Cours n^0= 7 : Les logiciels d'analyse

Table des matières

I. Introduction

La Bio-informatique est un champ de recherche multi-disciplinaire où travaillent de nombreux biologistes, informaticiens, mathématiciens et physiciens, dans le but de résoudre un problème scientifique posé par la biologie. Elle a pour objectif l'analyse et la modélisation des éléments biologiques en se basant sur diverses disciplines: l'informatique (algorithmique, complexité, combinatoire des mots, cryptographie, simulation, bases de données, analyse de données et méthodes de classification, calcul distribué et parallèle, développement de logiciels de recherche, etc.). Le terme bio-informatique peut également décrire (par abus de langage) toutes les applications informatiques résultant de ces recherches. Cela va de l'analyse du génome à la modélisation de l'évolution d'une population animale dans un environnement donné, en passant par la modélisation moléculaire, l'analyse d'image, le séquençage du génome et la reconstruction d'arbres phylogénétiques (phylogénie). Cette discipline constitue la « biologie *in silico* », par analogie avec *in vitro* ou *in vivo*. La bioinformatique est une science jeune, en pleine expansion et pluridisciplinaire avec de nombreux problèmes théoriques ouverts et des applications majeures en médecine, biotechnologie, pharmacie et agronomie.
La Bio-informatique est constituée par l'ensemble des concepts et des techniques nécessaires à l'interprétation de l'information génétique (séquences) et structurale (repliement 3D). C'est le décryptage de la « bio-information ». La bio-informatique est donc une branche théorique de la biologie[]. Il s'agit en fait d'analyser, modéliser ou prédire les informations issues d'activités de recherche.

II. Plateforme des logiciels bioinformatiques

La plateforme dispose d'un ensemble de logiciels via le web et en ligne de commande. Pour certains de ces logiciels, la plate-forme s'engage à assurer leur maintenance (version à jour) et leur bon fonctionnement (réalisation de tests fonctionnels).

1. Logiciels disponibles via le web

Logiciel	Description	Type
ncbi-blast	Recherche de similarité dans des banques de données, NCBI Blast	Recherche de similarités
SRS	Extraction de séquence (en cours d'installation)	Extraction de séquences
EMBOSS	Le Package EMBOSS (European Molecular Biology Open Software Suite) est une suite logicielle développée par l'EBI et l'institut SANGER. La suite comprend programmes, utilitaires et banques de séquences qui permettent de couvrir l'ensemble	Suite logicielle

	des besoins élémentaires dans le domaine de l'analyse et de l'exploitation des séquences biologiques.	
EuGène'Hom	EuGène'Hom est un logiciel de prédiction de génes pour les Eukariote basé sur l'analyse comparative	Outil d'annotation
FrameD	Prédiction de gènes dans des organisme Prokaryote ou de transcrit	Outil d'annotation
Gepeto	Gepeto est un outil de comparaison de prédictions de prédictions de gènes	Outil d'annotation
Multalin	Programme d'alignement multiple de séquences génomiques et protéiques	Alignement multiple

2. Logiciels disponibles en ligne de commande

Logiciel	Description	Type
clustalw	Programme d'alignement multiple de séquences protéiques ou nucléiques.	Alignement de séquences
Multalin	Alignement multiple de séquence	Alignement de séquences
erpin	(Easy RNA Profile IdentificatioN)	ARNnc
RNA Vienna Package	La suite Vienna RNA est dédiée à la prédiction et à la comparaison des structures secondaires des ARN.	ARNnc
tRNAscan-SE	Prédiction d'ARN de transfert	ARNnc
RNAmmer	Recherche d'ARN ribosomique	ARNnc
patscan	Recherche de motifs dans des séquences protéiques et nucléiques	Motif
InterProScan	Recherche de domaines protéiques dans la banque InterPro.	Protéines
wu-blast	Recherche de similarité dans des banques de données, Washington University Blast	Recherche de similarités
ncbi-blast	Recherche de similarité dans des banques de données, NCBI Blast	Recherche de similarités
EMBOSS	Le Package EMBOSS (European Molecular Biology Open Software Suite) est une suite logicielle développée par l'EBI et l'institut SANGER. La suite comprend programmes, utilitaires et banques de séquences qui permettent de couvrir l'ensemble des besoins élémentaires dans le domaine de l'analyse et de l'exploitation des séquences biologiques.	Suite logicielle
ReadSeq	Reformatage de sequence fasta	Utilitaire
Dialign	Alignement de séquences protéiques et nucléiques, basé sur une comparaison de segments de séquences.	Alignement de séquences

