection par plasmodium, et d'identifier ainsi des cibles utiles qui permettront de bloquer le développement du parasite dans l'insecte. Cette approche du transcriptome par la technique SAGE permet en outre de quantifier des transcrits dont les gènes n'ont pas encore été identifiés et donne une valeur de l'abondance des transcrits les uns par rapport aux autres. Cette technique est d'ailleurs souvent utilisée chez l'homme pour étudier le transcriptome des cellules cancéreuses.

Il apparaît donc nécessaire de continuer l'exploration du génome et du transcriptome de *l'Anopheles gambiae* dans le but de développer des outils de contrôle vectoriel et des stratégies enfin efficaces à grande échelle pour bloquer la transmission du parasite à l'homme. Ceci passe inévitablement par la caractérisation de cibles impliquées dans le cycle reproductif du moustique et donc par une meilleure compréhension de la biologie générale des interactions homme-plasmodium-moustique.

Approches génomiques de la biologie cellulaire des algues marines

Bernard KLOAREG

Les algues, apparues il y a environ 2 milliards d'années, se définissent comme l'ensemble des organismes capables de photosynthèse et n'appartenant pas au règne des plantes : il s'agit donc d'un ensemble polyphylétique regroupant à la fois des eucaryotes et certains procaryotes (cyanobactéries …).
Les algues actuelles sont le fruit d'une évolution caractérisée par plusieurs endosymbioses et dont la première a abouti à la formation de chloroplastes, suite à l'assimilation d'organismes procaryotes photosynthétiques (vraisemblablement de cyanobactéries) par les cellules hétérotrophes primitives. Les plastes, dont l'enveloppe contient 4 membranes, semblent apparaître plus tardivement et auraient pour origine une endosymbiose secondaire.

L'étude des algues a un double intérêt :

- Du point de vue de la recherche fondamentale, l'étude des algues permet de mieux appréhender un volet critique de l'évolution des êtres vivants : l'avènement de la « multicellularité » et l'évolution en parallèle des génomes.
- D'un point de vue finaliste, les algues représentent économiquement un poids important dans l'industrie agro-alimentaire (polysaccharides utilisés comme gélifiants, consommation brute en Asie, production industrielle d'iode …)

Certaines algues infectées par des pathogènes développent des moyens de défenses efficaces compte tenu de l'absence de système de défense élaboré. La génomique constitue alors un outil indispensable pour mieux comprendre ces mécanismes. Chez l'algue brune (*Laminaria digitana*) une réaction précoce à l'agression par pathogène est observée et se caractérise par l'émission de péroxyde d'oxygène (burst oxydant). Ce burst induit alors un afflux d'iodure, toxique et bactériostatique. L'algue rouge (*chondrus crispus*) adopte la même stratégie de défense par burst oxydatif, mais émet aussi des composés volatils halogénés. Après étude du génome de plastes et de thalles d'algues rouges, 152 protéines et plus de 1000 ESTs ont pu être associés à ce phénomène. Une étude plus approfondie de ceux-ci a été entreprise en vue de mettre en évidence les gènes impliqués dans le stress oxydant, ou dans la reconstruction de la paroi cellulaire.

Parmi les nombreuses protéines séquencées, on peut citer la myelopéroxydase, enzyme que l'on retrouve aussi chez les animaux, qui prend le relais des NADPH oxydases, dans la destruction du pathogène, via la synthèse de composés halogénés et la sécrétion d'acide hypochloreux (eau de Javel). D'autres protéines importantes dans le processus de reconstruction de la membrane de l'algue, après agression, ont également été séquencées (ex : protéine a domaines de Van Willebrand impliquée dans des interactions de type protéine-protéine). On peut remarquer, après comparaison d'ESTs de différentes espèces, que certaines protéines sont retrouvées chez l'algue et chez l'homme et non chez les plantes.

