UNIVERSITÉ DU QUÉBEC À MONTRÉAL

DES ALGORITHMES BIOINFORMATIQUES POUR LA RECHERCHE DES REGIONS GÉNOMIQUES RESPONSABLES D'UNE MALADIE

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN INFORMATIQUE

PAR

DUNAREL BADESCU

NOVEMBRE 2009

REMERCIEMENTS

Je présente mes remerciements au Dr Vladimir Makarenkov, mon directeur de recherche, pour son suivi, ses conseils et ses suggestions lors de la réalisation de ce projet de maîtrise. Je le remercie particulièrement de m'avoir donné la force de pousser mes limites personnelles au bon moment et d'aller jusqu'au bout de ce projet de maîtrise.

Je remercie également Dr Abdoulaye Baniré Diallo, mon codirecteur, pour m'avoir guidé sur ce chemin complexe et multidisciplinaire qui est la bioinformatique.

Qu'ils trouvent ici toute l'expression de ma gratitude et de ma profonde reconnaissance!

Je profite de ce mémoire pour adresser également mes remerciements au Dr Mathieu Blanchette, mes collègues Alix Boc, Alpha Boubacar Diallo, pour leurs points de vue et leur soutien qui ont été d'une grande utilité.

Mes remerciements s'adressent aussi à mon épouse Mihaela et mon fils Elian qui m'ont accordé leur soutien total durant ces deux dernières années.

Je ne saurais terminer sans manifester ma reconnaissance au *Conseil de recherches en sciences naturelles et en génie du Canada* (CRSNG) ainsi qu'au *Fonds québécois de la recherche sur la nature et les technologies* (FQRNT), qui ont contribué au financement de ce projet de maîtrise. À tous ceux qui ont contribué de près ou de loin à la réalisation de ce projet, qu'ils trouvent ici mes remerciements les plus sincères.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

ADENO	Adenocarcinome
ADN	Acide DésoxyriboNucléique
ARN	Acide RiboNucléique
GenBank	NIH genetic sequence database
HGT	Horizontal Gene Transfert
HIV	Human Imunodeficiency Virus
HMM	Hidden Markov Model
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
NJ	Neighbor-Joining
NNI	Nearest Neighbour Interchange
PARS	Parsimony Program
PHYLIP	PHYLogeny Inference Package
SPR	Subtree Pruning and Regrafting
SQUAM	Carcinome aux cellules squameuses
TBR	Tree Bisection and reconnection
VIH	Virus de l'imunodéficience humaine
VPH	Virus du papillome humain

RÉSUMÉ

L'évolution des espèces est régie par les modifications stochastiques qui ont eu lieu au niveau du code génétique - l'ADN - composé d'une suite de petites molécules (les nucléotides). Selon l'ampleur de ces évènements, il y a d'abord des modifications à petite échelle, impliquant quelques nucléotides – les insertions, délétions et substitutions. Due à l'impossibilité actuelle de différencier les insertions des délétions, on les appelle communément *indels*. D'un autre coté, il y a des modifications à grande échelle – impliquant parfois des grandes régions génomiques ou des chromosomes. Les modifications à grande échelle les plus fréquentes sont: les duplications, translocations, inversions et délétions.

Au cours de ce projet, nous avons développé une méthode de génomique comparée, capable de relier l'information épidémiologique, comme la carcinogenicité et l'invasivité des souches, aux séquences génomiques. Cette méthode permet de détecter des régions statistiquement significatives à analyser plus en détail par des biologistes, tout en étant capable de discriminer ce seuil à l'aide du calcul des p-values.

Nous avons utilisé cette méthode dans l'étude du virus du papillome humain et de la bactérie *Neisseria Meningitidis*, bactérie responsable de la méningite.

Pour le virus du papillome humain, notre méthode a été capable de détecter le domaine PDZ, une région du gène E6, qui est une *condition sine qua non* de la carcinogenicité du produit de ce gène.

Au cours des analyses phylogénétiques de cette famille nous avons trouvé une corrélation statistiquement significative entre les événements à petite échelle et les données épidémiologiques. Par la suite nous avons proposé une séquence de tests pour orienter l'analyse statistique de cette corrélation. Nous avons également remarqué que la carcinogenicité est généralement monophylétique, donc issue d'un ancêtre commun. L'arbre phylogénétique inféré est le premier basé sur les génomes entiers, ce qui a permis d'étudier la variabilité des topologies de gènes par rapport à celle du génome.

Pour la bactérie *Neisseria Meningitidis* nous avons montré qu'il est possible de syntoniser les fonctions de discrimination, pour établir la différence entre les régions responsables du maximum d'invasivité et celles qui ont un rôle structural dans ce processus, détection des structures moléculaires connues (i.e. les anses extra cellulaires, dans notre cas).

Les résultats de nos travaux ont permis la mise à la disposition de la communauté internationale de deux bases de données, pour le VPH[1] et le *Neisseria*[2], respectivement. Ces bases contiennent des régions candidates à être analysées en laboratoire par des biologistes.

Mots clés : Algorithme pour la détection des régions carcinogènes, événements évolutionnaires, analyse de redondance, arbre phylogénétique, conception de vaccin, mutations, invasivité, *Neisseria Meningitidis*, virus du papillome humain.

---

[1] http://ancestors.bioinfo.uqam.ca/articles/JCB2009/supplemental.zip

[2] http://adn.bioinfo.uqam.ca/~dunarel_b/neisseria_2009

INTRODUCTION

La bioinformatique est un domaine multidisciplinaire qui associe la biologie, l'informatique, les statistiques et les mathématiques. La génomique et la protéomique sont des domaines de la biologie moléculaire qui étudient, respectivement, les caractéristiques du code génétique et des protéines ainsi que leur fonction. Ces deux disciplines ont généré de grandes quantités de données, qui sont pour la plupart stockées dans des bases de données centralisées et publiques. Bien que leur disponibilité soit immédiate et des interfaces conviviales ont été développées, leur analyse et génération de connaissances sont loin d'être achevées.

La génomique comparée est un domaine plus récent et en plein essor. Elle utilise surtout des méthodes bioinformatiques pour collecter des connaissances à partir de la comparaison des données génomiques déjà stockées dans des bases de données. Son avènement élargit le chemin des études bioinformatiques car les jeux de données sont plus complexes, plus volumineux, mais aussi on dispose des données pour documenter plusieurs aspects d'un même problème. Ainsi aujourd'hui les annotations et les meta-données sont aussi disponibles et jouissent de la même précision que les séquences primaires. Les structures d'ordre supérieur des ARN et des protéines – incluant les différentes conformations, domaines et coordonnées tridimensionnelles complexifient la nature des calculs et les perspectives d'analyse.

Pour les modifications à petite échelle au niveau de l'ADN, des modèles statistiques d'évolution, basés sur la nature stochastique des évènements ont été développés, et des outils d'analyse pour la phylogénie, l'alignement et analyse de séquences, l'étude des modèles de variation des séquences, le groupage et d'autres, sont disponibles aujourd'hui à la communauté des chercheurs.

Bien que les études de biologie moléculaire soit extrêmement pointues, l'immense diversité biologique fait que la bioinformatique est capable de suggérer certains points chauds à explorer. Ceci réduit le temps et les ressources dépensées par la recherche en laboratoire. C'est le cas de notre méthode de génomique comparée, capable de relier l'information épidémiologique aux séquences génomiques pour détecter des régions significatives et intéressantes.

Nous avons travaillé sur des données génomiques et épidémiologiques.

Pour le VPH, les données génomiques ont été constituées par les séquences complètes de 83 types, comprenant les annotations de gènes, disponibles dans la base de données GenBank du NCBI (Benson et al., 2005). Les données épidémiologiques ont été constituées par une large étude internationale, comprenant 3,607 femmes diagnostiquées histologiquement pour un cancer du col de l'utérus, et provenant de 25 pays sur plusieurs continents (Muñoz et al., 2003, 2004).

En ce qui concerne le *Neisseria,* nous avons pris la séquence et les annotations de la souche H44/76 sur GenBank et les alignements multiples disponibles pour le gène fetA (frpB), sur le site spécialisé en la recherche de ce microorganisme (Neisseria Research Community Website, 2009; Thompson et al. 2003). Une étude sur les souches hyperinvasives rencontrées lors de plusieurs épidémies a constitué la base de nos données épidémiologiques (Urwin et al., 2004).

Pour mieux comprendre le résultat de nos travaux nous présentons d'abord au chapitre I une introduction des notions de biologie des microorganismes, et spécialement le virus du papillome humain ainsi que *Neisseria Meningitidis*, sur lesquels nos études ont été menées. Ensuite, dans le chapitre II, nous présentons les méthodes d'analyse employées dans nos articles ainsi que les efforts menés ailleurs pour la détection des séquences fonctionnelles.

Le travail effectué a mené à la rédaction de quatre articles, dont «An evolutionary study of the human papillomavirus genomes» (Badescu et al., 2008, publié par *Springer-Verlag* à la suite de la conférence Recomb-CG), qui ne sera pas présenté car il a été ensuite developpé, complété et publié dans *Journal of Computational Biology* (Diallo et al., 2009b).

Les deux autres articles ont été acceptés pour publication par, dans *Classification as a Tool for Research* (édité par *Springer-Verlag*), à la suite de la conférence IFCS 2009, et dans un volume special, suite à la conférence SFC-CLADAG 2008 (édité aussi par *Springer-Verlag*).

Ces articles font partie intégrante de ce mémoire de maîtrise et sont presentés au cours des chapitres III, IV et V, préfacés par un court résumé. Ce sont les articles suivants:

«Classification of the Human Papilloma Viruses» (Diallo et al., 2009a).

Cet article traite de la classification génomique du virus du papillome humain. Le calcul des scénarios d'insertions et délétion de nucléotides les plus probables ont été calculés pour étudier la distribution des indels sur les différentes branches de l'arbre et sa relation à la carcinogenicité.

«A whole genome study and identification of specific carcinogenic regions of the Human Papilloma Viruses» (Diallo et al., 2009b).

Cet article décrit une nouvelle méthode statistique pour la détection des régions potentiellement correlées au cancer et l'applique à l'étude du génome du virus du papillome humain.

«Identification of specific genomic regions responsible for the invasivity of *Neisseria Meningitidis*» (Badescu et al., 2010).

Ce dernier article présente plusieurs fonctions de discrimination en comparant les résultats aux structures moléculaires connues (comme les anses extracellulaires des protéines membranaires et les feuillets-β transmembranaires). Ainsi, nous avons étendu le champ de nos recherches sur d'autres types d'information épidémiologique (information reliée à la transmission et la rependue de la maladie dans une population), en l'occurrence sur l'invasivité (capacité du microorganisme d'envahir les tissus de l'hôte) de la bactérie *Neisseria Meningitidis*, responsable de la méningite.

Pour ces trois articles, ma cotribution est la suivante : j'ai participé à la conception et à l'implémentation des algorithmes et de la base de données des virus et des bactéries. J'ai

aussi participé à la mise en place des scripts pour l'automatisation des analyses. Enfin, j'ai contribué à l'analyse des résultats obtenus. J'ai aussi documenté les aspects biologiques et trouvé la relation au domaine protéique PDZ, ainsi qu'aux anses extracellulaires.

Ce mémoire contient également trois annexes. Les annexes fournissent le code source des méthodes développées ainsi que plusieurs résultats supplémentaires concernant l'analyse du virus du papillome humain et de la bactérie *Neisseria Meningitidis*.

CHAPITRE I

NOTIONS DE BIOLOGIE DES MICROORGANISMES

Dans ce chapitre nous présentons les définitions de base des notions biologiques que nous avons employées dans notre travail, suivies d'un court aperçu de la biologie du virus du papillome humain et de la bactérie *Neisseria Meningitidis,* les deux microorganismes sur lesquels nos recherches ont porté.

Nous rappelons ici que, par définition, un microorganisme est un organisme de taille microscopique ou ultramicroscopique – ceci inclut les virus et les bactéries (Merriam-Webster Online Dictionary, 2008).

1.1 Définitions de base

1.1.1 ADN

L'acide désoxyribonucléique (ADN) est une macromolécule de poids moléculaire élevé, formée de polymères de nucléotides (Saenger, 1984). Dans une cellule il renferme l'ensemble des informations nécessaires au développement et au fonctionnement de l'organisme. Il est transmis lors de la reproduction, constituant le support de l'hérédité - l'ensemble des caractéristiques et potentialités héritées d'un ancêtre (Merriam-Webster Online Dictionary, 2008). Chaque nucléotide est constitué de trois éléments : un groupe phosphate, un sucre (le désoxyribose) et une base azotée (Brown, 2002). Il se présente sous la forme d'une double chaîne hélicoïdale dont les deux brins sont complémentaires, comme

présenté sur la figure 1.1 (Watson et Crick, 1953). Les deux brins forment le grand et le petit sillon, dont la forme et la profondeur dépendent des angles formés entre ces diverses composantes. Les principales formes d'hélice sont appelées A, B et Z. Elles influencent la forme et l'accessibilité aux grands et petits sillons et par ce moyen influencent la spécificité de la liaison aux enzymes (Brown, 2002).

Il existe quatre types de nucléotides au niveau de l'ADN : l'adénine (A), la guanine (G), la cytosine (C) et la thymine (T), divisés en deux groupes : les purines -A et G - et les pyrimidines - T et C (Lodish et al. 2000). L'ADN est présent dans le noyau des cellules eucaryotes, dans le cytoplasme des cellules procaryotes, dans la matrice des mitochondries ainsi que dans les chloroplastes. Certains virus possèdent également de l'ADN dans la capside, structure protéique qui entoure et protège l'acide nucléique d'un virus.

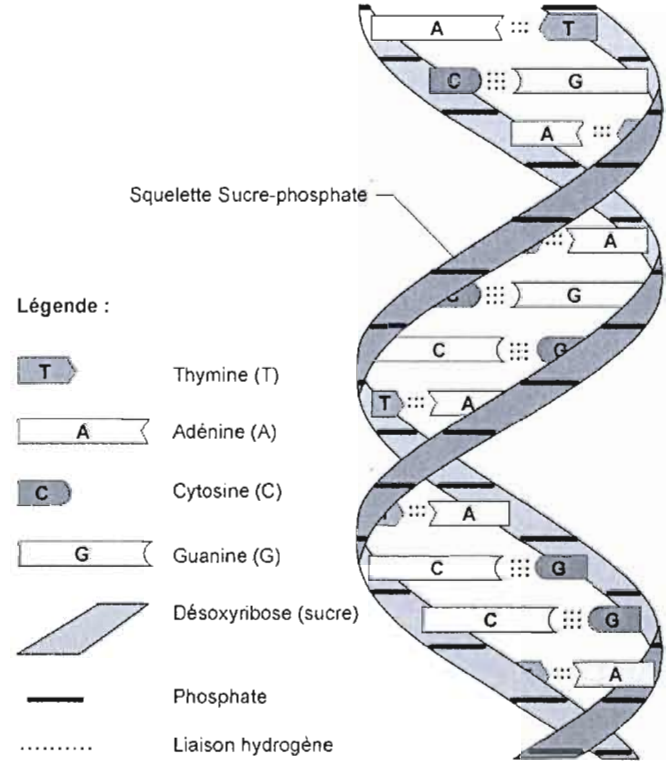

Figure 1.1 Complémentarité des brins d'ADN (Thibaut-Adrien, 2008).
Les interactions possibles sont: A-T et T-A, G-C et C-G

L'ADN se retrouve normalement sous cette forme stable de double brin.

Parfois la double hélice peut être ouverte afin de permettre la transcription ou la réplication. Cette ouverture est régie par des enzymes appelées *topoisomerases* (Champoux, 2001).

1.1.2 Théorie fondamentale de la biologie moléculaire

Francis Crick en 1958, annonce qu'une fois l'information codée dans des protéines, elle ne revient plus au niveau des acides nucléiques.

Le schéma général du transfert de l'information dans les systèmes biologiques est la suivante. L'ADN est copié en ADN (réplication), qui est copié en ARNm (transcription) qui sert de modèle pour la synthèse de protéines – traduction (Crick, 1970). Des exceptions à cette théorie générale sont bien connues à date comme la transcription inverse et la réplication de l'ARN.

Carl Richard Woese est l'auteur de la théorie du monde ARN. Conformément à cette théorie, il existait l'action catalytique au niveau de l'ARN. L'ADN serait apparu plus tard, pour prendre le rôle de dépositaire de l'information génétique. Aussi les protéines ont joué le rôle d'éléments catalytiques.

Dans son rôle actuel, l'ARN est un intermédiaire entre l'ADN et les protéines. Dans cette vision l'ARN ribosomal actuel reste un vestige (Woese, 1968; Gilbert, 1986). Il y a, à ce jour des virus qui stockent leur information au niveau de l'ARN.

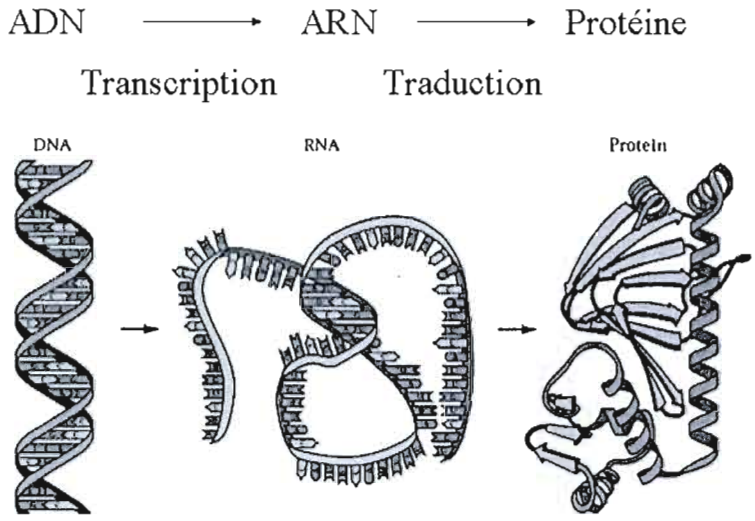

Figure 1.2. Théorie fondamentale de la biologie moléculaire (Le Guillou, 2009).

1.1.2.1 La transcription inverse

Il existe une famille de virus, appelés *rétrovirus*, dont le plus connu est le VIH, qui possèdent une enzyme, appelée, retrotranscriptase. Elle permet de transférer l'information codée au niveau du génome viral en ARN vers le génome humain en ADN.

Chez les eucaryotes, organismes plus complexes, dont les plantes et l'humain sont des exemples, on rencontre le cas des rétrotransposons. Les transposons sont des séquences d'ADN endogènes capables de se déplacer et de se multiplier dans le génome de l'hôte, sans pourtant être infectieux, à la différance des virus. À la différence des transposons, qui ont un intermédiaire ADN, les rétrotransposons ont un intermédiaire ARN. Chez le blé environ la moitié du matériel génétique est de cette origine (SanMiguel et Bennetzen, 1998).

La synthèse des extrémités des chromosomes – appelées télomères – se fait aussi par transcription inverse.

1.1.3 ARN

L'acide ribonucléique (ARN) est une macromolécule formée par la polymérisation de nombreux nucléotides, tout comme l'ADN. Les principales differences sont que la ribose remplace la désoxyribose dans l'ARN, qui n'est pas double brin comme l'ADN, mais bien simple brin. Les nucléotides sont l'adénine (A), la guanine (G), la cytosine (C) et l'uracile (U). L'uracile est la contrepartie de la thymine, et en vertu de cette relation bijective, les bases de données modernes présentent les séquences d'ARN codées comme celles d'ADN avec un T à la place du U.

L'ARN est présent dans le cytoplasme, les mitochondries ainsi que dans le noyau cellulaire. Il sert d'intermédiaire dans la synthèse des protéines. Il y a plusieurs types d'ARN correspondant à la fonction qu'ils ont dans la cellule.

L'ARN est une copie d'une région de l'un des brins de l'ADN. Les enzymes qui effectuent cette copie ADN→ARN s'appellent des ARN polymérases.

Tableau 1.1 Principaux types et fonction des ARN.

Nom	Acronyme	Fonction
ARN messager	ARNm	Assure le « plan de construction » d'une protéine.
ARN de transfert	ARNt	Apporte un acide aminé qui correspond à un codon.
ARN ribosomique	ARNr	Constitue le ribosome après la maturation et l'association à des protéines.
micro ARN	miARN	Entraînent le blocage de la traduction de certains ARNm par les ribosomes. Ils peuvent réguler l'expression de plusieurs gènes.

Par des appariements internes au sein d'une molécule simple brin l'ARN présente une structure secondaire. Cet ensemble d'appariements induit une topologie particulière, composée de régions en hélice (tiges) et de régions non-appariées (boucles) (Doty et al., 1959).

Les ARN non-codants sont impliqués principalement dans la régulation de l'expression des gènes (Projet ENCODE, 2007)

1.1.3.1 La réplication de l'ARN

Il existe plusieurs virus dont l'information génomique est stockée sous forme d'ARN. Ainsi la réplication du virus se fait au niveau de l'ARN. Les enzymes catalysant ce processus sont connues sous le nom d'ARN polymérases ARN dépendantes (Brown, 2002). Ces mêmes enzymes sont aussi impliquées dans l'inactivation de l'ARN chez les eucaryotes (Ahlquist, 2002).

1.1.4 Codon

Un codon est un triplet de nucléotides A, C, U ou G de l'ARN messager - ARNm. Il va se transcrire en un des 20 acides aminés naturels. Certains codons sont synonymes car plusieurs d'entre eux codent pour le même acide aminé.

1.1.5 Protéine

Une protéine est une macromolécule composée par une ou plusieurs chaînes d'acides aminés liés entre eux par des liaisons peptidiques (Brändén et Tooze, 1996). En fonction du poids moléculaire on les appelle peptides ou polypeptides en dessous de 10kDa, et protéines pour celles de plus grandes dimensions (Oliva, 2004).

Lors de la traduction l'ordre des aminoacides constitue la structure primaire de la protéine. Ensuite la protéine se replie sur elle-même à l'aide de liaisons hydrogènes pour former des structures secondaires, les plus importantes étant l'hélice alpha et le feuillet bêta. Les différentes structures secondaires s'agencent les unes par rapport aux autres pour donner la structure tertiaire. Les forces qui gouvernent ce repliement sont les forces physiques classiques. Il existe aussi une structure quaternaire qui décrit un ensemble d'unités peptidiques (Lodish et al., 2000).

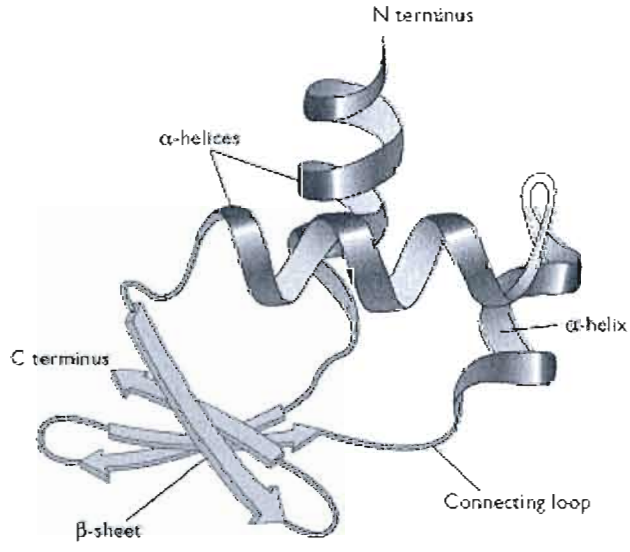

Figure 1.3 Structure d'une protéine. (D'après Turner et al., 1997, redessinée dans Brown, 2006).
Une protéine hypothétique qui comprend 3 hélices-alpha et 4 feuillets-bêta.

Les protéines constituent la base des fonctions cellulaires. Elles sont impliquées dans la catalyse de réactions chimiques, le transport, la communication, la signalisation et la

reconnaissance de signaux. Nombreuses protéines ont aussi un rôle structural comme par exemple celles qui forment la capside virale (Lodish et al., 2000).

1.1.6 Gène

Par définition, le gène est une séquence spécifique de nucléotides, localisée sur un chromosome, qui est une unité fonctionnelle de l'hérédité, contrôlant la transmission et l'expression d'un ou de plusieurs caractères, en spécifiant la structure d'un polypeptide particulier, d'habitude une protéine, ou en contrôlant une fonction, ou un autre matériel génétique (Merriam-Webster Online Dictionary, 2008).

Les termes «gène», ainsi que «phénotype» et «génotype», ont été introduits par le botaniste danois Wilhelm Johannsen (1905, 1909). Il appela les gènes, *unités facteurs*, *éléments* de l'hérédité. Ainsi les particules de Mendel, jusqu'alors des unités abstraites de l'hérédité deviennent des «gènes» (Johannsen, 1911). Morgan et ses collaborateurs, par cartographie des chromosomes, ont trouvé une réalité topographique en définissant des gènes-loci (Morgan et al, 1915; Le Guillou, 2009).

Le gène peut simplement être vu comme le modèle initial de la synthèse des ARN et protéines, comme l'indique la théorie centrale de la biologie moléculaire. L'aspect de contrôle de fonction ne peut pas être exclu de la définition à cause de l'importance du mécanisme régulateur des gènes (Jacob et Monod, 1961).

1.1.7 Cadre ouvert de lecture (ORF – open reading frame en anglais)

Le cadre *ouvert de lecture* est un gène potentiel - putatif. Il ne faut pas le confondre avec le *cadre de lecture,* qui est une séquence de triplets de nucléotides potentiellement translatable en un polypeptide, et qui est déterminé par le placement du codon d'initiation de la translation (Merriam-Webster Online Dictionary, 2008). Tous les cadres de lecture ne sont pas transcrits, il y en a 6 pour les 2 brins d'ADN.

Une séquence d'ADN, potentiellement transcrite en ARN débute par un codon d'initiation, celui qui correspond à la méthionine (Met), et se termine par un codon stop, celui-ci non transcrit. Entre ces deux codons, la phase ouverte de lecture contient un certain nombre de codons codant potentiellement une protéine. A l'aide de logiciels bioinformatiques on peut inférer la présence de ces cadres de lecture car aléatoirement sur 64 codons, dont 3 sont stop, la longueur moyenne d'une phase de lecture ne dépasse pas normalement une vingtaine de codons. Plus le cadre est long, plus on peut affirmer avec confiance qu'il sera transcrit.

1.1.8 Lignée

Une lignée est un regroupement de tous les ancêtres d'une même espèce.

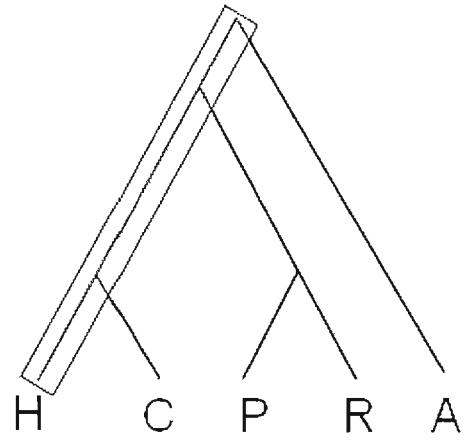

Figure 1.4 Lignée dans un arbre phylogénétique.
La figure représente un arbre phylogénétique. Les lettres représentent les symboles des espèces. Pour l'espèce *H*, Le rectangle gris représente la lignée.

1.1.9 Transition

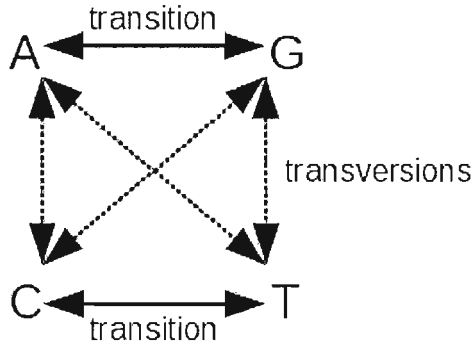

Figure 1.5 Transitions et Transversions.

La transition est une mutation au cours de laquelle le groupe est gardé le même. Une base purique est remplacée par une autre base purique ou une base pyrimidique par une autre pyrimidique (figure 1.5).

1.1.10 Transversion

La transversion est une mutation au cours de laquelle une base purique est remplacée par une base pyrimidique ou vice-versa, ainsi changeant de groupe (figure 1.5).

1.1.11 Espèce

Selon Mayr (1942): « Les espèces sont des groupes de populations naturelles, effectivement ou potentiellement interfécondes, qui sont génétiquement isolées d'autres groupes similaires. » L'espèce est donc une unité de population qui permet le flux génétique dans des conditions naturelles (de Queiroz, 2005; Mayr, 1996).

1.1.12 Souche

C'est une variante génétique, un sous-type de microorganisme. Dans certains contextes il peut être utilisé pour décrire des ancêtres d'une population.

1.1.13 Allèle

Un allèle est une version d'un même gène. Chaque allèle se différencie par une ou plusieurs différences au niveau de la séquence nucléotidique (composition et ordre des nucléotides).

1.1.14 Recombinaison dans les séquences virales

Comme on l'a déjà vu, les mutations constituent un mécanisme important de l'évolution. La recombinaison en est un autre. Il existe une recombinaison homologue où les segments échangés correspondent aux mêmes régions, dans les deux organismes ou hétérologue quand les segments échangés ne correspondent pas au même gène ou proviennent de grandes insertions-délétions.

1.2 Virus du Papillome Humain (VPH)

Le VPH est un virus à ADN d'une dimension d'environ 8kpb[3]. Le génome des virus génitaux code huit gènes et autant de protéines ainsi qu'une région régulatrice. Les gènes sont désignées avec les lettres E pour *Early – précoce* – et L – *Late* – tardif, en accord avec leur stage de différention de l'épithélium: E1, E2, E5, E6, et E7 sont exprimées tôt dans la différentiation, E4 est exprimée tout le long tandis-que L1 et L2 sont exprimées durant les stages finaux de différentiation (voir figure 1.6). Les protéines précoces sont exprimées à un niveau bas, ce qui peut expliquer les longues périodes de latence. L1 est la protéine majeure capsidale, L2 sert de lien avec l'ADN plasmidique (Shiffman et al., 2007 ; Doorbar, 2006). E1 et E2 sont des protéines régulatrices modulant la transcription et la réplication alors que E5, E6, et E7 modulent la transformation. Le rôle de la protéine E4 n'est pas encore complètement élucidé, plusieurs études indiquent qu'elle pourrait faciliter la réplication du génome et l'activation des fonctions tardives (Wilson et al., 2007), ainsi qu'elle pourrait être responsable de l'assemblage du virus (Prétet et al., 2007).

Il a été classifié en une première étape, ensemble avec les polyomavirus dans la famille des Papovaviridae grâce à leurs caractéristiques communes comme la capside non-envelopée, et le génome ADN à double brin (de Villiers et al., 2004). Le seul élément commun entre les deux familles est un domaine de la protéine E1 (de Villiers, 2004). Il code pour une hélicase qui est semblable à l'antigène T du virus simian 40 (SV40), un polyomavirus, à la protéine NSI des parvovirus et même à un élément extra-chromosomique dans un ver plat - *Girardia tigrina* (Rebrikov et al., 2002). L'antigène T du SV40 se lie au supresseur tumoral p53 et inhibe sa transcription (Dobbelstein et Roth, 1998). Plus de 200 génotypes de papillomavirus existent et plus de 100 ont été classifiés (ICTVdB, 2006).

---

[3] Kpb = kilo-paires de bases. Une mesure de la longueur d'une chaine d'acide nucléique qui équivaut à mille paires de bases (Merriam-Webster Online Dictionary, 2008). Comme il y a quatre paires de bases possibles, chaque position peut être codée sur quatre bits, bien que fréquemment dans les bases de données, elle est codée sur un octet. Ainsi en pratique, elle equivaut à un kilo-octet.

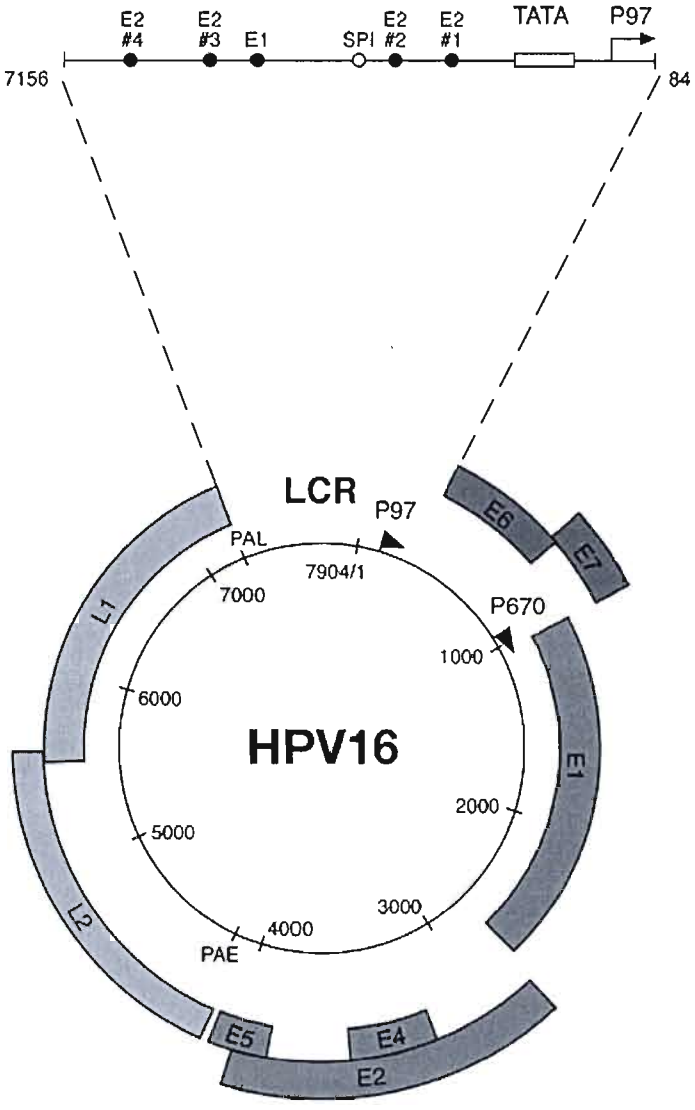

Figure 1.6 Structure du génome du VPH16.
Il a une taille de 7904 pb. Il est représenté comme un cercle noir, avec les promoteurs précoces y (p97) et tardifs (p670) en flèche noire. Les six ORF précoces (E1, E2, E4 et E5) E6 et E7, sont exprimées à partir de p97 ou p670 à de différents stages de la différentiation cellulaire, épithéliale. Les ORFs L1 et L2 sont eux aussi exprimés à partir de p670 en suivant un changement dans les modèles de splicing et un glissement dans l'usage des sites de polyadenilation, du précoce (PAE) au tardif (PAL). Tous les gènes viraux sont codés sur le même brin. La longue région de contrôle (LCR) de 7156 à 7184 est élargie pour permettre la visualisation des sites de liaison E2 et l'élément TATA du promoteur p97. La localisation des sites de liaison de E1 et SP1est aussi montrée. (Reproduit d'après Doorbar, 2006)

Traditionnellement, en se basant sur le tropisme on classifiait les VPH en trois grands groupes - cutané, muqueux, mixte (Segondy, 2008). Contrairement à plusieurs virus, la classification actuelle du VPH n'est pas basée sur des critères morphologiques mais plutôt sur des ressemblances génétiques (de Villiers et al., 2004). Elle a aussi changé au cours du temps étant toujours sujet de discussion. Une classification basée sur les ressemblances génomiques, la pathogénie et le potentiel à donner le cancer, divise les papillomavirus en Genres et Espèces comme Alpha, Bêta, Gamma-papillomavirus (de Villiers et al., 2004).

Le VPH est responsable de plusieurs infections sexuellement transmissibles les plus fréquentes. Certaines souches infectent les muqueuses génitales, d'autres se transmettent par contacts cutanés et infectent la peau. Les manifestations cliniques les plus connues sont les condylomes acuminés.

1.2.1 Carcinogenicité

Plusieurs souches du Virus du papillome humain seraient impliquées dans l'apparition du cancer du col de l'utérus (Schiffman et al., 2007). Les souches non-carcinogènes, tout comme le manque d'infection au VPH ne corrèlent pas ou corrèlent négativement aux modifications initiales du cancer du col de l'utérus (Castle et al., 2007). Cependant, la majeure partie n'est pas carcinogène, notamment les souches qui donnent des verrues vulgaires et les verrues plantaires. Plus de 40 génotypes infectent les muqueuses et parmi cela 13 à 18 types font partie de la catégorie haut-risque pour le cancer. Cette catégorie est considérée comme une pré-condition à l'apparition du cancer du col de l'utérus et, à forte probabilité, elle est impliquée dans la genèse d'une partie des cancers ano-génitaux et aéro-digestifs. Même les souches à moindre risque sont responsables d'une morbidité considérable et donnent des verrues génitales (Trottier et Franco, 2006).

Une étude dans 11 pays, sur 15 613 femmes âgées de 15 à 74 ans montre une prévalence très variable, allant de 1-4% en Espagne à 20 fois plus - 25·26% au Nigeria (Clifford et al., 2005). Cette forme de cancer est la deuxième plus fréquente chez les femmes au monde et la septième parmi toutes les formes de cancer.

Ceci est un problème global de santé publique, avec une estimation de 493,000 nouveaux cas et 274,000 décès au niveau de l'an 2002, partout au monde. Il est beaucoup plus fréquent dans les pays en voie de développement où 83% des cas apparaissent, avec 15% des cas nouveaux de cancer chez les femmes, comparativement à 3.6% dans les pays développés. Les différences s'expliquent par l'introduction des programmes de *screening* – détection précoce – en Europe, Amérique du Nord et Australie/Nouvelle Zélande, dans les années 1960-1970. Les fréquences sont en baisse nettement en Chine, pourtant en Afrique sous-Saharienne, Amerique Centrale, Asie du Sud et Centrale et en Mélanésie il demeure la principale forme de cancer chez les femmes (Parkin et al., 2005).

1.3 *Neisseria Meningitidis* (méningocoque)

Le méningocoque est une bactérie gram-négatif connue pour son rôle dans les méningites chez l'homme. Il se transmet par voie aérienne, par inhalation. Pour son invasivité, le contact avec des patients atteints de la maladie augmente le risque de transmission de 500 à 2000 fois (Peltola, 1983). Le site «neisseria.org» se met au service de la communauté de chercheurs pour centraliser l'information publique existante. Jusqu'à présent un nombre de 4 souches invasives et 3 asymptomatiques ont été sequencées. De très subtiles différences entre ces souches seraient responsables de leur virulence.

Le génome est constitué d'un chromosome circulaire avec une moyene d'environ 2.2 Mpb et un contenu en G+C d'environ 51%. Il y a en moyenne 1971 CDS, d'une longueur moyenne de 885 bp (Schoen et al., 2009). Vu le grand nombre de gènes, pour décrire l'ensemble commun, on utilise la notion de *pangénome*, qui dans ce cas serait étendu en moyenne sur 82% du génome (Tettelin et al., 2008; Schoen et al., 2009). Aussi pour définir l'ensemble de gènes pathogènes on définit le *pathogénome*. Les études de génomique comparée sur le *pathogénome* ne font pas unanimité (Snyder et al., 2005; Schoen et al., 2006). Apparemment les gènes potentiellement responsables de l'invasivité le sont pour la colonisation des tissus du naso-pharinx et moins pour l'invasivité des tissus hôtes (Schoen et al., 2009).

Les études de génomique comparative ont pourtant révelé toute une panoplie de mécanismes génétiques qui soutiendraient la flexibilité génomique. *Neisseria meningitidis* serait un paradigme pour les organismes qui utilisent la variabilité pour s'adapter à un environnement changeant et hostile (Schoen et al., 2009). Le génome se caractérise par l'abondance d'ADN répétitif qui constitue 20% de son génome (le quatrième rang dans une étude récente) et qui contribue à sa variabilité (Achaz et al., 2002). Dans une autre étude sur des familles de bactéries, qui utilise l'ordre des gènes comme mesure de stabilité *Neisseria meningitidis* s'est classé la 6-ème moins stable (Rocha, 2006). La recombinaison intra-génomique est utilisée pour générer la diversité phénotypique (Schoen et al., 2007, 2009). Plusieurs transferts horizontaux de gènes, originant dans la même ou d'autres espèces apparentées ont été identifiés (Maiden et al., 1996). Leur biologie est complexe, se divisant en éléments minimalement mobiles (Saunders et Snyder, 2002), îles d'ADN transféré horizontalement (Tettelin et al., 2000), îles génomiques canoniques (Hotopp et al., 2006) et phages deffectifs (Schoen et al., 2009). Par exemple, le seul facteur démontré à être associé à un type pathogène – la capsule polysacharidique – a été obtenu par transfert latéral (Elias et al., 2006).