Muscle	Programme permettant de créer des alignements multiples à partir de séquences nucléiques ou protéiques.	Alignement de séquences
Gth	GenomeThreader : Alignement épissé d'une banque de cDNA ou de protéine contre une banque génomique (Licence uniquement pour les académiques)	Alignement de séquences
Apollo	Environemment d'annotation experte	Annotation
Blat	The BLAST-Like Alignment Tool : Recherche de similarité dans des banques de données	Recherche de similarités
SnoGPS	Recherche de gènes snoRNA H/ACA dans une séquence génomique	ARNnc
PhrapCross_match	Assemblage	Contig, assemblage
Consed	Consed est un outil pour la visualisation, l'édition et la finition des séquences assemblées avec phrap.	Assemblage
Cap3	Un outil d'assemblage de séquences nucléiques en contig.	Assemblage
PCAP	Nettoyage de séquences avant assemblage	Assemblage
Phylip	PHYLIP (PHYLogeny Inference Package), est un package de 34 programmes dédiés à la reconstruction phylogénétique.	Phylogénie
RepeatMasker	Recherche de zones à faible complexité	Répétition
Phred/Phrap	Analyse des données issus du séquenceur	Assemblage
PolyPhred	Analyse des données issus du séquenceur	SNP
SignalP	Prédiction de site peptidique	Protéines
TMHMM	Prédiction d'hélices transmembranaire	Protéines
Amos	A Modular, Open-Source whole genome assembler	Assemblage
Carthagene	Logiciel d'ordonnancement de marqueurs plus particulièrement dédié au problème de construction de cartes jointes pour des pedigrees simples (backcross, intercross, outbreds et hybrides irradiés, populations multiples).	Carte génétique
R-2.8.1	Logiciel pour des calculs statistiques	Statistiques
MUMmer3.21	Alignement de génomes complet	Alignement de séquences
wgs-5.3	Whole-Genome Shotgun Sequencing	Assemblage

III. Les logiciels de comparaison avec les banques de séquences

La taille sans cesse croissante des banques de séquences a nécessité l'élaboration d'algorithmes spécifiques pour effectuer la comparaison d'une séquence avec une banque de données car les algorithmes standards de comparaison entre deux séquences sont généralement trop longs sur des machines classiques. Leur but est de filtrer les données de la banque en étapes successives car peu de séquences vont avoir des similitudes avec la séquence comparée.

1. Le logiciel FASTA

L'algorithme est basé sur l'identification rapide des zones d'identité entre la séquence recherchée et les séquences de la banque. Cette reconnaissance est primordiale car elle permet de considérer uniquement les séquences présentant une région de forte similitude avec la séquence recherchée. On peut ensuite, à partir de la meilleure zone de ressemblance, appliquer localement à ces séquences un algorithme d'alignement optimal. Le logiciel regroupe en fait deux programmes de recherche avec les banques de données. Le premier est le programme FASTA qui possède une version nucléique et protéique et le deuxième est le programme TFASTA qui recherche une séquence protéique avec les séquences d'une base nucléique traduite dans les 6 phases.

a. Les différentes étapes de l'algorithme

- Pour chaque séquence de la banque, l'algorithme se déroule en quatre étapes sélectives distinctes qui permettent de cibler rapidement et précisément les régions intéressantes pour l'alignement optimal.

- La première étape consiste à repérer les régions les plus denses en identités partagées par les deux séquences. La codification numérique des séquences est ici utilisée (voir la recherche de segments identiques) avec une longueur des segments codés (k-tuple) qui est généralement de 1 ou 2 pour les protéines et de 4 à 6 pour les acides nucléiques. Cette étape confère à l'algorithme l'essentiel de sa rapidité.