La génomique joue donc un rôle clé dans l'étude de l'évolution des espèces mais également dans l'identification de gènes d'intérêts en rapport avec le phénomène étudié. Elle se heurte néanmoins à un problème récurent : le manque de données biochimiques (tout particulièrement pour l'algue rouge) et surtout le manque de génomes connus chez les algues.
Il reste à présent aux spécialistes à valider les fonctions biologiques supposées des protéines potentiellement impliquées par une approche génétique et biochimique classique. Il devrait être alors possible d'établir des modèles biologiques fiables de phénomènes comme la régénération membranaire ou les stratégies de défense chez les algues.

Mécanismes d'évolution des génomes chez les Picocyanobactéries marines

Frédéric PARTENSKY

Le phytoplancton représente à lui seul 80% de la chlorophylle de l'océan et génère près de 20% de l'oxygène de la planète. Sur les nombreuses espèces qui le composent, 90% sont des picocyanobactéries. L'étude de ces organismes est donc très importante pour analyser l'impact des évolutions climatiques actuelles et pour mieux cerner les mécanismes de l'évolution chez les procaryotes. Parmi les picocyanobactéries 2 espèces ont intéressé l'équipe de la station biologique de Roscoff : *Synechococcus* et *Prochlorococcus*. Ces espèces, phylogénétiquement proches et à très petits génomes, diffèrent par leur répartition au sein des océans. Alors que la localisation de *Synechococcus* semble dépendante des gradients de concentration en sels nutritifs (variables selon la profondeur), celle de *Prochlorococcus* paraît beaucoup moins sensible à ces gradients. De plus, il a été observé que selon la disposition au sein de la colonne d'eau, différentes sous-populations de *Prochlorococcus*, appelées écotypes, pouvaient être identifiées. Ces écotypes, isolés par cytométrie en flux, se différencient par leur optimum lumineux de croissance et par leur taux photosynthétiques maximaux. *Prochlorococcus* a été proposé comme modèle d'étude pour l'analyse des génomes minimaux assurant la photosynthèse.

L'équipe a supervisé l'annotation du génome nouvellement disponible d'une souche *Prochlorococcus marinus* SS120. De part la faiblesse du GC% de ce génome, les origines de réplication ont pu être localisées via une approche AT skew. La prédiction des gènes a été réalisée à partir des programmes Critica, Genemarks et Glimmer : ~ 1900 ORFs ont été prédites après corrections. La fonction supposée des gènes a été déterminée par COGs (Clusters of Orthologue Groups) et PSI-BLAST. Ce travail a été facilité par la petite taille du génome qui accroît la probabilité d'appartenance aux différentes classes fonctionnelles des gènes. Au final, l'analyse de ce génome a apporté de nombreuses informations : fréquence génique très élevée et très peu de copies multiples, pas d'élément transposable ni de gènes impliqués dans l'assimilation des nitrates … : donc très faibles potentialités d'évolution a priori si un changement climatique important devait survenir.

L'analyse des séquences protéiques a souligné la forte proportion d'acides aminés codés à partir de séquences exoniques riches en A et T, en accord avec le faible GC%. Une telle distribution en acides aminés implique un impact important sur la fonction des protéines et résulterait d'une mutation des gènes codant pour les enzymes de réparation de AT vers GC avec pour conséquence une baisse d'activité de ces enzymes. Par alignement de séquences protéiques entre différentes souches, dont une supposée ancestrale, l'équipe a pu observer une hausse du taux global de mutation neutres chez SS120 quelque soit la catégorie fonctionnelle de la protéine. De plus, cet effet est observé en absence de pression de sélection, l'espèce étant quasiment seule à occuper sa niche écologique, sauf pour l'ADNr. Une telle évolution de génome n'a jamais été observée chez les espèces ni parasites ni symbiotiques.