La classification traditionnelle est sérologique. Elle est basée sur une combinaison de sérogroupe (différences dans la capsule), sérotype (porine – protéine majeure de membrane externe), sérosoutype (autres protéines de membrane externe), immunotype (lipooligozaharide). Les méningocoques isolés du sang et le liquide cérébro-spinal expriment plus souvent le sérogroupes A, B, C, Y et W-135, tandis-que ceux isolés à partir des porteurs asymptomatiques sont nongroupables ou ils expriment les sérogroupes capsulaires B, Y, X, Z, ou 29E (Tzeng et Stephens, 2000). À cause de la grande fréquence de variation des structures membranaires, le système sérologique de classification n'est pas assez discriminatif, pour distinguer les méningocoques (Tzeng et Stephens, 2000). D'autres méthodes basées sur la génomique ont été développées, comme: *multilocus enzyme electrophoresis typing* – ET, et *multilocus sequence typing* – MLST - (Maiden et al., 1998).

L'Afrique subsaharienne est la plus touchée par les épidémies (Tikhomirov et al., 1997; Pinner et al.,1992). La plus large épidémie enregistrée, a eu lieu en 1996-97, avec 30 000 décès sur 300 000 cas, due au serogroupe A (Tikhomirov et al., 1997). Pendant une

épidémie aux États-Unis en 1989-91, des taux d'attaque de la maladie au sérogroupe C, de 1% ont été enregistrées (Jackson et al., 1995). Des nombreuses épidémies ont eu lieu en Europe, États-Unis, Canada, Chine, Népal, Mongolie, Nouvelle Zélande, Cuba, Brésil, Chili, Arabie Saoudite, et en Afrique du Sud depuis 1980 (Tzeng et Stephens, 2000).

N. meningitidis est un agent commensal[4] et pathogène exclusif de l'homme. Entre les épidémies il demeure asymptomatique dans le nasopharynx de 5-10% des adultes (Greenfield et al.,1971). Cette colonisation est importante pour la création d'une immunisation protectrice. Ce sont des germes qui rarement deviennent pathogènes (Schoen et al., 2009). Le plus fréquemment, les personnes porteuses ne développent aucun symptôme et peuvent garder cet état pendant plusieurs mois ou années. Chez les enfants, le risque d'infection et de maladie grave peut être 20 fois plus grand que chez les adultes. Au cours des épidémies, ce profil se déplace vers les enfants plus grands, les adolescents et les adultes (Tzeng et Stephens, 2000).

Quelques groupes clonaux, génétiquement définis, donnent la majorité des infections, comme le complexe ET-37, ET-5 ou le cluster A4 (Maiden et al., 1998). Les groupes invasifs donnent plus de cas de maladie, surtout quand nouvellement introduits dans une population - le complexe ET-37 (Raymond et al., 1997) Par contre les clones rencontrés chez les porteurs nasofaringeans asymptomatiques rarement produisent la maladie, même en présence d'un fort pourcentage d'acquisition et transmission (Jones et al, 1998 ; Caugant et al., 1988). PubMLST (une base de données publique) regroupe l'information épidémiologique et la classification des souches déjà publiées dans des articles scientifiques. Récemment on a enregistré des souches résistantes à la pénicilline, qui ont causé des maladies au Canada

---

[4] Une bactérie commensale obtient des nutriments sans endommager ni profiter à son hôte (Merriam-Webster Online Dictionary, 2008).

(Blondeau et al., 1995) La résistance au chloramphénicol a aussi été rapportée en Asie du sud-ouest (Tzeng et Stephens, 2000).

CHAPITRE II

ÉVOLUTION ET ANALYSE DES SÉQUENCES

Dans ce chapitre, nous présentons d'abord la notion d'évolution des espèces qui a révolutionné l'étude du monde vivant. Des méthodes de classification la prenant en hypothèse se sont ensuite développées. Mais c'est avec l'avènement des données moléculaires massives et le développement d'outils bioinformatiques reposant sur de modèles statistiques que ce nouveau paradigme s'est définitivement imposé.

Nos algorithmes s'appliquent à la détection des séquences fonctionnelles (parties du génome dont l'existence, la composition et la structure ont une relation documentée à une fonction moléculaire). Différentes approches sont abordée à la fin du chapitre.

2.1 Notions d'évolution

2.1.1 La biologie évolutive et l'introduction de la phylogénie

Charles Darwin a initié les bases de cette discipline scientifique par son travail publié en 1859 sous le titre «On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life». Il a introduit une théorie de l'évolution des populations au cours des générations par un processus de sélection naturelle. Il a illustré la diversité de la vie, et a présenté des arguments pour un modèle de branchement de l'évolution et de descendance commune. Il y a inclus des arguments accumulés au cours de son voyage autour du monde à bord du Beagle en 1830, de même que ses travaux subséquents en recherche et expérimentation. Ainsi, il était également le promoteur de la

phylogénie comme modèle de classification. D'ailleurs la seule figure du livre, reproduite à la figure 2-1 est un arbre phylogénétique.

De son côté, Ernst Haeckel a formulé la théorie conformément à laquelle l'ontogénie – le développement des organismes individuels – récapitule la phylogénie – histoire évolutive des espèces. Bien que cette théorie soit contredite aujourd'hui, Haeckel a utilisé avec succès les arbres phylogénétiques pour décrire les relations entre les espèces. La figure 2.2 présente un modèle progressif, presque linéaire de l'évolution de l'humain, en discordance avec la vision de Darwin, ayant de nombreux branchements.

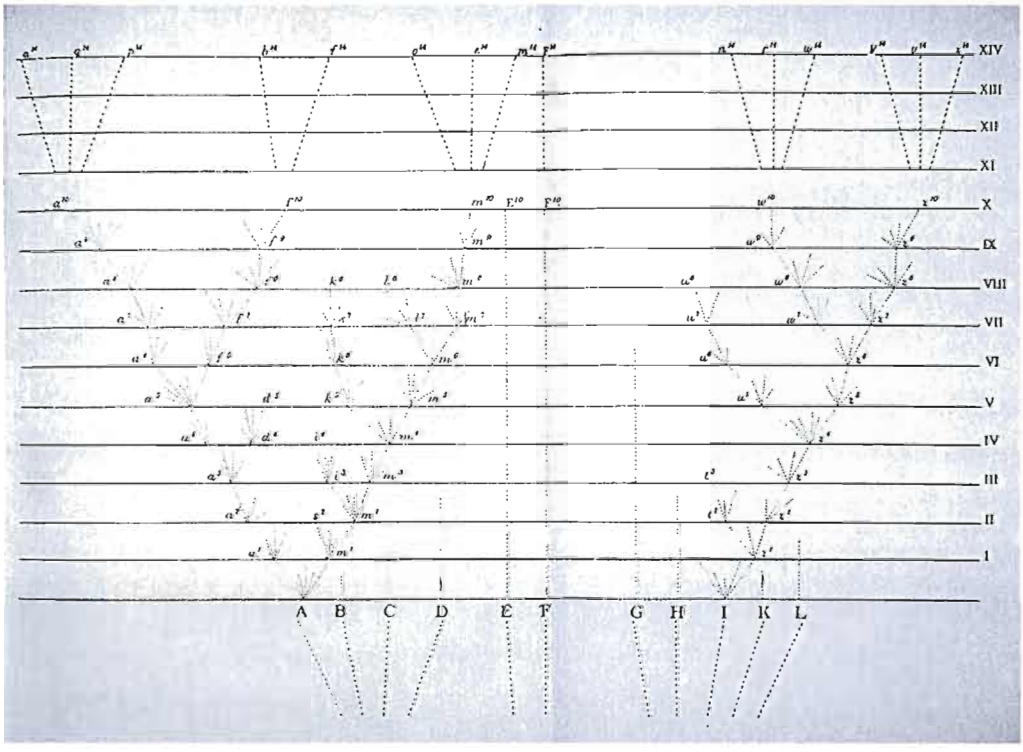

Figure 2.1 Arbre phylogénétique (Darwin, 1859).

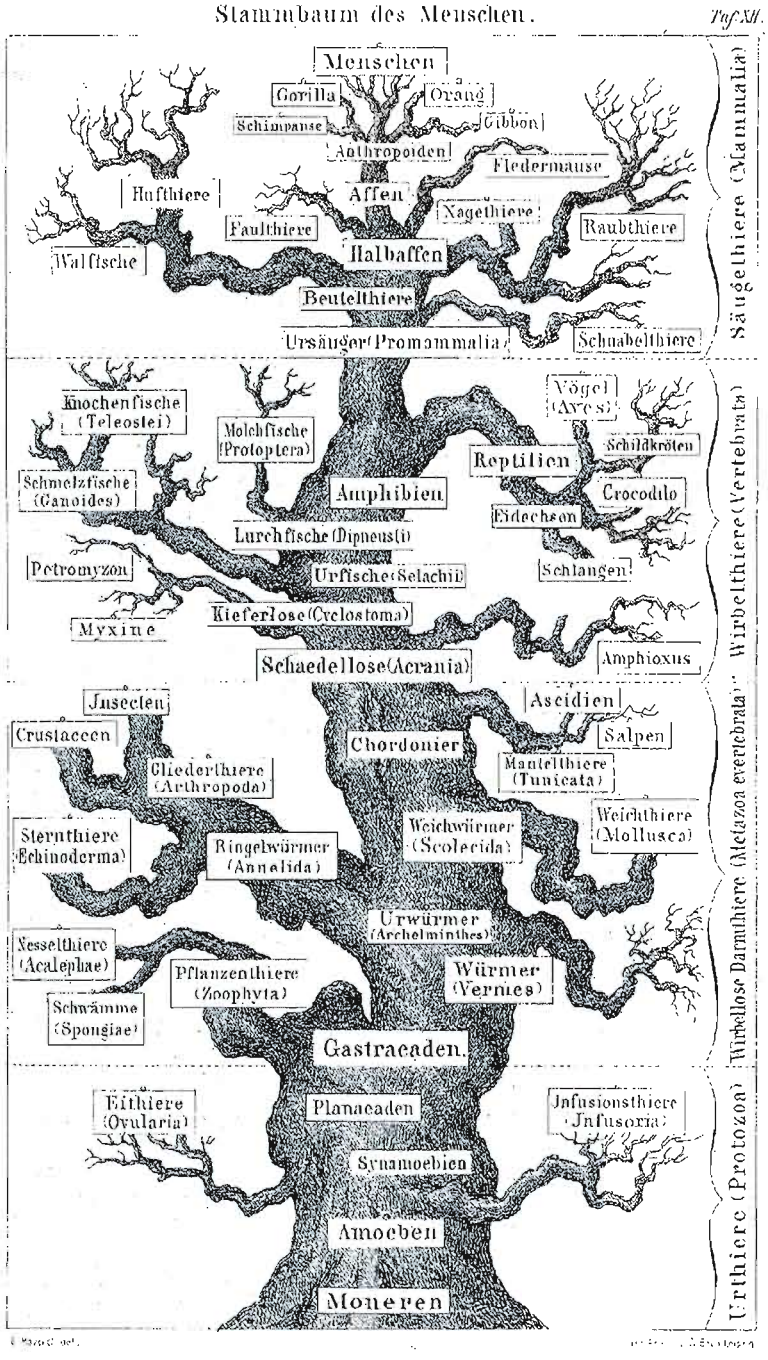

Figure 2.2 Généalogie de l'homme, lithographie (Haeckel, 1874).

2.1.2 L'arbre de vie

Bien que l'existence du concept d'arbre de vie soit bien connu, c'est à Charles Darwin que nous devons une interprétation en sciences (Darwin, 1872). Ernst Haeckel a présenté le premier arbre universel du vivant, représentant toutes les espèces et tous les groupes connus à l'époque. Durant plusieurs décennies, les seuls moyens de classification des espèces étaient l'ensemble des traits observables - d'abord les caractères anatomiques, morphologiques, physiologiques, éthologique, ensuite moléculaires, caractérisant un être vivant donné - ex: couleur des yeux, des cheveux, glycémie, etc.

Puis, avec la découverte du règne *Archaea* (Woese et al., 1978), l'utilisation de la taxonomie phylogénétique basée sur l'unité 16S de l'ARN ribosomal (Woese et Fox, 1977) est devenue la méthode standard. Woese apporta ainsi de sérieux arguments en faveur de la classification moléculaire du vivant (Woese et al., 1990).

De nos jours, avec le séquençage complet ou partiel de plusieurs génomes, les arbres phylogénétiques sont inférés à partir de segments génomiques couvrant plusieurs gènes, voir des génomes entiers. Ainsi nous nous dirigeons peu à peu vers l'inférence phylogénomique. Par ailleurs, la quasi-totalité des arbres phylogénétiques est inférée de nos jours à partir de données génomiques (Pierce, 2007). Ce type de données est donc recommandé pour reconstruire des phylogénies de nombreuses espèces sur une longue période de temps – âges géologiques et horloge moléculaire.

La classification phylogénétique du vivant reste un sujet complexe, avec encore de nombreuses incertitudes (Lecointre et al., 2009). Un des sujets les plus disputés reste l'enracinement de l'arbre que des nombreux transferts latéraux (i.e., horizontaux) rendent plus complexe (Becerra et al., 2007). Il est important de noter que pour éviter l'influence des transferts latéraux de gène lors de la reconstruction phylogénétique, les gènes anciens distribués sur toutes les lignées de la vie avec très peu ou pas de transfert, sont utilisés (ARN 16S et 12S).

Il existe actuellement un projet collaboratif sur la toile, appelé *Tree of Life Web Project* (ToL). Il regroupe les efforts des biologistes et enthousiastes de la nature de partout au monde, en compilant le travail de centaines d'experts et contributeurs amateurs (Maddison et Maddison, 1996; Maddison et Schulz, 2007).

2.1.3 Évolution moléculaire

L'évolution dépend de deux facteurs, la variabilité génétique et le changement de la fréquence des allèles dans la population au fil du temps (Duret, 2008). La source de cette variabilité est principalement constituée de mutations - incluant les insertions, substitutions, délétions, duplications, translocations et transferts horizontaux de gènes.

La sélection naturelle est en partie responsable du sort des mutations qui modifient la *valeur d'adaptation* des organismes. D'un coté les allèles qui confèrent plus de valeur ont la tendance d'augmenter leur proportion, jusqu'à la fixation, qui va remplacer l'allèle original dans la population, par le processus de sélection positive. Les allèles qui réduisent la valeur d'adaptation, par contre sont soumis au processus de sélection négative ou sélection épurante - *purifying selection*. Certaines allèles qui confèrent des avantages à l'état hétérozygote (une partie des allèles pour le même locus génétique) sont maintenus à une certaine proportion dans la population par le biais de la sélection balancée (Duret, 2008). Les mutations qui n'affectent pas la valeur d'adaptation des individus ne sont pas affectées par la sélection naturelle mais par la dérive génétique de la population (Duret, 2008).

Conformément à la théorie séléctionniste, la sélection naturelle joue le rôle primordial, les processus non-adaptatifs ne comptant que pour des contributions mineures à l'évolution. Ainsi les différences entre espèces sont données par des mutations fixées par sélection positive dans la population en conséquence de l'adaptation à l'environnement et le polymorphisme est le résultat de la sélection balancée (Duret, 2008).

Motoo Kimura a proposé la théorie de *l'évolution moléculaire neutre* (Kimura, 1968) soutenue par King et Jukes avec leur *évolution non-Darwinienne* (King et Jukes, 1969).

Elle affirme que l'immense majorité des changements moléculaires sont causés par la *fixation aléatoire* des mutants (due à la *dérive génétique* des populations de dimension finie) à valeur adaptative neutre sous le flux continu des mutations. Elle affirme aussi que les polymorphismes de l'ADN et protéines – qui forment la variabilité à l'intérieur d'une même espèce, sont sélectivement neutres ou presque neutres, et sont maintenues dans l'espèce par la balance entre les *entrée mutationnelles - mutational input*, et *l'extinction aléatoire - random extinction* (Kimura, 1991).

2.1.3.1 Mutations

Les mutations représentent la principale force de l'évolution, en créant des changements permanents au matériel génétique. Elles ont à l'origine, des erreurs de la division cellulaire, en particulier, de la réplication de l'ADN, mais aussi l'exposition aux radiations, substances chimiques et virus. Parfois elles sont engendrées de manière contrôlée au cours de la division des lignées cellulaires reproductives (méiose) ou l'hyper mutation nécessaire à la production des anticorps.

En fonction de l'effet produit, elles peuvent être défavorables (comme une interruption d'une fonction cellulaire importante), favorables ou neutres. Les mutations neutres ne modifient pas la capacité de l'organisme de survivre et se reproduire dans l'environnement et peuvent s'accumuler au fur du temps.

On différencie les mutations à petite échelle impliquant un ou quelques nucléotides – insertions, délétions et substitutions, et les mutations à grande échelle qui modifient de grandes régions génomiques, en général de la taille d'un gène ou d'un chromosome – translocations, duplications, délétions et le transfert horizontal de gènes.

2.1.4 Les causes de la variation dans la fréquence des allèles

Certains caractères phénotypiques sont le résultat d'une ou plusieurs conformations de gènes. Ici on s'intéresse au processus qui peut modifier la fréquence des allèles dans une population.

2.1.4.1 La dérive génétique

Elle provient des erreurs d'échantillonnage des allèles à la traversée des générations et barrières géographiques. Plus la population est petite, plus elle est importante due à l'impossibilité de maintenir la diversité des gènes dans la population originale.

Cette force est entièrement régie par la chance et mène à l'homogénéisation, en éliminant progressivement certaines allèles. Quand un allèle atteint une fréquence de 100%, on dit qu'il est *fixé,* et quand il atteint 0%, on dit qu'il est *perdu.*

2.1.4.2 Le flux génétique ou migration

Le transfert des allèles à une autre population agrandit la diversité intra-population et baisse celle inter-population, agissant comme un frein à la spéciation.

2.1.5 La sélection

La sélection naturelle en particulier est produite par une mortalité et fertilité différentielle. La sélection sexuelle est aussi importante, elle est composée d'un couple formé d'une caractéristique et d'une autre qui est l'attraction pour la première caractéristique.

2.2 Méthodes de classification

2.2.1 Classification phylogénétique - cladistique

Les temps modernes ont vu l'apparition de la classification phylogénétique, un système de classification des êtres vivants (Hennig, 1950). Elle se base sur les rapports de proximité évolutive entre espèces. Cette classification se veut plus objective, seuls les caractères empiriquement observables et propres au rang d'espèces sont retenus comme uniques témoins de l'héritage ancestral de chaque espèce (Lecointre et Guyader, 2002). Cette objectivité est pourtant fortement liée à la vision évolutive moderne.

La représentation schématique est donnée par un cladogramme, qui est un arbre phylogénétique non enraciné, contenant des nœuds et des feuilles (appelés taxons). Les groupes qui incluent un ancêtre commun et tous ses descendants sont appelés groupes monophylétiques (voir figure 2.3) (Hennig, 1975). Le taxon est une entité qui regroupe tous

les organismes vivants possédant en commun certains caractères taxonomiques ou diagnostiques bien définis (Allaby, 2009).

Pour réconcilier la notion de clade et celle de taxon il faut définir des noms pour les clades. Ainsi a été introduite la classification cladistique – systématique phylogénétique. Bien qu'elle apporte de puissants outils d'analyse, elle a eu beaucoup de difficultés à s'imposer (Hennig, 1975). La taxonomie moderne est régie par différents Codes Internationaux de Nomenclature (zoologique, botanique, bactérienne – ICZN (Allaby, 2009) et ICBN). Ce sont des classifications typologiques.

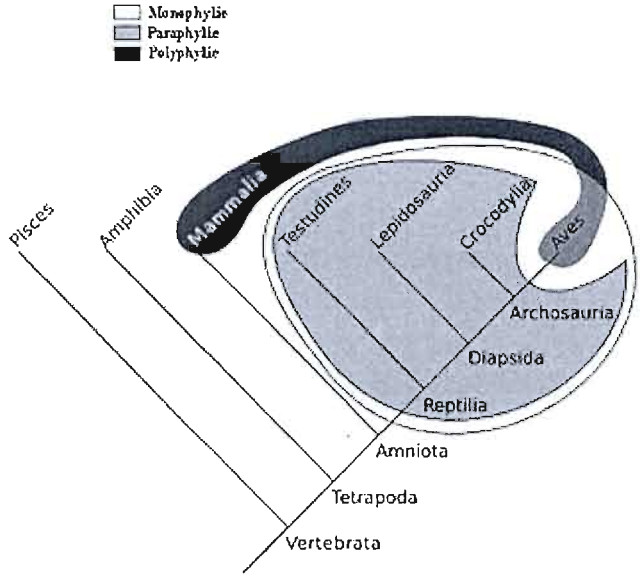

Figure 2.3 Différence entre monophylie, paraphylie et polyphilie (Kintaro, 2008).
Le groupe des « sauropsides » (ici en gris pale), regroupe les reptiles et oiseaux. Il est consideré comme monophyletique car tous les descendants d'une même ancêtre sont présents. Si on enlève les oiseaux, on obtient le groupe des reptiles (en gris), qui est un groupe paraphyletique car ils ne représente qu'une partie de cette descendance. Les mammifères et les oiseaux forment ensemble le groupe « animaux à sang chaud » (en gris foncé). Ceci est un groupe polyphiletique car ses membres n'ont pas les mêmes ancêtres.

Notre âge moderne, avec son paradigme évolutionnaire non-typologique, se retrouve plutôt dans une nouvelle classification phylogénétique appelée PhyloCode (Sluys et al., 2004). Malgré les nombreux arguments en sa faveur, elle n'a pas encore réussi à remplacer la classification traditionnelle (Pleijel et Rouse, 2003). Des différences subtiles entre holo-phylie et mono-phylie, ainsi qu'entre modèle et réalité substantielle, rendent ce sujet complexe et le débat encore ouvert (Envall, 2008).

2.2.2 Arbres phylogénétiques

2.2.2.1 Définition

Un arbre phylogénétique est une structure mathématique d'arbre - un cas particulier de graphe qui n'a qu'une seule source et aucun cycle.

Il contient quatre éléments principaux :

- La racine (au cas où l'arbre serait enraciné) qui indique l'ancêtre commun des espèces représentées dans l'arbre.
- Les nœuds externes (ou feuilles) représentent les espèces contemporaines pour lesquelles les informations ont été disponibles lors de la construction de l'arbre. Ils sont communément appelés taxons.
- Les nœuds internes qui représentent des ancêtres inférés, hypothétiques.
- Les branches (ou arêtes) de l'arbre qui montrent les relations de descendance entre les nœuds (i.e. taxons). Ces arêtes peuvent avoir des longueurs. Ces longueurs peuvent correspondre à plusieurs informations dont, le taux de mutation, la distance génomique, etc.

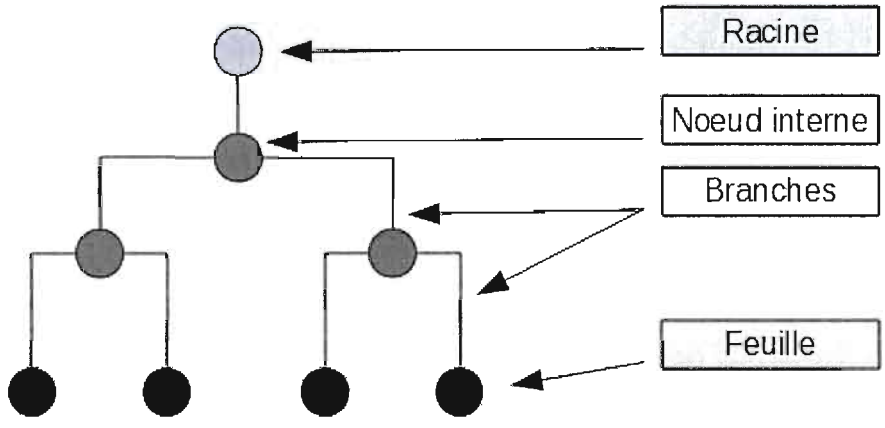

Figure 2.4 Exemple d'un arbre phylogénétique.

2.2.2.2 Caractéristiques

Le degré d'un nœud de l'arbre est défini comme le nombre de branches adjacentes à ce nœud. Tout nœud de degré supérieur à trois est appelé *non résolu,* sinon le nœud est *résolu.* Un arbre phylogénétique binaire *non-enraciné* ayant n espèces (feuilles) dont tous les nœuds internes sont résolus est composé de :

- $2n$-2 nœuds (n-2 nœuds internes et n feuilles) et
- $2n$-3 branches.

Lorsque l'ancêtre commun de toutes les espèces de l'arbre est identifié, l'arbre est enraciné. Il est orienté dans le sens du temps d'évolution des espèces. L'arbre enraciné permet de définir une relation de descendance entre deux nœuds successifs. Objectivement, il est impossible d'identifier l'origine de diversification des espèces mais la méthode de l'outgroup est souvent utilisée. Dans cette méthode, on ajoute aux séquences à traiter, avant le calcul de l'arbre, une espèce (i.e. séquence) très éloignée. Le nœud-racine père de cette séquence sera la nouvelle racine.

2.2.3 Méthodes de reconstruction d'arbres phylogénétiques

Il existe deux approches principales pour l'inférence des arbres phylogénétiques.

2.2.3.1 L'approche phénétique

La phénétique repose sur l'hypothèse que le degré de ressemblance est corrélé avec le degré de parenté. Ainsi elle essaye de quantifier la ressemblance entre les êtres vivants à classer. Elle se base sur le calcul de distances entre les espèces (i.e. matrice de dissimilarités) pour reconstruire un arbre phylogénétique, en effectuant un regroupement hiérarchique. L'arbre phylogénétique résultant est appelé phénogramme, où la longueur des branches dépend de la distance génétique.

Plusieurs algorithmes bioinformatiques très efficaces sont basés sur cette approche - par exemple Neighbour Joining (Saitou et Nei, 1987) et UPGMA (Sneath & Snokal, 1973). Elle tend à être moins utilisée car les méthodes employées relèvent plutôt de la classification traditionnelle, avec notamment l'analyse de similarité. En effet, avant l'approche cladistique et la découverte des données génomiques, la classification des espèces se faisait par comparaisons entre leur morphologie, leur comportement et leur répartition géographique. Cette ancienne méthode était peu pertinente en raison des analogies : certaines ressemblances entre êtres vivants ou taxons ne pouvant être attribuées à une ascendance commune.

Un exemple bien connu est celui des ailes des oiseaux et des chauves-souris dont les caractères analogues ne sont pas hérités d'un ancêtre commun ailé, mais qui ont évolué indépendamment vers une forme ailée due à la sélection naturelle.

Lorsque le nombre de caractères pris en compte est réduit, comme par le passé, il est impossible d'inférer la classification correcte. L'approche phénétique moderne a pris le nom de Phénétique Numérique, où un nombre important de caractères est pris en comte, des centaines si possible (Ridley, 2004).

L'ADN, l'ARN et les protéines sont des molécules polymères. Chaque résidu de la molécule (nucléotide pour l'ADN et l'ARN ou acide aminé pour la protéine) peut être considéré comme un caractère. Ainsi cette technique est très puissante lorsque appliquée au niveau moléculaire.

Il faut mentionner que si la méthode est employée comme phénétique, mais avec l'intention de déduire l'histoire évolutionnaire, elle devient une méthode de cladistique (Mayer, 1993).

2.2.3.2 L'approche cladistique

L'approche cladistique ne rapproche dans un arbre phylogénétique que les êtres vivants qui partagent des caractères homologues: lorsqu'une ressemblance entre deux taxons peut être attribuée à une ascendance commune, on parle d'homologie. Les êtres vivants possédant le caractère homologue descendent d'un ancêtre commun. Tous ceux qui ne possèdent pas le caractère homologue ne descendent pas d'ancêtres communs et sont jugés donc éloignés génétiquement. La cladistique repose sur l'identification de l'homologie des caractères. Ceci fait que la qualité de l'alignement des séquences ou d'autres méthodes de détection de l'homologie jouent un rôle plus important dans cette approche que dans l'approche phénétique.

La cladistique se base sur un modèle d'évolution et infère un arbre optimal en fonction de ce modèle et une évaluation des possibles ancêtres au niveau de chaque nœud. Cette approche basée sur la généalogie, va de pair avec la classification phylogénétique, tout en donnant des résultats d'une meilleure qualité putative. Les désavantages sont liés aux difficultés de calcul et aux nombreuses topologies d'arbres à investiguer.

Parmi les approches cladistiques nous avons : maximum de parcimonie, maximum de vraisemblance et la méthode bayésienne. Nous allons présenter l'approche de maximum de vraisemblance plus en détail. Cette approche est la plus utilisée de nos jours car elle est paramétrisable et donne de meilleurs résultats pour plusieurs modèles d'évolution (Kolaczkowski et Thornton, 2004). Pour ces mêmes raisons, elle a aussi été employée dans le cadre de nos travaux.

2.2.4 Le maximum de vraisemblance

Étant donné un échantillon de distribution d'une variable aléatoire, l'estimation du maximum de vraisemblance est une méthode statistique générale, permettant d'inférer les

paramètres de la distribution de probabilité subséquente, qui maximise la probabilité de l'échantillon.

Cette méthode a été développée par le statisticien et généticien Ronald Fisher entre 1912 et 1922. Depuis elle a été employée en bioinformatique par Edwards et Cavalli-Sforza (1963) dans le cadre de l'étude des données sur la fréquence des gènes. La première application d'une méthode de maximum de vraisemblance aux séquences moléculaires a été effectuée par le statisticien Jerzy Neyman (1971).

Étant donné une famille paramétrisée de fonctions de densité, où θ est le paramètre et x est le résultat de l'expériment :

$$x \mapsto f(x \mid \theta),$$

la *fonction de vraisemblance* est :

$$\theta \mapsto f(x \mid \theta),$$

écrite sous la forme :

$$L(\theta \mid x) = f(x \mid \theta),$$

où f(x | θ) est une *fonction de densité de probabilité*, lorsque x est la variable et θ est le paramètre et f(x | θ) est une *fonction de vraisemblance*, lorsque θ est la variable et x le paramètre.

2.2.5 Inférence d'arbres phylogénétiques basée sur le maximum de vraisemblance

L'application aux arbres phylogénétiques suppose un modèle d'évolution qui nous permet de calculer les probabilités de transition en fonction du temps écoulé - appliqué à la longueur des branches. Aussi on suppose que l'évolution est indépendante pour les différents sites et lignées. Pour trouver l'arbre le plus vraisemblable, les nucléotides de toutes les

séquences à chaque site sont considérés séparément, la probabilité étant le produit des probabilités individuelles. Bien que ce soit un avantage, de pouvoir prendre en compte des variations du taux de substitution, dans la pratique une variable continue est très difficile à calculer. Une bonne approximation est d'utiliser un nombre limité des classes de taux de variation. Si le modèle d'évolution employé est réversible, alors on obtient un arbre non enraciné.

1ère étape : *Générer des topologies d'arbres* :

Cette étape consiste à trouver la meilleure topologie d'arbre, mais ici la vraisemblance ne nous aide pas. La recherche heuristique se fait à partir d'une topologie de départ, reconnue pour donner rapidement des résultats assez fiables, comme NJ (Saitou et Nei, 1987), puis à l'aide d'algorithmes gloutons ou *branch-and-bound* on converge rapidement vers un minimum local. Pour éviter de tomber dans un minimum local on utilise des techniques de modifications des arbres comme :

- échange avec le voisin le plus proche - *Nearest Neighbour Interchange* (NNI),
- élagage et transplantation de sous-arbres - *Subtree Pruning and Regrafting* (SPR),
- bissection et reconnexion d'arbre - *Tree Bisection and reconnection* (TBR).

Comme le problème n'a pas été prouvé comme NP-complet, des alternatives à la recherche heuristique ont été développées. Un algorithme de maximisation de l'espérance mathématique qui garantit que les étapes itératives ne diminuent jamais le score, a été proposé (Friedman et al., 2002). Pour éviter de tomber dans un minimum local, une étape de recuit simulée a été rajoutée à la méthode. Dans certains cas les arbres obtenus par cette méthode sont les mêmes que ceux obtenus par la méthode de maximum de parcimonie. Les rapprochements des deux méthodes ont été analysés par Tuffley et Steel (1997).

2ème étape : *Optimiser la longueur des branches* :

À l'aide des méthodes d'optimisation numérique de type Newton-Raphson, pour une même topologie, les meilleures longueurs de branches sont inférées (Schadt et al., 1998;

Guindon et Gascuel, 2003). Des algorithmes génétiques ont aussi été employés (Shen et Heckendorn, 2004).

3ème étape : *Calculer la vraisemblance d'un arbre donné* :

Pour un arbre ayant des longueurs de branches précises, une fonction est employée pour évaluer un score de vraisemblance. Felsenstein (1981) a développé une méthode d'élagage qui permet de réduire le nombre de calculs en se basant sur les dépendances de données et la mise à jour des états déjà calculés. Cela est effectué en reconstruisant le dispositif ancestral à tous les nœuds de l'arbre considéré. Cette méthode ne garantit pas cependant l'obtention d'un arbre de vraisemblance maximale, le résultat dépendant de l'ordre de calcul des ancêtres. On peut essayer plusieurs ordres et choisir la version qui donne le meilleur score. Ces heuristiques permettent de maîtriser le temps de calcul.

Plus les modèles sont simples, et des restrictions sont imposées (comme par exemple des substitutions indépendamment identiquement distribuées aux différents sites, un processus homogène de substitution, processus identique au long des branches de l'arbre phylogénétique) plus les arbres obtenus sont semblables ou même identiques (Yang, 1994, 1996). D'autant plus proche de l'arbre réel est le résultat que le modèle choisi est plus proche de celui qui a généré les substitutions observées. Kuhner et Feslenstein, (1994) ont fait des simulations sur les différents modèles choisis et ont démontré que des différences importantes apparaissent dans les résultats quand les taux de substitution ne sont pas homogènes. Sur l'analyse de données réelles, les modèles plus complexes, comme HKY (Hasegawa et al., 1985), et le modèle général réversible GTR (Tavaré, 1986), ont donné de meilleurs résultats que les modèles très simples ou très complexes, étant un compromis en termes d'analysabilité / réalité.

Les comparaisons entre l'approche de maximum de vraisemblance (MV) et celle de maximum de parcimonie (MP) donnent des résultats variables en fonction des jeux de données analysés. Par exemple, Suzuki et Nei (2001), en évaluant les deux approches pour la détection de la sélection positive aux sites individuels d'aminoacides, trouvent l'inférence MP robuste à l'usage de différents modèles évolutionnaires et en général plus fiable que MV. Comme MV dépend des modèles et des paramètres initiaux, MP est à recommander quand le

nombre des séquences est relativement grand et la longueur des branches dans l'arbre phylogénétique est relativement petite. Par contraste Sorhannus (2003), trouve MV plus fiable que MP, dans la détection toujours, de la sélection positive dans les sites d'aminoacides, sur un jeu de données différent. Enfin, Zhang et Nei (1997) estiment que MV est plus fiable quand la divergence des séquences est grande et les branches de l'arbre sont longues. Un autre avantage de MV est que si la topologie de l'arbre est partiellement incorrecte, les résultats ne sont pas affectés pour la partie correcte.

Cette approche permet également d'appliquer les différents modèles d'évolution et d'estimer la longueur des branches en fonction des changements évolutifs.

Lors du développement initial des méthodes de maximum de vraisemblance le temps de calcul et la puissance des ordinateurs limitaient son utilisation, de sorte que l'Edwards et Cavalli-Sforza ont eu recours à la parcimonie comme approximation. Les algorithmes qui trouvent le score de vraisemblance font des recherches dans un espace multidimensionnel de paramètres, ce qui réduit leur applicabilité à des données à grande échelle. Ces algorithmes sont aussi très sensibles aux erreurs systématiques quand le modèle d'évolution utilisé ne reflète pas le processus actuel d'évolution, surtout pour les modèles simples (Zhang et Nei, 1997).

Plusieurs programmes sont disponibles dont DNAML (DNA maximum likelihood program) de Felsenstein (1981), ProtML (Adachi et Hasegawa, 1992), PUZZLE (Strimmer et Von Haeseler, 1996) et PHYML (Guindon et Gascuel, 2003). Cette dernière méthode permet de traiter un très grand nombre d'espèces en un temps raisonnable.

2.2.6 Critères d'évaluation des méthodes d'inférence d'arbres phylogénétiques

Plusieurs critères de qualité sont utilisés (Penny et al., 1992) :

- Efficacité (complexité asymptotique, espace mémoire).
- Puissance (bonne utilisation des données sans perte d'information).
- Consistance (convergence vers une même solution avec des données différentes et même structure du problème).
- Robustesse (bons résultats en dépit de la violation du modèle initial choisi).
- Falsifiabilité (possibilité de reconnaître quand les résultats de la méthode sont inappropriés).

2.2.7 Alignement multiple

L'alignement est une opération qui permet de comparer des séquences de nucléotides pour repérer les éléments correspondants, appelés homologues. L'homologie se réfère au concept d'évolution à partir d'un ancêtre commun. Cette opération est nécessaire pour toutes les comparaisons de séquences. Les espaces dans les alignements sont alors considérées comme des insertions-délétions de nucléotides. Cette opération est très gourmande en ressources. L'approche naïve est NP – complète. Il existe une solution exponentielle (Lipman et al., 1989). Elle a été implémentée dans le programme *MSA*, que Gupta et al. ont plus tard optimisé (1995).

Diverses heuristiques ont été développées, comme l'alignement progressif ou itératif par modèles de Markov cachés, par motifs et par la refonte dirigée. L'alignement progressif aligne une à une les séquences avec celles déjà alignées. Le plus souvent les séquences les plus similaires sont alignées en premier. Pour retrouver cet ordre on utilise fréquemment un arbre phylogénétique comme guide.

La méthode d'alignement progressif la plus connue est *ClustalW*, d'après le programme du même nom (Thompson et al., 1994). *T-Coffee* (Notredame, 2000) fait partie de la même catégorie de logiciels, plus lent que *ClustalW,* mais donne de meilleurs résultats pour des séquences moins apparentées. *DIALIGN* (Brudno et al., 2003) recherche des motifs de séquences qu'il aligne localement. *MUSCLE* (Edgar, 2004) dans sa dernière version, prétend avoir les meilleurs scores d'exactitude.

La résurgence des méthodes HMM – vient avec la version 3 de *HMMER* qui promet d'être tout aussi rapide que *BLAST* (Altschul et al., 1990), pourtant avec les avantages des méthodes probabilistes (Eddy, 2008). Cette méthode est pourtant plus difficile à utiliser car elle crée et gère des profils.

Figure 2.5 Exemple d'un alignement multiple (Miguel Andrade, 2006).
Les lignes représentent des séquences. Les caractères (nucléotides) dans les colonnes sont considérés comme homologues. Le tiret «-» est employé pour définir un indel (insertion ou délétion de nucléotides). Dépendamment du contexte, il peut aussi être interprété comme un nucléotide manquant.

2.2.8 Réseaux réticulés

Bien que les arbres phylogénétiques sont un bon support pour représenter l'évolution Darwinienne, certains phénomènes comme le transfert horizontal de gènes, qui est un mécanisme qui permet aux bactéries d'échanger des gènes (Sonea et Panisset, 1976; Sonea et Mathieu, 2000) ou la recombinaison génétique ne peuvent pas être modélisés à l'aide des arbres. Ainsi, les réseaux réticulés permettent de modéliser une descendance à partir de plusieurs ancêtres.

Pour construire un réticulogramme, ou un réseau réticulé, la première étape consiste à reconstruire l'arbre original représentant l'évolution des espèces, puis la seconde, à ajouter les réticulations ou branches supplémentaires - nécessaires. Un programme très fréquemment employé est T-Rex (Makarenkov, 2001) – voir figure 2.6.