- Dans une deuxième étape, on recalcule à l'aide d'une matrice de scores élémentaires un score pour les dix meilleurs régions d'identité trouvées dans l'étape précédente en considérant éventuellement des associations non exactes entre certains éléments des séquences. Pour les protéines, on utilisera ici une matrice de substitution comme la PAM250 de Dayhoff ou la BLOSUM50. Cette deuxième étape correspond donc à une recherche de similitudes sans insertion-délétion uniquement sur les régions de haute identité. Les scores obtenus correspondent à des régions initiales de premier ordre et l'on qualifie de score init1 celui qui représente la région de plus fort score parmi les dix analysées.

- La troisième étape essaie de joindre les régions définies à l'étape précédente, bien entendu s'il en existe au moins deux et si chacune de celles-ci possède un score supérieur à un score seuil prédéfini. Ce seuil correspond en fait à un score moyen attendu pour des séquences non apparentées. On réunira ces régions initiales à chaque fois que la somme de leurs scores diminuée d'une pénalité de jonction est supérieure ou égale au score init1. Ce score s'il existe est appelé initn et correspond à une région initiale de deuxième ordre.

- La quatrième étape consiste à effectuer l'alignement optimal de la séquence recherchée avec la séquence de la banque en considérant uniquement les parties des séquences délimitées par la meilleure région initiale de score initn (qui est égale à init1 s'il n'y a pas eu de jonction à l'étape 3). On obtient alors un score optimal dénommé opt. Cet alignement est effectué uniquement pour un nombre limité de séquences fixé par l'utilisateur. Ce sont les séquences qui correspondent aux plus hauts scores initiaux initn.

Une estimation statistique est donnée en traçant l'histogramme des meilleures scores obtenus pour chaque séquence de la banque avec le calcul de la moyenne et de l'écart type liés à cette distribution. Cette estimation utilise la théorie selon laquelle les similarités locales d'une séquence comparée avec une banque de données suit une distribution de valeurs extrêmes (voir par exemple Altschul et al.,1994).

b. Les qualités de l'algorithme

L'algorithme possède une bonne sensibilité du fait qu'il prend en compte les insertions-délétions. Ceci est fait en minimisant les explorations entre les deux séquences puisqu'on ne considère que les séquences potentiellement intéressantes pour effectuer l'étape de programmation dynamique, en ciblant de plus, les régions où l'on doit effectuer la recherche d'alignement. L'étape ultime d'alignement optimal est réalisée uniquement sur la meilleure région de haute similitude même si d'autres régions possèdent un score suffisant pour l'effectuer. Cela permet d'éviter en partie le bruit de fond dû à des motifs non significatifs et intrinsèques à la séquence recherchée mais a l'inconvénient de ne pas pouvoir considérer de grandes insertions durant l'alignement des séquences. Cette lacune est maintenant évitée dans la dernière version du logiciel (Octobre 1995) pour l'alignement des séquences protéiques. En effet celle-ci considère la totalité des séquences pour effectuer l'algorithme d'alignement local de Smith et Waterman (1981) plutôt que d'effectuer l'alignement global de Needleman et Wunsch (1970) uniquement sur des portions de séquences protéiques. L'édition des résultats est maintenant triée en fonction des scores opt contrairement aux premières versions qui considéraient les scores initiaux (initn), ce qui rendait parfois difficile la détection d'un alignement dont le score optimal est bon mais dont le score initial initn est médiocre. Enfin Pearson (1990) explique que lorsque le score opt est plus faible que le score initn, alors la similitude est souvent inintéressante.

L'estimation statistique est faite à partir des scores obtenus avec l'ensemble des séquences de la banque. Cependant, le logiciel fournit également des programmes d'estimation statistique basés sur une méthode de Monte Carlo (cf. l'évaluation des résultats) pour estimer la validité d'un score opt particulier entre une séquence de la banque et la séquence recherchée. Il s'agit des programmes PRDF et PRSS qui considèrent une distribution de valeurs extrêmes pour les scores aléatoires et qui sont directement inspirés du programme PRDF2 (Pearson, 1990) qui regroupe les séquences en courts segments pour effectuer les simulations. Le programme PRDF produit des simulations selon l'algorithme de Needleman et Wunsch appliqué localement pour l'étape d'alignement optimal alors que le programme PRSS utilise l'algorithme complet de Smith et Waterman entre deux séquences protéiques.