La finalité de la réduction de la taille du génome semble donc différente dans le cas de *Prochlorococcus* : baisse de la consommation énergétique, diminution du volume cellulaire et hausse du rapport surface/volume pour favoriser la photosynthèse, diminution de l'effet d'auto-ombrage qui gène la photosynthèse, … Ce phénomène s'expliquerait donc par une hausse du taux de mutation global entraînant une évolution des séquences protéiques mais où seulement les mutations neutres seraient conservées. Cette évolution aura probablement été facilitée par la colonisation d'une niche écologique en absence de concurrence. Certaines limites sont cependant apparues car pour une cellule libre, le %GC et la taille du génome semblent atteindre une valeur critique sous laquelle la vie dans ces conditions environnementales apparaît impossible.

L'approche génomique de l'évolution chez *Prochlorococcus* aura, en revanche, été décevante concernant l'identification de gènes spécifiques d'écotypes. Seuls quelques gènes codant pour des transporteurs ou des protéines impliquées dans des mécanismes de transport ont pu être découverts. Les perspectives restent néanmoins encourageantes de part les connaissances qu'apportera le séquençage d'autres espèces voisines (actuellement en cours de finalisation) et l'utilisation des puces à ADN pour l'analyse des transcriptomes entre souches *Prochlorococcus* sauvages et mutantes.

Analyse multi échelle des génomes et corrélation à longue portée dans les Génomes

Claude THERMES

L'analyse multi-échelle des génomes vise à répondre à l'une des questions posées par les généticiens et les biologistes : existe-t-il une relation entre deux nucléotides distants l'un de l'autre ? L'avis général sur la question veut qu'il existe une corrélation à longue distance mais uniquement au niveau des séquences introniques. Pour conforter cette hypothèse, les spécialistes prennent pour exemple les phénomènes liés à la dynamique des génomes (les duplications, insertions ...).

Ces notions de corrélation à longue portée sont quantifiables et parfois indépendantes de l'échelle à laquelle on se situe : on parle d'invariance d'échelle. Concrètement, si on progresse le long d'une séquence nucléotidique et que l'on trace une barre noire verticale chaque fois que l'on rencontre une purine (par exemple) nous transformons notre séquence nucléotidique en un signal qui ressemble à un code barre. Il est alors possible d'élargir la « boite » c'est-à-dire la longueur de la séquence analysée et si le signal correspondant ne change pas, il y a invariance d'échelle.

Admettons à présent que l'on adopte un outil de quantification de la corrélation longue distance autre que celui du code barre : par exemple la mesure de la proportion de GC ou de AT le long d'une séquence nucléotidique. La corrélation de distance sera « mesuré » par un coefficient : le coefficient de corrélation. Pour calculer celui-ci, il convient de suivre les étapes suivantes :

→ Lorsque l'on identifie un GC le long de notre séquence, on note +1, versus -1 pour AT.

→ La séquence nucléotidique est alors transformée en un signal numérique dont on peut tracer le profil

→ De ce profil on calcule ensuite la variance σ, puis l'on trace une courbe :

$$\log(\sigma) = \log(\text{différence d'échelle})$$

→ On déduit de cette courbe le coefficient de Pears H qui représente un indice de corrélation.

Ce coefficient ne peut être déterminé qu'à partir d'une courbe logarithmique homogène, ce qui n'est pas toujours le cas car il est fréquent d'observer des courbes en paliers. Pour limiter le problème, une ou plusieurs dérivations doivent être effectuées à partir de la courbe initiale ; on obtient *in fine* un profil exploitable à partir duquel on peut déduire le coefficient de corrélation. Si $H < 0,6$, il n'y a pas de corrélation longue portée (versus $H>0,6$).