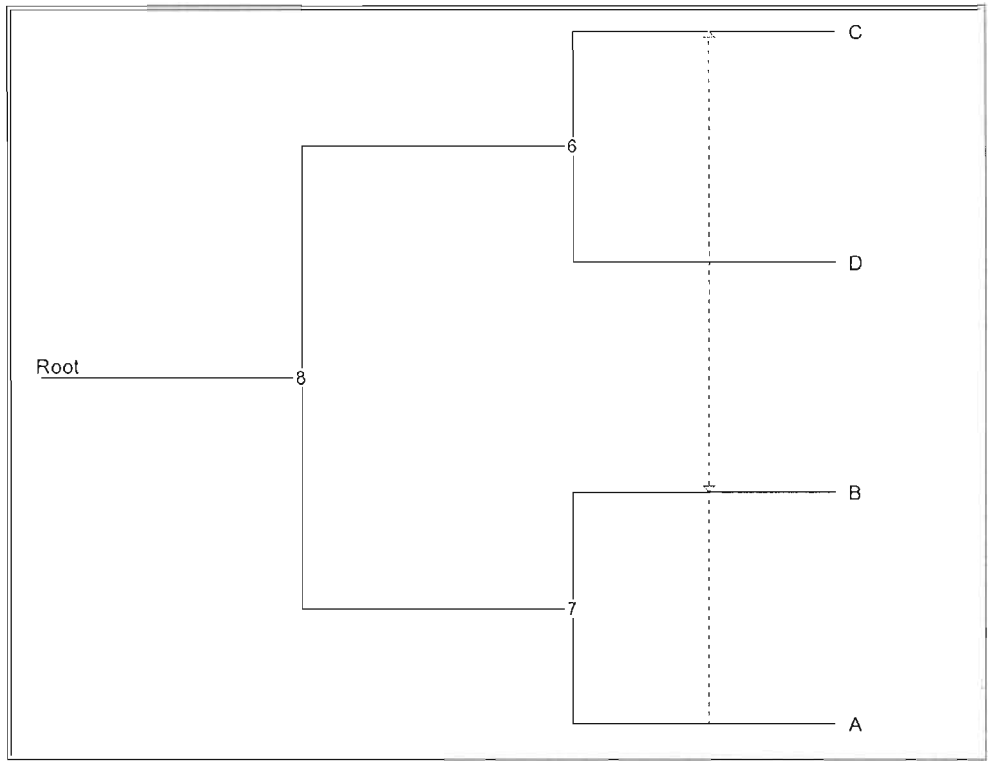

Figure 2.6 Exemple d'un réseau phylogénétique inféré par T-Rex (Makarenkov, 2001).

2.3 Détection de séquences fonctionnelles

La détection des régions biologiques fonctionnelles et la caractérisation de leur fonction est un pas essentiel dans la compréhension du fonctionnement du génome. Une façon de procéder à cette détection est de rechercher les régions qui évoluent à différentes vitesses et qui ont des patrons d'évolution différents par rapport à l'évolution naturelle génétique. Ceci est normalement validé par le calcul des *p*-values significatifs et les tests de permutation.

Cette quête de régions fonctionnelles putatives présente aussi un aspect économique, car à la dimension du génome les analyses exhaustives en laboratoire humide de biologie moléculaire sont très chères. Le criblage computationnel permet de concentrer les recherches sur des régions plus restreintes.

Il existe trois approches principales :

1) Détection des séquences conservées sur plusieurs lignées. (Boffelli et al., 2003; Margulies et Blanchette, 2003 ; Cooper et al., 2005; Siepel et al. 2005).

2) Identification de régions géniques qui ont un rapport d_N/d_S [5]anormalement elevé (Nielsen et Yang, 1998; Yang et Nielsen, 2002; Clark et al. 2003; Forsberg et Christiansen, 2003; Guindon et al. 2004; Nielsen et al, 2005).

3) Détection de la sélection lignée-spécifique (Sergei et al., 2005; Siepel et al., 2006).

L'approche 1 propose de détecter les régions conservées, celles qui sont soumises à la pression évolutive négative. C'est le cas des régions essentielles pour le fonctionnement de la cellule, fonctions, qui une fois perdues, entraînent la mort ou l'incapacité de se reproduire.

---

[5] d_N/d_S - rapport substitutions non-synonymes / synonymes.

L'approche 2 propose de détecter des régions qui sont soumises aux pressions évolutives positives, séquences qui apportent à la cellule des fonctions nouvelles, lui donnant un avantage de survie ou de reproduction. Le principe est d'établir le rapport entre des substitutions synonymes et non-synonymes au niveau des codons des régions codantes. Plusieurs combinaisons de 3 nucléotides (codons), correspondent aux mêmes acides aminés dans les protéines. Il existe donc certaines mutations au niveau de l'ADN qui ne changent pas l'aminoacide correspondant, et c'est pour cette raison qu'elles sont appelées synonymes. Quand les mutations dans l'ADN résultent dans un changement au niveau des aminoacides (et donc dans la protéine correspondante), elles sont appelées non-synonymes. Les scores peuvent être calculés comme moyenne sur la longueur d'un gène pour chaque espèce, pour comparer les espèces entre elles ou bien sûr un même site, à l'aide d'un arbre phylogénétique pour comparer les sites. Il existe des mesures de neutralité des séquences comme le z-test - basé sur la distribution normale ou le LRT (likelihood ratio test) – test du rapport de vraisemblance – basé sur la distribution χ^2. Des méthodes d'analyse par fenêtre coulissante ont également été développées (Clark et Kao, 1991). Ces méthodes permettent une sélection spécifique à certaines lignées, mais elles doivent être connues à l'avance. Elles nécessitent des limites de gènes, données par des annotations au niveau du génome.

Les méthodes de la catégorie 3 se basent sur la variation des conditions de sélections entre les lignées et ne nécessitent pas d'annotations. Elles se concentrent, tout comme la catégorie 1, sur la sélection négative.

Une approche inverse à celle de la catégorie 3, appelée méthode des signatures, en se basant sur les annotations de gènes, leur orthologie et leur phylogénie, détermine la vitesse de la sélection sur une lignée spécifique. En étudiant les covariations à travers les gènes et génomes on peut mieux définir les similarités dans les fonctions des gènes (Shapiro et Alm, 2008).

CHAPITRE III

CLASSIFICATION DES VIRUS DU PAPILLOME HUMAIN

3.1 Résumé en français de Diallo et al. (2009a)

Sommaire. Le virus du papillome humain est bien connu pour son potentiel à générer le cancer du col de l'utérus. Cet article traite de la relation entre les insertions et les délétions de nucléotides au niveau de l'ADN des souches de la famille du papillome et le développement d'un cancer. Les scénarios les plus probables ont été calculés et une analyse de redondance linéaire et polynomiale ont été effectuées. Les p-values ont été significatives pour la majorité des gènes investigués. Ceci ouvre la voie à une analyse plus poussée, que les biologistes pourraient mener par la suite.

3.1.1 Introduction

Le virus du papillome humain est une famille de virus bien connue pour sa diversité génétique et pour son potentiel de causer le cancer du col de l'utérus. Ce sont des virus à ADN d'une taille génomique d'environ 8 Kpb. La taille réduite du génome a permis le typage de plus de 100 souches et le séquençage de plus 80. Le typage se fait sur une partie de la séquence du gène L1, gène structural très variable. A partir des alignements de séquences et des arbres phylogénétiques un type nouveau est déclaré lorsqu'une souche diffère de plus de 10% de son voisin le plus proche. Les classifications antérieures utilisaient plutôt le degré de

risque ou la prédominance cutanée ou muqueuse. Des recherches ont été faites basées sur des gènes individuels comme le E6 et E7 qui sont impliqués dans la suppression de mécanismes anti-cancéreux comme le facteur p53 ou le gène du rétinoblastome (Van Ranst, Kaplan et Burk, 1992). Cette étude phylogénétique est basée sur le génome au complet de 83 VPH, et comme données épidémiologiques, les diagnostiques de 3607 femmes atteintes du cancer du col de l'utérus et provenant de 25 pays. Plus de 89% ont développé une forme microscopique de cancer appelé carcinome aux cellules squameuses – SQUAM – et 5% une autre forme appelé adénocarcinome - ADENO. La distribution du jeu de données se retrouve à la figure 3.1. Dans plus de la moitié des cas les types 16 et 18 sont impliqués dans le cancer. Nous avons utilisé la régression linéaire et polynomiale ainsi que l'analyse de la redondance pour documenter la relation entre la carcinogenicité et les insertions, délétions et conservations au long des branches de l'arbre phylogénétique représentant l'évolution de ces virus.

3.1.2 Inférence de l'histoire des événements évolutionnaires

Les séquences génomiques ont été alignées avec *ClustalW* et l'arbre phylogénétique de la figure 3.2 à été inféré avec la méthode de maximum de vraisemblance PHYML, selon le modèle d'évolution HKY. Pour enraciner l'arbre, le virus du papillome bovin type 1 a été utilisé comme *outgroup*. Les scores de bootstrap [6] ont été calculés pour évaluer la robustesse de l'arbre. En général l'arbre obtenu est en concordance avec celui officiel de NCBI/ICTV basé sur le gène L1. En effet, l'évolution du gène L1 reflète celle du génome entier. Les souches les plus dangereuses se retrouvent dans les sous-arbres enracinés par les nœuds 16 et 18. Le sous-arbre enraciné par le nœud A (figure 3.2), ont un pourcentage de caractères

---

[6] Le bootstrap est une méthode de rééchantillonnage des séquences utilisées pour inférer un arbre phylogénétique. Le pourcentage d'identité des branches des bootstrappés par rapport à l'arbre original, est calculée (Felsenstein, 1985). Le bootstrap constitue une mesure de robustesse principale en analyse phylogénétique.

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

conservés qui est très élevé. La conclusion est que la carcinogenicité pourrait être héritée à partir de cet ancêtre commun.

3.1.3 Trouver des relations entre les deux types de cancer et la distribution des indels/conservations dans les gènes du VPH

La régression linéaire et polynomiale ont été appliquées séparément sur les huit gènes dans la table 3.2 avec des p-values significatives. Nous avons appliqué la méthode de Makarenkov et Legendre (2002), pour estimer quelle type de régression donne les meilleurs résultats avec les données présentes. Ainsi la signification de la différence suggère que le modèle polynomial soit plus approprié à ce cas. La table 3.2 indique que les gènes E4 et L2 fournissent une meilleure correlation. Pour mieux représenter les corrélations, un type spécial de dessin appelé Biplot a été employé, voir Makarenkov et Legendre (2002). Ainsi les angles entre les vecteurs reflètent leur corrélation. La projection de différents types de VPH au long des 6 variables réponse correspondantes aux colonnes de la table 3.2 donne la corrélation correspondante (voir figure 3.4). On observe que les souches carcinogènes présentent plus de conservations et de délétions que celles non-carcinogènes, qui à leur tour présentent plus d'insertions. Le type SQUAM de cancer est fortement positivement corrélé avec les délétions dans le gène L2 et les deux types de cancer sont corrélés négativement avec le nombre d'insertions.

3.1.4 Conclusions

Premièrement nous avons trouvé que les groupes VPH sont monophylétiques, ce qui est en accord avec la classification officielle actuelle. Ensuite nous avons inféré le scénario le plus probable d'insertions et de délétions, pour chacun des gènes. Nous avons ainsi trouvé que les caractères sont conservés à plus de 90% durant l'évolution. La régression linéaire et polynomiale a trouvé des p-values significatives pour tous les gènes mis à part E5, avec la plus forte relation pour les gènes E4 et L2. L'analyse de la redondance a permis pour chaque gène à part d'avoir plus de détail sur chaque combinaison de types d'évènement et de cancer.

3.2 Classification of the Human Papilloma Viruses

Abdoulaye Baniré Diallo[1,2], Dunarel Badescu[1], Mathieu Blanchette[2] and Vladimir Makarenkov[1]

1 Département d'informatique, Université du Québec à Montréal, C.P. 8888, Succursale Centre-Ville, Montréal (Québec), H3C 3P8, Canada
2 McGill Centre for Bioinformatics and School of Computer Science, McGill University, 3775 University Street, Montréal, Québec, H3A 2B4, Canada

Abstract. In this study we present a whole-genome phylogenetic classification of the Human Papilloma Viruse (HPV) family. We found that all groups based on single gene classification are monophyletic and high risk of carcinogenicity taxa are clustered together. The most likely insertion and deletion (indel) scenarios of HPV nucleotides were computed to study the distribution of indels on different edges. We also searched for relationships between the number of indels which occurred during the evolution of the HPV family and the degree of carcinogenicity of considered taxa. Linear and polynomial redundancy analyses (RDA) were carried out to relate the HPV carcinogenicity with the number of insertions, deletions and conservations.

Keywords: Human Papilloma Viruses (HPV), insertions and deletions (indels) of nucleotides, phylogenetic tree, redundancy analysis (RDA), virus evolution

3.2.1 Introduction

Human Papilloma Viruses (HPV) form a family of viruses that are well-known for their genomic diversity (Antonsson et al. 2000) and potential to cause cervical cancer. Nowadays, about a hundred of HPVs have been identified and the whole genomes of more than eighty of them have been sequenced (ICTVdB, 2006). They are double-stranded, circular DNA genomes with sizes close to 8 Kbp with complex evolutionary relationships and a small set of genes. In general, the genes E5, E6, and E7 modulate the transformation process, the two regulatory proteins, E1 and E2, modulate the transcription and replication and the two structural proteins, L1 and L2, compose the viral capsid. A new HPV is recognized as a new HPV type if its complete genome has been cloned and the DNA

sequence of the gene L1 differs by more than 10% from the closest known HPV type (Munoz et al. 2003 and 2004; De Villiers et al. 2004).

Older classifications grouped HPVs according to their higher or lower risk of cutaneous or mucosal diseases. Most of the studies were based on a single gene (usually E6 or E7) analysis. The latter genes are predominantly found in cancer cells due to the binding of their products to the p53 tumour suppressor protein and the retinoblastoma gene product, respectively (Van Ranst, Kaplan and Burk, 1992).

Diagnostics of 3,607 women with cervical cancer from 25 countries revealed that about 89% of them had squamous cell carcinoma (SQUAM cancer) and about 5% had adenosquamous carcinoma (ADENO cancer) (Munoz et al. 2003; see also Figure 3.1 below). It is worth noting that more than the half of the infection cases are due to the types 16 and 18 of HPV (Chan et al. 1995; figure 3.1).

Here we studied a whole genome phylogenetic classification of the HPVs and the insertion and deletion (indel) distribution among HPV lineages leading to the different types of cancer. Multiple linear and polynomial regressions and redundancy analyses were used to relate the taxa carcinogenicity with the number of insertions, deletions and conservations, which, in this study, include both conservations and mutations of nucleotides, and estimate the significance of the obtained relationships.

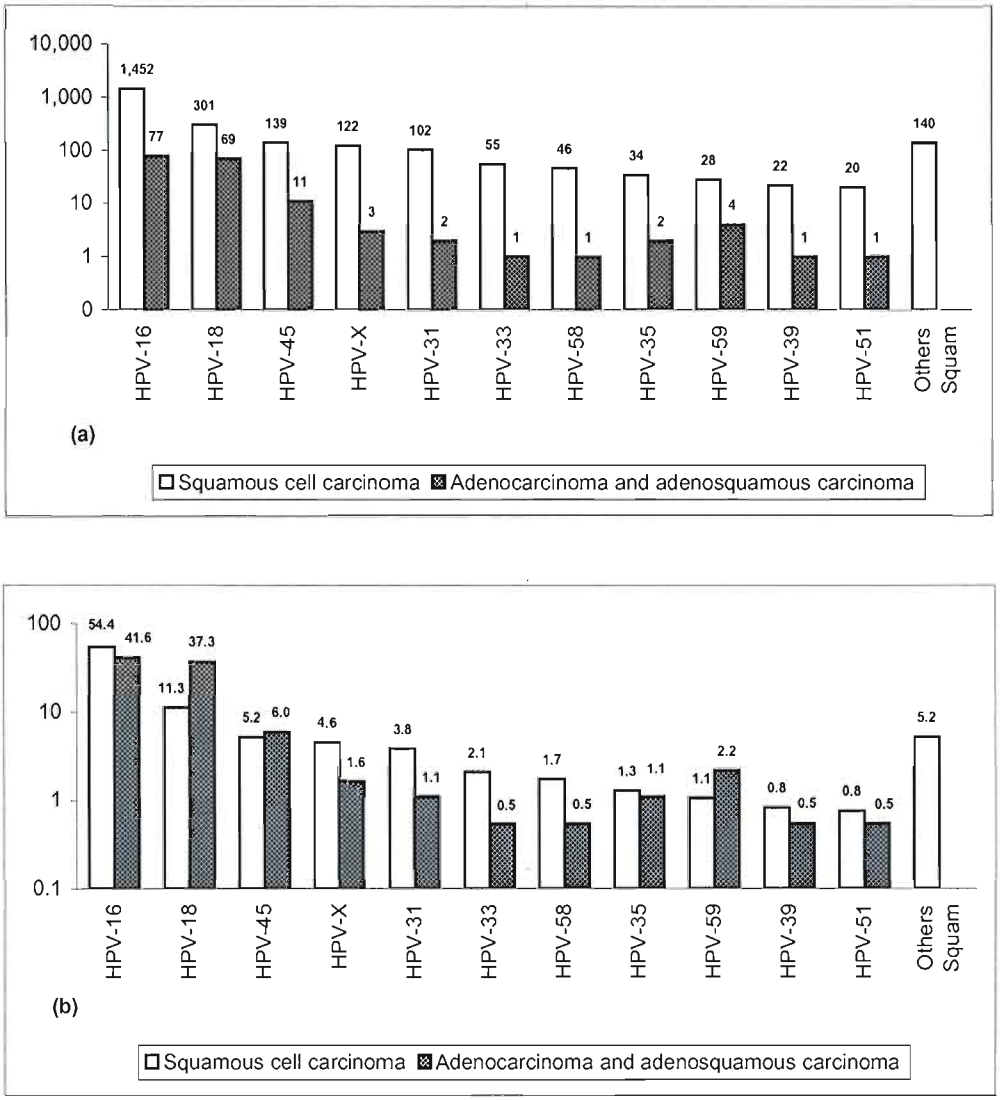

Figure 3.1 Distribution of 11 carcinogenic HPVs in terms of the SQUAM and ADENO cancers (drawn using the data from Munoz et al. 2003).

Total numbers (a) and percentages of cases (b) are represented on a logarithmic scale. Category "Other Squam" is composed of HPV types being found only for the Squamous cell carcinoma and accounting less than 3% of cases, namely: HPVs-52, 56, 73, 68, 82, 26, 66, 11, 6, HR, 53, 55, 81 and 83. HPVs-35, HR, 68 and X were not considered in this study because their complete genomes were not yet available.

3.2.2 Inferring the history of evolutionary events

Available genomes of HPVs identified by the ICTV (ICTVdB, 2006) were downloaded and aligned using ClustalW (Thompson et al. 1994). The alignment length was 10,426 bp. The phylogenetic tree of 83 HPVs (figure 3.2) was inferred using the PHYML method (Guindon and Gascuel, 2003) with the HKY model of evolution. As suggested in Van Ranst, Kaplan and Burk (1992), the bovine PV of type 1 was used as an outgroup to root the phylogenetic tree.

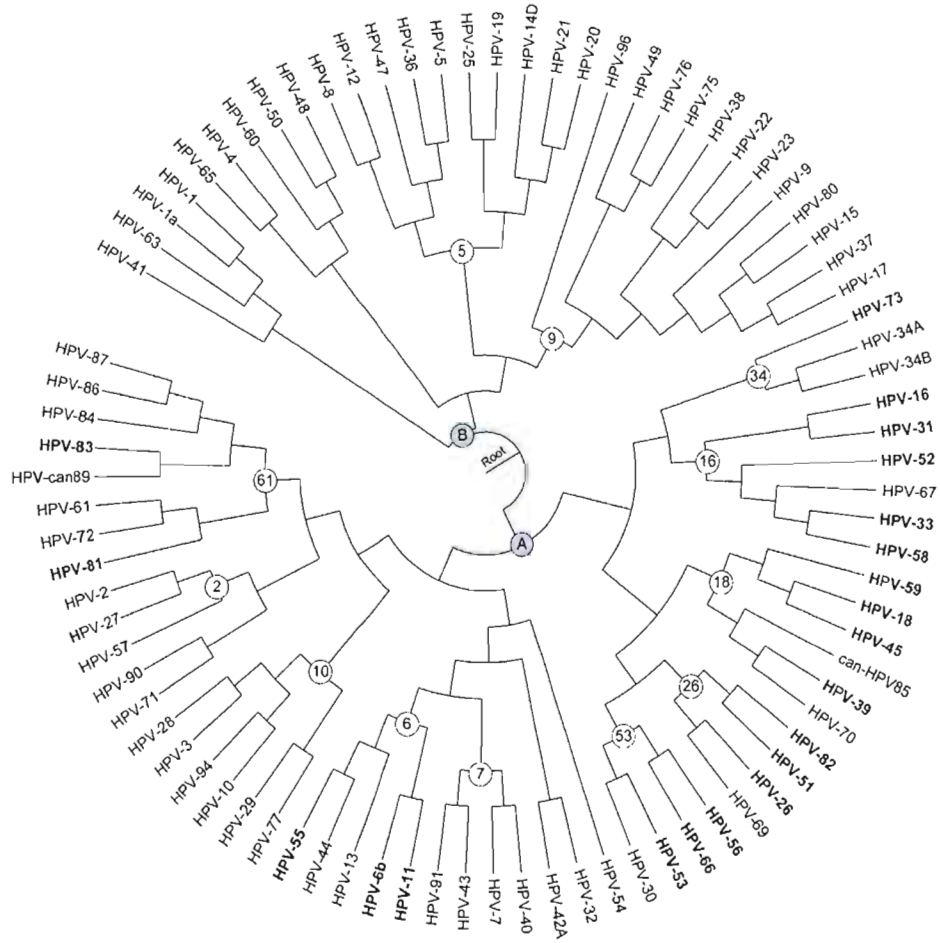

Figure 3.2 Phylogenetic tree of 83 HPVs obtained using the PHYML method.
The white labelled nodes identify the exiting HPV groups according to the NCBI taxonomy browser and the shaded ones (A and B) distinguish between the non carcinogenic and carcinogenic families. The 21 carcinogenic HPVs are indicated in bold (see figure 3.1).

The bootstrap scores were computed to assess the robustness of the edges. For clarity, they are not indicated in Figure 3.2; mention that they are higher than 80% for most of the edges. In the obtained tree, most of the HPV groups (denoted by numerated nodes) are in agreement with the NCBI/ICTV classifications based on the gene L1. Thus, the evolution of the gene L1 to classify those taxa reflects the whole genome evolution. The most dangerous HPV taxa (see figure 3.2) are located in the subtrees rooted by the nodes 16 and 18.

To quantify the indel distribution, the most likely indel scenarios were computed using a heuristic algorithm described in Diallo, Makarenkov and Blanchette (2007). Table 3.1 presents the distribution of the predicted indel and conservation events for all HPV genes.

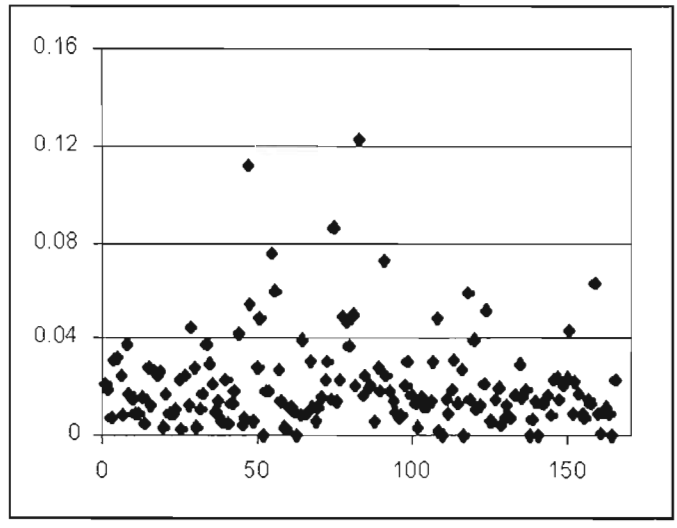

Figure 3.3 Indel distribution along the tree edges.
The big values usually correspond to the edges leading to non carcinogenic taxa. The abscissa axis represents the edge number (not shown in Figure 2) and the ordinate represents the indel percentage along this edge.

It appears that most of the genes have more than 90% of the characters conserved throughout the evolution (figure 3.3).

The indel frequencies are higher in the subtrees rooted by the node 61, where only low-risk-carcinogenicity HPVs are located (figure 3.2). The groups located in the subtree rooted by the node A usually have high percentage of conserved characters on each edge. One can

conclude that the organisms of this subtree inherited their carcinogenicity from their least common ancestor. The detailed analysis of the edges of this subtree should be carried out but this goes beyond the scope of this article.

Table 3.1 For each of the 15 genes of HPV, this table reports the numbers of the Conserved, Inserted and Deleted regions (and the percentages of nucleotides in these regions) in all lineages of the tree in figure 3.2.

Evol. events / Genes	Cons.	Ins.	Del.	%Cons.	%Ins.	%Del.
E1	12111	601	2774	91.8	0.3	1
E1A	1784	509	320	91.8	1.4	0.6
E2	13304	306	3460	85.2	0.1	2.2
E4	6318	195	2117	85.1	0.1	3.8
E5	1688	356	503	73.1	2.1	3.1
E5A	208	162	68	79.3	8.2	1.3
E5B	101	31	19	16.3	7.7	0.2
E6	7323	613	1529	89.0	0.2	1.1
E7	3457	0	1393	59.4	0	3.9
E8	84	0	0	52.6	0	0
L1	9664	314	2751	92.7	0.1	1.0
L2	21716	494	5138	92.3	0.4	2.6
X	484	0	230	43.7	0	1.8
Y	1457	54	679	83.2	0.3	2.6
Z	0	0	6	0	0	0.4

3.2.3 Finding relationships between the two types of cancer and the indel/conservation distributions in the HPV genes

We carried out linear and polynomial regressions to establish relationships between the explanatory variables (conservations, insertions and deletions in our case) and response variables (cancer/no cancer outcomes for the SQUAM and ADENO cancers, respectively). To perform the regression, we considered the eight most important HPV genes for the group

of 83 HPV viruses (Table 3.2). The numbers of conserved, inserted and deleted regions as well as the percentages of characters involved in these evolutionary events, reported in Table 3.1, formed the matrix of explanatory variables **X**. Two binary variables, consisting of the SQUAM and ADENO cancer outcomes, formed the matrix of response variables **Y**.

Table 3.2 Percentages of variance accounted for by the linear and polynomial regression for the 8 most important HPV genes and for the whole genomes.

p-values of the linear and polynomial regressions as well as of their difference are reported. The genes, *E4* and *L2*, for which the best results were obtained, are highlighted. The numbers of taxa available for each gene are shown between the parentheses in the first column.

Statistics / Genes	% of variance for Lin. Regr.	% of variance for Pol. Regr.	Lin. Regr. p-value	Pol. Regr. p-value	Difference p-value
E1(81)	24.89	41.02	0.01	0.01	0.03
E2(81)	24.49	41.70	0.01	0.01	0.02
E4(57)	32.12	58.47	0.01	0.01	0.01
E5(20)	39.84	64.98	0.49	0.72	0.71
E6(81)	31.80	43.42	0.01	0.01	0.08
E7(81)	30.89	38.36	0.01	0.01	0.17
L1(83)	24.74	33.38	0.01	0.01	0.30
L2(83)	42.55	47.54	0.01	0.01	0.64
All genes	27.57	36.15	0.02	0.03	0.65

If a HPV organism can initiate the SQUAM cancer (21 of such HPV organisms were considered) the corresponding value of the first response variable was set to 1, and if it can initiate the ADENO cancer (9 of such HPV organisms were considered) the value of the second response variable was set to 1, otherwise they were set to 0.

Linear and polynomial regressions were carried out separately for the eight genes in Table 3.2 and for the whole genomic sequences. Generally, both linear and polynomial models were significant: most of the *p*-values for the linear and polynomial regressions were smaller than 0.05 (Table 3.2). We also performed the test of the difference between the polynomial and linear regressions (last column of Table 3.2) according to the method

discussed in Makarenkov and Legendre (2002). This test allows one to estimate the possibility of overfitting the data by polynomial regression. If both polynomial and linear models are significant, the significance of the difference between them suggests that the polynomial model is more appropriate in this case, otherwise it suggests the overfitting by polynomial regression and the linear model is preferable. The results in Table 3.2 indicate that for the genes E4 and L2 the presence and absence of the SQUAM and ADENO cancers correlate the best with the considered evolutionary events. These two genes should be further analysed by virologists interested by studying the carcinogenic human papilloma viruses. In this way, we carried out linear (for the gene L2 because the difference between the polynomial and linear regressions for this gene was not significant, see Table 3.2) and polynomial (for gene the E4 because this difference was significant) redundancy analysis (RDA) to find the detailed relationship between the carcinogenic HPVs and the insertion, deletion and conservation (including real conservations and mutations) events they underwent. RDA (Rao, 1973) allows one to model relationships between the explanatory variables (conservations, insertions and deletions), response variables (cancer/no cancer outcomes for the SQUAM and ADENO cancers) and considered group of species (83 HPV organisms). For instance, the polynomial RDA, introduced in Makarenkov and Legendre (2002), allows for modeling non-linear relationships between the explanatory and response variables.

The correlation biplot (see Makarenkov and Legendre, 2002) was used in this study to represent the relationships between the variables in **X** and **Y**. In such a biplot the angles between the variables from sets **X** and **Y** reflect their correlations; projecting a HPV type (denoted by a triangle in Figure 3.4) at right angle on a response variable **y** approximates the value of this HPV type along this variable; projecting a HPV type at right angle on an explanatory variable **x** approximates the value of this HPV type along this variable. In total, 6 response variables corresponding to the columns of Table 3.1 for both genes L2 (Figure 3.4a) and E4 (Figure 3.4b), and 2 combined variables *% Ins.* x *% Del* and *Ins.* x *% Del* for the gene E4 only (Figure 3.4b) were depicted. The two represented combined variables were chosen among all available combined variables because they provided the strongest positive correlations with the SQUAM cancer arrow (Figure 3.4b); all other combined variables are

not represented in polynomial biplot because they don't correlate strongly, either positively or negatively, with the two response variables depicted by dashed arrows. The lengths of the arrows representing the response and explanatory variables in both diagrams were multiplied by 20 and 15, respectively; this does not change the interpretation of the diagrams.

While observing the biplot ordination diagram drawn for the gene L2 (Figure 3.4a), the following main trends can be noticed: both types of carcinogenic HPVs have a greater number of conserved and deleted nucleotides compared to the non carcinogenic HPVs, whereas the non carcinogenic HPVs usually have a higher number of insertions. Also, the presence of the SQUAM cancer is strongly positively correlated with the percentage of deleted nucleotides in the lineages of the gene L2, and both SQUAM and SQUAM cancer types are strongly negatively correlated with the number of insertions. As to the gene E4 (Figure 3.4b), the presence of the SQUAM cancer is strongly positively correlated with the percentage of inserted nucleotides as well as with the two depicted combined variables consisting of the products of the percentages of inserted and deleted nucleotides and of the number of insertions and percentage of deletions. Also, the SQUAM cancer HPVs are strongly negatively correlated to the percentages of conserved and deleted nucleotides. Finally, for the gene E4 both types of carcinogenic HPVs have a higher number of insertions compared to the non carcinogenic ones.

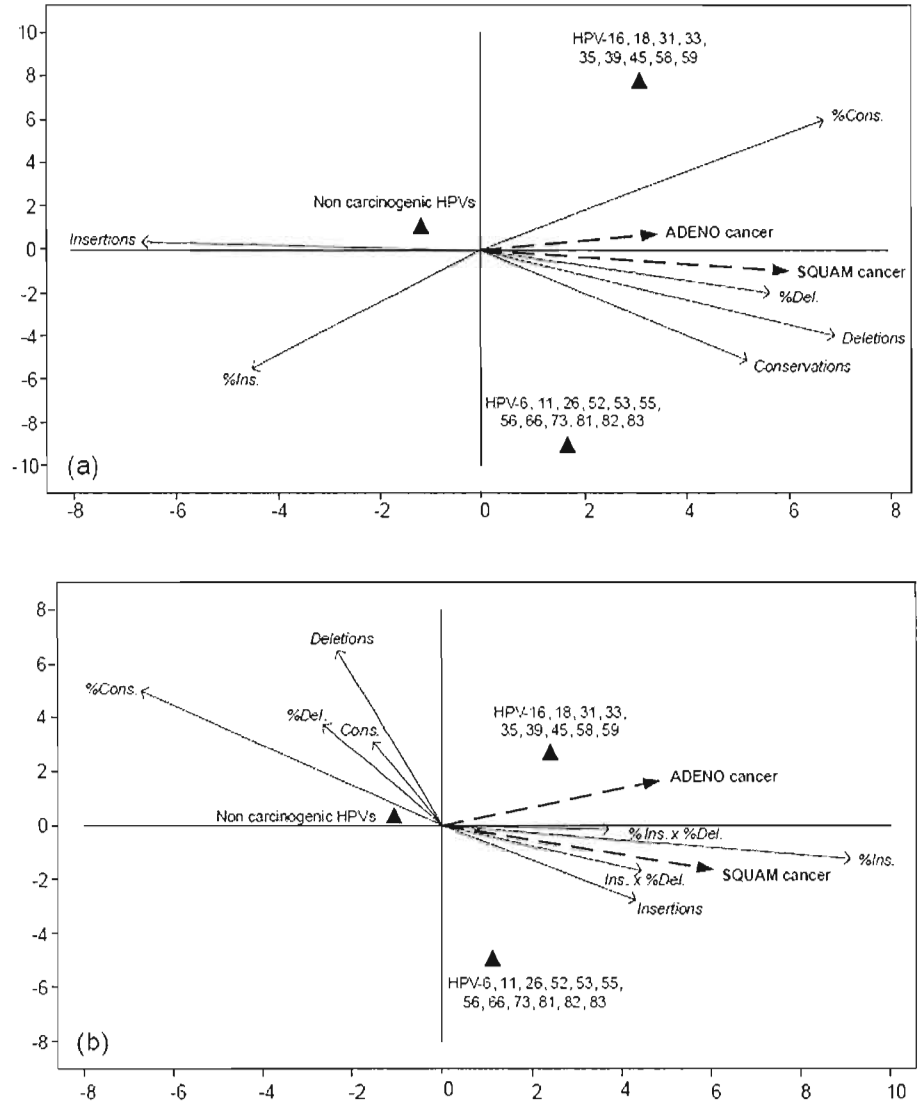

Figure 3.4 Linear (case a - for the gene L2) and polynomial (case b - for the gene E4) RDA biplots for the 83-taxa HPV dataset.

Triangles represent the three types of HPVs: viruses causing both types of cancer: HPVs-16, 18, 31, 33, 35, 39, 45, 58 and 59; viruses causing only the SQUAM cancer: HPVs-6, 11, 26, 52, 53, 55, 56, 66, 73, 81, 82 and 83; and, non carcinogenic HPVs. Note that all HPVs causing the ADENO cancer also cause the SQUAM cancer. Dashed arrows represent two binary response variables: SQUAM and ADENO cancers. Solid arrows represent the numbers of conserved, inserted and deleted regions and the corresponding percentages of the conserved, inserted and deleted nucleotides.

3.2.4 Conclusion

In this article we studied the classification of the Human Papilloma Viruses (HPV) presumed to be the main cause of the cervical cancer. First, we inferred the PHYML phylogenetic tree (Guindon and Gascuel, 2003) of the 83 available HPV organisms (Figure 3.2) on the basis of the whole genome phylogenies. We found that all HPV groups (see the 12 HPV subtrees denoted by white nodes in Figure 3.2) are monophyletic (i.e., compatible with the current NCBI/ICTV classifications). Then, we inferred the most likely insertion and deletion scenarios for each of the 15 considered HPV genes (see Table 3.1) and found that most of them have more than 90% of the characters conserved throughout the evolution (Figure 3.3). Multiple linear and polynomial regressions were carried out in order to establish relationships between the conservation, insertion and deletion events and cancer/no cancer outcomes for the SQUAM and ADENO cancers. We found that the presence and absence of both types of cancer correlated the best with the considered evolutionary events in the genes *E4* and *L2*, and the only gene for which the regression p-values were not significant was the gene *E5*. This result warranted additional investigations of the genes *E4* and *L2* consisting of the linear (Rao, 1973) and polynomial (Makarenkov and Legendre, 2002) RDA conducted for them. RDA biplots drawn for these two genes, shown in Figure 3.4, present the detailed relationships between the SQUAM and ADENO cancers, 3 types of HPV groups and 6 selected evolutionary events. Further investigations should be conducted by virologists based on the findings on this study.

3.2.5 References

ANTONSSON, A., FORSLUND, O., EKBERG, H., STERNER, G. and HANSSON, B. G. (2000): The Ubiquity and impressive genomic diversity of human skin papillomaviruses suggest a commensalic nature of these viruses. J. of Virology, 74, 11636-11641.

CHAN, S.Y., DELIUS, H., HALPERN, A.L. and BERNARD, H.U. (1995): Analysis of genomic sequences of 95 PV types: uniting typing, phylogeny, and taxonomy. J. of Virology, 69, 3074–3083.

DE VILLIERS, E.M., FAUQUET, C., BROKER, T.R., BERNARD, H.U. and ZUR HAUSEN, H. (2004): Classification of papillomaviruses. Virology, 324, 17–27.

DIALLO, A. B., MAKARENKOV, V. and BLANCHETTE, M. (2007): Exact and heuristic algorithms for the Indel Maximum Likelihood Problem. J. of Computational Biology, 14, 446-461.

GUINDON, S. and GASCUEL, O. (2003): A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Systematic Biology, 52, 696-704.

ICTVdB Management (2006): The Universal Virus Database. New York, Büchen-Osmond, C., Columbia University.; website: http://www.ncbi.nlm.nih.gov/ICTVdb/.

MAKARENKOV, V. and LEGENDRE, P. (2002): Nonlinear redundancy analysis and canonical correspondence analysis based on polynomial regression. Ecology, 83, 1146-1161.

MUÑOZ, N., BOSCH, F.X., DE SANJOSE, S., HERRERO, R., CASTELLSAGUE, X. and SHAH, K.V. et al. (2003): Epidemiologic classification of human papillomavirus types associated with cervical cancer. New England J. of Medicine, 348, 518–527.

MUÑOZ, N., BOSCH, F.X., CASTELLSAGUÉ, X., DIAZ, M., DE SANJOSE, S. and HAMMOUDA, D. et al. (2004): Against which human papillomavirus types shall we vaccinate and screen? The international perspective. International J. of Cancer, 111, 278–285.

RAO, C.R. (1973): Linear statistical inference and its applications. Wiley, New York, New York, USA.

THOMPSON, J.D., HIGGINS, D.G. and GIBSON, T.J. (1994): CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research, 22, 4673–4680.

VAN RANST, M., KAPLAN, J.B. and BURK, R.D. (1992) : Phylogenetic Classification of Human Papillomaviruses: Correlation With Clinical Manifestations. J. of General Virology, 73, 2653-2660.

CHAPITRE IV

UNE ETUDE DU GENOME ENTIER ET L'IDENTIFICATION DE REGIONS SPECIFIQUES CARCINOGENES DU VIRUS DU PAPILLOME HUMAIN

4.1 Résumé français de Diallo et al. (2009b)

Sommaire. Dans cet article, accepté pour publication dans *Journal of Computatioinal Biology*, nous décrivons un nouvel algorithme pour l'analyse du contenu informationnel des alignements de séquences multiples en relation aux données épidémiologiques de carcinogenicité pour identifier des régions nécessitant des analyses plus poussées. Cet algorithme est basé sur une procédure de fenêtre coulissante et le calcul des p-values. 83 souches du virus du papillome humain ont été analysées. Ceci nous a permis de détecter des régions qui pourraient être influencées par des insertions, délétions ou substitutions. L'article comprend aussi une analyse phylogénétique de la famille du virus du papillome humain.

4.1.1 Introduction

Le virus du papillome humain a un rôle causal dans la genèse du cancer du col de l'utérus, une maladie grave qui frappe un demi-million de personnes chaque année. Sa diversité génomique et la petite taille de son génome le font un bon sujet des analyses

phylogénétiques. D'ailleurs sa classification est entièrement génétique. Les protéines des gènes L1 et L2 composent la capside virale et ainsi ont été étudiés en premier. Les gènes E5, E6, E7 modulent la transformation et E1, E2 la transcription et la réplication. Un rôle de E4 dans l'assemblage du virus et l'activation des fonctions virales lui ont été récemment attribuées. Pour cette étude nous avons utilisé les données de Munoz et al. (2003,2004), (voir table 4.1) une étude de grande ampleur, et de grande qualité, ne sois-ce que par le test de typage PCR centralisé. La centralisation des tests permet une homogénéité des techniques, des standards et des conditions de travail, ainsi que des équipements et du personnel hautement qualifié. Une analyse phylogénétique de l'arbre génomique et une comparaison des arbres générés pour chaque gène individuellement à l'aide de la distance de Robinson et Foulds ont été effectuées. Ensuite nous avons développé un algorithme pour analyser le contenu informationnel des alignements multiples pour identifier des régions responsables de la carcinogenicité. Cet algorithme est basé sur une formule qui prend en compte la similarité entre les souches carcinogènes et la dissimilarité par rapport aux souches non-carcinogènes à l'intérieur d'une fenêtre coulissante. Pour décider sur la relevance des régions identifiées nous avons calculé les p-values correspondantes.