2. Le logiciel BLAST

L'intérêt de l'algorithme est que sa conception est basée sur un modèle statistique. Celui-ci a été établi d'après les méthodes statistiques de Karlin et Altschul (1990 ; 1993) qui s'appliquent aux comparaisons de séquences sans insertion-délétion. L'unité fondamentale de BLAST est le HSP (High-scoring Segment Pair). C'est un couple de fragments identifiés sur chacune des séquences comparées, de longueur égale mais non prédéfinie, et qui possède un score significatif. En d'autres termes, un HSP correspond à un segment commun, le plus long possible, entre deux séquences qui correspond à une similitude sans insertion-délétion ayant au moins un score supérieur ou égal à un score seuil. Un deuxième score MSP (Maximal-scoring Segment Pair) a été défini comme étant le meilleur score obtenu parmi tous les couples de fragments possibles que peuvent produire deux séquences. Les méthodes statistiques de Karlin et Altschul sont appliquées pour déterminer la signification biologique des MSPs et par extrapolation la signification des scores HSPs obtenus lors de la comparaison. Ce logiciel possède en fait quatre programmes distincts de comparaison avec les bases de données. BLASTN (séquence nucléique contre base nucléique), BLASTP (séquence protéique contre base protéique), BLASTX (séquence nucléique traduite en 6 phases contre base protéique), et

TBLASTN (séquence protéique contre base nucléique traduite en 6 phases).

a. Les différentes étapes de l'algorithme

La stratégie de la recherche consiste à repérer tous les HSPs (fragments similaires) entre la séquence recherchée et les séquences de la base. Pour déterminer un HSP, des mots de longueur fixe sont identifiés dans un premier temps entre la séquence recherchée et la séquence de la banque. Dans le cas des acides nucléiques, cela revient à des recherches d'identité entre les deux séquences sur des segments de longueur fixe (généralement 11). Par contre dans le cas des protéines, on effectue d'abord une liste de mots similaires pour chaque mot de longueur fixe (généralement 3) de la séquence recherchée et l'on repère ensuite dans la banque les séquences qui possèdent au moins un de ces mots. Un mot similaire est un mot qui, comparé avec un mot de la séquence recherchée, obtient un score supérieur à un score seuil lorsque l'on considère une matrice de substitution. Dans un deuxième temps, on cherche à étendre la similitude dans les deux directions le long de chaque séquence, à partir du mot commun, de manière à ce que le score cumulé puisse être amélioré.

L'extension s'arrêtera dans les trois cas suivants:

- Si le score cumulé descend d'une quantité x donné par rapport à la valeur maximale qu'il avait atteint.

- Si le score cumulé devient inférieur ou égal à zéro.

- Si la fin d'une des deux séquences est atteinte.

La signification des segments similaires obtenus est ensuite évaluée statistiquement. Celle-ci est faite en fonction de la longueur et de la composition de la séquence, de la taille de la banque et de la matrice de scores utilisée. Cette estimation donne en fait la probabilité que l'on a d'observer au hasard une similitude de ce score à travers la banque de séquences considérée. Lorsque plusieurs HSPs sont trouvées sur la même séquence, le programme utilise alors une méthode de « somme statistique » (Karlin et Altschul, 1993) qui considère que la signification statistique d'un ensemble de HSPs doit être plus élevée que n'importe quel HSP appartenant à cet ensemble. Les HSPs, dont la signification statistique satisfait une valeur seuil désignée par l'utilisateur sont ensuite édités.

b. Les qualités de l'algorithme

Le principal avantage est que le fondement de l'algorithme s'appuie avant tout sur des critères statistiques. Un autre point intéressant de la méthode (essentiellement pour les protéines) est que la première étape de reconnaissance des similarités ne recherche pas uniquement des zones d'identité mais accepte la présence de similitudes en considérant une matrice de scores. Ceci permet d'intégrer dès le début de la recherche les critères biologiques compris dans la matrice. De plus, les résultats peuvent être édités selon plusieurs tris possibles et en particulier selon leur signification statistique et non suivant la valeur de leur score. On retrouvera donc les segments les plus probables en début de liste. Ce logiciel a été très optimisé dans son écriture, notamment par une précodification de la banque, ce qui lui vaut d'être un des

plus rapides tout en conservant une sensibilité satisfaisante. De plus, il possède des versions qui s'exécutent sur machines parallèles.