Il a été mis en évidence une corrélation longue distance également au niveau éxonique chez certains organisme eucaryotes, ce qui est en contradiction avec le « dogme » du dynamisme génomique. Cette nouvelle corrélation a été justifiée par l'organisation de l'ADN en nucléosomes. L'hypothèse avancée pour justifier cette nouvelle corrélation est que l'ADN des eucaryotes possède des propriétés de séquence favorisant son enroulement autour des histones (contrairement aux bactéries qui ne présentent alors pas de corrélation à longue distance). Pour tester l'hypothèse, les spécialistes se basent sur deux « tables de corrélations » : l'une décrivant l'enroulement de l'ADN autour du nucléosome et l'autre se basant sur la coupure de l'ADN par la DNAse. Ces deux tables de codage permettent de tracer un profil de courbure intrinsèque à partir duquel il est possible de mettre en évidence une éventuelle corrélation à longue distance.

Une corrélation existe chez tous les génomes. Suivant les génomes, un codage spécifique est associé au LCR (séquence permettant la compaction et l'enroulement de l'ADN). L'objectif sera de trouver un codage qui fasse « disparaître » les LCR chez certains organismes et pas chez d'autres afin de trouver de nouvelles propriétés associées à des phénomènes biologiques.

Physique de l'ADN et analyses génomiques

Edouard YERAMIAN

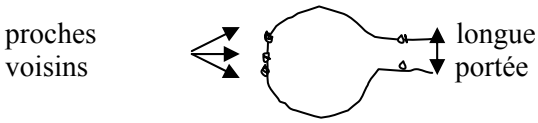
Pour mieux comprendre les molécules biologiques, des modèles moléculaires globaux ont été développés. Des modèles structuraux (atomistiques locaux et simplifiés à grande échelle) sont notamment utilisés pour les analyses génomiques. Au vu des avancées de la biophysique et de la bioinformatique, il est désormais intéressant d'établir une intersection entre ces deux domaines et de comparer les analyses génomiques faites à partir de ceux-ci.

Biophysique

La question posée par les biophysiciens lors d'analyses génomiques est de déterminer s'il existe une relation entre la physique de base et les messages codés par le génome. Les biophysiciens utilisent un modèle de transition simple dit en "hélice-pelote". Il s'agit d'étudier les interactions entre nucléotides en considérant :

- Les interactions dites à longue portée : indispensables pour établir des modèles structuraux mais sont à l'origine de toutes les difficultés rencontrées pour les calculs. Ces effets dépendent de la longueur de la chaîne (boucles...)
- Les interactions proches voisins : étude des nucléotides proches entre eux (courbure de l'ADN)
- Les interactions générales longue portée

Les temps de calculs sont les suivants : en seconde pour les interactions proches voisins, n^k (heures à année) pour les interactions longue portée et exp(n) pour les interactions général longue portée.



La solution pour réduire le temps de calcul est d'utiliser des algorithmes qui permettent de se rapprocher du temps de calcul dans le cas des interactions proches voisins pour les interactions à longue portée. La fonction de partition Z en statistique permet d'énumérer toutes les configurations possibles de l'ADN et l'on attribue un poids à chaque configuration grâce à une constante chimique : Z = K1 + K2+ ... Pour une configuration, on a 2 valeurs : l'une thermodynamique qui dépend du proche voisin (part de l'hélice) et l'autre qui dépend de la longueur de l'ADN (part de la boucle). Une matrice constituée de toutes les configurations est calculée avec des codes 0 et 1.

Bioinformatique

Pour le modèle bioinformatique, il s'agit de calculer des gaps et des scores là aussi grâce à une matrice. On réalise des alignements d'une génome contre un autre pour déterminer la longueur des gaps – qui se révèle être la difficulté - et l'on peut noter que ce calcul n'est pas linéaire mais exponentiel, comme le modèle biophysique.

Applications à la génomique

Pour étudier un génome, on calcul tout d'abord le pourcentage de GC, puis on établit un modèle physique complet. On peut même étudier au fur et à mesure l'ouverture de l'hélice le long d'un chromosome entier en augmentant d'un degré la température, ce qui a amené les physiciens à voir une relation entre la segmentation de la physique et la segmentation du génome, tout comme la relation hélice ouverte / non ouverte et codant / non codant, donc à délimiter des gènes. Bien sûr, cette relation n'est pas universelle, et la comparaison avec un modèle bioinformatique vise à conforter (ou non) les résultats obtenus par l'approche biophysique.