4.1.2 Analyse des insertions-délétions dans les génomes du VPH et la réconciliation des arbres de gènes

À partir des données de 83 génomes de VPH, un arbre phylogénétique a été inféré. La robustesse de l'arbre est en général de 80% en score bootstrap, pour les branches – voir figure 4.1. Pour plus d'informations, prière de se référer au chapitre précédent. La table 4.2 présente le nombre prédit de conservations, insertions et délétions au long des lignées, selon un scénario le plus vraisemblable, inférés par la méthode heuristique de Diallo et al. (2006, 2007). Les plus grandes fréquences des indels se retrouvent dans les sous-arbres enracinés au node 61, qui ne contient que des organismes à bas-risques de carcinogenicité. Pour comparer les topologies d'arbres nous avons inféré des arbres phylogénétiques, selon la même méthodologie, pour chaque gène, selon les limites annotées dans les génomes séquencées, et nous avons calculé les distances de Robinson et Foulds. Comme le nombre des feuilles des arbres n'est pas le même, nous avons du normaliser les résultats. Les gènes E4 et E5 diffèrent

le plus et le gène E2 réconcilie mieux avec les autres topologies. En moyenne chaque arbre diffère des autres à 32%. Ces résultats confirment des suppositions antérieures que les VPH subissent une fréquente recombinaison.

4.1.3 Algorithme pour l'identification des régions carcinogènes putatives

L'algorithme est basé sur l'hypothèse que les séquences des régions responsables du cancer ont plus de similarité entre elles que par rapport au séquences non carcinogènes. Nous avons construit trois jeux de données – Haut-Risque, SQUAM et ADENO- avec respectivement les types de VPH associés aux données épidémiologiques. Ensuite nous avons balayé tous les alignements de séquences, pour chaque gène, à l'aide d'une fenêtre coulissante d'une largeur de 3 à 20 nucléotides (voir figure 4.3). Un extrait des résultats se retrouve à la table 4.3, avec ces *régions touche* –(*hit regions*). Les images des régions non-superposables se retrouvent aux figures 4.4 à 4.8). Trois analyses séparées ont été faites pour chaque type de données. Les formules 4.1 et 4.2 donnent la variabilité à l'intérieur des séquences carcinogènes et entre les séquences carcinogènes et non-carcinogènes respectivement, pour une certaine fenêtre. La fonction de distance employée a été celle de Hamming, pour aussi prendre en compte les insertions et délétions, ignorées ou pénalisées par les transformations de distances. D'ailleurs les transformations ne s'appliquent pas aux courtes séquences. La formule 4.3 est une fonction d'agrégation et prend en compte tous ces arguments. Nous présentons un schéma de l'algorithme 4.1, d'une complexité asymptotique de $O(ln^2w)$ – l étant la longueur de l'alignement multiple, n – le nombre de souches et w – la largeur de la fenêtre coulissante. Pourtant la complexité alghorithmique peut être réduite à $O(ln^2)$, en évitant de recalculer la distance de Hamming entre les régions voisines qui se superposent. Pour une fenêtre non-superposante la complexité est de $O(ln^2)$. Si la largeur de la fenêtre varie, on multiplie la complexité par la différence entre la plus grande et la plus petite taille de celle-ci. Pour définir une *région touche,* il faut la comparer à un seuil. Pourtant, il est difficile de déterminer cette constante seuil. Pour ceci nous avons utilisé une approche de calcul de p-values. Pour calculer ces p-values nous avons utilisé une méthode Monte-Carlo de permutations de colonnes à une dimension de la fenêtre. Un million d'échantillons a été tiré et les valeurs Q correspondantes calculés. Pour chaque région le

nombre de fois que la valeur de Q est plus grande que celle de l'échantillon est prise en compte. En tant que seuil nous avons utilisé la valeur de 0.001.

4.1.4 Résultats, discussion et conclusion

Un extrait des meilleurs scores pour chaque gène et jeu de données est présenté dans la table IV-3. Les plus grandes valeurs de Q, pour des tailles plus grandes de la fenêtre se retrouvent dans les gènes E2 et E6 à l'exception d'une région spécifique du gène L1 et les VPH à Haut-Risque qui donnent la meilleure valeur de $Q = 0.55$. Pour les tailles réduites de la fenêtre les plus grandes valeurs de Q se retrouvent dans le gène E4 et la catégorie Haut-Risque. Il est intéressant de noter que les résultats dépendent de la dimension de la fenêtre. Toutes les régions présentées dans la table 4.3 ont une p-valeur de 0. Les figure 4.5 et 4.6 montrent des régions avec de p-values en dessous de 0.001, d'une longueur de 40 et 60 nucléotides dont le domaine PDZ (Choongho et Laimonis, 2004), une région bien documentée comme indispensable à l'apparition du cancer (Tohru et al., 1997). Le meilleur résultat de la fonction Q dans le gène L1, un gène structural, montre une autre possible utilisation de la méthode, la recherche d'épitopes dans les régions antigéniques, et la recherche des vaccins, comme nous allons le voir dans le chapitre V.

4.2 A whole genome study and identification of specific carcinogenic regions of the Human Papilloma Viruses

Abdoulaye Baniré Diallo[1,2,‡,*], Dunarel Badescu[1,*], Mathieu Blanchette[2], and Vladimir Makarenkov[1]

1 Département d'informatique, Université du Québec à Montréal, C.P. 8888, Succursale Centre-Ville, Montréal (Québec), H3C 3P8, Canada
2 McGill Centre for Bioinformatics and School of Computer Science, McGill University, 3775 University Street, Montréal, Québec, H3A 2B4, Canada
‡ Corresponding author

---

* The two first authors contributed equally to the work and should be considered as joint first authors.

Abstract. In this article, we undertake a study of the evolution of Human Papillomaviruses (HPV), whose potential to cause cervical cancer is well known. First, we found that the existing HPV groups are monophyletic and that the high-risk of carcinogenicity taxa are usually clustered together. Then, we present a new algorithm for analyzing the information content of multiple sequence alignments in relation to epidemiologic carcinogenicity data to identify regions that would warrant additional experimental analyses. The new algorithm is based on a sliding window procedure and a p-value computation to identify genomic regions that are specific to HPVs causing disease. Examination of the genomes of 83 HPVs allowed us to identify specific regions that might be influenced by insertions, deletions, or simply by mutations, and that may be of interest for further analyses.

keywords: Algorithm for carcinogenic region detection; Human Papilloma Viruses; Evolutionary events; Phylogenetic trees

4.2.1 Introduction

Human papillomaviruses (HPV) have a causal role in cervical cancer with almost half a million new cases identified each year (Angulo and Carvajal Rodriguez, 2007; Bosch et al., 1995;Muñoz, 2000). The HPV genomic diversity is well known (Antonsson et al., 2000). About one hundred HPV types are identified, and the whole genomes of more than eighty of them are sequenced (see the latest Universal Virus Database report by International Committee on Taxonomy of Viruses (ICTV)).

A typical HPV genome is a double-stranded, circular DNA genome of size close to 8 Kbp, with complex evolutionary relationships and a small set of genes. In general, the E5, E6, and E7 genes modulate the transformation process, the two regulatory proteins, E1 and E2, modulate transcription and replication, and the two structural proteins L1 and L2 compose the viral capsid. Protein E4 has an unclear function in the HPV life cycle, however, several studies indicate that it could facilitate the viral genome replication and the activation

of viral late functions (Wilson et al., 2007), and it could also be responsible for virus assembly (Prétet et al., 2007).

A HPV is considered to belong to a new HPV type if both its complete genome has been cloned and the DNA sequence of the gene L1 differs by more than 10% from the closest known HPV type. The comparison of HPV genomes, conducted by ICTV, is based on nucleotide substitutions only (Muñoz et al., 2003; de Villiers et al., 2004). Older HPV classifications were built according to their higher or lower risk of cutaneous or mucosal diseases. Most of the HPV studies were based on single gene (usually E6 or E7) analyses. The latter genes are predominantly linked to cancer due to the binding of their products to the p53 tumor suppressor protein and the retinoblastoma gene product pRb (Van Ranst et al., 1992).

To define carcinogenic types, we used epidemiologic data from a large international survey on HPVs in cervical cancer and from a multicenter case-control study conducted on 3,607 women with incident, histologically confirmed cervical cancer recruited in 25 countries (Muñoz et al., 2003, 2004). HPV DNA detection and typing in cervical cells or biopsies were centrally done using PCR assays which attest for the quality of the study (Muñoz et al., 2003). More than 89% of patients them had squamous cell carcinoma (i.e. Squam cancer) and about 5% had adenosquamous carcinoma (i.e. Adeno cancer) see Table 4.1 adapted from (Muñoz et al., 2003). More than half of the infection cases are due to the types 16 and 18 of HPV, which are thus referred to as high-risk HPVs (Chan et al., 1995).

Table 4.1 Distribution of carcinogenic HPVs for the Squam and Adeno types of cancer.
Complete genomic sequence data is not available yet for HPVs-35, HR, 68, and X.

	Squamous cell carcinoma		Adenocarcinoma and adenosquamous carcinoma	
HPV types	**Number**	**% positive**	**Number**	**% positive**
HPV-16	1.452	54.38	77	41.62
HPV-18	301	11.27	69	37.3
HPV-45	139	5.21	11	5.95
HPV-31	102	3.82	2	1.08
HPV-52	60	2.25		
HPV-33	55	2.06	1	0.54
HPV-58	46	1.72	1	0.54
HPV-56	29	1.09		
HPV-59	28	1.05	4	2.16
HPV-39	22	0.82	1	0.54
HPV-51	20	0.75	1	0.54
HPV-73	13	0.49		
HPV-82	7	0.26		
HPV-26	6	0.22		
HPV-66	5	0.19		
HPV-6	2	0.07		
HPV-11	2	0.07		
HPV-53	1	0.04		
HPV-81	1	0.04		
HPV-55	1	0.04		
HPV-83	1	0.04		
Total	**2.293**	**85.89**	**168**	**90.37**

In this paper, we first studied a whole genome phylogenetic classification of the HPV and the insertion and deletion (indel) distribution among HPV lineages leading to the different types of cancer. First, we inferred a phylogenetic tree of 83 HPVs based on whole HPV genomes. We found that the evolution of the L1 gene, used by ICTV to establish the HPV classification, generally reflects the whole genome evolution. Second, we compared the gene trees built for the most important HPV genes (E1, E2, E4, E5, E6, E7, L1 and L2) using the normalized Robinson and Foulds topological distance (Robinson and Foulds, 1981). Then, we described a new algorithm for analyzing the information content of multiple sequence alignments in order to identify regions that may be responsible for the carcinogenicity.

This algorithm is based on a new formula taking into account the sequence similarity among carcinogenic taxa and the sequence dissimilarity between the carcinogenic and non-carcinogenic taxa, computed for a genomic region bounded by the position of the sliding window. To facilitate the identification of relevant regions, we compute p-values for the different regions according to their score obtained with our new formula. Using the new technique we developed, we examined all available genes in 83 HPV genomes and identified the specific genomic regions that would warrant interest for future biological studies.

4.2.2 Indel analysis of HPV genomes and reconciliation of HPV gene trees

The 83 completely sequenced HPV genomes (all identified by the ICTV) were downloaded and aligned using ClustalW (Thompson et al., 1994), producing an alignment with 10426 columns. The phylogenetic tree of 83 HPVs (Figure 4.1) was inferred using the PHYML program (Guindon and Gascuel, 2003) with the HKY substitution model.

Bootstrap scores were computed to assess the robustness of the edges using 100 replicates. Most branches obtain support above 80%, but for a better readability, they are not represented in Figure 4.1. However, they are given in the supplemental materials 3. As suggested in (Van Ranst et al., 1992), the bovine PV of type 1 was used as outgroup to root this phylogeny.

To the best of our knowledge, the constructed phylogenetic tree is the first whole genome phylogenetic tree of HPVs. Our analysis revealed the presence of 12 known monophyletic HPV groups that are denoted by numerated nodes, labeled according to the ICTV annotation, in Figure 4.1. The other monophyletic groups obtained were not depicted by numbers. The whole-genome phylogeny obtained usually corresponds to the HPV classification provided by ICTV on the basis of the L1 gene. Most of the dangerous HPVs (see Table 4.1) can be found in the sister subtrees rooted by the nodes 16 and 18. As carcinogenicity may be introduced into a HPV by an insertion or deletion (indel) of a group of nucleotides, we first addressed the problem of indel distribution in the evolution of HPV.

Thus, the most likely indel scenario was inferred using a heuristic method described in (Diallo et al., 2006, 2007). Such a scenario includes the distribution of the predicted indel and base conservation events for all HPV genes. Table 4.2 reports, for each of the 8 main genes of HPV, the total number of conservations, insertions and deletions of nucleotides that occurred during their evolution. Genes E1, L1 and L2 show more than 90% conservation at the nucleotide level, E2, E4 and E6 between 80 and 90%, and E5 and E7 respectively 73% and 59%. [7]

[7] Supplemental materials are available at:
http://ancestors.bioinfo.uqam.ca/articles/JCB2009/supplemental.zip

Table 4.2 Numbers of Conservations, Insertions and Deletions.

For each of the 8 main HPV genes, this table reports the numbers (and average numbers) of Conservations (including substitutions), Insertion and Deletions of nucleotides that occurred during evolution.

Variable/Gene	Conservation	Insertion	Deletion	Avg. Cons.	Avg. Ins.	Avg. Del.
E1	12111	601	2774	0.918	0.003	0.010
E2	13304	306	3460	0.852	0.001	0.022
E4	6318	195	2117	0.851	0.001	0.038
E5	1688	356	503	0.731	0.021	0.031
E6	7323	613	1529	0.890	0.002	0.011
E7	3457	0	1393	0.594	0.000	0.039
L1	9664	314	2751	0.927	0.001	0.010
L2	21716	494	5138	0.923	0.004	0.026

The highest indel frequencies are in the subtrees rooted by the node 61 where there are only low risks of carcinogenicity (Figure 4.1). The groups included in the subtree A have low percentage of indels on in each branch. It is likely that the organisms of this subtree inherited their carcinogenicity from their closest common ancestor.

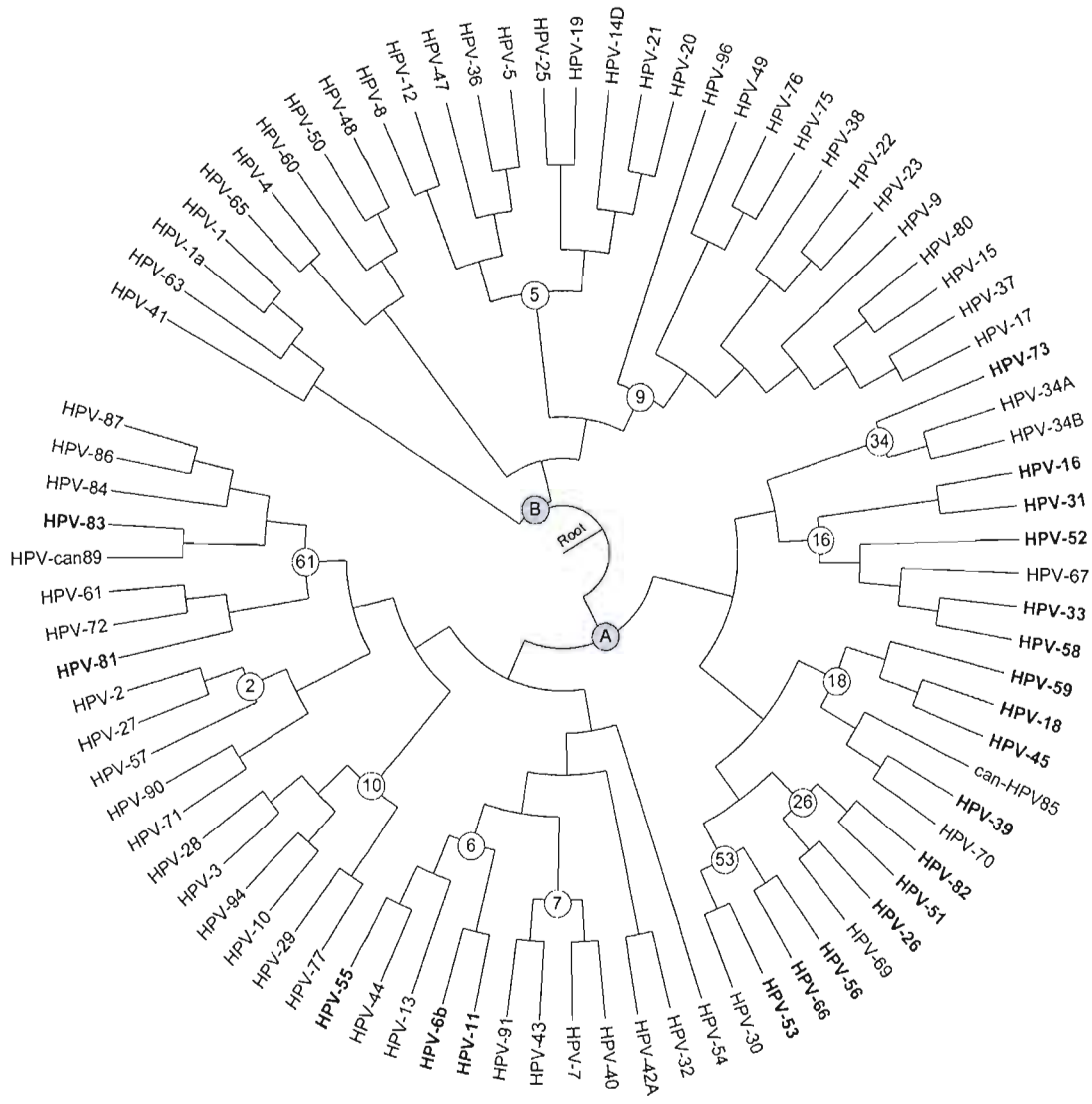

Figure 4.1 Phylogenetic tree of 83 HPVs obtained with PHYML.
The 21 carcinogenic HPV are shown in bold. The white nodes identify the existing HPV groups according to the ICTV and NCBI taxonomic classifications; the shaded nodes (A and B) distinguish between the non-carcinogenic and carcinogenic families. Bootstrap scores are above 80% for most of the branches; for a better readability, they are not represented. The HPVs 1 and 34 are present in two copies, (1 and 1a) and (34A and 34B), respectively.

We also carried out an analysis intended to compare the topologies of the gene phylogenies built for the 8 main HPV genes. Thus, we first aligned, using ClustalW (Thompson et al., 1994), the HPV gene sequences, separately for each gene, and inferred 8

gene phylogenies using the PHYML program (Guindon and Gascuel, 2003) with the HKY model. In order to measure their degree of difference, we computed the Robinson and Foulds (RF) topological distances between each pair of gene trees (Robinson and Foulds, 1981). As the number of tree leaves varied from 70 to 83 (due to the non-availability of some gene sequences for a few HPVs), we reduced the size of some trees prior to this pairwise topological comparison and normalized all distances by the largest possible value of the RF distance, which is $2n - 6$ for two binary trees with n leaves. Figure 4.2 shows the results obtained, with RF distances are depicted as stacked rectangles.

The results suggest that the trees representing the evolution of the E4 and E5 genes differ the most, on average, from the other gene phylogenies, whereas the phylogeny of E2 reconciles the most the topological differences of this group of gene trees. Two HPV gene phylogenies differ from each other by about 32%, on average. In the future, it might also be interesting to compare the gene trees we obtained using Maximum Likelihood tests such as Shimodaira-Hasegawa (Shimodaira and Hasegawa, 1999) or Kishino-Hasegawa (Kishino and Hasegawa, 1989) and to assess the confidence of phylogenetic tree selection using program such as CONSEL (Shimodaira and Hasegawa, 2001).

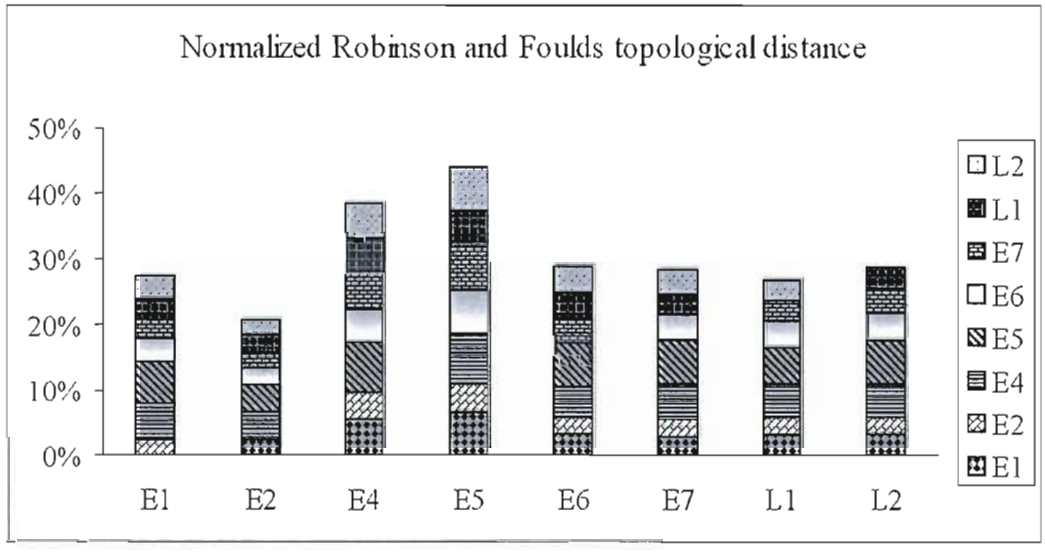

Figure 4.2 Average normalized Robinson and Foulds topological distance for each of the 8 main HPV genes.

Each column of the diagram represents a gene and consists of the stacked rectangles whose heights are proportional to the values of the normalized Robinson and Foulds topological distances between the phylogeny of this gene and those represented by the stacked rectangles. The column heights depict the total average distance. For the sake of presentation the percentage values on the ordinate axis were divided by 7 (which is the number of pairwise comparisons made for each gene tree).

These results confirm the hypothesis made in a number of HPV studies (see for instance (Narechania et al., 2005; Varsani et al., 2006)), that most HPV genes undergo frequent recombination events. Uncritical phylogenetic analyses performed on recombinant sequences could lead to the impression of novel, relatively isolated branches. Recently, Angulo and Carvajal-Rodriguez (2007) have provided new support to the recent evidence of recombination in HPV. They found that the gene with recombination in most of the groups is L2 but the highest recombination rates were detected in L1 and E6. Gene E7 was recombinant only within the HPV16 type. The authors concluded that this topic deserves further study because recombination is an important evolutionary mechanism that could have a high impact both in pharmacogenomics and for vaccine development.

4.2.3 Algorithm for the identification of putatively carcinogenic regions

This section describes a new algorithm intended for finding genomic regions that may be responsible for HPV carcinogenicity. The algorithm is based on the hypothesis that sequence regions responsible for cancer are likely to be more similar among carcinogenic HPVs than between carcinogenic and non-carcinogenic HPVs. The following procedure was adopted. First, 83 available HPV genomes were downloaded and inserted into a relational database along with the clinical information regarding identified HPV types and histological type of cancer occurrences (Muñoz et al., 2003, 2004).

We constructed three HPV Types Datasets: "High-Risk", containing HPVs16 and 18, "Squamous", containing HPV types responsible for Squamous Cell Carcinoma (HPV-6, 11, 16, 18, 26, 31, 33, 39, 45, 51, 52, 53, 55, 56, 58, 59, 66, 73, 81, 82, 83) and Adeno with types responsible for Adenocarcinoma (HPV-16, 18, 31, 33, 35, 39, 45, 51, 58, 59). See Table 4.1

for more details. HPV types with incomplete genome information or without annotations were excluded from the dataset. As previously, we used the gene sequences aligned separately for each gene. Then, we scanned all gene sequence alignments using a sliding window of a fixed width (in our experiments the window width ranged from 3 to 20 nucleotides, see Figure 4.3).

First, a detailed scan of each gene with increments of 1 nucleotide was performed to identifying the regions with a potential for causing carcinogenicity (the main results are reported in Table 4.3), and called here hit 10 regions. Second, a non-overlapping windows of width 20 nucleotides was carried out for plotting Figures 4.4, 4.5, 4.6, 4.7 and 4.8. Three separate analyses were made for the three above-described carcinogenic families: High-Risk, Squamous and Adeno HPVs.

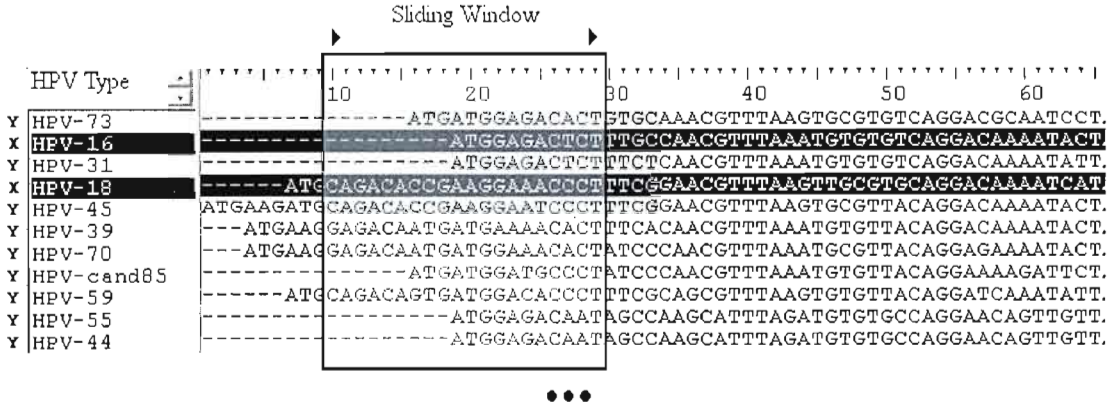

Figure 4.3 A sliding window of a fixed width was used to scan each HPV gene separately.
The sequences in black belong to the set X (carcinogenic HPVs; in this example HPVs 16 and 18), all other sequences belong to the set Y (non-carcinogenic HPVs). The organism is indicated in the column on the extreme left.

Once the window position is fixed and the taxa are assigned to the sets X (carcinogenic HPVs) and Y (non-carcinogenic HPVs), the hit region identification function, denoted here as Q, can be computed. This function is defined as a difference between the means of the squared distances computed among the sequence fragments (bounded by the sliding window

position) of the taxa from the set X and those computed only between the sequence fragments from the distinct sets X and Y. The mean of the squared distances computed among the sequence fragments of the carcinogenic taxa from the set X, and denoted here $V(X)$, is computed as follows:

$$V(X) = \frac{1}{(N(X)(N(X)-1)/2)} \sum_{\{x_1, x_2 \in X \mid x_1 \neq x_2\}} dist_h^2(x_1, x_2) \tag{4.1}$$

and the mean of the squared distances computed only between the sequence fragments from the distinct sets X and Y, and denoted here as $D(X, Y)$, is computed as follows:

$$D(X,Y) = \frac{1}{N(X)N(Y)} \sum_{\{x \in X, y \in Y\}} dist_h^2(x, y), \tag{4.2}$$

where $N(X)$ and $N(Y)$ are the cardinalities of the sets X and Y, respectively, and $dist_h(x_1, x_2)$ is the Hamming distance between the sequence fragments corresponding to the taxa x_1 to x_2. Then, the hit region identification function Q is defined as follows:

$$Q = \ln(1 + D(X, Y) - V(X)). \tag{4.3}$$

The larger the value of this function for a certain genomic region, the more distinct are the carcinogenic taxa from the non-carcinogenic ones. The use of the Hamming distance instead of the well-adapted sequence to distance transformations such as the Jukes-Cantor (1969), Kimura 2-parameter (1980) or Tamura-ei (1993) corrections, is justified by the two following facts: first, often the latter transformation formulae are not applicable to short sequences (remember that in our experiments the sequence lengths, equal to the sliding window width, varied from 3 to 20 nucleotides), and second, most of the well-known transformation models either ignore gaps or assign a certain penalty to them. As the carcinogenicity of HPVs can be related to an insertion or deletion of a group of nucleotides, the gaps should not be ignored but rather considered as valid characters, with the same

weight as the other nucleotides, when computing the pairwise distances between the genomic regions.

The time complexity of this algorithm executed with overlapping sliding windows of a fixed width, and advancing one alignment site by step, is $O(l \times n^2 \times w)$, where l is the length of the multiple sequence alignment, n the number of taxa, and w the window width. However, this complexity can be reduce to $O(n^2 \times l)$ if we avoid recomputing the Hamming distance for neighbouring overlapping windows. This can be done by only removing the value of the left column of the sliding window while taking into account the value of added column in the Hamming distance of the sliding window. For a non-overlapping sliding window, the time complexity is $O(n^2 \times l)$. If the width of the sliding window varies, as it was the case in our experiments, the time complexity should be obviously multiplied by the difference between the maximum and minimum window widths.

The detailed algorithmic scheme is presented below. To identifiy a region as a hit, one might use a measure to determine whether the given region has a value of Q higher than a given threshold. However, it is unclear what will be the best value of threshold, since the distribution of values of Q might be different in function of the alignment.

One possibility could be to rank the Q values and choose a set of highest ones. Moreover, an approach involving the computation of p-values could be implemented to determine the regions that have a value of Q that is different from the normal Q values of the alignment. Here, we used the mentioned different approaches to choose the relevant regions according to their value of Q.

To compute the p-value for each given regions W_i with a Q value Q_i, random sampling of the alignment columns according to the window size has been done. One million samples were generated and their Q values computed. For each given region the number of time that Q from the sample is higher than Q_i is counted. It is worth noting that, one would expect most of the region with value of Q to have a p-value less than 0.001.

Algorithm 4.1 Algorithmic scheme(MSA, MSA L,X, N(X), Y, N(Y), WIN MIN, WIN MAX, S, TH)

Require: MSA: Multiple sequence alignment (considered as a matrix),
MSA_L: Length of MSA,
X: Set of carcinogenic taxa,
N(X): Cardinality of the set X,
Y: Set of non-carcinogenic taxa,
N(Y): Cardinality of the set Y,
WIN_MIN: Minimum sliding window width,
WIN_MAX: Maximum sliding window width,
S: Sliding window step,
TH: Minimum Q value for Hit (i.e., hit threshold).

Ensure: Set of Hit Regions: (*win width, idx, Q*), where
win_width : Current sliding window width,
idx : Hit Index (i.e., its genomic position),
Q : Value of the hit region identification function.

1: **for** *win_width* **from** *WIN MIN* **to** *WIN MAX* **do**
2: **for** *idx* **from** 0 **to** *MSA L−win_width* **with step** *S* **do**
3: $MSA\_X \leftarrow \mathrm{MSA}[X][idx..idx + win\_width]$
4: $\mathrm{MSA\_Y} \leftarrow \mathrm{MSA}[Y][idx..idx + win\_width]$
5: $V(X) \leftarrow D(X,Y) \leftarrow 0$
6: **for all distinct** $i, j \in X$ **do**
7: $V(X) \leftarrow V(X) + \mathrm{dist}^2{}_{\mathrm{h}}(MSA\_X[i], MSA\_X[j])$
8: **end for**
9: $V(X) \leftarrow 2 \times V(X)/(N(X) \times (N(X) - 1))$
10: **for each** $i \in X$ **and** $j \in Y$ **do**
11: $D(X,Y) \leftarrow D(X,Y) + \mathrm{dist}^2{}_{\mathrm{h}}(MSA\_X[i], MSA\_Y[j])$
12: **end for**
13: $D(X,Y) \leftarrow D(X,Y) / (N(X) \times N(Y))$
14: $Q \leftarrow \ln(1 + D(X,Y) - V(X))$
15: **if** $Q > TH$ **then**
16: *identify the current region (win width, idx, Q) as a hit region*
17: **end if**
18: **end for**
19: **end for**

4.2.4 Results, discussion and conclusion

The procedure for identifying hit regions in the 83 available HPV genomes was carried out twice: first, with overlapping windows of width w ($w = 3..20$), advancing one alignment site by step, and second, with non-overlapping windows of width 20. The 8 most important HPV genes (see Table 4.3) were scanned in such a way. The scan based on the overlapping windows provided over 35,000 values of Q bigger than 0.25. From the best 100 results obtained for each gene, we manually selected (see Table 4.3) the longest contiguous regions (up to 20 nucleotides) corresponding to the largest values of the hit region identification function Q. The values of Q were dependent on the window width, with better results usually associated with small windows.

For instance (see Table 4.3), for larger window sizes, the largest values of Q were found during the scans of genes E2 and E6 for all types of HPVs, with the exception of the overall best score obtained during the scan of the gene L1 for the High-Risk HPV types (the value of 0.574 for a 14-nucleotide region starting with the index 241, see Table 4.3). For windows of small width, the largest values of Q were observed during the scan of the gene E4 for the High-Risk HPV category but in Table 4.3 we show only the best results for the longer contiguous regions of size 13 to 20 nucleotides. All the regions presented in Table 4.3 have a p-value of 0.

Figure 4.4 depicts the progressive results obtained during the scan of the L1 gene and the High-Risk HPVs (HPVs-16 and 18) with the non-overlapping windows of size 20 nucleotides. The highest score, for the non-overlapping windows of size 20 among all genes and all types of HPV-caused cancer, of the Q function ($Q = 0.55$) was obtained for this gene.

Table 4.3 Selected high-scoring regions with respect to the values of the hit region identification function Q.

The best results for the contiguous regions of size 13 to 20 are reported. The best entry by HPV type (High-Risk, Squam, Adeno) and by gene is presented. The largest values of Q are in bold.

Dataset	Gene	Q	Index	Window width	D(X,Y)	V(X)
High-Risk	E1	0.417	695	16	0.74	0.22
Squam	E1	0.345	575	14	0.50	0.08
Adeno	E1	0.353	307	20	0.52	0.09
High-Risk	**E2**	**0.553**	**1289**	**13**	**0.76**	**0.02**
Squam	E2	0.385	613	16	0.47	0.00
Adeno	E2	0.415	1265	20	0.66	0.14
High-Risk	E4	0.480	606	17	0.62	0.00
Squam	E4	0.373	1035	15	0.46	0.01
Adeno	E4	0.395	549	15	0.49	0.00
High-Risk	E5	0.339	88	13	0.41	0.01
Squam	E5	0.401	72	16	0.50	0.00
Adeno	E5	0.363	72	16	0.44	0.00
High-Risk	E6	0.496	725	17	0.69	0.05
Squam	**E6**	**0.531**	**725**	**17**	**0.76**	**0.06**
Adeno	**E6**	**0.521**	**725**	**17**	**0.75**	**0.06**
High-Risk	E7	0.258	206	13	0.34	0.05
Squam	E7	0.263	445	16	0.38	0.08
Adeno	E7	0.262	110	16	0.40	0.10
High-Risk	**L1**	**0.574**	**241**	**14**	**0.79**	**0.02**
Squam	L1	0.294	1159	15	0.34	0.00
Adeno	L1	0.302	1181	17	0.56	0.20
High-Risk	L2	0.310	1751	14	0.65	0.28
Squam	L2	0.320	1916	15	0.38	0.00
Adeno	L2	0.313	1914	17	0.37	0.00

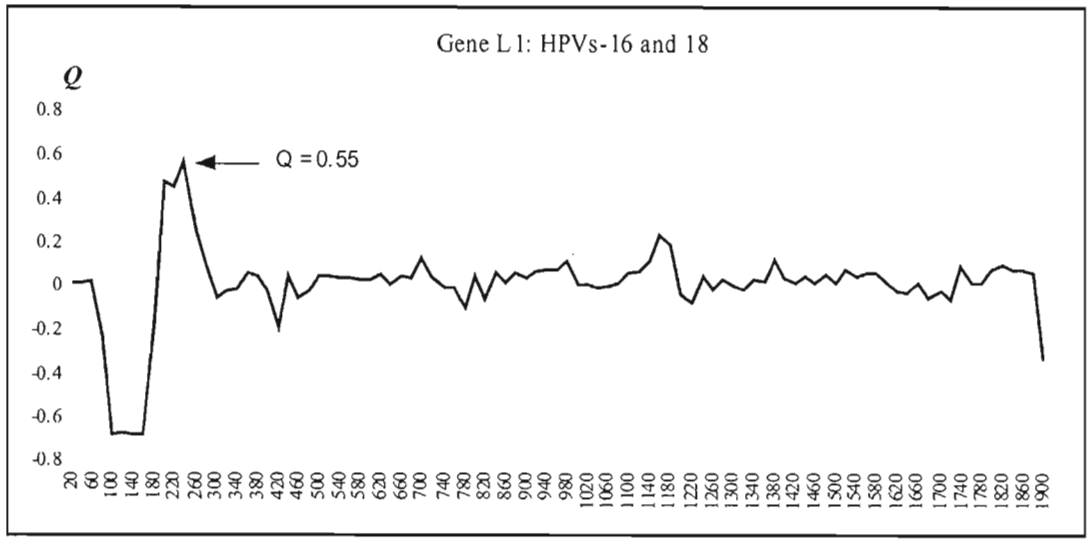

Figure 4.4 The variation of the hit identification function Q for the High-Risk HPVs (HPVs-16 and 18) obtained with the non-overlapping sliding widow of width 20 during the scan of the L1 gene.
The abscissa axis represents the window position.

As most of the largest values of Q were obtained for the genes E2 and E6, we also present in Figure 4.7 and 4.8 the progressive results diagrams illustrating the scan of these genes with the non-overlapping windows of size 20. The largest values of the hit region identification function Q are usually found during the scan of the genes E2 and E6. Moreover, we found that in these two genes the number of regions obtaining p-values less than 0.001 is the largest. For instance, in gene E6, three large regions of size between 40 nucleotides and 60 nucleotides have a p-value less than 0.001 (Figure 4.5 and 4.6). The last region of figure of E6 surprisingly corresponds to a PDZ domain-binding motif (-X-T-X-V) at the carboxy terminus of the protein, which is essential for targeting PDZ proteins for proteasomal degradation. Such proteins include hDlg, hScrib, MAGI-1, MAGI-2, MAGI-3, and MUPP1 (Choongho and Laimonis, 2004). The interaction between the E6 protein and hDLG or other PDZ domain-containing proteins could be an underlying mechanism in the development of HPV-associated cancers (Tohru et al., 1997) .

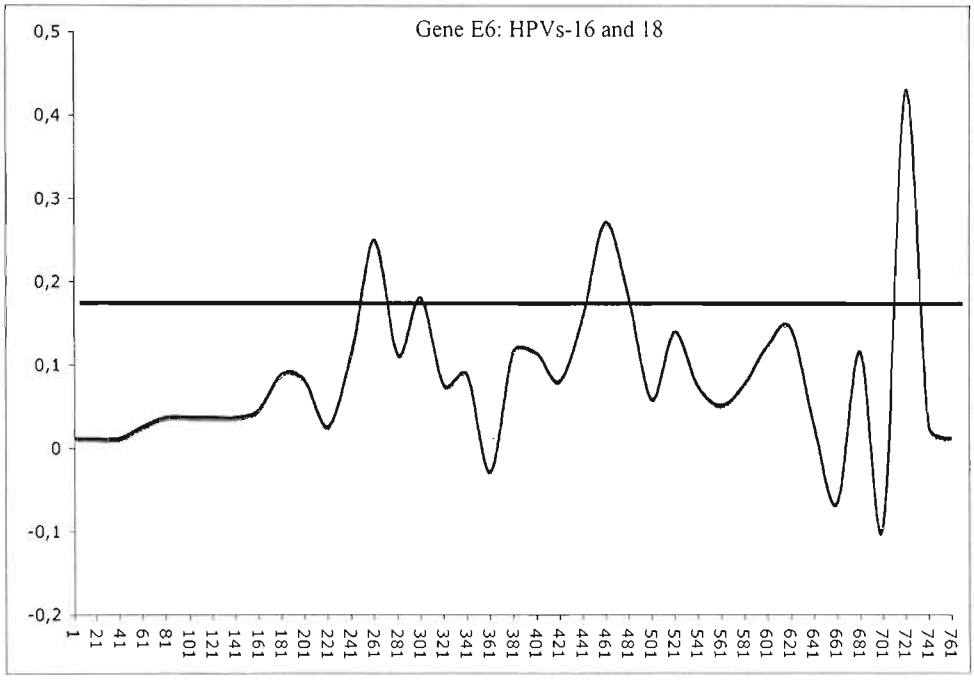

Figure 4.5 The variation of the hit identification function Q for the High-Risk HPVs (HPVs-16 and 18) obtained with the non-overlapping sliding widow of width 20 during the scan of the E6 gene.
The horizontal line cutting the graph represents the threshold of p-value less than 0.001. The abscissa axis represents the window position.