Malgré tous les paramètres mis en jeu, le modèle biophysique est relativement robuste et a permis de découvrir de nombreux gènes comme le gène de la résistance à la chloroquine de *P. falciparum* ainsi que d'autres gènes de *P. vivax* ou encore *C. elegans*. Les annotations combinées des chromosomes des diverses méthodes existantes ne permettent pas encore d'isoler tous les gènes. Il reste quelques difficultés à résoudre notamment avec les petits exons.

Colonisation du génome par de l'ADN mitochondrial

Miria RICHETTI

Les mitochondries sont responsables de la respiration, donc de la production d'énergie, au sein des cellules. Localisées dans le cytoplasme, donc hors du noyau, elles contiennent leur propre ADN dit "**ADN mitochondrial**" ("ADNmit") stocké sous forme de petites boucles. Chez l'homme, cet ADNmit comprend **37 gènes** répartis sur **16 569 paires de base**. La taille du génome mitochondrial est donc très faible par rapport à celle du génome nucléaire (qui compte environ 3 milliards de pb) et cette différence s'est accentuée au cours de l'évolution par le transfert d'un certain nombre de gènes de la mitochondrie vers le noyau. A l'inverse il n'a été observé que de rarissimes transferts de gènes du noyau vers la mitochondrie.

Les chercheurs de l'Institut Pasteur se sont donnés pour objectif de répondre via une approche de génomique à la question suivante : pourquoi y a-t-il si peu de transfert de gènes du noyau vers la mitochondrie et comment co-évoluent ces deux génomes ? Les spécialistes pensent que la faible fréquence de transfert de gènes nucléaires vers la mitochondrie peut s'expliquer par les conditions requises pour ce transfert au niveau de la structure des gènes avec notamment l'absence d'introns indispensable au sein du génome mitochondrial. Une hypothèse complémentaire est souvent apportée précisant que la forte dérive génétique des gènes mitochondriaux (haploïdes, petite taille, héritage maternel, absence de recombinaison …) serait un facteur clé poussant ces gènes à gagner le noyau dans un souci de stabilité sur le long terme. D'autre part, les mécanismes de ce transfert particulier de gènes au sein d'une même cellule sont encore largement méconnus.

Par comparaison des génomes nucléaires et mitochondriaux chez les plantes (*Arabidobsis* et plusieurs angiospermes) et les levures, il a été mis en évidence plusieurs mécanismes potentiels de transfert : duplication et transfert de pseudo-gènes mitochondriaux devenant actif dans le noyau, cassures double-brin suivie de réparations imparfaites créant ainsi un motif appelé signature d'insertion … Ces signatures ont été particulièrement étudiées par l'équipe et celles-ci semblent, de par leur nature très variable, constituer un mécanisme de polymorphisme insertionnel supplémentaire. A partir de cette découverte, l'équipe a interrogé les bases de données humaines pour voir si ces polymorphismes d'insertion d'origine mitochondriale étaient présents chez l'homme. Non seulement ces séquences sont présentes, mais il semble qu'elles constituent le mécanisme principal de transfert de gène mitochondrial vers le noyau chez l'homme. Tous les groupes ethniques présentent environ 200 de ces séquences appelées NUMTs (*nuclear pseudogenes of mitochondrial origin*) caractérisées par une grande variation de fréquences alléliques. La répartition des NUMTs est homogène le long du génome et il a été identifié 27 NUMTs spécifiques de l'espèce humaine. Chez l'homme, les NUMTs possèdent des sites d'intégration préférentielle au niveau des exons ce qui explique la rareté du phénomène et qui souligne l'importance de ces séquences en pathologie humaine. En particulier, les NUMTs ont été très récemment impliquées dans une forme de déficience sévère en facteur VIII, dans le désordre affectif bipolaire ou encore le syndrome de Pallister-Hall (pathologie liée à la catastrophe de Tchernobyl).

Cette approche génomique initialement orientée vers la compréhension d'un mécanisme génétique chez les plantes a permis de découvrir un acteur très important du polymorphisme en évolution chez l'homme avec un impact non négligeable en pathologie. De telles études démontrent l'impact que peut avoir une approche génomique à grande échelle dans la mise en relation de phénomènes a priori distincts dans leurs modèles respectifs.

Application des polymorphismes humains pour l'étude des retards mentaux de l'enfant

Laurence COLLEAUX

Le génome humain compte de 3.10^9 pb, mais seulement 5% de cet ADN correspond à des séquences codantes, et codera pour ~ 30000 gènes. Sur l'ensemble du génome, on peut retrouver jusqu'à 1% de variabilité génétique entre deux individus : ce polymorphisme constitue un outil indispensable en génétique humaine pour l'étude de nombreuses pathologies. Différents polymorphismes d'intérêts sont retrouvés le long du génome :

- Les RFLPs, polymorphismes de fragment de restriction
- Les VNTRs, répétitions de bases en tandem (comme les microsatellites et les mini-satellites)
- Les SNPs, variations d'une seule base dans la séquence, très fréquentes

La 1ère application de ces polymorphismes, en génétique humaine, a été la construction de carte génétique, grâce à l'étude de la ségrégation des marqueurs microsatellites au cours des générations. Ces marqueurs microsatellites sont un outil de choix, car ils sont très informatifs et repartis de façon homogène sur le génome. Le génotypage de ces marqueurs permet la recherche d'anomalies chromosomiques, de faire des études familiales, d'effectuer les diagnostics pré-natal et pré-implantatoire, des tests de paternité et de nombreux tests en criminologie (police scientifique).

Cette étude du polymorphisme, a été appliquée à l'étude des retards mentaux chez l'enfant. Les handicaps mentaux, touchent près de 3% de la population générale, et l'importante hétérogénéité étiologique (origines très diverses : génétique, métabolique, forme syndromique…), rend cette étude très complexe. Le génotypage des marqueurs microsatellites a permis d'analyser la contribution génétique dans différents cas de retards mentaux, et ainsi de pouvoir prédire le risque pour une nouvelle naissance. Différentes causes génétiques ont pu être mises en évidence : disomies uniparentales (conséquence de la présence de deux copies d'un locus provenant d'un même parent), remaniements sub-télomériques (cause de nombreux retards mentaux) …

L'emploi des polymorphismes a permis l'identification de nouveaux syndromes et il devrait être rapidement possible de corréler génotype et phénotype concernant ces syndromes offrant ainsi l'espoir d'identifier des gènes candidats. Des études de gènes candidats et des analyses de cartographie par homozygotie ont été entreprises pour identifier des régions d'homozygoties partagées entre individus atteints. L'analyse de gènes candidats impliquerait :

→ De vérifier que la mutation n'est pas un variant allélique présent dans la population générale
→ De faire une étude fonctionnelle, des tests physiologiques
→ D'analyser la conservation de ce gène entre espèces pour connaître son importance

L'avenir, dans l'analyse des polymorphismes humains, sera l'étude des SNPs, peu informatifs mais très fréquents et uniformément distribués. Les études d'associations, pourront permettre de détecter un gène d'effet mineur impliqué dans une pathologie par déséquilibre de liaison. L'étude pangénomique par puces à ADN permettra de découvrir de nombreux SNPs d'intérêts. Enfin, l'approche pharmacogénétique des SNPs devrait permettre de définir des profils de réactivité à un traitement ou de sensibilité à certains effets secondaires.