It is worth noting that according to recent findings the high expression of E6 and disruption of E2 might play an important role in the development of HPVinduced cervical cancer (Wang et al., 2007). As result of E6 high expression, the immune system is potentially evaded (P. et al., 2008). Disruption of the gene E2 was observed in invasive carcinomas (Chan et al., 2007) and in high-grade lesions (Graham and Herrington, 2000). Surprisingly, the overall largest value of Q was obtained for a specific region of the L1 gene. This underlines the possible use of our method for investigating particular regions of capsidal proteins in relation with vaccine design. It has been shown that linear epitopes within the protein L1 that induce neutralizing antibodies exist (Combita et al., 2002).

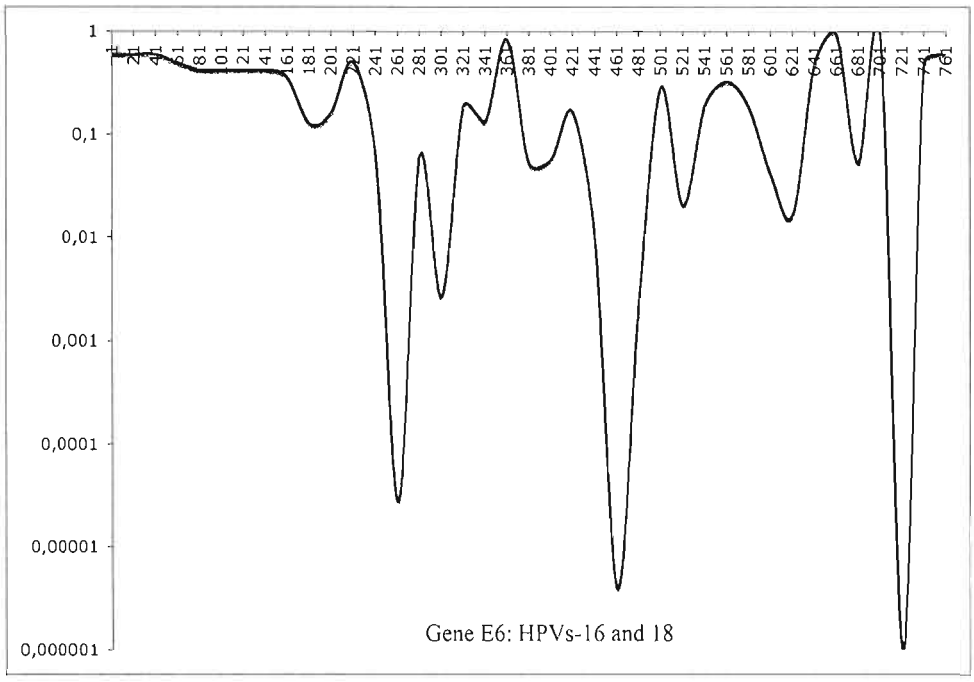

Figure 4.6 The variation of the p-value in the different region of the alignment for the High-Risk HPVs (HPVs-16 and 18) obtained with the non-overlapping sliding widow of width 20 during the scan of the E6 gene.
The abscissa axis represents the window position.

We observed that the results obtained depend on the window width. As substitutions affect individual sites whereas indels often involve several consecutive nucleotides, small window sizes will tend to favor the former. However, the use of the Hamming distance, which does not ignore gaps in calculation, and variable window width allows us to account for both substitution and indel events. In the future, it would be interesting to study in more detail, in collaboration with virologists, all genomic regions providing the highest scores of the hit region identification function Q (particular attention should be paid to the E2, E6 and L1 genes), and to determine, for each selected region, the evolutionary events (substitutions or indels) responsible for the observed differences in the carcinogenic and non-carcinogenic

HPVs, and then establish at which level (i.e. on which branch) of the associated gene phylogeny this event has occurred. It may also be interesting to consider merging our results to those given by methods for detecting sequences under lineage-specific selection such as DLESS (Siepel et al., 2006). Next, we plan to compare this work with other approaches on the computational virology, which used some simpler methods, such as signatures, to analyze other viruses. Another interesting development would be to design more sophisticated statistical tests allowing one to measure the statistical significance of the obtained results.

Acknowledgement B.D. is an NSERC fellow. We thank Alix Boc and Emmanuel Mongin for their useful comments.

Additionnal materials related to this study are available at: <http://ancestors.bioinfo.uqam.ca/articles/JCB2009/supplemental.zip>. These materials contain the data used and the whole results for all scanned genes with different window width.

4.2.5 Bibliography

Angulo, M. and Carvajal Rodriguez, A., 2007. Evidence of recombination within human alpha-papillomavirus. Virology Journal 4, 33.

Antonsson, A., Forslund, O., Ekberg, H., Sterner, G., and Hansson, B., 2000.The ubiquity and impressive genomic diversity of human skin papillomaviruses suggest a commensalic nature of these viruses. Journal of Virology 74(24), 11636–1164.

Bosch, F., Manos, M., Muñoz, N., Sherman, M., Jansen, A., Peto, J., Schiffman, M., Moreno, V., Kurman, R., and Shan, K., 1995. Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. international biological study on cervical cancer (ibscc) study group. Journal of the National Cancer Institute 87(11), 796–802.

Chan, P., Cheung, J., Cheung, T., Lo, K., Yim, S., Siu, S., and Tang, J., 2007. Profile of viral load, integration, and e2 gene disruption of hpv58 in normal cervix and cervical neoplasia. Journal of Infectious Diseases 196(6), 868–875.

Chan, S., Delius, H., Halpern, A., and H.U., B., 1995. Analysis of genomic sequences of 95 papillomavirus types: uniting typing, phylogeny, and taxonomy. Journal of Virology 69(5), 3074–3083.

Choongho, L. and Laimonis, A., 2004. Role of the pdz domain-binding motif of the oncoprotein e6 in the pathogenesis of human papillomavirus type 31.Journal of Virology 78(22), 12366–12377.

Combita, A.-L., Touz´e, A., Bousarghin, L., Christensen, N., and Coursaget, P., 2002. Identification of two cross-neutralizing linear epitopes within the l1 major capsid protein of human papillomaviruses. Journal of Virology 76(13), 6480–6486.

de Villiers, E., Fauquet, C., Broker, T., Bernard, H., and Zur Hausen, H., 2004. Classification of papillomaviruses. Virology 324(1), 17–27.

Diallo, A., Makarenkov, V., and Blanchette, M., 2007. Exact and heuristics methods to indel maximum likelihood problem. Journal Computational Biology 14, 446–461.

Diallo, A., Makarenkov, V., and M., B., 2006. Finding maximum likelihood indel scenarios. In Comparative Genomics, 171–185. Springer. LNCS vol. 4205. Graham, D. and Herrington, C., 2000. Hpv-16 e2 gene disruption and sequence variation in cin 3 lesions and invasive squamous cell carcinomas of the cervix: relation to numerical chromosome abnormalities. Molecular Pathology 53, 201–206.

Guindon, S. and Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52(5), 696–704.

Kishino, H. and Hasegawa, M., 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from dna sequence data, and the branching order in hominoidea. Journal of Molecular Evolution 29, 170–179.

Muñoz, N., 2000. Human papillomavirus and cancer: the epidemiological evidence.Journal of Clinical Virology 19(1-2), 1–5.

Muñoz, N., Bosch, F., Castellsagu´e, X., Daz, M., de Sanjose, S., Hammouda, D., Shah, K., and Meijer, C., 2004. Against which human papillomavirus types shall we vaccinate and screen? the international perspective. International Journal of Cancer 111, 278–285.

Muñoz, N., Bosch, F., de Sanjos´e, S., Herrero, R., Castellsagu´e, X., Shah, K., Snijders, P., and Meijer, C., 2003. Epidemiologic classification of human papillomavirus types associated with cervical cancer. New England Journal of Medecine 384, 518–527.

Narechania, A., Chen, Z., DeSalle, R., and Burk, R., 2005. Phylogenetic incongruence among oncogenic genital alpha human papillomaviruses. Journal of Virology 79, 15503–15510.

P., C., Gillan, V., Bratlie, S., Bouvard, V., Banks, L., Tommasino, M., and Campo, M., 2008. The e6e7 oncoproteins of cutaneous human papillomavirus type 38 interfere with the interferon pathway. Virology 377(2), 408–418.

Prétet, J., Charlot, J., and Mougin, C., 2007. Virological and carcinogenic aspects of hpv. Bulletin Academic National de Medecine 191(3), 611–613.

Robinson, D. and Foulds, L., 1981. Comparison of phylogenetic trees. Mathematical Biosciences 53, 131–147.

Shimodaira, H. and Hasegawa, M., 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Molecular Biology and Evolution 16, 1114–1116.

Shimodaira, H. and Hasegawa, M., 2001. Consel: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17, 1246–1247.

Siepel, A., Pollard, K., and Haussler, D., 2006. New methods for detecting lineage-specific selection. In Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB 2006), 190–205.

Thompson, J., Higgins, D., and Gibson, T., 1994. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research 22, 4673–4680.

Tohru, K., H., A., Masatoshi, F., H., Y., A., T., and I., M., 1997. Binding of high-risk human papillomavirus e6 oncoproteins to the human homologue of the drosophila discs large tumor suppressor protein. pnas 94(21), 11612–11616.

Van Ranst, M., Kaplanlt, J., and Burk, R., 1992. Phylogenetic classification of human papillomaviruses: Correlation with clinical manifestations. Journal of General Virology 73, 2653–2660.

Varsani, A., Van der Walt, E., Heath, L., Rybicki, E., Williamson, A., and Martin, D., 2006. Evidence of ancient papillomavirus recombination. Journal of General Virology 87, 2527–2531.

Wang, J., Ding, L., Gao, E., and Cheng, Y., 2007. Analysis on the expression of human papillomavirus type 16 e2 and e6 oncogenes and disruption of e2 in cervical cancer. Zhonghua Liu Xing Bing Xue Za Zhi 28(10), 968–971.

Wilson, R., Ryan, G., Knight, G., Laimins, L., and Roberts, S., 2007. The full-length e1^e4 protein of human papillomavirus type 18 modulates differentiation-dependent viral dna amplification and late gene expression. Virology 362(2), 453–460.

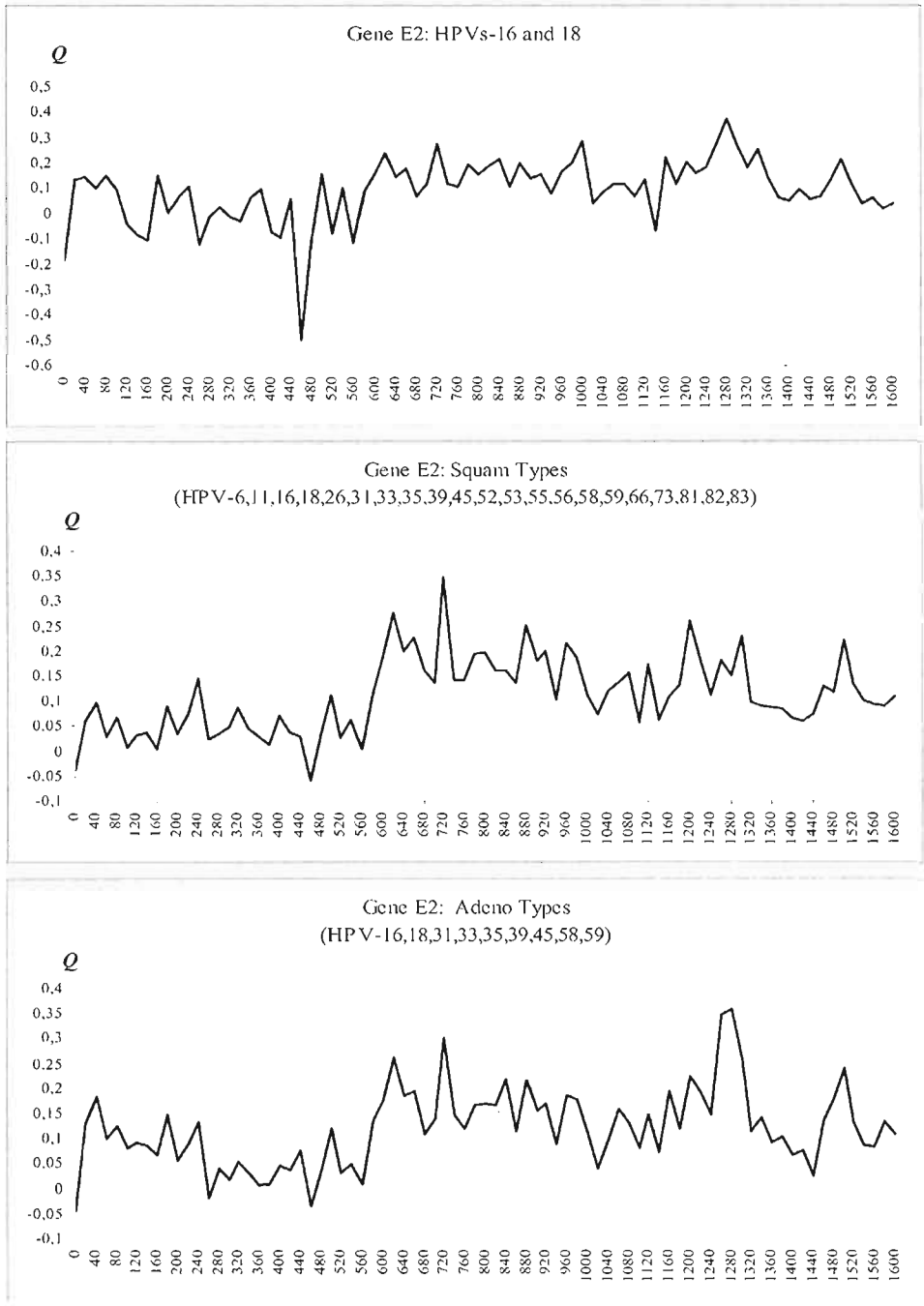

Figure 4.7 The variation of the hit identification function Q for: High-Risk HPVs (HPV-16 and 18), (b) Squam cancer causing HPVs, and c) Adeno cancer causing HPVs obtained with the non-overlapping sliding widow of width 20 during the gene E2 scan.

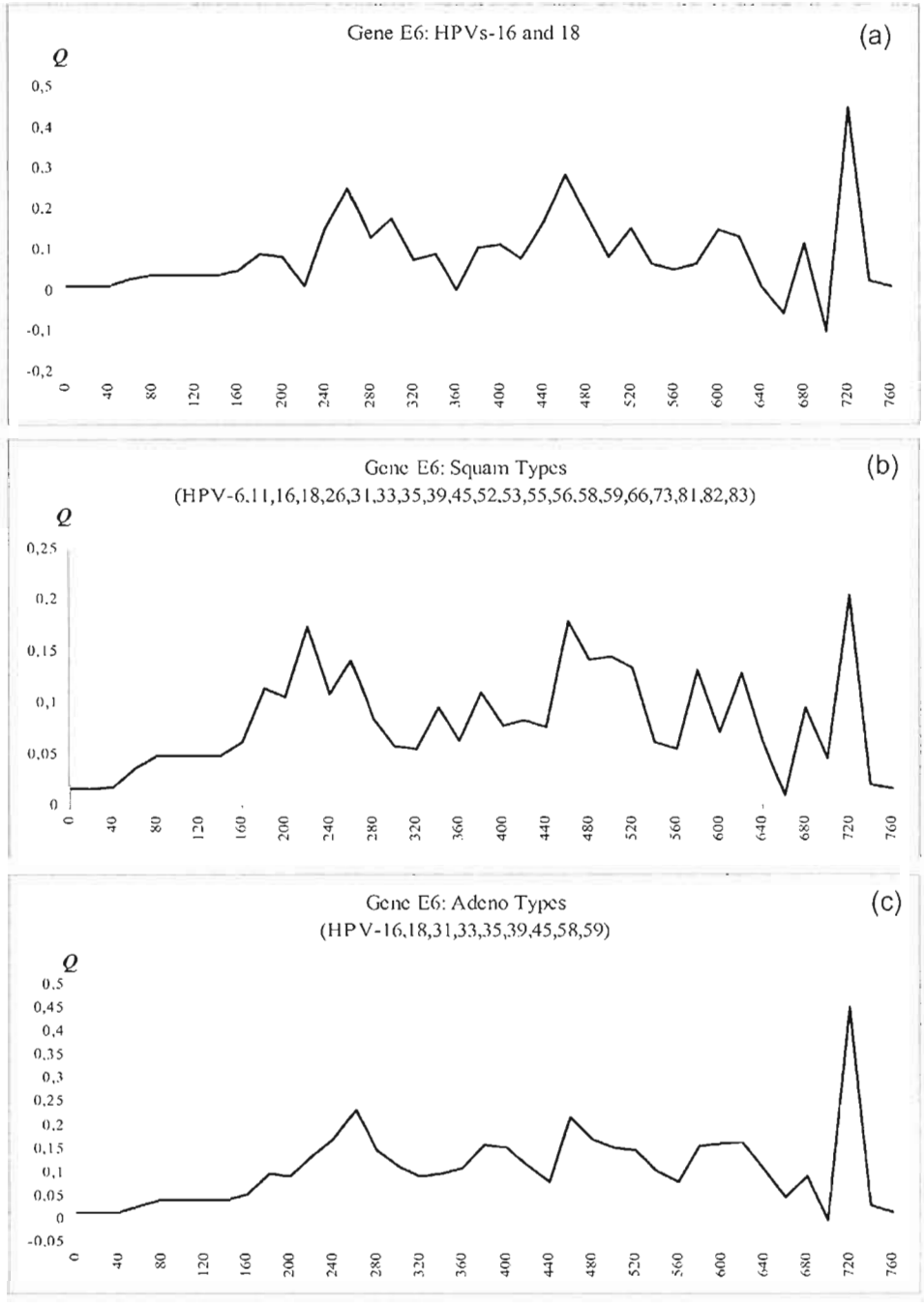

Figure 4.8 The variation of the hit identification function Q for: (a) High-Risk HPVs (HPV-16 and 18), (b) Squam cancer causing HPVs, and c) Adeno cancer causing HPVs obtained with the non-overlapping sliding widow of width 20 during the gene E6 scan.

CHAPITRE V

IDENTIFICATION DES REGIONS GENOMIQUES SPECIFIQUES RESPONSABLES DE L'INVASIVITE DU *NEISSERIA MENINGITIDIS*

5.1 Résumé français de Badescu et al. (2010)

Sommaire. Cet article présente quatre fonctions de discrimination pour l'identification de segments génomiques distinctif pour deux groups de données. Elles sont conçues dans le but de détecter les régions génomiques responsables de maladie. L'une d'entre elles a déjà été employée dans l'analyse du VPH en relation avec la carcinogenicité. Cet article améliore l'algorithme de Badescu et al. (2008) pour l'analyse du contenu informationnel d'alignements multiples en relation avec les données épidémiologiques. *Neisseria Meningitidis* est une bactérie mondialement responsable d'une forme grave de méningite et de septicémie. Au cours de cette étude, ces fonctions ont été utilisées pour l'identification des régions génomiques spécifiques, responsables de son hyperinvasivité. Cette étude suggère que les fonctions testées permettent d'identifier des régions relevantes et des structures moléculaires connues.

5.1.1 Introduction

L'évolution des bactéries se fait au rythme des événements génomiques à petite échelle comme les insertions, délétions et substitutions ou à grande échelle comme le transfert horizontal de gène, la duplication de segments, les transpositions, etc.. Sur une échelle de

temps réduite les allèles du même organisme divergent très peu, quelques rares événements sont d'habitude présents. L'alignement des séquences de ces allèles assure que les nucléotides placés dans la même région partagent une histoire évolutionnaire commune. Sous une pression sélective, les modifications moléculaires conduiront à des comportements épidémiologiques différents. Une stratégie employée pour la détection de ces régions de séquences, qui par rapport au comportement ont un impact sur l'espèce en question est la détection des régions conservées. Ceci peut être fait avec des arbres phylogénétiques, de modèles de Markov cachés, la détection des régions sous sélection lignée-spécifique, détection de motifs, et autres méthodes comme l'utilisation des signatures. Ces méthodes analysent une famille à la fois et ne peuvent pas prendre en compte plusieurs catégories de données. Cependant leur complexité algorithmique exponentielle réduit leur applicabilité à un nombre très limité de taxons.

5.1.2 *Neisseria Meningitidis* et les protéines *FrpB*

Neisseria Meningitidis est une bactérie Gram négative ayant une variabilité importante, 7300 nouveaux membres étant identifiés dans la base de données PubMLST. Pour cette raison elle constitue un bon candidat aux études de génomique comparée. FrpB est une protéine exprimée lorsque la bactérie est placée dans un milieu pauvre en fer. Ceci est le cas dans l'organisme humain, où un important mécanisme de défense antibactérienne lie l'élément Fe libre aux protéines. Une topologie putative a été proposée avec 26 feuillets-β et 11 anses exposées à la surface et accessibles au système immunitaire. Cette étude se consacre à la détection de ces régions, hôtes de la lutte bactérie-système immunitaire.

5.1.3 Algorithme pour la détection des régions génomiques responsables de la maladie

Le flux de l'algorithme est présenté à la figure 5.1, ainsi qu'au chapitre précédent. Les formules 5.1, 5.2 et 5.3 sont utilisées pour la variabilité intra- et inter-groupes. Quatre fonctions d'agrégation (5.4-5.7) sont présentées et testées. Les souches invasives et non-invasives se retrouvent dans les ensembles X et Y respectivement.

5.1.4 Résultats et discussion

Nous avons balayé l'alignement multiple du gène FrpB avec les quatre versions de la fonction d'agrégation et une taille de la fenêtre de 10 nucleotides. Les plus grandes valeurs correspondent aux régions grises des figures 5.2 et 5.3. Les régions grises représentent les anses exposées à la surface de la cellule (L1-L12). Pour la fonction Q_1, les grandes valeurs positives représentent des régions conservées dans l'ensemble X (des souches invasives). Ces régions sont hautement divergentes en cas des petites valeurs de Q_1. La fonction Q_2 relève les mêmes relations pour l'ensemble Y (des souches non-invasives). La fonction Q_3 est positive quand la distance entre les ensembles X et Y est plus grande que sa variabilité. Elle corrèle moins avec les zones grises excepté pour le gène L3. Pour comparer les différentes fonctions, nous avons procédé à une normalisation de moyenne zéro et variance unitaire. Les résultats sont présentés à la table 5.1. La fonction Q_4 représente la distance entre les deux groupes, et se distingue comme la meilleure en terme de la détection des anses extra-cellulaires.

5.1.5 Conclusion

La fonction Q_4 corrèle mieux avec la détection des anses extra-cellulaires que les fonctions Q_1, Q_2 et Q_3 introduites dans ce papier. Ceci suggère que notre algorithme est capable de détecter des structures moléculaires connues en relation avec les données épidémiologiques.

5.2 Identification of specific genomic regions responsible for the invasivity of Neisseria Meningitidis

Dunarel Badescu[1], Abdoulaye Baniré Diallo[1,2], and Vladimir Makarenkov[1]

1 Département d'informatique,Université du Québec à Montréal, C.P. 8888, Succursale Centre-Ville, Montr´eal (Québec), H3C 3P8, Canada
{badescu.dunarel,diallo.abdoulaye,makarenkov.vladimir}@uqam.ca
2 McGill Centre for Bioinformatics and School of Computer Science, McGill University, 3775 University Street, Montréal (Québec), H3A 2B4, Canada

Summary. In this article, we present four distance-based discrimination functions for the identification of relevant genomic segments that distinguish between two groups of data. These discrimination functions are designed for the detection of genomic regions responsible for disease. One of them was previously employed for the analysis of the Human Papilloma Virus family in relation to carcinogenicity. Here, we used an improved version of the algorithm of Badescu et al. (2008) for analyzing the information content of a multiple sequence alignments (MSA) in relation to epidemiologic data [2]. In this study, those functions have been applied to identify specific genomic regions responsible for the hyperinvasivity of *Neisseria Meningitidis*. *Neisseria Meningitidis* is a major causal agent of meningitis and septicaemia worldwide. This study suggests that the tested functions permit to identify relevant regions and known molecular features. We found that one of the new functions tested is specifically well correlated with surface-exposed loops, regions important in vaccine design.

Key words: algorithm, distance functions, hit identification, *Neisseria Meningitidis*, invasivity, vaccine design.

5.2.1 Introduction

The evolution of bacteria is driven by several small scale evolutionary events such as substitutions, insertions and deletions and of nucleotides, and large scale mutations such as horizontal gene transfer, duplication of nucleotide segments, etc. However, on a small scale

time frame, alleles from the same bacteria organisms diverge little. For instance, when looking into a single gene, only small scale evolutionary changes are commonly present. Alignment of those allele sequences ensure that nucleotides placed on the same region (i.e. same site position), impart the same evolutionary history. Under selective pressure, these molecular modifications will lead to different epidemiological behaviors. One big issue in comparative genomics is the identification of these molecular modifications through the sequences conservation. One of the well known strategies for identifying genomic sequence regions that have high impact on the given species according to a specific behavior, consists of detecting sequence regions that are conserved accross species. Highly conserved regions specific to a family of organisms might have an important role on the common functions of this group [11]. Several methods for finding unusual hyper conserved genomic segments have been designed. Most of them are based on phylogenetic trees. They identify hyper conserved genomic segments using hidden Markov model such as Phastcons [11], detect sequences under lineage-specic selection such as DLESS [10] or detect nearly exact motifs using phylogenetic footprinting [3]. Other simpler methods such as signatures or exact motif finding are also used but they have little application. It is important to notice that, the latter methods analyze a single family at once, and cannot take into account different data categories. Finally due to their exponential time complexity, they are limited to small number of taxa. Being able to classify a family of organisms into a few categories can be an important clue for the detection of their common features. Statistically analyzing the intra- and inter-population variability between two categories can help finding quickly the DNA regions responsible for the difference between the observed categories. In this paper, we tested four distance-based functions for the identification of such differences. They are integrated into an improved version of the algorithm of [2] for analyzing the information content of a MSA in relation to epidemiological data. The proposed functions have been applied to the detection of DNA regions related to the hyperinvasivity of the *Neisseria Meningitidis*. The results presented here suggest that the new functions have a good correlation with known molecular features involved in immunological conflict, and responsible for hyperinvasivity.

5.2.2 *Neisseria Meningitidis* and the *FrpB* proteins

Neisseria Meningitidis is a Gram negative bacterium with a high medical importance and very large family. It has small genomic size with 2.2 Mbp. At the time of writing, more than 7300 genetically distinct known members of *Neisseria species* were listed into the PubMLST database [5]. The latter factor makes it well suited for carrying out for comparative genomic studies. However, bacteria grown under iron starvation express several proteins, FrpB being the most abundant one. It is a 70kDa outer membrane protein (OMP), expressed in large amounts in all strains, and antibodies against this protein appear to be bactericidal. Since iron limitation is a condition met in the body, proteins expressed under this condition are considered as a potential vaccine component [8]. A putative FrpB protein topology was proposed [9] with a 26-stranded β-barrel and a 22-stranded β-barrel with 11 surface-exposed loops. It is these loops that are accessible to the host immune system. Natural antibodies are generated against these regions and bacteria express variability in order to evade this defence mechanism. Also these 11 surface-exposed loops are also a favorite place of guest-host interaction. This study will focus on the detection of these surface-exposed loop regions under the knowledge of the organism categories (invasive and non−invasive alleles).

5.2.3 Algorithm for detection of genomic regions responsible for disease

This section describes the steps used for finding genomic regions that responsible for the invasivity of *Neisseria Meningitidis*. The algorithm tests several hypothesis such as whether sequence regions responsible for invasivity are likely to be more similar among invasive strand, or not. The algorithm takes as input a multiple sequence alignment (MSA) of nucleotides, and a set of organisms, clustered into two different groups (i.e. categories) according to their invasivity: X (invasive) and Y (non−invasive). We scanned the sequence alignment using an overlapping sliding window of a fixed width (in our experiments the window width ranged from 5 to 20 nucleotides). Once the window position in MSA is fixed and the organisms are assigned to the groups X and Y, various discrimination functions can be defined. The different steps of our procedure are described below. Figure 5.1 presents the

algorithmic flow of the hit identification, followed by the description of the different steps of this algorithm.

Step 1: Collection and annotation of the MSA of the FetA alleles: MSA of the FetA allele sequences are available from the Neisseria Research Community databank [12, 7].We annotated the MSA using the information on the surface-exposed loops, beta-sheets and periplasmic loops, as was explained in [6]. Identification and presentation were carried out on the H44/76 strain, with the GenBank accession number X89755.1 [9].

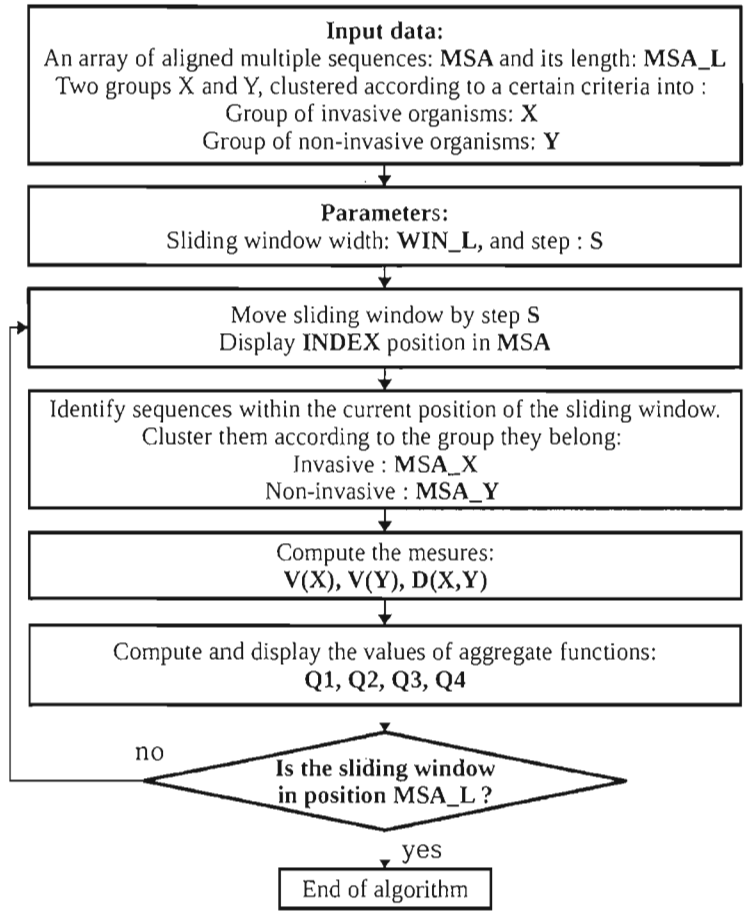

Figure 5.1 Algorithmic flow of the hit identification function Q, using pluggable functions Q_1, Q_2, Q_3, Q_4.

Step 2: Classification of taxa as invavise or non-invasive: To form groups X and Y on an invasivity basis we used a list of identified hyperinvasive meningococci [13]. We built a list of uniques FetA sequence tags carried by these alleles. With a local BLAST we searched for the presence of those tags into the distinct sequences belonging to the MSA [1].We classified as belonging to the X category any allele that has a perfect hit with at least one of the selected invasive tags. All others were put in the non-invasive category Y.

Step 3: Computating the detection functions Q values: For a fixed alignment window position the hit region identication functions (i.e. hit region is a region responsible for disease), denoted here as Q_1, Q_2, Q_3 and Q_4 are computed as follows. These functions are defined as a difference between the means of the squared distances computed among the sequence fragments (bounded by the sliding window position) of the taxa from the set X and those computed only between the sequence fragments from the distinct sets X and Y. To compute the function values, the variability of groups X and Y respectively $V(X)$ and $V(Y)$ is computed as well as the distance between X and Y, denoted $(D(X,Y))$. The variability of the group X corresponds to the mean squared distance computed among the sequence fragments of the invasive organisms. This variability is computed as follows:

$$V(X) = \frac{1}{(N(X)(N(X)-1)/2)} \left\{ \sum_{x_1, x_2 \in X \mid x_1 \neq x_2} dist_h^2 (x_1, x_2) \right\} \qquad (5.1)$$

The variability of class Y corresponds to the mean of the squared distance computed among the sequence fragments of the non-invasive organisms. This variability is computed as follows:

$$V(Y) = \frac{1}{(N(Y)(N(Y)-1)/2)} \left\{ \sum_{y_1, y_2 \in Y \mid y_1 \neq y_2} dist_h^2 (y_1, y_2) \right\} \qquad (5.2)$$

The distance between the groups X and Y corresponds to the mean of the squared distances computed among the sequence fragments from X and Y .This distance is computed as follows:

$$D(X,Y) = \frac{1}{N(X)N(Y)} \sum_{\{x \in X, y \in Y\}} dist_h^2(x,y), \qquad (5.3)$$

where $N(X)$, $N(Y)$, $\text{dist}_h^2(x, y)$ are respectively the cardinalities of the groups X and Y, and the square of the Hamming distance between the sequences x and y. We propose to examine four different hit identification functions allowing one to detect DNA zones responsible for disease. The function Q_1 focuses on the specific regions of the alignment that are either well-conserved within the invasive set X, when it is positive, or highly divergent, when it is negative. The function Q_2 focuses on the specific regions of the alignment that are either well-conserved within the non-invasive set Y, when it is positive, or highly divergent, when it is negative. The function Q_3 is positive when the distance between the taxa in X and Y is higher than the variability between the groups. The last function Q_4 consists only of the mean squared distances between the two groups:

$$Q_1 = D(X,Y) - V(X), \qquad (5.4)$$

$$Q_2 = D(X,Y) - V(Y), \qquad (5.5)$$

$$Q_3 = 2D(X,Y) - V(X) - V(Y), \qquad (5.6)$$

$$Q_4 = D(X,Y). \qquad (5.7)$$

Step 4: Identify hit regions. To identifiy a region as a hit, one might use a measure to determine whether the given region has a value of Q higher than a predefined threshold. However, it is necessary to normalize the obtained results given by Q_1, Q_2, Q_3 and Q_4 prior to compare them. We compare the trends of the different function according to the known regions of surface-exposed loops of *FetA* alleles. One can also determine the hit regions computing the p-values of the proposed functions [4].

5.2.4 Results and discussion

We scanned the MSA of the *FrpB* gene using the algorithm described in the previous section, with the four versions of the aggregate discrimination function Q, and window of size 10 nucleotides. Larger window sizes were less discriminative in terms of regionality, smaller sizes introduced more noise. One can notice (see Figures 5.2 and 5.3) that the high values of the four tested functions usually correspond to the gray zones (regions supposed to be responsible for the invasivity) of the graphics. The values of the function Q_2 are generally lower than those of the other functions. The latter means that the the divergence in Y is almost similar to the divergence between the alleles from X and Y (Figure 5.2(b)). The high values of the function Q_1 (above 0.05) in Figure 5.2(a) can be induced by highly conserved features in the X. The function Q_3 (Figure 5.3(a)) correlates less with the gray zone, except for L3. Furthermore, the trends of the function Q_4 suggests that for almost all gray zones (except L7), the genomic segments are different between the groups X and Y . In order to compare the values of the four competing functions, we carried out a zero mean and unit variance normalization. Table 5.1 presents the maximum values obtained for to the eleven gray zones in Figures 5.2 and 5.3. More than the half of the maximum values on this table are above the fixed threshold of 2. This result shows that the function Q_4 is the best one in terms of extracellular loops (gray zone) detection. The same conclusion can be drawn when observing the graphics in Figures 5.2 and 5.3.

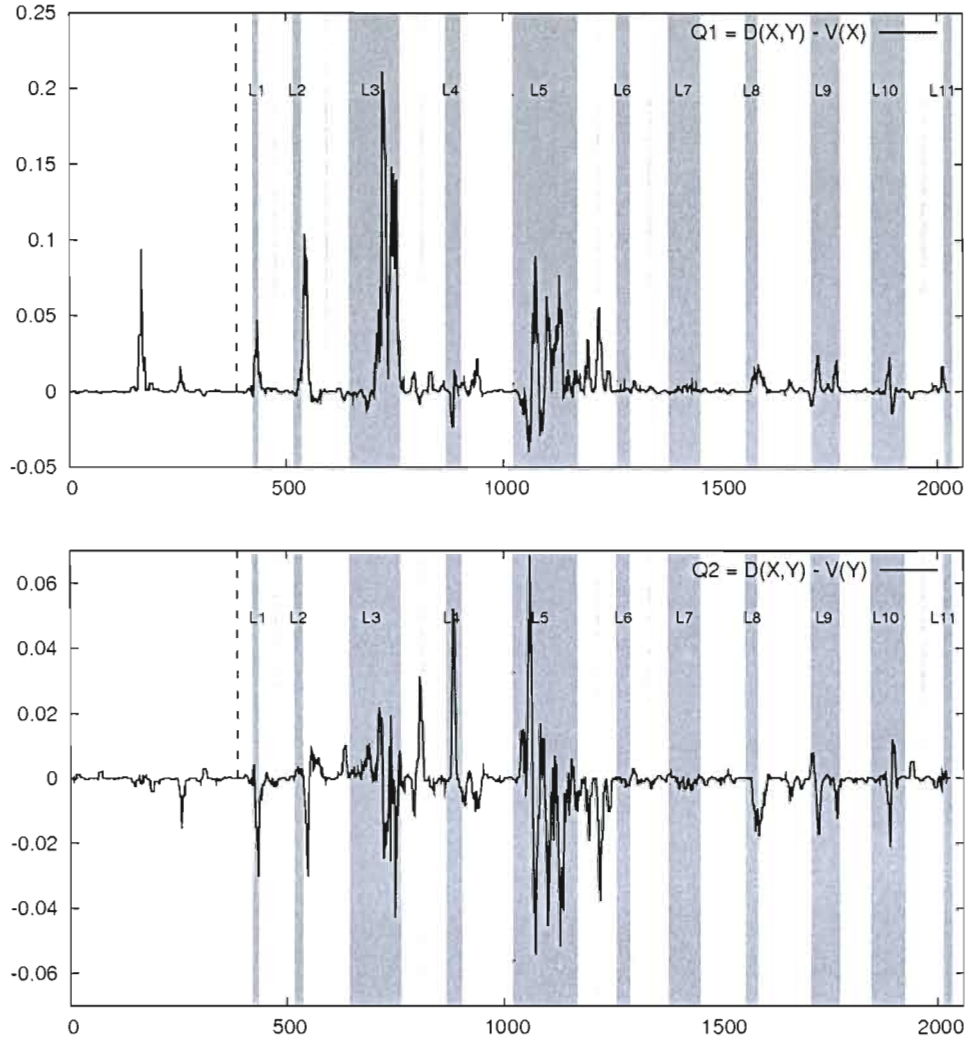

Figure 5.2 The variation of the hit identification functions Q_1 and Q_2 for the *Neisseria Meningitidis* containing invasive sequence tags obtained with a non-overlapping sliding window of size 10 during the gene *FrpB* scan.

The abscissa axis represents the window position. Gray zones are the positions of the surface-exposed loops.