Les virus à ARN emprisonnés par leur génome

Simon WAIN-HOBSON

On connaît actuellement des milliers de virus à ADN et ARN et parmi ces espèces, les bactériophages à ARN sont les plus simples représentants. Le génome du bactériophage M52 a été le premier séquencé en 1976. Depuis, un grand nombre de virus ont vu leur génomes séquencés, et on classe les virus généralement en 2 catégories : 80% appartiennent à la classe des virus à ARN (polarité +/-, mono ou double brins, expriment la reverse transcriptase), le reste aux virus à ADN (polarité +/-, mono ou double brins). Les virus à ARN présentent deux particularités par rapport aux virus ADN : un taux de mutation très élevé et l'absence des mécanismes de correction des erreurs de séquence (les variations de séquences sont définitives !).

Le HIV est un virus à ARN et son génome est soumis à des exigences apparemment contradictoires : il doit produire des copies fidèles de son génome tout en se protégeant d'un environnement très hostile. Cette protection passe par des mécanismes de variation de séquence impliquant notamment la génération d'erreurs lors de la réplication. Les mutations (environs 0,25 par génome) et hypermutations (700 par génome) jouent un rôle important dans l'évolution des rétrovirus en général et du HIV en particulier. La reverse transcriptase du HIV ne "relit" aucune des copies qu'elle produit (ARN à ADN) et (ADN à ADN) : elle commet en moyenne une erreur sur 10^4 et, le génome du VIH comprenant environ 10.000 pb, il existera en moyenne au moins une différence de séquence entre deux virus frères d'une même génération.

Lorsque l'on comprend mieux l'aspect variable du génome de tels virus, il est fondamental de s'interroger sur la fréquence de cellules infectées *in vitro* présentant des copies de virus « génomiquement » différents. Une approche par PCR et séquençage dans des cellules T CD4+ a permis la mise en évidence de plusieurs copies d'ADN intégrées (ou provirus) au sein d'une même cellule. Ainsi,toute cellule infectée *in vitro* subit en réalité une infection par de multiples variants. Ce phénomène devrait être encore plus marqué *in vivo*. De plus, il a été observé qu'une cellule contient en moyenne 4 provirus mais 9 séquences d'ADN différentes. Il y'a donc plus de génomes que de provirus intégrés car il existe des formes de virus non intégrés : les cellules subissent donc non pas une infection mais de très nombreuses infections après une primo-infection !

Les spécialistes se sont alors demandés combien de virus totaux pouvaient être retrouvés en moyenne au sein d'une cellule infectée : après séquençage, près de 700 molécules d'ADN différentes ont été identifiées ... Connaissant le nombre intégré et non intégré de virus, il a été estimé qu'environ 4000 virus sont nécessaires pour présenter un tel profil d'infection *in vivo* étant donné le potentiel de variabilité du génome du HIV.

Les chercheurs se sont ensuite demandés si ce potentiel de variabilité du génome dans le temps était à l'origine d'une évolution réelle ou de simples modifications de séquence. Dans le cas du HIV chez qui le taux d'erreur est très élevé, la force de sélection est insuffisante pour maintenir l'identité de la population qui reste hétérogène. Il a été montré que la variabilité d'une protéine était conservée entre génomes de virus différents et que très peu de substitutions étaient nécessaires à l'évolution et à l'adaptation des virus. En comparant les génomes des virus à ARN par rapport à d'autres virus en étudiant le pourcentage d'acides aminés variables entre des protéines autologues comme les protéases, il a été mis en évidence que les virus à ARN évoluaient peu et qu'ils était ainsi très conservés.

Au final, il apparaît que les virus à ARN ne peuvent voir leur génome s'étendre car ils sont « emprisonnés » par la taille de leur génome et changent donc plus qu'ils ne s'adaptent.

Du génome à la cellule : comment prédire la fonction des gènes

Antoine DANCHIN

Il existe 2 grandes approches des systèmes biologiques : du phénotype au génome ou du génome au phénotype. Depuis le développement des méthodes de séquençage et de la génomique, une quantité sans cesse croissante de génomes sont disponibles et l'étude du vivant tend de plus en plus vers cette 2ème approche. S'il on essaye de répondre à la question « Qu'est-ce que la vie ? », trois composantes apparaissent déterminantes : un métabolisme, une compartimentation et un transfert d'information. On peut remarquer que cette définition exclut les virus !

Le métabolisme compte au moins 800 molécules (substrats, enzymes et co-enzymes) indispensables au développement de la vie. La compartimentation se fait selon 2 modes : procaryotes (milieux interne et externe) et eucaryotes (uni et pluricellulaires, complexité croissante). Enfin, le transfert d'information met en jeu 3 lois : la réplication (loi de complémentarité), la transcription (loi de complémentarité) et la traduction (loi du code génétique).

Une cellule peut être comparée à « un ordinateur qui fabrique des ordinateurs » ; dans un tel modèle, le génome est le disque dur que la cellule lit et exploite grâce à des dispositifs aboutissant ainsi à des programmes adaptés. Selon le principe de la machine de Turning, un tel ordinateur, capable d'agir selon un programme défini et de changer de programme au besoin en fonction d'une situation, est capable de surmonter tous les obstacles de la vie.

Pour comprendre un génome, il faut acquérir les données de séquence, les analyser puis les gérer. Ces 3 étapes utilisent la bio-informatique mais celle-ci est toujours encadrée par la question biologique pertinente liée à ce génome. Le concept central à mettre en évidence dans les approches génomiques est le concept de fonction, auquel se rattachent des idées de biologie et d'informatique (contrôle, codage, information) et des idées de physique-chimie (espace, temps, matière, énergie).

Comment trouver la fonction des gènes ?

Le triplet variation-sélection-amplification est essentiel dans l'émergence de la fonction. Concrètement, l'évolution crée la fonction qui capture la structure, elle-même codée qu sein de la séquence. En conséquence les analyses inter-espèces grande échelle et les analyses dites de voisinage des génomes permettent d'orienter vers la fonction des gènes. Il faut cependant rester prudent car la fonction observée d'un gène ou d'une protéine n'est pas nécessairement sa fonction première : tout dépend du contexte biologique ! De plus, de telles erreurs d'approximations de la fonction ne sont pas corrigeables par l'analyse des séquences.

Analyse globale des données

L'analyse grande échelle des génomes repose soit sur des comparaisons de ce qui est présent par rapport à ce qui est absent, soit sur ce qui co-varie. Une telle approche est toujours réalisée par comparaison à des modèles qu'il faut judicieusement établir et qui n'existent même pas dans certaines situations. Quand la fonction est difficile à établir, les analyses de voisinage peuvent être très utiles car le plus souvent, la réponse existe déjà quelque part mais est perdue dans un flot d'informations très diverses ! Ce problème tend à s'aggraver avec l'accumulation des données issues des séquençages. L'analyse de voisinage peut être de lancer une recherche multidisciplinaire sur une protéine inconnue, de réaliser une étude synthénique, … L'annotation est l'étape fondamentale de l'organisation des données ; il est donc primordial de bien comprendre les logiciels utilisés ainsi que les concepts à leur origine afin de cerner au mieux l'information pertinente.

Organisation des génomes : exemple de l'étude du génome d'*E. coli*

Cette organisation apparaît si rigide que la pression de sélection sur l'ADN est visible dans le texte génomique qui distingue le brin direct et le brin retardé. Sur chacun de ces gènes, l'organisation est-elle régie par le hasard ? Une approche statistique par un simple test de Fisher permet d'affirmer de manière fiable que chez *E. coli*, c'est l'essentialité des gènes qui détermine leur organisation sur le brin le chromosome. Ainsi, une majorité de gènes vitaux, dont de nombreux intervenant dans le métabolisme général, se retrouvent localisés au niveau du brin direct à proximité d'une origine de réplication de manière à favoriser leur transcription.