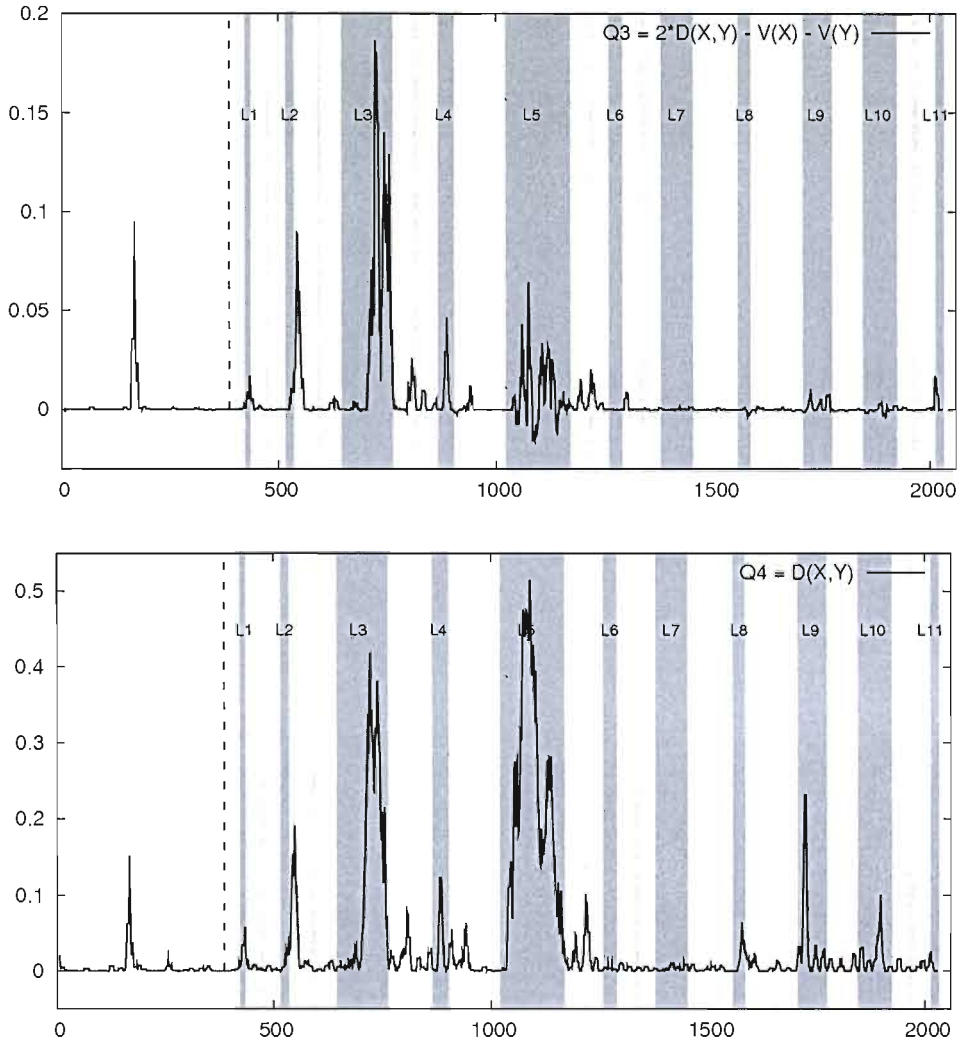

Figure 5.3 The variation of the hit identification functions Q_3 and Q_4 for the *Neisseria Meningitidis* containing invasive sequence tags obtained with a non-overlapping sliding window of size 10 during the gene *FrpB* scan.

The abscissa axis represents the window position. Gray zones are the positions of the surface-exposed loops.

Table 5.1 Normalized maximum values of the functions Q_1, Q_2, Q_3, Q_4 in each gray region.

Higher values for each region are highlighted. Last column shows the number of detected regions scoring over the threshold 2.0.

MAX	1	2	3	4	5	6	7	8	9	10	11	Gray zones detected
Q1	4.51	6.92	21.34	0.98	8.9	0.07	0.15	1.57	2.19	2.06	0.24	6
Q2	-0.57	0.9	4.56	10.56	13.87	0.74	0.34	0.17	1.71	2.5	0.57	4
Q3	2.12	8.56	26.46	6.42	9	0.52	-0.05	-0.12	0.84	0.2	0.75	5
Q4	2.51	5.19	21.56	5.99	26.66	0	0.05	2.87	11.72	4.82	0.05	8

5.2.5 Conclusion

In this paper we considered four different functions for detecting hit regions responsible for disease. We found that the function Q_4 correlate the best with the surface exposed loops (a feature of the secondary structure of OMPs) of the *Neisseria Meningitidis* of the *FrpB* gene. This suggests that our algorithm is able to detect known regions of interest in respect to given epidemiological criteria. Another interesting development would be to design a statistical test allowing one to measure the signicance of the obtained results such as computing p-values. It will be also important to test the four functions considered in this study, in the other context where the information about the species can be grouped into categories according to specific features such as biological functions, phenotypic differences or behavioral changes.

5.2.6 References

1. S. F. Altschul , W. Gish , W. Miller , E. W. Myers, and D. J. Lipman. Basic local alignment search tool. J. Mol. Biol., 215(3):403–410, 1990.

2. D. Badescu, A. B. Diallo, M. Blanchette, and V. Makarenkov. An evolution study of the human papillomavirus genomes. In Proceedings of RECOMB Comparative Genomics 2008, Springer, Lecture Notes in Bioinformatics Series, Paris,128–140

3. M. Blanchette and M. Tompa. FootPrinter: a program designed for phylogenetic footprinting Nucleic Acids Research, 31(13), 3840-3842, 2003.

4. A. B. Diallo, D. Badescu, V. Makarenkov, and M. Blanchette, A whole genome study and identification of specific carcinogenic regions of the Human Papilloma Viruses. Journal of Computational Biology, To appear, 2009.

5. K. Jolley, M.-S. Chan, and M. Maiden. mlstdbnet - distributed multi-locus sequence typing (mlst) databases. BMC Bioinformatics, 5(1):86, 2004.

6. J. Kortekaas, A. Pettersson, J. van der Biezen, V. E. Weynants, P. van der Ley, J. Poolman, M. P. Bos, and J. Tommassen. Shielding of immunogenic domains in Neisseria Men. FrpB by the major variable region. Vaccine, 25(1), 72-84, 2007.

7. http://www.neisseria.org Neisseria Research Community Website, 2009.

8. A. Pettersson, J. T. Poolman, P. van der Ley, and J. Tommassen. Response of Neisseria Men. to iron limitation. Antonie van Leeuwenhoek, 71, 129-136, 1997.

9. A. Pettersson, A. Maas, D. van Wassenaar, P. van der Ley, and J. Tommassen.Molecular characterization of FrpB, the 70-kilodalton iron-regulated outer membrane protein of *Neisseria Meningitidis*. Infect. Immun., 63(10):4181–4184, 1995.

10. A. Siepel, K. S. Pollard and D. Haussler. New methods for detecting lineage-specic selection. Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB 2006), 190-205, 2006.

11. A. Siepel, G. Bejerano, J. S. Pedersen, et al. Evolutionarily cons. elements in vertebrate, insect, worm, and yeast genomes. Genome Res., 15, 1034-1050, 2005.

12. E. A. L. Thompson, I. M. Feavers, and M. C. J. Maiden. Antigenic diversity of meningococcal enterobactin receptor FetA, a vaccine component. Microbiology, 149(7):1849–1858, 2003.

13. R. Urwin, J. E. Russell, E. A. L. Thompson, et al. Distribution of Surface Protein Variants among Hyperinvasive Meningococci: Implications for Vaccine Design. Infect. Immun., 72(10):5955–5962, 2004.

CHAPITRE VI

CONCLUSION ET PERSPECTIVES

Dans ce mémoire, nous avons premièrement présenté une analyse bioinformatique de la relation entre l'apparition des insertion-délétions dans le génome du virus du papillome humain et son degré de carcinogénicité, ainsi que le type histologique de cancer du col de l'utérus. Nous avons trouvé une corrélation statistiquement significative entre ces événements et nous avons proposé une séquence de tests pour orienter l'analyse statistique de cette corrélation (Diallo et al., 2009a). Ces résultats sont issus d'une reconstruction du scénario d'insertions et délétions au niveau génomique du virus du papillome humain, reconstruction qui nous a permis de localiser les insertions-délétions au niveau de l'arbre phylogénétique (Diallo et al., 2007). Nous avons également remarqué que la carcinogenicité est généralement monophylétique, donc pourrait être issue d'un ancêtre commun (Badescu et al., 2008 ; Diallo et al., 2009b). Ces résultats pourraient permettre de mieux placer les nouveaux types de virus du papillome dès leur séquençage dans un groupe à risque spécifique. En analysant leur composition au niveau des évènements élémentaires nous pourrons prédire avec une certaine confiance à quel type de cancer les nouveaux types de virus sont plus enclins.

Ensuite nous avons développé un algorithme de type fenêtre coulissante pour retrouver les régions spécifiques qui sont reliées à la carcinogenicité du virus du papillome (Badescu et al., 2008; Diallo et al., 2009b). L'algorithme permet de discriminer le seuil à partir duquel les données peuvent être considérées comme statistiquement significatives à l'aide du calcul des p-values (Diallo et al., 2009b).

Si nous le comparons aux algorithmes décrits au chapitre II, *Détection des Séquences Fonctionnelles*, il teste aussi bien la sélection négative - conservation - ou positive - diversité. Comme cet algorithme requiert un alignement multiple des séquences en entrée, il peut aussi bien fonctionner au niveau génomique, sans besoin d'annotations de gènes. Mais comme il peut y avoir des inversions ou autres événements à grande échelle, les annotations permettent de mieux établir l'homologie, et donc il est préférable de les prendre en compte quand elles sont disponibles. Toutefois, l'algorithme n'est pas limité aux séquences codantes seulement.

Par rapport à la détection de la sélection lignée-spécifique, nous pouvons le considérer comme une forme particulière où les lignées doivent être divisées en deux classes *à priori*. Plus les classes sont équilibrées en terme de nombre de membres, plus les résultats seront d'une meilleure qualité.

La division des membres en deux classes, nous a permis d'étendre l'investigation sur un autre critère épidémiologique, l'invasivité d'une bactérie, *Neisseria Meningitidis* (Badescu et al., 2010).

La méthode reste limitée par la disponibilité des séquences pour diverses souches, allèles de microorganismes et celle des informations épidémiologiques sur les infections, les épidémies. Plus la maladie est importante, de meilleures informations épidémiologiques sont disponibles.

Quand le microorganisme en question a un comportement ambivalent, avec de nombreuses souches virulentes ou très peu virulentes, la répartition dans les deux classes (invasive, non-invasive) se fait d'une manière équilibrée et les résultats obtenus sont de meilleure qualité.

Finalement, nous nous sommes attaqués aux fonctions de discrimination, tout en essayant de faire la différence entre les régions responsables de maximum d'invasivité et celles qui ont un rôle structural dans ce processus (Badescu et al., 2010). Ainsi on a trouvé que la fonction Q_4 est optimale à detecter les structures moléculaires connues (i.e. les anses extra cellulaires, dans notre cas).

Dans le futur, nous aimerions étendre notre méthode en l'associant à une autre, indépendante pour détecter les structures moléculaires et retrouver les régions responsables d'invasivité cachés à l'intérieur de régions plus profondes. Cette approche aidera la conception des vaccins, car les recherches en laboratoire de biologie moléculaire sont toujours guidées par des informations *à priori*.

D'autres critères de groupement des séquences comme fonctions biologiques, différences phénotypiques, changement comportemental, pourraient être envisagé, ainsi que l'expansion du nombre des groupements.

La version actuelle du programme est basée sur l'analyse de nucléotides. Elle a été capable de détecter le domaine PDZ, une région du gène E6, qui est une *condition sine qua non* de la carcinogénicité du produit de ce gène (Diallo et al., 2009b). Il nous reste à développer une version acide aminée du programme et des fonctions de distance correspondantes et comparer les résultats des deux implémentations. Certains bénéfices sont envisagés vu la meilleure qualité des alignements protéiques et la capacité de prendre en compte les codons.

Nous proposons pour la suite du projet d'intégrer notre méthode avec l'inférence du scénario d'insertions et de délétions au niveau de l'arbre phylogénétique dans le but de mieux localiser les branches et les événements individuels – indels et substitutions – responsables de carcinogénicité.

À l'avenir, il serait également intéressant de développer un procédé similaire basé sur une approche bayésienne ou maximum de vraisemblance.

On se propose également de tester l'hypothèse de cycle de vie des fonctionnalités avec l'apparition par insertions, la stabilisation par substitution et disparition par délétion.

La version actuelle du programme prend en entrée des informations de structure primaire de l'ADN. Nous espérons aussi pouvoir étendre notre algorithme à l'analyse des structures protéiques en trois dimensions.

Le travail a été décrit dans les trois publications incluses dans le présent mémoire qui sont les suivantes :

1) Classification des Virus du Papillome Humain (Diallo et al., 2009a).

2) Une étude du génome entier et l'identification de régions spécifiques carcinogènes du Virus du Papillome Humain (Diallo et al., 2009b)

3) Identification des régions génomiques spécifiques responsables de l'invasivité du *Neisseria Meningitidis* (Badescu et al., 2010)

Il y a aussi des aspects pratiques qui pourraient être exploités, comme la production de sondes d'hybridation et la conception d'un test de détection des risques de carcinogénicité sur du matériel récolté chez des patients pas encore diagnostiqués de cancer. Ceci ne peut être fait qu'en collaboration avec des entreprises canadiennes oeuvrant dans le domaine de la biologie moléculaire et pharmacie.

APPENDICE A

CODE SOURCE POUR LE CHAPITRE III

Nous présentons dans cette annexe le code source utilisé pour l'analyse des évènements élémentaires sur les différentes lignées.

Ce code a été implanté pour une base de données *Oracle* en *SQL* et *PL/SQL*. L'analyse syntaxique des fichiers de sortie du programme *Ancestors*, programme qui calcule les séquences ancestrales a été faite en *Java* et *Ruby*.

A.1 Chargement des alignements multiples

La classe *ParseAlignMult* insère les alignements multiples en format *Fasta* dans la base de données.

```
package indels2netbeans;

import java.io.BufferedReader;
import java.io.FileNotFoundException;
import java.io.FileReader;
import java.io.IOException;
import java.sql.Connection;
import java.sql.PreparedStatement;
import java.sql.SQLException;
import java.sql.Statement;
import oracle.jdbc.pool.OracleDataSource;
```

```
/**
 *
 * Analyse syntaxique et insertion dans une BD
 */
public class ParseAlignMult {

    private Connection conn;
    private PreparedStatement pstmt;

    /** Creates a new instance of ParseAlignMult */
    public ParseAlignMult() throws ClassNotFoundException,
InstantiationException, IllegalAccessException, SQLException {
        connecteOracle();
    }

    /*
     * Connexion à Oracle
     */
    public void connecteOracle() throws SQLException {
        OracleDataSource ods = new OracleDataSource();
        String url =
"jdbc:oracle:thin:@//arnt.bioinfo.uqam.ca:1521/xe";
        ods.setURL(url);
        ods.setUser("papil");
        ods.setPassword("papil");
        conn = ods.getConnection();

        Statement stmt = conn.createStatement();
        stmt.execute("DELETE FROM ALIGN_MULT");
        conn.commit();
    }

    /*
     *  Desallocation de ressources
     */
    public void deconnecteOracle() throws SQLException {

        conn.close();
    }
    /*
     * Insère une ligne d'un alignement multiple
     */
    public void insertAlignMult(String noeud, String seq_gap)
    throws FileNotFoundException, IOException, SQLException {
        //pour la base de données
        pstmt = conn.prepareStatement("insert into ALIGN_MULT " +
                "(NOEUD,SEQ_GAP) " +
                "values " +
                "(?,?)");

        //on fait uppercase
        pstmt.setString(1,noeud.toUpperCase());
        pstmt.setString(2,seq_gap);
        pstmt.executeUpdate();
        conn.commit();


    }
```

```
        /*
         * Analyseur syntaxique simple d'un fichier en format FASTA
         */
        public void isoleEnregFasta() throws FileNotFoundException,
IOException, SQLException, ClassNotFoundException, InstantiationException,
IllegalAccessException {
            //on va charger les lignes en ce buffer pour en faire un String
            StringBuffer sb = new StringBuffer();
            BufferedReader br = new BufferedReader(new
FileReader("files//clustal.fasta"));

            String line="";

            while ((line = br.readLine()) != null) {

                sb.append(line);
                //si la ligne contient le descripteur fasta on l'isole
                if (line.indexOf('>')!=-1 && sb.length()!=0) {
                    sb.append('\n');
                }

            }

            String fichier =sb.toString();

          //on isole la ligne de description des sequences
            String[] words = fichier.split(">");  // one or spaces.

            String word="";
            for (int i=0; i<words.length; i++) {
                word=words[i];

                if (word.length()!=0) {
                    //on envoie l'enregistrement
                    isoleDescriptSeq(word);
                }

            }

            deconnecteOracle();

        }

        /*
         * Isole la description de la séquence
         */
        public void isoleDescriptSeq(String entree) throws
FileNotFoundException, IOException, SQLException, ClassNotFoundException,
InstantiationException, IllegalAccessException {

            String ant="";

            String[] words = entree.split("\n");  //

            String word="";
            for (int i=0; i<words.length; i++) {
```

```
                word=words[i];
                if (i%2==0){
                    ant=word;
                } else {
                    //on envoie les séquences pour les mots
                    //dans la base de données
                    insertAlignMult(ant,word);
                    //on affiche pour le contrôle
                    System.out.println("ant:"+ant);
                    System.out.println("seq:"+word);
                }


            }

        }

    }
```

A.2 Analyse syntaxique des fichiers de sortie d'*Ancestors*

La classe *ParseModif* se charge de l'analyse syntaxique des fichiers de sortie d'*Ancestors*. Elle insère dans une base de données normalisée les modifications élémentaires et les relations parent – enfant des nœuds de l'arbre phylogénétique.

```
package indels2netbeans;

import java.io.BufferedReader;
import java.io.FileNotFoundException;
import java.io.FileReader;
import java.io.IOException;
import java.io.StreamTokenizer;
import java.sql.Connection;
import java.sql.PreparedStatement;
import java.sql.SQLException;
import java.sql.Statement;
import java.util.Vector;
import oracle.jdbc.pool.OracleDataSource;

/**
 *
 * Analyse syntaxique des états ancestraux
 */
public class ParseModif {

    private Connection conn;
    private PreparedStatement pstmt;
    int dbg = 0;

    /** Creates a new instance of ParseModif */
    public ParseModif() throws ClassNotFoundException,
InstantiationException, IllegalAccessException, SQLException {
```

```
            connecteOracle();


        }

        public void connecteOracle() throws SQLException {
            OracleDataSource ods = new OracleDataSource();
            String url =
"jdbc:oracle:thin:@//arnt.bioinfo.uqam.ca:1521/xe";
            ods.setURL(url);
            ods.setUser("papil");
            ods.setPassword("papil");
            conn = ods.getConnection();

            Statement stmt = conn.createStatement();
            stmt.execute("DELETE FROM ancetres_noeuds");
            conn.commit();

            stmt.execute("DELETE FROM modif_elem");
            conn.commit();

        }

        public void deconnecteOracle() throws SQLException {

            conn.close();
        }

        /*
         * Isole les enregistrements
         */
        public void isoleEnregFasta() throws FileNotFoundException,
IOException, SQLException {
            //on va charger les lignes en ce buffer pour en faire un String
            StringBuffer sb = new StringBuffer();
            BufferedReader br = new BufferedReader(new
FileReader("files//clustal.edgeState"));

            String line = "";

            while ((line = br.readLine()) != null) {

                sb.append(line);
                //si la ligne contient le descripteur fasta on l'isole
                if (line.indexOf('>') != -1 && sb.length() != 0) {
                    sb.append('\n');
                }
            }

            String fichier = sb.toString();
            //on isole la ligne de description des séquences
            String[] words = fichier.split(">");  // one or spaces.

            String word = "";
            for (int i = 0; i < words.length; i++) {
                word = words[i];
                if (word.length() != 0) {
```

```
                    isoleDescriptSeq(word);
                }
            }
          }

        /*
         * Isole les descriptions (entête)
         */
        public void isoleDescriptSeq(String entree) throws SQLException {

            String ant = "";

            String[] words = entree.split("\n");  //

            String word = "";
            for (int i = 0; i < words.length; i++) {
                word = words[i];
                if (i % 2 == 0) {
                    ant = word;
                } else {
                   //on envoie les séquences pour les mots
                    isoleMots(ant, word);
                    //isole les ancetres et les insere dans la BD
                    isoleAncetres(ant);
                }
            }
        }

        /*
         * Isole les mots (séquences de lettres identiques
         */
        public void isoleMots(String ant, String entree) throws
SQLException {
            String etat = "";
            String car = "";
            int index = 0;
            int noChar = 0;

            for (int i = 0; i < entree.length(); i++) {

                car = new Character(entree.charAt(i)).toString();

                if (!car.equals(etat)) {
                    //A chaque changement
                    //Affiche les etats finis et corects
                    if (etat != "") {
                        System.out.println(ant + "\t" + etat + "\t" + index
+ "\t" + noChar + "\t" + (index + noChar - 1));
                        insereModifElem(ant, etat, index, (index +
noChar));
                    }
                    etat = car;
                    //index normal pas java
                    index = i + 1;
                    //parce qu-on compte aussi le changement
                    noChar = 1;
                } else {
                    noChar++;
```

```
                }
            }
            //et on flush
            //Affiche les états finis et corrects
            if (!etat.equals("")) {
                System.out.println(ant + "\t" + etat + "\t" + index + "\t"
+ noChar + "\t" + (index + noChar));
                insereModifElem(ant, etat, index, (index + noChar));
            }
        }

        /*
         * Identifie les relations ancêtre - descendants
         * et les insère dans une table de corrélation
         * dans la BD
         */
        public void isoleAncetres(String anc) throws SQLException {
            pstmt = conn.prepareStatement("insert into ancetres_noeuds " +
                    "(noeud_anc,noeud_feu) " +
                    "values " +
                    "(?,?)");

            if (anc.indexOf("+") != -1) {
                String[] words = anc.split("\\+");  //
                for (int i = 0; i < words.length; i++) {
                    pstmt.setString(1, anc);
                    pstmt.setString(2, words[i]);
                    pstmt.executeUpdate();
                    System.out.println(words[i]);
                }
            } else {
                pstmt.setString(1, anc);
                pstmt.setString(2, anc);
                pstmt.executeUpdate();
            }
            pstmt.close();
            conn.commit();
        }

        /*
         * Insertion d'une modification élémentaire
         * dans une table normalisée
         *
         * Une ligne par modification
         */
        public void insereModifElem(String noeudAnc,
                String typeModif,
                Integer idxAlignDeb,
                Integer idxAlignFin) throws SQLException {

            pstmt = conn.prepareStatement("insert into modif_elem " +
                    "(noeud_anc,type_modif,IDX_ALIGN_DEB,IDX_ALIGN_FIN) " +
                    "values " +
                    "(?,?,?,?)");

            pstmt.setString(1, noeudAnc);
            pstmt.setString(2, typeModif);
            pstmt.setInt(3, idxAlignDeb);
```

```
            pstmt.setInt(4, idxAlignFin);
            pstmt.executeUpdate();
            pstmt.close();


        }
    }
```

A.3 Exemple d'utilisation des classes *ParseAlignMult* et *ParseModif*

```
    //efface l'alignement multiple
      ParseAlignMult pam = new ParseAlignMult();
    //parse et insère l'alignement multiple
      pam.isoleEnregFasta();
      pam.deconnecteOracle();

      ParseModif pm=new ParseModif();
      pm.isoleEnregFasta();
      pm.deconnecteOracle();
```

A.4 Extraction des limites des gènes

Exemple d'utilisation suivi de la classe GeneLimits :

```
require 'rubygems'
require 'gene_limits'

gl = GeneLimits::GeneLimits.new
gl.geneLimitsDatabase

module GeneLimits

  require 'bio'
  require 'bio/io/flatfile'
  require 'bio/db'
  require 'oci8'

  class GeneLimits

    def geneLimitsDatabase

      conn = OCI8.new('papil', 'papil','arnt')

      cursor = conn.parse("insert into
gene_limits_2(type,gene,index_deb,index_fin)
                        values (:type,:gene,:index_deb,:index_fin)")

      ff = Bio::FlatFile.open(Bio::GenBank,
'/home/dunarel/NetBeansProjects/alignement/sequences/sequences.gbwithparts')
      ff.each_entry do |x|
        x.each_cds() do |feature|
            gene = ''
```

```
                product = ''
                note = ''
                debut = ''
                fin = ''

                feature.each do |qualifier|
                  if qualifier.qualifier == 'product'
                    product= qualifier.value
                  end

                  if qualifier.qualifier == 'gene'
                    gene= qualifier.value
                  end

                  if qualifier.qualifier == 'note'
                    note= qualifier.value
                  end
                  end

                matched = /[ELXYZ]\d?\w?/.match(gene).to_s
                if matched == ''
                  matched = /[ELXYZ]\d?\w?/.match(product).to_s
                end

                if matched == ''
                  matched = /[ELXYZ]\d?\w?/.match(note).to_s
                end

                matched.capitalize()
                position= case feature.position
                when /join\((\d+).*\.\.(\d+)\)/:  $1 +'..' + $2
                else feature.position
                end
                debut = position.split("..")[0]
                debut = debut.gsub('<','')

                fin = position.split("..")[1]
                fin = fin.gsub('>','')
                cursor.bind_param(':type', x.accession) # bind by name
                cursor.bind_param(':gene', matched) # bind by name
                cursor.bind_param(':index_deb', debut) # bind by name
                cursor.bind_param(':index_fin', fin) # bind by name

                cursor.exec()
        puts x.gi + "\t" + x.accession + "\t" + debut  + "\t" + fin + "\t" +
matched

      end

            cursor.close
            conn.commit
            conn.logoff
          end

        end #end class
      end #end module
```

A.5 Somme et moyenne des évènements – insertion, délétion, conservation, absence d'événements – sur les différentes lignées.

A.5.1 Verification de la validité des données.

Seuls les nœuds qui correspondent à une souche et gène valide sont pris en compte.

```
      CREATE OR REPLACE FORCE VIEW "PAPIL"."LIGNEE_ESPECE_GENE_EXIST"
    ("NOEUD", "GENE", "HPV_TYPE", "ACCESSION", "SQUAM", "ADENO", "SUM_C",
"SUM_I", "SUM_D", "SUM_N", "AVG_T_C", "AVG_T_I", "AVG_T_D", "AVG_T_N") AS
      select
"NOEUD","GENE","HPV_TYPE","ACCESSION","SQUAM","ADENO","SUM_C","SUM_I","SUM_D
","SUM_N","AVG_T_C","AVG_T_I","AVG_T_D","AVG_T_N"
    from lignee_espece_gene
    where (noeud,gene) in (select type,gene
                           from gene_limits gl)
    order by noeud,gene;
```

A.5.2 Vue subséquente qui calcule les statistiques sur les lignées.

```
    CREATE OR REPLACE FORCE VIEW "PAPIL"."LIGNEE_ESPECE_GENE" ("NOEUD",
"GENE", "HPV_TYPE", "ACCESSION", "SQUAM", "ADENO", "SUM_C", "SUM_I",
"SUM_D", "SUM_N", "AVG_T_C", "AVG_T_I", "AVG_T_D", "AVG_T_N") AS
      select lgm.noeud_feu as noeud,
             lgm.gene,
             aat.hpv_type,
             aat.accession,
             aat.squam,
             aat.adeno,
           --lgm.gene,
           sum(lgm.sum_cons) as sum_c,
           sum(lgm.sum_ins) as sum_i,
           sum(lgm.sum_del) as sum_d,
           sum(lgm.sum_na) as sum_n,
           avg(lgm.sum_t_cons) as avg_t_c,
           avg(lgm.sum_t_ins) as avg_t_i,
           avg(lgm.sum_t_del) as avg_t_d,
           avg(lgm.sum_t_na) as avg_t_n
    from lignees_genes_modif lgm
     join acces_asoc_tab aat on trim(aat.noeud)=trim(lgm.noeud_feu)
    group by lgm.noeud_feu,
             lgm.gene,
             aat.hpv_type,
             aat.accession,
             aat.squam,
             aat.adeno
     ;
```

A.6 Cueillette de tous les événements au long des branches de l'arbre pour une lignée.

Cette vue distribue les événements aux descendants d'un nœud interne.

```
       CREATE OR REPLACE FORCE VIEW "PAPIL"."LIGNEES_GENES_MODIF"
("NOEUD_FEU", "GENE", "SUM_CONS", "SUM_INS", "SUM_DEL", "SUM_NA",
"SUM_T_CONS", "SUM_T_INS", "SUM_T_DEL", "SUM_T_NA") AS
       select an.noeud_feu,
            bmst.gene,
            bmst.sum_cons,
            bmst.sum_ins,
            bmst.sum_del,
            bmst.sum_na,
            bmst.sum_t_cons,
            bmst.sum_t_ins,
            bmst.sum_t_del,
            bmst.sum_t_na
     from branches_modif_stats_tab bmst
      join ancetres_noeuds an on an.noeud_anc=bmst.noeud_anc
      ;
```

A.7 Calcul des statistiques – somme des événements – pour une branche.

Pour des raisons d'efficacité cette table à été crée pour contenir les données de la vue suivante. Elle fonctionne comme une vue matérialisée.

```
CREATE TABLE "PAPIL"."BRANCHES_MODIF_STATS_TAB"
   (    "NOEUD_ANC" VARCHAR2(1000 BYTE) NOT NULL ENABLE,
  "GENE" VARCHAR2(20 BYTE),
  "IDX_ALIGN_DEB_MIN" NUMBER,
  "IDX_ALIGN_FIN_MAX" NUMBER,
  "SUM_CONS" NUMBER,
  "SUM_INS" NUMBER,
  "SUM_DEL" NUMBER,
  "SUM_NA" NUMBER,
  "SUM_T_CONS" NUMBER,
  "SUM_T_INS" NUMBER,
  "SUM_T_DEL" NUMBER,
  "SUM_T_NA" NUMBER
   );
```

Vue qui calcule les statistiques – somme des événements – sur une branche :

```
      CREATE OR REPLACE FORCE VIEW "PAPIL"."BRANCHES_MODIF_STATS"
("NOEUD_ANC", "GENE", "IDX_ALIGN_DEB_MIN", "IDX_ALIGN_FIN_MAX", "SUM_CONS",
"SUM_INS", "SUM_DEL", "SUM_NA", "SUM_T_CONS", "SUM_T_INS", "SUM_T_DEL",
"SUM_T_NA") AS
      select BRANCHES_MODIF_ELEM_CATEG.noeud_anc,
            gene,
            IDX_ALIGN_DEB_MIN,
           IDX_ALIGN_FIN_MAX,
            sum(cons) as sum_cons,
            sum(ins) as sum_ins,
            sum(del) as sum_del,
            sum(na) as sum_na,
            sum(t_cons) as sum_t_cons,
            sum(t_ins) as sum_t_ins,
            sum(t_del) as sum_t_del,
            sum(t_na)  as sum_t_na
    from BRANCHES_MODIF_ELEM_CATEG
    group by noeud_anc,
                    gene,
                    IDX_ALIGN_DEB_MIN,
                    IDX_ALIGN_FIN_MAX
     ;
```

A.8 Denormalisation des données – événements élémentaires.

Pour compter les événements on procède à une dénormalistation par type de modification, et à un pivot de la vue des événements élémentaires.

```
     CREATE OR REPLACE FORCE VIEW "PAPIL"."BRANCHES_MODIF_ELEM_CATEG"
("NOEUD_ANC", "GENE", "IDX_ALIGN_DEB_MIN", "IDX_ALIGN_FIN_MAX", "CONS",
"INS", "DEL", "NA", "T_CONS", "T_INS", "T_DEL", "T_NA") AS
      select BRANCHES_MODIF_ELEM.NOEUD_ANC,
            BRANCHES_MODIF_ELEM.GENE,
            BRANCHES_MODIF_ELEM.IDX_ALIGN_DEB_MIN,
            BRANCHES_MODIF_ELEM.IDX_ALIGN_FIN_MAX,
            case
                when BRANCHES_MODIF_ELEM.TYPE_MODIF='C' then 1
                else 0
            end as cons,
            case
                when BRANCHES_MODIF_ELEM.TYPE_MODIF='I' then 1
                else 0
            end as ins,
            case
                when BRANCHES_MODIF_ELEM.TYPE_MODIF='D' then 1
                else 0
            end as del,
            case
                when BRANCHES_MODIF_ELEM.TYPE_MODIF='B' then 1
                else 0
```

```
            end as na,
            case
                when BRANCHES_MODIF_ELEM.TYPE_MODIF='C' then modif_gene_dim
                else 0
            end as t_cons,
            case
                when BRANCHES_MODIF_ELEM.TYPE_MODIF='I' then modif_gene_dim
                else 0
            end as t_ins,
            case
                when BRANCHES_MODIF_ELEM.TYPE_MODIF='D' then modif_gene_dim
                else 0
            end as t_del,
            case
                when BRANCHES_MODIF_ELEM.TYPE_MODIF='B' then modif_gene_dim
                else 0
            end as t_na
     from BRANCHES_MODIF_ELEM
      ;
```

A.9 Prise en compte des événements sur une branche entre les limites des gènes correspondants dans le génome.

```
       CREATE OR REPLACE FORCE VIEW "PAPIL"."BRANCHES_MODIF_ELEM"
("NOEUD_ANC", "GENE", "IDX_ALIGN_DEB_MIN", "IDX_ALIGN_FIN_MAX",
"IDX_DEB_MIN", "IDX_FIN_MAX", "GENE_DIM", "TYPE_MODIF", "IDX_ALIGN_DEB",
"IDX_ALIGN_FIN", "MODIF_DIM", "MODIF_GENE_DIM") AS
       select modif_elem.noeud_anc,
              ANC_GENE_LIMITS.GENE,
              ANC_GENE_LIMITS.IDX_ALIGN_DEB_MIN,
              ANC_GENE_LIMITS.IDX_ALIGN_FIN_MAX,
              anc_gene_limits.idx_deb_min,
              anc_gene_limits.idx_fin_max,
              (anc_gene_limits.idx_fin_max-anc_gene_limits.idx_deb_min+1) as
gene_dim,
              modif_elem.type_modif,
              modif_elem.idx_align_deb,
              modif_elem.idx_align_fin,
              (modif_elem.idx_align_fin-modif_elem.idx_align_deb+1) as
modif_dim,
              (modif_elem.idx_align_fin-
modif_elem.idx_align_deb+1)/(anc_gene_limits.idx_fin_max-
anc_gene_limits.idx_deb_min+1) as modif_gene_dim
      from modif_elem
       join anc_gene_limits on anc_gene_limits.noeud_anc=modif_elem.noeud_anc
and
            modif_elem.idx_align_deb>anc_gene_limits.idx_align_deb_min and
            modif_elem.IDX_ALIGN_FIN<ANC_GENE_LIMITS.IDX_ALIGN_FIN_MAX

      ;
```

A.10 Les limites des gènes dans les ancêtres.

La limite dans un ancêtre est considérée le plus court intervalle qui contient tous les caractères des descendants actuels.

```
       CREATE OR REPLACE FORCE VIEW "PAPIL"."ANC_GENE_LIMITS" ("NOEUD_ANC",
"GENE", "IDX_DEB_MIN", "IDX_FIN_MAX", "IDX_ALIGN_DEB_MIN",
"IDX_ALIGN_FIN_MAX") AS
       select ancetres_noeuds.noeud_anc,
            gene_limits.gene,
            min(gene_limits.index_deb) as idx_deb_min,
            max(gene_limits.index_fin) as idx_fin_max,
            min(gene_limits.index_align_deb) as idx_align_deb_min,
            max(gene_limits.index_align_fin) as idx_align_fin_max
     from ancetres_noeuds
      join gene_limits on gene_limits.type=ancetres_noeuds.noeud_feu
     group by ancetres_noeuds.noeud_anc,
              gene_limits.gene

      ;
```

A.11 Limites des gènes dans les alignements multiples et dans les séquences non alignées.

```
       CREATE TABLE "PAPIL"."GENE_LIMITS"
        (    "TYPE" VARCHAR2(20 BYTE) NOT NULL ENABLE,
       "GENE" VARCHAR2(20 BYTE) NOT NULL ENABLE,
       "INDEX_DEB" NUMBER,
       "INDEX_FIN" NUMBER,
       "INDEX_ALIGN_DEB" NUMBER,
       "INDEX_ALIGN_FIN" NUMBER
        );
```

A.12 Tous les descendants de chaque nœud interne.

```
       CREATE TABLE "PAPIL"."ANCETRES_NOEUDS"
        (    "NOEUD_ANC" VARCHAR2(1000 BYTE) NOT NULL ENABLE,
       "NOEUD_FEU" VARCHAR2(100 BYTE) NOT NULL ENABLE,
        CONSTRAINT "ANCETRES_NOEUDS_PK" PRIMARY KEY ("NOEUD_ANC",
"NOEUD_FEU")
       USING INDEX PCTFREE 10 INITRANS 2 MAXTRANS 255 COMPUTE STATISTICS
       STORAGE(INITIAL 65536 NEXT 1048576 MINEXTENTS 1 MAXEXTENTS 2147483645
       PCTINCREASE 0 FREELISTS 1 FREELIST GROUPS 1 BUFFER_POOL DEFAULT)
       TABLESPACE "SYSTEM"  ENABLE
        ) PCTFREE 10 PCTUSED 40 INITRANS 1 MAXTRANS 255 NOCOMPRESS LOGGING
       STORAGE(INITIAL 65536 NEXT 1048576 MINEXTENTS 1 MAXEXTENTS 2147483645
       PCTINCREASE 0 FREELISTS 1 FREELIST GROUPS 1 BUFFER_POOL DEFAULT)
       TABLESPACE "SYSTEM" ;
```

A.13 Gènes, les limites et annotations extraites – données brutes.

```
CREATE TABLE "PAPIL"."GENES_SPECIES"
 (    "ID" NUMBER,
"SPECIES" VARCHAR2(30 BYTE),
"ACCESSION" VARCHAR2(30 BYTE),
"GENE" VARCHAR2(30 BYTE),
"IDX_DEB" VARCHAR2(30 BYTE),
"IDX_FIN" VARCHAR2(30 BYTE),
"IDX_DEB1" VARCHAR2(30 BYTE),
"IDX_FIN1" VARCHAR2(30 BYTE),
"IDENT" VARCHAR2(30 BYTE),
"OBS" VARCHAR2(255 BYTE),
"MODIF" VARCHAR2(255 BYTE),
"TRANSLATION" VARCHAR2(4000 BYTE),
 CONSTRAINT "GENES_SPECIES_PK" PRIMARY KEY ("ID")
 USING INDEX PCTFREE 10 INITRANS 2 MAXTRANS 255 COMPUTE STATISTICS
 STORAGE(INITIAL 65536 NEXT 1048576 MINEXTENTS 1 MAXEXTENTS 2147483645
 PCTINCREASE 0 FREELISTS 1 FREELIST GROUPS 1 BUFFER_POOL DEFAULT)
 TABLESPACE "SYSTEM"  ENABLE
  );


 CREATE OR REPLACE TRIGGER "PAPIL"."bi_GENES_SPECIES"
 before insert on "GENES_SPECIES"
 for each row
begin
 for c1 in (
   select "GENES_SPECIES_SEQ".nextval next_val
   from dual
 ) loop
   :new."ID" :=  c1.next_val;
 end loop;
end;



/
ALTER TRIGGER "PAPIL"."bi_GENES_SPECIES" ENABLE;
```

A.14 Données épidémiologiques sur le degré de carcinogenicité des souches pour les types de cancer SQUAM et ADENO.

```
CREATE TABLE "PAPIL"."HPV_CANCERO"
 (    "SQUAM" NUMBER,
"ADENO" NUMBER,
"HPV_TYPE" VARCHAR2(50 BYTE)
 ) ;
```

A.15 Associations entre le numéro d'accession – type de virus

Les articles scientifiques descriptifs ainsi que les données épidémiologiques utilisent les types génériques de virus (ex : HPV-18).

Les alignements multiples et les annotations se retrouvent pour des séquences spécifiques, indexées dans les bases de données publiques par une clé primaire.

Cette table fait les associations necessaires.

```
CREATE TABLE "PAPIL"."HPV_TYPES"
  (    "ACCESSION" VARCHAR2(20 BYTE),
"HPV_TYPE" VARCHAR2(50 BYTE)
  );
```

A.16 Carte des régions intergéniques, basée sur la vue subséquente.

```
CREATE TABLE "PAPIL"."IGEN_LIMITS_TAB"
  (    "TYPE" VARCHAR2(20 BYTE),
"GENE" VARCHAR2(20 BYTE),
"INDEX_DEB" NUMBER,
"INDEX_FIN" NUMBER
  );

CREATE OR REPLACE FORCE VIEW "PAPIL"."IGEN_LIMITS_V" ("TYPE", "REG",
"NB", "INDEX_DEB", "INDEX_FIN") AS
  select type,reg,nb,index_deb,index_fin
from table(indel_operations.igen_limits(cursor(select distinct type
from gene_limits)))
order by type,nb,reg,index_deb,index_fin

;
```

A.17 Code Pl/SQL pour la détection des régions intergéniques.

```
create or replace PACKAGE "INDEL_OPERATIONS" as

  TYPE igen_limits_rec is record(
  type varchar(20),
  reg varchar(20),
  nb number,
  index_deb number,
  index_fin number);

  type gene_type_rec is record(
  type varchar(20));
```

```
  TYPE igen_limits_tab IS TABLE OF igen_limits_rec;

  type gene_type is ref cursor return gene_type_rec;

  TYPE strong_refcur_t IS REF CURSOR RETURN igen_limits_rec;
  TYPE refcur_t IS REF CURSOR;


 genome_indexes clob;

  function igen_limits(p_gene_type gene_type)
 return igen_limits_tab
 pipelined;


end;

create or replace PACKAGE BODY "INDEL_OPERATIONS" as

 --working buffer
 buff varchar2(4000);

 --input variables
 v_in gene_type_rec;
 --intergenic extraction variables
  etat varchar(1);
  car varchar(1);
  no_char integer;
  idx integer;

  gene_order integer;

  --pipeline
  v_row igen_limits_rec;




 function igen_limits(p_gene_type gene_type)
 return igen_limits_tab
 pipelined as
  align_mult_long integer;
    cursor c1(gene_type varchar) is
    select gene,index_deb,index_fin
    from gene_limits
    where type=gene_type;

  begin


   LOOP

-- Fetch from cursor variable
  FETCH p_gene_type INTO v_in;
  EXIT WHEN p_gene_type%NOTFOUND; -- exit when last row is fetched

   --initialisation
```

```
        etat:=' ';
        car:=' ';
        no_char:=0;
        idx:=0;
        gene_order:=0;
        dbms_lob.open(genome_indexes, DBMS_LOB.LOB_READWRITE);
        dbms_lob.trim(genome_indexes,0);

        --longueur du génome
        select max(length(seq_gap))
        into align_mult_long
        from align_mult;

        dbms_output.PUT_LINE(to_char(align_mult_long));



        for i in 1..align_mult_long loop
        buff:='_';

          dbms_lob.writeAppend(genome_indexes, length(buff), buff);

        end loop;

        for item in c1(v_in.type)
        loop
         buff:=lpad('*', (item.index_fin-item.index_deb+1), '*');
         dbms_output.put_line(to_char(item.index_deb)|| ',' ||
to_char(item.index_fin) || ',' || to_char(length(buff)));
         dbms_lob.write(genome_indexes,length(buff),item.index_deb,buff);

        end loop;



        --extraction des régions intergéniques


        for i in 1..length(genome_indexes)
        loop
          car:= dbms_lob.substr(genome_indexes,1,i);
           if car<>etat then
            --A chaque changement
            --Affiche les états finis et corrects
               if etat<>' ' then
              --output the row
              gene_order:=gene_order+1;
              v_row.type:=v_in.type;
              v_row.reg:=case when etat = '_' then 'I'
                              when etat = '*' then 'G'
                              end ;
              v_row.nb:=gene_order;
              v_row.index_deb:=idx;
              v_row.index_fin:=idx+no_char-1;
              pipe row(v_row);
```

```
            --dbms_output.put_line(etat || ',' || idx || ',' || no_char ||
',' || (idx+no_char-1));
            end if;


            etat := car;
               --index normal pas java
               idx:=i;
               --parce qu-on compte aussi le changement
                   no_char:=1;
          else
                   no_char:=no_char+1;
          end if;

          end loop;

            --et on flush
            --Affiche les états finis et corrects
             if etat<>' ' then
            --output the row
            gene_order:=gene_order+1;
            v_row.type:=v_in.type;
            v_row.reg:=case when etat = '_' then 'I'
                            when etat = '*' then 'G'
                            end ;
            v_row.nb:=gene_order;
            v_row.index_deb:=idx;
            v_row.index_fin:=idx+no_char-1;
            pipe row(v_row);

                 dbms_output.put_line(etat || ',' || idx || ',' || no_char
|| ',' || (idx+no_char-1));
            end if;


        dbms_lob.close(genome_indexes);

       end loop;


       --close p_gene_type;

       return;
      end;


      begin
        dbms_lob.createtemporary(genome_indexes, TRUE);

     end;
```

A.18 Modifications élémentaires issues de l'analyse de la procédure de reconstruction des ancêtres.

```
CREATE TABLE "PAPIL"."MODIF_ELEM"
```

```
  (    "NOEUD_ANC" VARCHAR2(1000 BYTE) NOT NULL ENABLE,
 "TYPE_MODIF" VARCHAR2(10 BYTE) NOT NULL ENABLE,
 "IDX_ALIGN_DEB" NUMBER,
 "IDX_ALIGN_FIN" NUMBER
  );
```

A.19 Les noms des nœuds sont parfois suivis du numéro de version.

```
  CREATE OR REPLACE FORCE VIEW "PAPIL"."ACCES_ASOC" ("ACCESSION",
"NOEUD") AS
  select distinct gl2.type as accession,am.noeud
from gene_limits_2 gl2
 join align_mult am on instr(am.noeud,gl2.type)>0
 ;
```

APPENDICE B

CODE SOURCE POUR LES CHAPITRES IV ET V

Cette annexe contient le code source utilisé pour l'implantation des fonctions Q, Q_0 pour les VPH et $Q_{1\text{-}4}$ pour le *Neisseria Meningitidis*.

L'implantation de la classe principale – *HitFunctionQ*, qui contient le calcul des fonctions, est en *Java*. Elle offre des méthodes d'interface pour les analyseurs syntaxiques écrits en *JRuby*. Pour des raisons d'efficacité l'implantation utilise des *ByteArray* – des tableaux – pour gérer la mémoire.

B.1 Classe principale *HitFunctionQ*

```
package q_func_java;

import java.io.BufferedReader;
import java.io.BufferedWriter;
import java.io.File;
import java.io.FileNotFoundException;
import java.io.FileReader;
import java.io.FileWriter;
import java.io.IOException;
import java.io.PrintWriter;
import java.util.ArrayList;
import java.util.Collections;
import java.util.HashMap;
import java.util.List;
import java.util.Map;

import java.util.Arrays;

/**
 *
 * Calcul des fonctions Q
 * Implantation avec ByteArray
```

```
 */
public class HitFunctionQ {

    //io - valeurs par defaut
    File infile = new File("files/gene_align_seqs_E1_.yaml"),
            outfile = new File("files/q_e1_java_opti.txt");
    Map<String, String> align_mult_types = new HashMap<String,
String>();
    List<String> cancero_squam = new ArrayList<String>();
    List<String> cancero_types = new ArrayList<String>();
    List<String> non_cancero_types = new ArrayList<String>();
    List<String> non_exist_cancero_types = new ArrayList<String>();
    List<String> all_types = new ArrayList<String>();
    int align_length = 0;
    int nx = 0;
    int ny = 0;
    byte[] seqs_x, seqs_y;

    public HitFunctionQ() {

    }

    /*
     * Initialisation par procédure
     */
    public void initializeCanceroSquam() {
        String[] canc_s = {"HPV-16", "HPV-18", "HPV-11", "HPV-26",
"HPV-31", "HPV-33", "HPV-35", "HPV-39", "HPV-45", "HPV-52", "HPV-53",
                "HPV-55", "HPV-56", "HPV-58", "HPV-59", "HPV-6", "HPV-66",
"HPV-73", "HPV-81", "HPV-82", "HPV-83"};

        cancero_squam = new ArrayList<String>(Arrays.asList(canc_s));

        System.out.println("cancero squam size: " +
cancero_squam.size());

    }

    /* Directionnement des entrées
     */
    public void setInputFile(String in) {
        infile = new File(in);
    }

    /* Directionnement des sorties
     */
    public void setOutputFile(String out) {
        outfile = new File(out);
    }

    /* Interface pour JRuby,
     * Spécifie les types cancéreux, ou invasifs
     */
    public void setCanceroTypes(ArrayList<String> ct) {

        this.cancero_types = ct;
        this.nx = cancero_types.size();
```

```
            for (String elem : cancero_types) {
                System.out.println("cancero_types_array_list: " + elem);
            }

        }

        /* Interface pour JRuby,
         * Spécifie les types non-cancereux, ou non-invasifs
         */
        public void setNonCanceroTypes(String[] nct) {
            this.non_cancero_types = new
ArrayList<String>(Arrays.asList(nct));
            this.ny = non_cancero_types.size();

            for (String elem : non_cancero_types) {
                System.out.println("non_cancero_types: " + elem);
            }

        }
        /* Interface externe(JRuby) pour charger un alignement multiple
         */
        public void setAlignMultTypes(Map<String, String> amt) {
            this.align_mult_types = amt;
            for (String s : align_mult_types.keySet()) {
                align_length = Math.max(align_length,
align_mult_types.get(s).length());
                //System.out.println(s + " : " + align_mult_types.get(s));

            }
            System.out.println("align_length = " + align_length);

        }

        /* Analyseur syntaxique pour alignements multiples en
         * format Yaml (Ruby), séparateur ":"
         */
        public void parseMsa()
                throws FileNotFoundException, IOException {

            BufferedReader reader = new BufferedReader(new
FileReader(infile));
            all_types.clear();
            align_mult_types.clear();

            String s = null;
            int line_no = 0;
            while ((s = reader.readLine()) != null) {
                if (line_no != 0) {

                    String[] sa = s.split(": ");

                    all_types.add(sa[0]);
                    align_mult_types.put(sa[0], sa[1]);
                    align_length = Math.max(align_length, sa[1].length());

                }
                line_no++;
```

```
            }
            reader.close();

            //System.out.println("all_types size:" + all_types.size());
            //System.out.println("HPV-18 pos:" + all_types.indexOf("HPV-
18"));
        }

        /* Longueur de l'alignement
         */
        public void setAllignLength(int al) {
            this.align_length = al;
        }

        /* Certains types sont cancéreux ou invasifs en général,
         * pourtant on ne détient pas des données.
         * L'algorithme travaille sur les types existants.
         *
         * Élimination des types non-existants dans les alignements
         */
        public void eliminateNonUsed() {

            cancero_types = new ArrayList<String>(cancero_squam);

            non_cancero_types = new ArrayList<String>(all_types);
            non_cancero_types.removeAll(cancero_types);


            non_exist_cancero_types = new ArrayList<String>(cancero_types);
            non_exist_cancero_types.removeAll(all_types);
            System.out.println("non_exist_cancero_types: " +
non_exist_cancero_types.size());
            System.out.println(non_exist_cancero_types.toString());

            cancero_types.removeAll(non_exist_cancero_types);

            nx = cancero_types.size();
            ny = non_cancero_types.size();

            Collections.sort(cancero_types);
            //Collections.sort(non_cancero_types);
            System.out.println("cancero_types" + cancero_types.toString());
            System.out.println(non_cancero_types.toString());

        }

        /* Disposition en mémoire, dans un ByteArray,
         * pour éfficacité de calcul
         */
        public void layoutSeqs() {

            System.out.println("align_length: " + align_length);
            System.out.println("nx: " + nx);
            System.out.println("ny: " + ny);


            seqs_x = new byte[align_length * nx];
            seqs_y = new byte[align_length * ny];
```

```
            int pos;


            for (String s : align_mult_types.keySet()) {
                if (cancero_types.contains(s)) {
                    pos = cancero_types.indexOf(s);

                    System.arraycopy(align_mult_types.get(s).getBytes(), 0,
                            seqs_x, pos * align_length, align_length);

                } else {
                    pos = non_cancero_types.indexOf(s);
                    //sa1 = new byte[sa[1].length()];

                    System.arraycopy(align_mult_types.get(s).getBytes(), 0,
                            seqs_y, pos * align_length, align_length);

                }
            }
        }

        /* Calcul des 5 fonctions,
         *  Q0 pour les VPH,
         *  Q1-4 pour le Neisseria Meningitidis
         */
        public void calculate() throws IOException {
            BufferedWriter wr = new BufferedWriter(new
FileWriter(outfile));
            PrintWriter pr = new PrintWriter(wr);
            pr.format("win_length,index,dXY,vX,vY,q0,q1,q2,q3,q4 \n");

            // Taille de la fenêtre 3-20 nucleotides
            for (int win_length = 20; win_length >= 3; win_length--) {
                for (int x = 0; x < (align_length - (win_length - 1)); x++)
{
    ////////////////////////////////////////////////////////////////////
///
                    //vX
                    double vX = 0.0;

                    for (int i = 0; i < nx; i++) {
                        for (int j = 0; j < nx && j != i; j++) {

                            double d_seq = 0.0;
                            double d_seqn = 0.0;

                            for (int h = 0; h < win_length; h++) {


                                //d_seq += (seqs_x.get(i * align_length + x
+ h) == seqs_x.get(j * align_length + x + h)) ? 0.0 : 1.0;
                                d_seq += (seqs_x[i * align_length + x + h]
== seqs_x[j * align_length + x + h]) ? 0.0 : 1.0;

                                //dx += (seqs_x.get(h) == seqs_y.get(x)) ? 0.0
: 1.0;
                                }
                            d_seqn = d_seq / win_length;
```

```
                            vX += (d_seqn * d_seqn);


                        }
                    }

                    vX /= (nx * (nx - 1) / 2.0);
    ////////////////////////////////////////////////////////////////////////
                    //vY
                    double vY = 0.0;

                    for (int i = 0; i < ny; i++) {
                        for (int j = 0; j < ny && j != i; j++) {

                            double d_seq = 0.0;
                            double d_seqn = 0.0;

                            for (int h = 0; h < win_length; h++) {


                                //d_seq += (seqs_x.get(i * align_length + x
+ h) == seqs_x.get(j * align_length + x + h)) ? 0.0 : 1.0;
                                d_seq += (seqs_y[i * align_length + x + h]
== seqs_y[j * align_length + x + h]) ? 0.0 : 1.0;

                            //dx += (seqs_x.get(h) == seqs_y.get(x)) ? 0.0
: 1.0;
                            }
                            d_seqn = d_seq / win_length;
                            vY += (d_seqn * d_seqn);


                        }
                    }

                    vY /= (ny * (ny - 1) / 2.0);

    ////////////////////////////////////////////////////////////////////////
                    // dXY
                    double dXY = 0.0f;

                    for (int i = 0; i < nx; i++) {
                        for (int j = 0; j < ny; j++) {
                            double d_seq = 0.0;
                            double d_seqn = 0.0;

                            //dy = dist_2seq(x, win_length, seqs_x, i,
seqs_y, j);
                            for (int h = 0; h < win_length; h++) {
                                //d_seq += (seqs_x.get(i * align_length + x
+ h) == seqs_y.get(j * align_length + x + h)) ? 0.0 : 1.0;
                                d_seq += (seqs_x[i * align_length + x + h]
== seqs_y[j * align_length + x + h]) ? 0.0 : 1.0;
                            // dy += (seqs_x.get(h) == seqs_y.get(x)) ? 0.0
: 1.0;

                            }
                            d_seqn = d_seq / win_length;
```

```
                            dXY += (d_seqn * d_seqn);

                        }
                    }
                    dXY /= (nx * ny);

                    double q0 = Math.log(1 + dXY - vX);
                    double q1 = dXY - vX;
                    double q2 = dXY - vY;
                    double q3 = (2 * dXY) - vX - vY;
                    double q4 = dXY;

                    pr.format("%d,%d,%f,%f,%f,%f,%f,%f,%f,%f\n",
win_length, x, dXY, vX, vY, q0, q1, q2, q3, q4);
                }
            }
            pr.close();
        }
    }
```

B.2 Exemple d'appel JRuby – pour VPH

L'analyse syntaxique est implantée en JRuby. Le calcul se fait en Java dans la classe principale.

```
require 'rubygems'
require 'yaml'
require 'bio'
require 'bio/io/flatfile'
require 'csv'

require '../../q_func_java_ba/dist/q_func_java_ba.jar'

require 'java'

import "q_func_java.HitFunctionQ"


hfq = HitFunctionQ.new();

puts "q_func_jruby_orig "

hfq.output_file = "../files/q_e1_jruby_ba_direct.txt"

ct = ['HPV-11', 'HPV-16', 'HPV-18', 'HPV-26', 'HPV-31', 'HPV-33', 'HPV-
35', 'HPV-39', 'HPV-45', 'HPV-52', 'HPV-55', 'HPV-58', 'HPV-59', 'HPV-6',
'HPV-66', 'HPV-73', 'HPV-81', 'HPV-82', 'HPV-83']
ct_al = java.util.ArrayList.new(ct)

hfq.cancero_types = ct_al

nct = ['HPV-54', 'HPV-75', 'HPV-76', 'HPV-12', 'HPV-77', 'HPV-13',
'HPV-34', 'HPV-14D', 'HPV-57', 'HPV-15', 'HPV-36', 'HPV-80', 'HPV-37', 'HPV-
60', 'HPV-61', 'HPV-17', 'HPV-40', 'HPV-38', 'HPV-1', 'HPV-41', 'HPV-2',
'HPV-20', 'HPV-84', 'HPV-cand85', 'HPV-21', 'HPV-19', 'HPV-42', 'HPV-63',
'HPV-3', 'HPV-cand86', 'HPV-22', 'HPV-43', 'HPV-4', 'HPV-cand87', 'HPV-44',
'HPV-65', 'HPV-5', 'HPV-23', 'HPV-cand89', 'HPV-cand90', 'HPV-7', 'HPV-
cand91', 'HPV-25', 'HPV-67', 'HPV-9', 'HPV-47', 'HPV-70', 'HPV-71', 'HPV-
27', 'HPV-50', 'HPV-48', 'HPV-69', 'HPV-30', 'HPV-28', 'HPV-49', 'HPV-10',
'HPV-29', 'HPV-94', 'HPV-cand96', 'HPV-32']
hfq.non_cancero_types = nct.to_java(:string)

#parseur des alignements multiples
msa = {}
Bio::FlatFile.open(Bio::FastaFormat,
  '../files/gene_align_seqs_E1.tfa') { |ff|

  ff.each_entry {|x|
    msa.store(x.entry_id,x.seq)
  }
}


h = java.util.HashMap.new(msa)
```

```
hfq.align_mult_types = h

hfq.layout_seqs();

t=Time.now

hfq.calculate();

t0=Time.now-t
puts "temps calcul: #{t0}"
```

B.3 Exemple d'appel Java – pour VPH

L'analyse syntaxique est intégrée à la classe principale.

```
package q_func_java;

import java.io.FileNotFoundException;
import java.io.IOException;

public class Main {

    /**
     * @param args the command line arguments
     */
    public static void main(String[] args) throws
FileNotFoundException, IOException {
        HitFunctionQ hfq = new HitFunctionQ();
        System.out.println("start");
        hfq.setInputFile("files/gene_align_seqs_E1_.yaml");
        hfq.setOutputFile("files/q_e1_java_ba.txt");
        hfq.initializeCanceroSquam();
        hfq.parseMsa();
        hfq.eliminateNonUsed();
        hfq.layoutSeqs();

        long t, t0;
        t = System.currentTimeMillis();
        hfq.calculate();

        t0 = (System.currentTimeMillis() - t) / 1000;
        System.out.println("temps calcul : " + t0);
    }
}
```

B.4 Exemple d'appel JRuby – pour Neisseria Meningitidis

L'analyse syntaxique est implantée en *JRuby*. Le calcul se fait en *Java* dans la classe principale.

```
require 'rubygems'
require 'yaml'
require 'bio'
require 'bio/io/flatfile'
require 'csv'

require '../../q_func_java_ba/dist/q_func_java_ba.jar'

require 'java'

import "q_func_java.HitFunctionQ"

hfq = HitFunctionQ.new();
```

```
puts "q_func_jruby_neisseria "

hfq.output_file = "../files/q_jruby_ba_neisseria.csv"

ct = ["fetA-76", "fetA-13", "fetA-34", "fetA-55", "fetA-77", "fetA-56",
"fetA-80", "fetA-15", "fetA-36", "fetA-57", "fetA-16", "fetA-37", "fetA-38",
"fetA-59", "fetA-17", "fetA-40", "fetA-20", "fetA-18", "fetA-41", "fetA-39",
"fetA-19", "fetA-01", "fetA-22", "fetA-43", "fetA-02", "fetA-44", "fetA-03",
"fetA-24", "fetA-45", "fetA-67", "fetA-04", "fetA-46", "fetA-05", "fetA-26",
"fetA-47", "fetA-27", "fetA-48", "fetA-06", "fetA-49", "fetA-07", "fetA-30",
"fetA-08", "fetA-29", "fetA-52", "fetA-10", "fetA-09", "fetA-53", "fetA-11",
"fetA-32", "fetA-33", "fetA-54"]
ct_al = java.util.ArrayList.new(ct)

hfq.cancero_types = ct_al

nct = ["fetA-14", "fetA-35", "fetA-78", "fetA-81", "fetA-79", "feta-
69", "fetA-60", "fetA-58", "fetA-61", "fetA-62", "fetA-42", "fetA-63",
"fetA-21", "fetA-64", "fetA-23", "fetA-65", "fetA-66", "fetA-25", "fetA-70",
"fetA-68", "fetA-71", "fetA-50", "fetA-72", "fetA-28", "fetA-51", "fetA-73",
"fetA-31", "fetA-74", "fetA-75", "fetA-12"]

hfq.non_cancero_types = nct.to_java(:string)

msa = {}
Bio::FlatFile.open(Bio::FastaFormat,
  '../files/fetA_a11e1es.tfa') { |ff|

  ff.each_entry {|x|
    msa.store(x.entry_id,x.seq)
  }
}

h = java.util.HashMap.new(msa)
hfq.align_mult_types = h


hfq.layout_seqs();

t=Time.now

hfq.calculate();

t0=Time.now-t
puts "temps calcul: #{t0}"
```

B.5 Conversion entre les formats *Yaml - Fasta*

```
#conversion format Yaml - Fasta
require 'rubygems'
require 'bio'
require 'bio/io/flatfile'
require 'csv'
```

```
#read hash from yaml file
seqs = Hash.new
File.open( "../files/gene_align_seqs_E1_.yaml" ) {
  |yf| seqs=YAML::load( yf )

}

#fill an alignment object
oa=Bio::Alignment::OriginalAlignment.new
seqs.each { |key,value|
  oa.add_seq(value,key)

}

#output the alignment to fasta
File.open("../files/gene_align_seqs_E1.tfa","w") {|f|
  f.puts  oa.output_fasta
}
```

APPENDICE C

CODE SOURCE POUR LE CHAPITRE V

Nous présentons dans la section suivante le code utilisé pour la mise en place d'une base de données pour le gène FrpB du *Neisseria Meningitidis*, ainsi que pour l'identification et la translation des index relatifs aux anses extracellulaires.

Le travail a été fait sur une base de données PostgreSQL à l'aide de JRuby et sa technologie ActiveRecord.

C.1 Scripts de création de la base de données – Migrations ActiveRecord.

C.1.1 Valeurs de la fonction Q

```
class MQFunctionGen < ActiveRecord::Migration
  def self.up
    create_table :q_function_gen do |t|
      t.integer :idx
      t.integer :win_length
      t.float :q_val, :null => true
      t.float :d_xy
      t.float :v_x
      t.string :obs
      t.text :data

    end


  end

  def self.down
```

```
    drop_table :q_function_gen
  end
end
```

C.1.2 Données de base sur l'alignement multiple – séquence avec trous - ainsi que les détails sur la séquence en question

```
class MMsaDna < ActiveRecord::Migration
  def self.up
    create_table :msa_dna do |t|
      t.integer :length
      t.text :seq
    end
  end

  def self.down
    drop_table :msa_dna
  end

end

class MMsaDnaPos < ActiveRecord::Migration

  def self.up
    create_table :msa_dna_pos do |t|
      t.integer :msa_dna_id
      t.integer :idx
      t.string :symbol
    end
  end

  def self.down
   drop_table :msa_dna_pos
  end
end
```

C.1.3 Séquence de nucléotides

```
class MSeqDna < ActiveRecord::Migration
  def self.up
    create_table :seq_dna do |t|
      t.integer :length
      t.text :seq
    end
  end

  def self.down
    drop_table :seq_dna

  end

end
```

```
class MSeqDnaPos < ActiveRecord::Migration
  def self.up


    create_table :seq_dna_pos do |t|
      t.integer :seq_dna_id
      t.integer :idx
      t.string :symbol
      t.integer :seq_aa_pos_id

    end

  end

  def self.down
   drop_table :seq_dna_pos
  end

end
```

C.1.4 Séquences d'acides aminés

```
class MSeqAa < ActiveRecord::Migration
  def self.up
    create_table :seq_aa do |t|
      t.integer :length
      t.text :seq
    end


  end

  def self.down
    drop_table :seq_aa

  end

end

class MSeqAaPos < ActiveRecord::Migration
  def self.up

    create_table :seq_aa_pos do |t|
      t.integer :seq_aa_id
      t.integer :idx
      t.string :symbol

    end

  end

  def self.down
   drop_table :seq_aa_pos
  end

end
```

C.1.5 Associations entre index nucléotides alignées (avec indels) et non-alignées (sans indels)

```
class MMsaDnaPosSeqDnaPos < ActiveRecord::Migration
  def self.up


    create_table :msa_dna_pos_seq_dna_pos do |t|
      t.integer :msa_dna_pos_id
      t.integer :seq_dna_pos_id
      t.string :align

    end

  end

  def self.down
   drop_table :msa_dna_pos_seq_dna_pos
  end

end
```

C.1.6 Anses exposées à la surface

```
class MSurfaceLoopsExposed < ActiveRecord::Migration
  def self.up


    create_table :surface_loops_exposed do |t|
      t.integer :position
      t.integer :idx_begin
      t.integer :idx_end


    end

  end

  def self.down
   drop_table :surface_loops_exposed
  end

end
```

C.1.7 Associations index acide aminé, séquence ADN non-alignée, séquence alignée pour chaque anse extracellulaire

```
class MIdxGraphMsa < ActiveRecord::Migration

#  def init
#    @t = :surface_loops_exposed
#    @ty = :integer
#
 # end

  def self.up
   # init
   @t = :surface_loops_exposed
    @ty = :integer
    #old
    [:idx_begin,:idx_end].each { |c|
      remove_column @t,c
    }
    #new
    [:idx_graph_aa_begin, :idx_graph_aa_end,
     :idx_seq_aa_begin,  :idx_seq_aa_end,
     :idx_msa_dna_begin,  :idx_msa_dna_end
    ].each { |c|
      add_column @t,c,@ty

    }

  end

  def self.down
    @t = :surface_loops_exposed
    @ty = :integer


    [:idx_graph_aa_begin, :idx_graph_aa_end,
      :idx_seq_aa_begin, :idx_seq_aa_end,
      :idx_msa_dna_begin, :idx_msa_dna_end
    ].each { |c|
      remove_column @t,c

    }

    [:idx_begin,:idx_end].each { |c|
      add_column @t,c,@ty
    }


  end

end
```

C.1.8 Positions des anses extracellulaires en indexes protéiques :

```
class MPeriplasmicLoopsExposed < ActiveRecord::Migration
  def self.up


    create_table :periplasmic_loops_exposed do |t|

      t.integer :position
      t.integer :idx_graph_aa_begin
      t.integer :idx_graph_aa_end
      t.integer :idx_seq_aa_begin
      t.integer :idx_seq_aa_end
      t.integer :idx_msa_dna_begin
      t.integer :idx_msa_dna_end


    end

  end
```

C.2 Modèles *ActiveRecord* – relations entre les tables

Définition du modèle objet - équivalent à des clés étrangères dans une base de données SQL :

```
require 'init.rb'
#require 'list.rb'

ActiveRecord::Base.pluralize_table_names = false

class QFunctionGen < ActiveRecord::Base

end


class MsaDna < ActiveRecord::Base

  has_many :msa_dna_pos,
    :class_name => "MsaDnaPos",
    :dependent => :nullify


end

class MsaDnaPos < ActiveRecord::Base
  belongs_to :msa_dna
  has_many :msa_dna_pos_seq_dna_pos ,
    :class_name => "MsaDnaPosSeqDnaPos"

end
```

```
class SeqDna < ActiveRecord::Base
  has_many :seq_dna_pos,
    :class_name => "SeqDnaPos",
    :dependent => :nullify

end

class SeqDnaPos < ActiveRecord::Base
  belongs_to :seq_dna
  belongs_to :seq_aa_pos,
    :class_name => "SeqAaPos"
  has_many :msa_dna_pos_seq_dna_pos,
    :class_name => "MsaDnaPosSeqDnaPos"
  has_many :msa_dna_pos,
    :through => :msa_dna_pos_seq_dna_pos,
    :class_name => "MsaDnaPos"

end


class SeqAa < ActiveRecord::Base
  has_many :seq_aa_pos,
    :class_name => "SeqAaPos",
    :dependent => :destroy

end

class SeqAaPos < ActiveRecord::Base
  belongs_to :seq_aa
  has_many :seq_dna_pos,
    :class_name => "SeqDnaPos",
    :dependent => :nullify


end

class MsaDnaPosSeqDnaPos < ActiveRecord::Base
  belongs_to :msa_dna_pos
  belongs_to :seq_dna_pos

end

class SurfaceLoopsExposed < ActiveRecord::Base
   acts_as_list :column => :position


end

class PeriplasmicLoopsExposed < ActiveRecord::Base
   acts_as_list :column => :position


end
```

C.3 Classe *ArUtils*, gère la connexion à la base de données, fait les migrations – création de tables

```
class ArUtils

 #connexion
  def connect(dolog)

    if dolog
     @logger = Logger.new $stderr
    else
      @logger = Logger.new nil
    end

    ActiveRecord::Base.logger = @logger
    ActiveRecord::Base.colorize_logging = false

    @config =
YAML.load_file(File.join(File.dirname(__FILE__),'database.yml'))
    #ActiveRecord::Base.establish_connection(@config["development"])
    ActiveRecord::Base.establish_connection(@config["cruby"])

    ActiveRecord::Base.pluralize_table_names = false
    #ActiveRecord::Schema.verbose = false


  end

  def migrate(version)
   #version = nil
   #version = nil
   #ActiveRecord::Migration.verbose = false
   ActiveRecord::Migrator.migrate("migrate",version)
  end



end
```

C.4 Le module *ArNeisseria* contient deux classes, *Initialize* et *Calculate*.

C.4.1 Classe *Initialize*

Initialize se charge de l'initialisation de la base de données avec les informations des articles déjà parus et les séquences officielles ADN et protéiques.

```
require 'ar_models'
require 'faster_csv'

module ArNeisseria

  class Initialize

    def init_msa_dna
      seq = <<END
GTACTGGATACCGTTACTGTAAAAGGCGACCGCCAAGGCAGCAAAATCCGTACCAACATCGTTACGCTGCA
ACAAAAA
GACGAAAGCACCGCAACCGATATGCGCGAACTCTTAAAAGAAGAGCCGTCCATCGATTTCGGCGGCGGCAA
CGGCACG
TCCCAATTCCTGACGCTGCGCGGCATGGGTCAGAACTCTGTCGACATCAAGGTGGACAACGCCTATTCCGA
CAGCCAA
ATCCTTTACCACCAAGGCAGATTTATTGTCGATCCCGCTTTGGTTAAAGTCGTTTCCGTACAAAAAGGCGC
GGGTTCC
GCCTCTGCCGGTATCGGCGCGACCAACGGCGCGATCATCGCCAAAACCGTCGATGCCCAAGACCTGCTCAA
AGGCTTG
GATAAAAACTGGGGCGTGCGCCTCAACAGCGGCTTTGCCAGCAACGAAGGCGTAAGCTACGGCGCAAGCGT
ATTCGGA
AAAGAGGGCAACTTCGACGGCTTGTTCTCTTACAACCGCAACGATGAAAAAGATTACGAAGCCGGCAAAGG
TTTCCGC
AAT---
GTCAACGGCGGCAAAACCGTACCGTACAGCGCGCTGGACAAACGCAGCTACCTCGCCAAAATCGGAACAACC
TTCGGCGACGACGACCACCGCATCGTGTTGAGCCACATGAAAGACCAACACCGGGGCATCCGCACTGTGCG
TGAAGAA
TTTACCGTCGGCGACAAAAGTTCACGGATAAAT---
ATTGACCGCCAAGCCCCTGCTTACCGCGAAACTACCCAATCC
AACACCAACTTGGCGTACACGGGTAAAAACCTGGGCTTTGTCGAAAAACTGGATGCCAACGCCTATGTGTT
GGAAAAA
GAACGCTATTCCGCCGATGACAGCGGCACCGGCTACGCAGGCAATGTAAAAGGCCCCAACCATACCCGAAT
CACCACT
CGTGGTGCGAACTTCAACTTCGACAGCCGCCTTGCCGAACAAACCCTGTTGAAATACGGTATCAACTACCG
CCATCAG
GAAATCAAACCGCAAGCATTTTTGAACTCGAAATTCTCCATCCCGACGACAGAAGAG------AAAAAC--
-GGTCAA
AAAGTCGATAAACCGATGGAACAACAAATGAAAGACCGTGCAGATGAAGACACTGTTCACGCCTACAAACT
TTCCAAC
CCGACCAAAACCGATACCGGCGTATATGTTGAAGCCATTCACGACATCGGCGATTTCACGCTGACCGGCGG
GCTGCGT
TACGACCGCTTCAAGGTGAAAACCCATGACGGCAAAACCGTTTCAAGCAGCAACCTTAACCCGAGTTTCGG
TGTGATT
TGGCAGCCGCACGAACACTGGAGCTTCAGCGCGAGCCACAACTACGCCAGCCGCAGCCCGCGCCTGTATGA
CGCGCTG
CAAACCCACGGTAAACGCGGCATCATCTCGATTGCCGACGGCACAAAAGCCGAACGCGCGCGCAATACCGA
AATCGGC
TTCAACTACAACGACGGCACGTTTGCCGCAAACGGCAGCTACTTCTGGCAGACCATCAAAGACGCGCTTGC
CAATCCG
CAAAACCGCCACGACTCT---
GTCGCCGTCCGTGAAGCCGTCAATGCCGGTTACATCAAAAACCACGGTTACGAATTG
GGCGCGTCCTACCGCACCGGCGGCCTGACTGCCAAAGTCGGCGTCAGCCACAGCAAACCGCGCTTTTAC--
----GAT
ACGCACAAAGACAAGCTGTTGAGCGCGAATCCTGAATTTGGCGCACAAGTCGGCCGCACTTGGACGGCCTC
CCTTGCC
TACCGCTTCCAAAATCCGAATCTGGAAATCGGCTGGCGCGGCCGTTATGTTCAAAAAGCTACGGGTTCGAT
ATTGGCG
```

```
     GCAGGTCAAAAAGAC---
CGCAAAGGCAACTTGGAAAACGTTGTACGCAAAGGTTTCGGTGTGAACGATGTCTTCGCC
     AACTGGAAACCGCTGGGCAAAGACACGCTCAATGTCAATCTTTCGGTTAACAACGTGTTCAACAAGTTCTA
CTATCCGCACAGC
     END
          seq.gsub! "\n",""
          puts seq.length

          #test same length as seq_dna
          y = seq.gsub("-","")
          puts "_stripped: #{y.length}"
          sleep 20

          #erase msa_dna
          del_msa_dna = MsaDna.find :all
          del_msa_dna.each { |x| x.destroy }



          #insert msa_dna
          my_msa_dna = MsaDna.new
          my_msa_dna.seq = seq
          my_msa_dna.length = seq.length

          #details of msa_dna
          for idx in 0..my_msa_dna.seq.length-1
            my_msa_dna_pos = MsaDnaPos.new
            my_msa_dna_pos.idx = idx
            my_msa_dna_pos.symbol = my_msa_dna.seq[idx..idx]

            my_msa_dna.msa_dna_pos << my_msa_dna_pos
          end


          #save master with detail
          my_msa_dna.save
          puts "inserted #{my_msa_dna.id}"


        end

        def init_seq_dna

          seq = <<END
     GTACTGGATACCGTTACTgtaaaaggcgaccgccaaggcagcaaaatccgtacc
     aacatcgttacgctgcaacaaaaagacgaaagcaccgcaaccgatatgcgcgaactctta
     aaagaagagccgtccatcgatttcggcggcggcaacggcacgtcccaattcctgacgctg
     cgcggcatgggtcagaactctgtcgacatcaaggtggacaacgcctattccgacagccaa
     atcctttaccaccaaggcagatttattgtcgatcccgctttggttaaagtcgtttccgtacaa
     aaaggcgcgggttccgcctctgccggtatcggcgcgaccaacggcgcgatcatcgcc
     aaaaccgtcgatgcccaagacctgctcaaaggc
     TTGGATAAAAACTGGGGCGTGCGCCTCA
     ACAGCGGCTTTGCCAGCAACGAAGGCGTAA GCTACGGCGC AAGCGTATTC GGAAAAGAGG
     GCAACTTCGACGGCTTGTTCTCTTACAACC GCAACGATGA AAAAGATTAC GAAGCCGGCA
     AAGGTTTCCGCAATGTCAACGGCGGCAAAA CCGTACCGTA CAGCGCGCTG GACAAACGCA
     GCTACCTCGCCAAAATCGGAACAACCTTCG GCGACGACGA CCACCGCATC GTGTTGAGCC
     ACATGAAAGACCAACACCGGGGCATCCGCA CTGTGCGTGA AGAATTTACC GTCGGCGACA
     AAAGTTCACGGATAAATATTGACCGCCAAG CCCCTGCTTA CCGCGAAACT ACCCAATCCA
```

```
ACACCAACTTGGCGTACACGGGTAAAAACC TGGGCTTTGT CGAAAAACTG GATGCCAACG
CCTATGTGTTGGAAAAAGAACGCTATTCCG CCGATGACAG CGGCACCGGC TACGCAGGCA
ATGTAAAAGGCCCCAACCATACCCGAATCA CCACTCGTGG TGCGAACTTC AACTTCGACA
GCCGCCTTGCCGAACAAACCCTGTTGAAAT ACGGTATCAA CTACCGCCAT CAGGAAATCA
AACCGCAAGCATTTTTGAACTCGAAATTCT CCATCCCGAC GACAGAAGAG AAAAACGGTC
AAAAAGTCGATAAACCGATGGAACAACAAA TGAAAGACCG TGCAGATGAA GACACTGTTC
ACGCCTACAAACTTTCCAACCCGACCAAAA CCGATACCGG CGTATATGTT GAAGCCATTC
ACGACATCGGCGATTTCACGCTGACCGGCG GGCTGCGTTA CGACCGCTTC AAGGTGAAAA
CCCATGACGGCAAAACCGTTTCAAGCAGCA ACCTTAACCC GAGTTTCGGT GTGATTTGGC
AGCCGCACGAACACTGGAGCTTCAGCGCGA GCCACAACTA CGCCAGCCGC AGCCCGCGCC
TGTATGACGCGCTGCAAACCCACGGTAAAC GCGGCATCAT CTCGATTGCC GACGGCACAA
AAGCCGAACGCGCGCGCAATACCGAAATCG GCTTCAACTA CAACGACGGC ACGTTTGCCG
CAAACGGCAGCTACTTCTGGCAGACCATCA AAGACGCGCT TGCCAATCCG CAAAACCGCC
ACGACTCTGTCGCCGTCCGTGAAGCCGTCA ATGCCGGTTA CATCAAAAAC CACGGTTACG
AATTGGGCGCGTCCTACCGCACCGGCGGCC TGACTGCCAA AGTCGGCGTC AGCCACAGCA
AACCGCGCTTTTACGATACGCACAAAGACA AGCTGTTGAG CGCGAATCCT GAATTTGGCG
CACAAGTCGGCCGCACTTGGACGGCCTCCC TTGCCTACCG CTTCCAAAAT CCGAATCTGG
AAATCGGCTGGCGCGGCCGTTATGTTCAAA AAGCTACGGG TTCGATATTG GCGGCAGGTC
AAAAAGACCGCAAAGGCAACTTGGAAAACG TTGTACGCAA AGGTTTCGGT GTGAACGATG
TCTTCGCCAACTGGAAACCGCTGGGCAAAG ACACGCTCAA TGTCAATCTT TCGGTTAACA
ACGTGTTCAACAAGTTCTACTATCCGCACAGC
END
      seq.gsub! "\n",""
      puts seq.length
      seq.gsub! " ",""
      puts seq.length
      seq.upcase!
      puts seq

      #erase seq_dna
      del_seq_dna = SeqDna.find :all
      del_seq_dna.each { |x| x.destroy }

      #insert seq_dna
      my_seq_dna = SeqDna.new
      my_seq_dna.seq = seq
      my_seq_dna.length = seq.length

      #details of seq_dna
      for idx in 0..my_seq_dna.seq.length-1
        my_seq_dna_pos = SeqDnaPos.new
        my_seq_dna_pos.idx = idx
        my_seq_dna_pos.symbol = my_seq_dna.seq[idx..idx]

        my_seq_dna.seq_dna_pos << my_seq_dna_pos
      end


      #save master with detail
      my_seq_dna.save
      puts "inserted #{my_seq_dna.id}"


    end

    def init_seq_aa
      seq = <<END
vldtvtvkgdrqgskirtnivtlqqkdestatdmrell
```

```
keepsidfgggngtsqfltlrgmgqnsvdikvdnaysdsq
ilyhqgrfivdpalvkvvsvqkgagsasagigatngaiia
ktvdaqdllkg
LDKNWGVRLNSGFASNEGVSYGASVFGKEGNFDGLFSYNRNDEKDYEAGKGFRNVNGGKT
VPYSALDKRSYLAKIGTTFGDDDHRIVLSHMKDQHRGIRTVREEFTVGDKSSRINIDR
QAPAYRETTQSNTNLAYTGKNLGFVEKLDANAYVLEKERYSADDSGTGYAGNVKGPNH
TRITTRGANFNFDSRLAEQTLLKYGINYRHQEIKPQAFLNSKFSIPTTEEKNGQKVDK
PMEQQMKDRADEDTVHAYKLSNPTKTDTGVYVEAIHDIGDFTLTGGLRYDRFKVKTHD
GKTVSSSNLNPSFGVIWQPHEHWSFSASHNYASRSPRLYDALQTHGKRGIISIADGTK
AERARNTEIGFNYNDGTFAANGSYFWQTIKDALANPQNRHDSVAVREAVNAGYIKNHG
YELGASYRTGGLTAKVGVSHSKPRFYDTHKDKLLSANPEFGAQVGRTWTASLAYRFQN
PNLEIGWRGRYVQKATGSILAAGQKDRKGNLENVVRKGFGVNDVFANWKPLGKDTLNV
NLSVNNVFNKFYYPHS
END

      seq.gsub! "\n",""
      puts seq.length
      seq.gsub! " ",""
      puts seq.length
      seq.upcase!
      puts seq

      #erase seq_aa
      del_seq_aa = SeqAa.find :all
      del_seq_aa.each { |x| x.destroy }

      #insert seq_dna
      my_seq_aa = SeqAa.new
      my_seq_aa.seq = seq
      my_seq_aa.length = seq.length

      #details of seq_dna
      for idx in 0..my_seq_aa.seq.length-1
        my_seq_aa_pos = SeqAaPos.new
        my_seq_aa_pos.idx = idx
        my_seq_aa_pos.symbol = my_seq_aa.seq[idx..idx]

        my_seq_aa.seq_aa_pos << my_seq_aa_pos

        #update corresponding seq_dna_pos
        (0..2).each { |i|
          idx_dna = my_seq_aa_pos.idx * 3 + i
          my_seq_dna_pos = SeqDnaPos.find_by_idx idx_dna
          #update the link (fk)
          my_seq_dna_pos.seq_aa_pos = my_seq_aa_pos
          my_seq_dna_pos.save
        }
      end
      #save master with detail
      my_seq_aa.save
      puts "inserted #{my_seq_aa.id}"

    end

    def init_surface_loops_exposed

      #erase surface_loops_exposed
      del = SurfaceLoopsExposed.find :all
```

```
            del.each { |x| x.destroy }
            #
            puts `pwd`

            i=0
           FasterCSV.foreach("migrate/surface_loops_exposed.csv") do |row|
             i+=1
              next if i==1;

             # all indexes are one based
              SurfaceLoopsExposed.create(
              :position => row[0],
              :idx_graph_aa_begin => row[1],
              :idx_graph_aa_end => row[2],
              #aa indexes are 0 based
              :idx_seq_aa_begin => row[1].to_i + 129 -1,
              :idx_seq_aa_end => row[2].to_i + 129 -1
              )

              #aa indexes are 0 based
              tab = SeqAaPos.find_by_idx(row[1].to_i + 129 -1)
              puts "pos: #{row[1]}, symbol: #{tab.symbol}"

           end

          end

          def init_periplasmic_loops_exposed


            #erase surface_loops_exposed
            del = PeriplasmicLoopsExposed.find :all
            del.each { |x| x.destroy }
            #
            puts `pwd`

            i=0
           FasterCSV.foreach("migrate/periplasmic_loops_exposed.csv") do
|row|
             i+=1
              next if i==1;

             # all indexes are one based
              PeriplasmicLoopsExposed.create(
              :position => row[0],
              :idx_graph_aa_begin => row[1],
              :idx_graph_aa_end => row[2],

              #aa indexes are 0 based
              #il y a un décalage de 129 dans les graphiques des articles
face a ceux
              #de toute la région protéique
              :idx_seq_aa_begin => row[1].to_i + 129 -1,
              :idx_seq_aa_end => row[2].to_i + 129 -1
              )

              #aa indexes are 0 based
              tab = SeqAaPos.find_by_idx(row[1].to_i + 129 -1)
```

```
        puts "pos: #{row[1]}, symbol: #{tab.symbol}"

      end

    end

  end
```

C.4.2 Classe *Calculate*

Calculate fait la conversion des index.

```
  class Calculate

    def assign_msa_to_dna
      del_asoc = MsaDnaPosSeqDnaPos.find :all
      del_asoc.each { |x| x.delete }

      md = MsaDna.find(:first)
      #puts md.seq
      #puts md.length
      j=0
      seq = md.seq
      for i in 0..seq.length-1
        puts "i: #{i}, j: #{j}"
        mdp = MsaDnaPos.find_by_idx(i)
        sdp = SeqDnaPos.find_by_idx(j)

        MsaDnaPosSeqDnaPos.create(:msa_dna_pos => mdp,
          :seq_dna_pos => sdp,
          :align => 'R')

      #on ne compte pas les trous pour la séquence non-alignée
      j = seq[i..i] != '-'? (j+1) : j

      end

    end


    # Intervalle des valeurs dans l'alignement multiple qui
correspondent à une
    # position dans la séquence protéique
    def min_max_msa_from_aa(idx)

      #Infinity = 1.0/0
      all_indexes = []

      aa_zero = SeqAaPos.find_by_idx idx
      #puts aa_zero.seq_aa.inspect

      aa_zero.seq_dna_pos.each { |my_seq_dna_pos|

        my_seq_dna_pos.msa_dna_pos.each { |my_msa_dna_pos|
          puts my_msa_dna_pos.inspect
```

```
        all_indexes << my_msa_dna_pos.idx
      }


    }
    #puts "min: #{all_indexes.min}, max: #{all_indexes.max}"

    return {:min => all_indexes.min,
      :max => all_indexes.max
    }

  end

  #limites des anses extracellulaires
  def assign_surface_loops_exposed_limits

    table = SurfaceLoopsExposed.find(:all)
    table.each { |row|
      min_min = min_max_msa_from_aa(row.idx_seq_aa_begin)[:min]
      row.idx_msa_dna_begin = min_min
      max_max = min_max_msa_from_aa(row.idx_seq_aa_end)[:max]
      row.idx_msa_dna_end = max_max
      row.save
    }


  end

  #calcul des translations
  def assign_periplasmic_loops_exposed_limits

    table = PeriplasmicLoopsExposed.find(:all)
    table.each { |row|
      min_min = min_max_msa_from_aa(row.idx_seq_aa_begin)[:min]
      row.idx_msa_dna_begin = min_min
      max_max = min_max_msa_from_aa(row.idx_seq_aa_end)[:max]
      row.idx_msa_dna_end = max_max
      row.save
    }

    #tab = SeqAaPos.find_by_idx(row[1].to_i + 129 -1)
    #  puts "pos: #{row[1]}, symbol: #{tab.symbol}"

  end
end
end
```

C.5 Programme principal

```
require 'rubygems'
require 'active_record'
#require '/usr/share/java/postgresql-jdbc-8.3.604.jar'


#require 'active_record/connection_adapters/postgresql_adapter'
#require 'active_record/connection_adapters/jdbc_adapter'
#require 'jdbc-postgres'

require 'ar_utils'

au = ArUtils.new
au.connect(true)

au.migrate(nil)

require 'ar_neisseria'

m_init = ArNeisseria::Initialize.new
m_init.init_msa_dna
m_init.init_seq_dna
m_init.init_seq_aa
m_init.init_surface_loops_exposed
m_init.init_periplasmic_loops_exposed


m_calc = ArNeisseria::Calculate.new

m_calc.assign_msa_to_dna
m_calc.assign_surface_loops_exposed_limits
m_calc.assign_periplasmic_loops_exposed_limits
```

C.5 Génération des graphiques

Script *Ruby* pour générer les codes sources en format *GnuPlot*. Les valeurs de la fonction Q sont prises dans un fichier en format *CSV*, généré par le code de l'annexe précédente. Les ajustements pour le milieu des intervalles gris se font directement à l'aide de constantes dans la source *GnuPlot*, ainsi que les positions des anses extracellulaires.

```
require 'erb'

windows = ["5","10","20"]
win = ""

middles = [2,5,10]
middle = 0

f_names = ["Q0","Q1","Q2","Q3","Q4"]
f_name = ""

y_ranges = [[-1,1,0.2],
            [-0.05,0.25,0.2],
            [-0.07,0.07,0.05],
            [-0.03,0.2,0.15],
            [-0.05,0.55,0.45],
]
y_range_min = 0
y_range_max = 0
y_label = 0

f_indexes = [6,7,8,9,10]
f_idx = 0

f_titles = [
  "log (1 + D(X,Y) - V(X))",
  "D(X,Y) - V(X)",
  "D(X,Y) - V(Y)",
  "2*D(X,Y) - V(X) - V(Y)",
  "D(X,Y)"
]
f_title = ""

template = ERB.new <<END
#
#
set terminal postscript eps noenhanced defaultplex \
   leveldefault color colortext \
   dashed dashlength 3.0 linewidth 1.0 butt \
   palfuncparam 2000,0.003 \
   "Helvetica" 16
set output "../eps_s/neisseria_<%= win %>_<%= f_name %>.eps"

#
set size 1,0.75
```

```
set size ratio 0.5
set lmargin 2
set bmargin 1
set rmargin 0
set tmargin 0

#
set datafile separator ","
set xrange [0:2060]
set yrange [<%= y_range_min %>:<%= y_range_max %>]
set style line 1 bgnd
set style line 2 lt rgb "cyan"
set style rect fc lt -1 fs solid 0.25 noborder

#
set obj rect from 423, -1 to 437,1
set label1 at 423-10, <%= y_label1 %> "L1" front font "Helvetica,12"

set obj rect from 516, -1 to 536,1
set label at 516-10, <%= y_label %> "L2" front font "Helvetica,12"

set obj rect from 645, -1 to 761,1
set label at 645+30, <%= y_label %> "L3" front font "Helvetica,12"

set obj rect from 864, -1 to 902,1
set label at 864-5, <%= y_label %> "L4" front font "Helvetica,12"

set obj rect from 1023, -1 to 1172,1
set label at 1023+40, <%= y_label %> "L5" front font "Helvetica,12"

set obj rect from 1260, -1 to 1292,1
set label at 1260-5, <%= y_label %> "L6" front font "Helvetica,12"

set obj rect from 1380, -1 to 1454,1
set label at 1380+15, <%= y_label %> "L7" front font "Helvetica,12"

set obj rect from 1557, -1 to 1586,1
set label at 1557-5, <%= y_label %> "L8" front font "Helvetica,12"

set obj rect from 1707, -1 to 1775,1
set label at 1707+10, <%= y_label %> "L9" front font "Helvetica,12"

set obj rect from 1845, -1 to 1925,1
set label at 1845+5, <%= y_label %> "L10" front font "Helvetica,12"

set obj rect from 2013, -1 to 2033,1
set label at 2013-30, <%= y_label %> "L11" front font "Helvetica,12"

set obj rect from 387,-1 to 389,1 fs solid 0.10
set obj rect from 471,-1 to 479,1 fs solid 0.10
set obj rect from 588,-1 to 599,1 fs solid 0.10
set obj rect from 807,-1 to 818,1 fs solid 0.10
set obj rect from 960,-1 to 974,1 fs solid 0.10
set obj rect from 1215,-1 to 1223,1 fs solid 0.10
set obj rect from 1332,-1 to 1343,1 fs solid 0.10
set obj rect from 1497,-1 to 1502,1 fs solid 0.10
set obj rect from 1659,-1 to 1661,1 fs solid 0.10
set obj rect from 1812,-1 to 1814,1 fs solid 0.10
```

```
      set obj rect from 1959,-1 to 1976,1 fs solid 0.10
      #
      set style line 3 lt 1 lc rgb "dark-blue" lw 3
      set style line 8 lt 3 lc rgb "black" lw 3
      #vertical delimiter
      set parametric
      const=386
      set trange [0:1]
      #
      plot  "../files/q_func_neisseria_<%= win %>.csv" using ($2+<%= middle
%>):<%= f_idx %> title '<%= f_name %> = <%= f_title %>' with lines ls 3, \
      const,t notitle with lines ls 8
      #
      END

      (0..2).each {|win_idx|
        win = windows[win_idx]
        middle = middles[win_idx]

        (0..4).each { |name_idx|
          f_name = f_names[name_idx]
          f_idx = f_indexes[name_idx]
          f_title = f_titles[name_idx]
          y_range_min= y_ranges[name_idx][0]
          y_range_max= y_ranges[name_idx][1]
          y_label = y_ranges[name_idx][2]

          #puts "win: #{win}, middle: #{middle}, f_name: #{f_name}, f_idx:
#{f_idx}, f_title: #{f_title}"
          t= template.result(binding)
          #puts t

          f = File.new("../gnu_plot/ruby_neisseria.plt", "w")
           f.puts t

          f.close


          #puts `gnuplot -persist ../gnu_plot/ruby_neisseria.plt`
          puts `gnuplot ../gnu_plot/ruby_neisseria.plt`
          filename = "neisseria_#{win}_#{f_name}"
          cmd = "../gnu_plot2/eps2png -f ../eps_s/#{filename}.eps"
          puts `#{cmd}`

        }

      }

      #copy results to article folder
       puts `pwd`
       puts `rm -rf
/home/dunarel_b/PROJETS/article_conf_ifcs_2009_v3/eps_s/*`
       puts `cp -r /home/dunarel_b/PROJETS/q_func_jruby_ba/eps_s/*
/home/dunarel_b/PROJETS/article_conf_ifcs_2009_v3/eps_s/`
```

RÉFÉRENCES

Achaz, G., Rocha, E.P., Netter, P. et Coissac, E. 2002. «Origin and fate of repeats in bacteria», *Nucleic Acids Res.*, 30(13):2987–2994.

Adachi, J. et Hasegawa, M. 1992. «Amino acid substitution of proteins coded for in mitochondrialDNA during mammalian evolution». *Jpn. J. Genet.*, 67:187-197.

Ahlquist, P. 2002. «RNA-dependent RNA polymerases, viruses, and RNA silencing». *Science,* 296(5571):1270–3.

Allaby, M. 2009. «*ICZN. A Dictionary of Zoology. 1999*». Encyclopedia.com.

Altschul, S.F., Gish, W., Miller, W., Myers, E. W. et Lipman, D.J. 1990. «Basic local alignment search tool». *J. Mol. Biol.*, 215(3):403–410.

Badescu, D., Diallo, A. B., Blanchette, M. et Makarenkov. V. 2008. «An evolution study of the human papillomavirus genomes». In *Proceedings of RECOMB Comparative Genomics 2008, Springer, Lecture Notes in Bioinformatics Series*, Paris, 128–140.

Badescu, D., Diallo, A.B. and Makarenkov, V. 2010. «Identification of specific genomic regions responsible for the invasivity of *Neisseria Meningitidis*». In *Classification as a Tool for Research*, Locarek-Junge, H. and Weihs, C. eds, proceedings of IFCS 2009. Studies in Classification, Data Analysis, and Knowledge Organization, Springer Berlin-Heidelberg-New York (à paraitre).

Becerra, A., Delaye, L., Islas, S. et Lazcano, A. 2007. «The very early stages of biological evolution and the nature of the last common ancestor of the three major cell domains». *Annual Review of Ecology, Evolution, and Systematics*, 38:361-379.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. et Wheeler, D.L. 2005. «GenBank», *Nucleic Acids Res.*, 33(Database Issue): D34–D38.

Blondeau, J.M., Ashton, F.E., Isaacson, M., Yaschuck, Y., Anderson, C. et Ducasse, G. 1995. «*Neisseria meningitidis* with decreased susceptibility to penicillin in Saskatchewan, Canada». *J. Clin. Microbiol.,* 13:1784–1786.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L. et Rubin, E.M. 2003. «Phylogenetic shadowing of primate sequences to find functional regions of the human genome». *Science*, 299:1391–1394.

Brändén, C.I. et Tooze, J. 1996. «*Introduction à la structure des protéines». (trad. Bernard Lubochinsky, préf. Joël Janin)*. De Boeck Université, Bruxelles.

Brown, T. A. 2006. *Genomes, Third Edition*. Garland Science.

Brown, T.A. 2002. *Genomes,* Department of Biomolecular Sciences, BIOS Scientific Publishers Ltd.

Brudno., M., Chapman, M., Göttgens, B., Batzoglou et S. et Morgenstern, B. 2003. «Fast and sensitive multiple alignment of large genomic sequences». *BMC Bioinformatics,* 4:66.

Castle, P.E., Stoler, M.H., Solomon, D., et Schiffman, M. pour the ALTS Group. 2007. «The Relationship of Community Biopsy-Diagnosed Cervical Intraepithelial Neoplasia Grade 2 to the Quality Control Pathology-Reviewed». *AJCP,* 127:805-815.

Caugant, D.A., Kristiansen, B.E., Froholm, L.O., Bovre, K. et Selander, R.K. 1988. «Clonal diversity of *Neisseria meningitidis* from a population of asymptomatic carriers». *Infect. Immun.* 94:2060–2068.

Champoux, J. 2001. «DNA topoisomerases: structure, function, and mechanism». *Annu. Rev. Biochem,* 70:369–413.

Clark, A.G. et Kao, T.H. 1991. «Excess nonsynonymous substitution of shared polymorphic sites among self-incompatibility alleles of Solanaceae». *Institute of Molecular Evolutionary Genetics*, 88(21):9823-9827.

Clark, A.G., Glanowski, S., Nielsen, R., Thomas, P.D., Kejariwal, A., Todd, M.A., Tanenbaum, D.M., Civello, D., Lu, F., Murphy, B. et al. 2003. «Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios». *Science*, 302:1960–1963.

Clifford, G.M. et al. 2005. «Worldwide distribution of human papillomavirus types in cytologically normal women in the International Agency for Research on Cancer HPV prevalence surveys: a pooled analysis». *The Lancet*, 366(9490):991-998.

Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., et Sidow, A. 2005. «Distribution and intensity of constraint in mammalian genomic sequence». *Genome. Res.*, 15:901–913.

Crick, F. 1970. «Central Dogma of Molecular Biology». *Nature*, 227:561-563.

Crick, F.H.C. 1958. «On Protein Synthesis». *Symp. Soc. Exp. Biol.,* XII, 139-163.

Darwin, C. 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.*

Darwin, C. 1872. *The Origin of Species. Sixth Edition.* The Modern Library, New York. 170–171.

de Queiroz, K. 2005. «Ernst Mayr and the modern concept of species». *Proc. Natl. Acad. Sci. U.S.A.*, 102 Suppl 1: 6600–7.

de Villiers, E.M., Fauquet, C., Broker, T.R., Bernard, H.U. et zur Hausen, H. 2004. «Classification Of Papillomaviruses». *Virology*, 324, 17–27.

Diallo, A. B., Makarenkov, V. et Blanchette, M. 2007. «Exact and heuristic algorithms for the Indel Maximum Likelihood Problem», *J. of Computational Biology*. 14:446-461.

Diallo, A.B., Badescu, D., Blanchette, M. and Makarenkov, V. 2009a. «Classification of the Human Papilloma Viruses». In *Proceedings of SFC-CLADAG 2008*, Springer Verlag, 8 pages (à paraitre).

Diallo, A.B., Badescu, D., Blanchette, M. and Makarenkov, V. 2009b. «A whole genome study and identification of specific carcinogenic regions of the Human Papilloma Viruses», *Journal of Computational Biology*, 16(10):1461-1473.

Dobbelstein, M. et Roth, J. 1998. «The large T antigen of simian virus 40 binds and inactivates p53 but not p73». *Journal of General Virology*, 79:3079–3083.

Doorbar, J. 2006. «Molecular biology of human papillomavirus infection and cervical cancer». *Clinical Science*, 110(5):525-541.

Doty P., Boedtker H., Fresco J. R. , Haselkorn R. et Litt M. 1959. «Secondary Structure in Ribonucleic Acids», In *Proc. Natl. Acad. Sci. U.S.A.*, 45:482-499.

Duret, L. 2008. «Neutral Theory: The Null Hypothesis of Molecular Evolution». *Nature Education*, 1:803-806.

Eddy, S.R. 2008. «A Probabilistic Model of Local Sequence Alignment that Simplifies Statistical Significance Estimation». *PLoS Comput. Biol.*, 4:e1000069

Edgar, R.C. 2004. «MUSCLE: multiple sequence alignment with high accuracy and high throughput». *Nucleic Acids Research,* 32(5):1792-97.

Edwards, A.W.F et Cavalli-Sforza, L.L. 1963. «The reconstruction of evolutionary trees. Pp.67-76 in Phenetic and Phylogenetic Classification», *Ed. V.H. Heywood and J.McNeill. Systematics Association Publ. No. 6, London.*

Elias, J., Harmsen, D., Claus, H., Hellenbrand, W., Frosch, M. et Vogel, U. 2006. «Spatiotemporal analysis of invasive meningococcal disease, Germany», *Emerg Infect Dis.*, 12(11):1689–1695.

ENCODE Project Consortium 2007. «Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project».

Envall, M. 2008. «On the difference between mono-, holo-, and paraphyletic groups: a consistent distinction of process and pattern». *Biological Journal of the Linnean Society, 94:217–220.*

Felsenstein, J. 1981. «Evolutionary trees from DNA sequences: A maximum likelihood approach». *Journal of Molecular Evolution*, 17:368-376.

Forsberg, R. et Christiansen, F.B. 2003. «A codon-based model of host-specific selection in parasites, with an application to the influenza A virus». *Mol. Biol. Evol.*, 20:1252–1259.

Friedman, N., Ninio, M., Peer, I. et Pupko, T. 2002. «A Structural EM Algorithm for Phylogenetic Inference». *Journal of Computational Biology*, 9(2):331-353.

Gilbert, W. 1986. «The RNA World». *Nature*, 319:618.

Greenfield, S., Sheehe, P.R. et Feldman, H.A. 1971. «Meningococcal carriage in a population of 'normal' families». *J. Infect. Dis.*, 5 :67–73.

Guindon, S., et Gascuel, O. 2003. «A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood». *Systematic Biology,* 52: 696-704.

Guindon, S., Rodrigo, A.G., Dyer, K.A. et Huelsenbeck, J.P. 2004. «Modeling the site-specific variation of selection patterns along lineages». *Proc. Natl. Acad. Sci. U.S.A*, 101:12957–12962.

Gupta, S.K., Kececioglu, J. et Schäffer, A.A. 1995. «Improving the practical space and time efficiency of the shortest-paths Approach to sum-of-pairs multiple sequence alignment». *J Comput. Biol*, 2:459-472.

Haeckel, E. 1874. *Anthropogenie oder Entwickelungsgeschichte des Menschen.* Gemeinverständliche wissenschaftliche Vorträge über die Grundzüge der menschlichen Keimes- und Stammes-Geschichte. Leipzig: Engelmann (scanned from Tafel XII by Hanno in 2002).

Hasegawa, M., Kishino, H., et Yano, T. 1985. «Dating of the human-ape splitting by a molecular clock of mitochondrial DNA». *J. Mol. Evol.*, 22:160–174.

Hennig, W. 1950. *Grundzüge einer Theorie der phylogenetischen Systematik*, Deutscher Zentralverlag, Berlin.

Hennig, W. 1975. «Cladistic Analysis or Cladistic Classification?: A Reply to Ernst Mayr». *Systematic Zoology*, 24(2):244-256.

Hotopp, J.C., Grifantini, R., Kumar, N., Tzeng, Y.L., Fouts, D. et Frigimelica, E. *et al.* 2006. «Comparative genomics of *Neisseria meningitidis*: core genome, islands of horizontal transfer and pathogen-specific genes». *Microbiology*, 152(12):3733–3749.

ICTVdB Management 2006. *The Universal Virus Database.* New York, Büchen-Osmond, C., Columbia University, website: http://www.ncbi.nlm.nih.gov/ICTVdb/.

Jackson, L.A., Schuchat, A., Reeves, M.W. et Wenger, J.D. 1995. «Serogroup C meningococcal outbreaks in the United States. An emerging threat». *JAMA*, 83:383–389.

Jacob, F. et Monod, J. 1961. «Genetic regulatory mechanisms in the synthesis of proteins». *Journal of molecular biology*, 3:318-56.

Johannsen, W. 1905. *Arvelighedslærens elementer (The Elements of Heredity).* Copenhagen.

Johannsen, W. 1909. *Elemente der exakten Erblichkeitslehre.* Gustav Fischer, Jena.

Johannsen, W. 1911. «The Genotype Conception of Heredity». *American Naturalist,* 45(531):129-159.

Jones, G.R., Christodoulides, M., Brooks, J.L., Miller, A.R., Cartwright, K.A. et Heckels, J.E. 1998. «Dynamics of carriage of *Neisseria meningitidis* in a group of military recruits: subtype stability and specificity of the immune response following colonization». *J. Infect. Dis.,* 56 :451–459.

Kimura, M. 1968. «Evolutionary rate at the molecular level». *Nature,* 217: 624–626.

Kimura, M. 1991. «The neutral theory of molecular evolution: a review of recent evidence». *Jpn. J. Genet.*, 66(4):367-86.

King, K.L. et Jukes, T.H. 1969. «Non-Darwinian Evolution». *Science*, 164 (3881):788–798.

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

Kuhner, M.K. et Felsenstein, J. 1994. «A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates». *Mol. Biol. Evol.,* 11(3):459-68.

Le Guillou, C. 2009. *Génétiques.* Portail Sciences de la vie et de la terre de l'académie Aix Marseille, http://www.svt.ac-aix-marseille.fr/expoconf/genetiques/index.htm

Lecointre, G. et Le Guyader, H. 2002. *Classification phylogénétique du vivant, page 13, 2e édition*, Belin, Paris

Lecointre, G., Le Guyader, H., Visset, D. et Mccoy, K. 2009. *The Tree of Life: A Phylogenetic Classification.*

Lipman, D.J., Altschul, S.F. et Kececioglu, J.D. 1989. «A tool for multiple sequence alignment». *Proc. Natl. Acad. Sc.i U.S.A.*, 86:4412-4415.

Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D. et Darnell, J. 2000. *Molecular Cell Biology (4ème edition)*, Freeman & Co., New York, NY.

Maddison, D. R. et Schulz, K.-S. 2007. *The Tree of Life Web Project*. http://tolweb.org.

Maddison, D.R. et Maddison, W.P. 1996. *The Tree of Life Project*. http://phylogeny.arizona.edu/tree/phylogeny.html

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M. et Spratt, B.G. 1998. «Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms». In *Proc. Natl. Acad. Sci. USA,* 176:3140–3145.

Maiden, M.C., Malorny, B. et Achtman, M. 1996. «A global gene pool in the neisseriae». *Mol. Microbiol.,* 21(6):1297–1298.

Makarenkov, V. 2001. «T-Rex: reconstructing and visualizing phylogenetic trees and reticulation networks». *Bioinformatics*, 17: 664-668.

Margulies, E.H. et Blanchette, M., NISC Comparative Sequencing Program, Haussler, D., Green, E.D. 2003. «Identification and characterization of multi-species conserved sequences». *Genome Res.*, 13: 2507–2518.

Mayer, G.C. 1993. «Darwin-L Message Log 3:73». Academic Discussion on the History and Theory of the Historical Sciences. *Archives of Darwin-L (1993–1997), a professional discussion group on the history and theory of the historical sciences.*

Mayr, E. 1942 *Systematics and the Origin of Species*. Columbia Univ. Press, New York.

Mayr, E. 1996. «What Is a Species and What Is Not?». *Philosophy of Science*, 63:262–77.

Merriam-Webster Online Dictionary. 2008. *Merriam-Webster Online*. http://www.merriam-webster.com/dictionary.

Morgan, T. H., Sturtevant, A. H., Muller, H. J. et Bridges, C. B. 1915. *The Mechanism of Mendelian Heredity*. Henry Holt and Company.

Muñoz, N., Bosch, F., Castellsagué, X., Daz, M., de Sanjose, S., Hammouda, D., Shah, K., et Meijer, C. 2004. «Against which human papillomavirus types shall we vaccinate and screen? the international perspective». *International Journal of Cancer,* 111:278–285.

Muñoz, N., Bosch, F., de Sanjosé, S., Herrero, R., Castellsagué, X., Shah, K., Snijders, P., et Meijer, C. 2003. «Epidemiologic classification of human papillomavirus types associated with cervical cancer». *New England Journal of Medecine,* 384:518–527.

Neisseria Research Community Website, 2009. http://www.neisseria.org.

Neyman, J. 1971. «Molecular studies of evolution: A source of novel statistical problems». Pp 1-27 in *Statistical Decision Theory and Related Topics, ed. S.S.Gupta and J. Yackel.* Academic Press, New York.

Nielsen, R. et Yang, Z. 1998. «Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene». *Genetics*, 148:929–936 .

Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J., et al. 2005. «A scan for positively selected genes in the genomes of humans and chimpanzees». *PloS Biol.*, 3:e170.

Notredame, C., Higgins, D.G. et Heringa, J. 2000. «T-Coffee: A novel method for fast and accurate multiple sequence alignment». *J Mol Biol.*, 302(1):205-217.

Oliva, A, Fariña, J.B. et Llabrés, M. 2004. «Measurement of uncertainty in peptide molecular weight determination using size-exclusion chromatography with multi-angle laser light-scattering detection and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry». *Analytica Chimica Acta*, 512:103–110.

Parkin, D.M., Bray, F., Ferlay, J. et Pisani, P. 2005. «Global cancer statistics, 2002». *CA Cancer J. Clin.*, 55:74–108.

Peltola, H. 1983. «Meningococcal disease: still with us». *Rev. Infect. Dis.,* 95(1983):71–91.

Penny, D., Hendy, M.D. et Steel, M.A. 1992. «Progress with methods for constructing evolutionary trees». *Trends in Ecology and Evolution*, 7:73-79.

Pierce, B.A. 2007. *Genetics: A conceptual Approach (3rd ed.).* W. H. Freeman.

Pinner, R.W., Onyango, F., Perkins, B.A., Mirza, N.B., Ngacha, D.M., Reeves, M., Dewitt, W., Njeru, N.N., Agata, E. et Broome, C.V. 1992. «Epidemic meningococcal disease in Nairobi, Kenya, 1988. The Kenya/Centers for Disease Control (CDC) Meningitis Study Group». *J. Infect. Dis.,* 273 :359–364.

Pleijel, F. et Rouse, G.W. 2003. «Ceci n'est pas une pipe: names, clades and phylogenetic nomenclature». *J. Zool. Syst. Evol. Research*, 41:162–174.

Prétet, J., Charlot, J., et Mougin, C. 2007. «Virological and carcinogenic aspects of hpv». *Bulletin Academic National de Medecine*, 191(3):611–613.

Raymond, N.J., Reeves, M., Ajello, G., Baughman, W., Gheesling, L.L., Carlone, G.M., Wenger, J.D. et Stephens, D.S. 1997. «Molecular epidemiology of sporadic (endemic) serogroup C meningococcal disease». *J. Infect. Dis.,* 178:1277–1284.

Rebrikov, D.V., Bogdanova, E.A., Bulina, M.E. et Lukyanov, S.A. 2002. «A new planarian extrachromosomal virus-like element revealed by subtractive hybridization». *Mol. Biol.*, 36:813–820.

Ridley, M. 2004. *Evolution, 3rd Edition.* Blackwell Science.

Rocha, E.P. 2006. «Inference and analysis of the relative stability of bacterial chromosomes». *Mol. Biol. Evol.*, 23(3):513–522.

Saenger, W. 1984. *Principles of Nucleic Acid Structure.* New York: Springer-Verlag.

Saitou, N. et Nei., M. 1987. «The neighbor-joining method: a new method for reconstructing phylogenetic trees». *Mol. Biol. Evol.*, 4(4):406-425.

SanMiguel, P. et Bennetzen, J.L. 1998. «Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotranposons». *Annals of Botany,* 82(Suppl A):37–44.

Saunders, N.J. et Snyder, L.A. 2002. «The minimal mobile element». *Microbiology,* 148(12):3756–3760.

Schadt, E.E., Sinsheimer, J.S. et Lange, K. 1998. «Computational advances in maximum likelihood methods for molecular phylogeny». *Genome research,* 8(3):222-33.

Schiffman, M., Castle, P.E., Jeronimo, J., Rodriguez, A.C. et Wacholder, S. 2007. «Human papillomavirus and cervical cancer», *Lancet*, 370:890-907.

Schoen, C., Claus, H., Vogel, U. et Frosch, M. 2006. «Genomes of pathogenic Neisseria species». In: J. Hacker and U. Dobrindt, Editors, *Pathogenomics*, Wiley-VCH, Weinheim (2006), 231–255.

Schoen, C., Joseph, B., Claus, H., Vogel, U. et Frosch, M. 2007. «Living in a changing environment: insights into host adaptation in *Neisseria meningitidis* from comparative genomics». *Int. J. Med. Microbio.,* 297(7–8):601–613.

Schoen, C., Tettelin, H., Parkhill, J. et Frosch, M. 2009. «Genome flexibility in *Neisseria meningitides*». *Vaccine,* 27 Suppl 2:B103-11. Epub.

Segondy, M. 2008. «Classification des papillomavirus (HPV)». *Revue Francophone des Laboratoires*, 405:23-25.

Sergei, L., Kosakovsky, P. et Simon Frost, D.W. 2005. «A Genetic Algorithm Approach to Detecting Lineage-Specific Variation in Selection Pressure», *Molecular Biology and Evolution*, 22(3):478-485.

Shapiro, B.J. et Alm, E.J. 2008. «Comparing Patterns of Natural Selection across Species Using Selective Signatures». *PLoS Genet*, 4(2): e23. doi:10.1371/journal.pgen.0040023

Shen, J., Heckendorn, R.B. 2004. «Discrete Branch Length Representations for Genetic Algorithms in Phylogenetic Search». In *Applications of Evolutionary Computing.*, 3005/2004: 94-103. Springer Berlin / Heidelberg.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. «Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes». *Genome Res.*, 15:1034–1050.

Siepel, A., Pollard K.S. et Haussler, D. 2006. «New methods for detecting lineage-specific selection». In *Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB 2006)*, 190-205.

Sluys, R., Martens, et K., Schram, F.R. 2004. «The PhyloCode: naming of biodiversity at a crossroads». *Trends Ecol Evol.*, 19(6):280-1.

Sneath, P.H.A. et Snokal, R.R. 1973. *Numerical taxonomy,* W. H. Freeman, San Francisco, California, USA.

Snyder, L.A., Davies, J.K., Ryan, C.S. et Saunders, N.J. 2005. «Comparative overview of the genomic and genetic differences between the pathogenic Neisseria strains and species». *Plasmid*, 54(3):191–218.

Sonea, S. et Mathieu, L. G. 2000. *Prokaryotology- A coherent view.* Presses de l'Université de Montréal, Montréal.

Sonea, S. et Panisset, M. 1976. «Pour une nouvelle bactériologie». *Revue Canadienne de Biologie*, 35:103-167.

Sorhannus, U. 2003. «The Effect of Positive Selection on a Sexual Reproduction Gene in Thalassiosira weissflogii (Bacillariophyta): Results Obtained from Maximum-Likelihood and Parsimony-Based Methods». *Molecular Biology and Evolution*, 20(3):1326-1328.

Strimmer, K., et von Haeseler, A. 1996. «Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies». *Mol. Biol. Evol,.* 13: 964-969.

Suzuki, Y. et Nei, M. 2001. «Reliabilities of Parsimony-based and Likelihood-based Methods for Detecting Positive Selection at Single Amino Acid Sites». *Molecular Biology and Evolution*, 18:2179-2185.

Tavaré S. 1986. «Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences». *American Mathematical Society: Lectures on Mathematics in the Life Sciences*, 17:57–86.

Tettelin, H., Riley, D., Cattuto, C. et Medini, D. 2008. «Comparative genomics: the bacterial pan-genome». *Curr. Opin. Microbiol.,* 11(5):472–477.

Tettelin, H., Saunders, N.J., Heidelberg, J., Jeffries, A.C., Nelson, K.E. et Eisen, J.A. *et al.* 2000. «Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58». *Science,* 287(5459):1809–1815.

Thibaut-Adrien, M. 2008. «Constitution de l'acide désoxyribonucléique». http://toutsurladn.blogspot.com/2008/10/constitution-de-lacide.html.

Thompson, E. A. L., Feavers I. M. et Maiden M. C. J. 2003. «Antigenic diversity of meningococcal enterobactin receptor FetA, a vaccine component». *Microbiology*, 149(7):1849–1858.

Thompson, J.D., Higgins, D.G. et Gibson, T.J. 1994. «CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice». *Nucleic Acids Res.,* 22: 4673-4680.

Tikhomirov, E., Santamaria, M. et Esteves, K. 1997. «Meningococcal disease: public health burden and control». *World Health Stat. Quart.,* 50:170-6.

Trottier, H. et Franco, E. L. 2006. «The epidemiology of genital human papillomavirus infection». *Vaccine,* 24S1 S1/4–S1/15.

Tuffley, C. et Steel, M. 1997. «Links between maximum likelihood and maximum parsimony under a simple model of site substitution». *Bull. Math. Biol.*, 59:581-607.

Turner, P.C., McLennan, A.G., Bates, A.D. et White M.R.H. 1997. *Instant Notes in Molecular Biology.* BIOS Scientific Publishers, Oxford.

Tzeng, Y.-L. et Stephens, D.S. 2000. «Epidemiology and pathogenesis of *Neisseria meningitides». Microbes and Infection.,* 2(6):687-700.

Urwin, R., Russell, J. E., Thompson, E. A. L. et al. 2004. «Distribution of Surface Protein Variants among Hyperinvasive Meningococci: Implications for Vaccine Design». *Infect. Immun.*, 72(10):5955–5962.

Watson J.D. et Crick F.H.C. 1953. «A Structure for Deoxyribose Nucleic Acid». *Nature,* 171: 737–738.

Wilson, R., Ryan, G., Knight, G., Laimins, L., et Roberts, S. 2007. «The full-length e1^e4 protein of human papillomavirus type 18 modulates differentiation-dependent viral dna amplification and late gene expression». *Virology*, 362(2):453–460.

Woese, C. 1968. *The Genetic Code.* Harper & Row.

Woese, C. et Fox, G. 1977. «Phylogenetic structure of the prokaryotic domain: the primary kingdoms». *Proc. Natl. Acad. Sci. USA,* 74(11):5088–90.

Woese, C., Kandler, O. et Wheelis, M. 1990. «Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya».. *Proc Natl Acad Sci USA,* 87 (12): 4576–9.

Woese, C., Magrum, L. et Fox, G. 1978. «Archaebacteria». *J. Mol. Evol.*,11(3):245–51..

Yang, Z. 1994. «Estimating the pattern of nucleotide substitution». *Journal of Molecular Evolution*, 39:105-111.

Yang, Z. 1996. «Among-site rate variation and its impact on phylogenetic analyses». *Trends in Ecology and Evolution*, 11:367-372.

Yang, Z. 1996. «Maximum likelihood models for combined analyses of multiple sequence data». *Journal of Molecular Evolution*, 42:587-596.

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

Zhang, J. et Nei, M. 1997. «Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods». *Journal of Molecular Evolution*, S139-S146.