Comme la recherche dans la banque de données est basée sur l'identification de segments, le bruit de fond est plus présent dans ce type d'approche. Il est généralement du à des qualités intrinsèques de la séquence recherchée comme la présence de régions répétées internes, ou la présence de segments de basse complexité non spécifiques d'une caractéristique biologique mais communs à plusieurs familles de protéines, par exemple les segments basiques ou acides. Des logiciels complémentaires qui opèrent comme filtres peuvent être utilisés comme paramètres dans les programmes BLAST pour remédier à ces problèmes. Ainsi, le programme SEG (Wootton et Federhen, 1993) masque des régions de faible complexité et le programme XNU (Claverie et States, 1993) cache des régions répétées de courte périodicité.

3. Logiciel de criblage de banque, AUTOMAT. (1991-2005)

a. Problématique biologique :

En 1991, cet axe de recherche portait sur l'infection par le VIH-1 : comment le virus fait-il pour interférer avec le bon fonctionnement du système immunitaire ? Pour répondre à cela, les concepteurs du logiciel se sont inspirés du fait que les rétrovirus, comme le VIH-1, peuvent prélever des fragments d'ADN ou d'ARN de leur hôte au cours de leur cycle de réplication dans l'hôte et ils ont supposé que ces séquences jouaient un rôle important dans l'effet nocif du virus sur le fonctionnement de l'immunité.

b. Traduction bioinformatique :

Le problème revenait donc à retrouver des séquences humaines au sein des séquences virales ADN ou protéines. En 1992, le web n'existait pas et BLAST n'était pas connu. Les concepteurs du logiciel ont développé leur propre logiciel de criblage de banques, appelé AUTOMAT, qui est aussi puissant que BLAST pour les analyses de protéines (Cantalloube et al, Bioinformatics, 1994 ; ibid, 1995), et plus puissant pour les analyses d'ADN .
Ils ont repris récemment les études sur AUTOMAT, et le logiciel est aujourd'hui disponible pour la communauté scientifique sur le serveur RPBS. Le logiciel Automat est utilisable directement sur le serveur Web http://bioserv.rpbs.jussieu.fr/RPBS/html/fr/T0_Home.html

4. Logiciel d'analyse génomique, SCAGEN. (1995-2005)

a. Problématique biologique :

Les concepteurs ont développé à partir de 1995 une approche génomique sur cohorte pour mieux comprendre les mécanismes de pathogenèse de l'infection par VIH-1 et par conséquent, pour pouvoir développer de manière rationnelle de nouvelles stratégies thérapeutiques ou diagnostic.

Ils ont en effet observé qu'il y avait des sujets séropositifs depuis plus de dix ans qui ne présentaient aucune altération de santé tant au niveau clinique qu'au niveau de leurs paramètres biologiques. Ces sujets ont été appelés non progresseurs. A l'opposé, ils ont pu observer qu'il y avait des sujets qui sont tombés malades moins de trois ans après avoir été contaminés, et qui ont été appelés progresseurs rapides. Ces sujets à profil extrême correspondent à 1% des

patients séropositifs. Pour expliquer ces différences de profil, il y a 3 types de facteurs : les facteurs génétiques du virus infectant, les facteurs environnementaux (autres infections, habitudes de vie), les facteurs génétiques de l'hôte. Étant en France où les modes de vie sont très semblables et les souches virales sont toutes de type B, il était clair que ces différences de profil étaient surtout liées aux facteurs génétiques de l'hôte.
Cette étude génomique sur cohorte (appelée projet GRIV) revenait donc à comparer la distribution des variations gènétiques dans les populations extrêmes, non progresseurs et progresseurs rapides. Ils ont constitué la plus grande cohorte au monde de patients extrêmes du SIDA avec 300 sujets non progresseur et 100 sujets progressseurs rapides. Grâce à la collaboration avec le Centre National de Génotypage à Evry, ils ont déjà pu analyser plusieurs milliers de variations génétiques au niveau des gènes de ces patients.

b. Traduction bioinformatique :

Afin d'exploiter la banque de données génétiques ainsi générée, ils ont développé le logiciel bioinformatique SCAGEN qui s'appuie à la fois sur des outils d'Informatique, de Génétique, et de Statistique, pour identifier les gènes qui interviennent dans ces profils de progression extrême (Hendel et al ; J Immunol, 1999 ; Flores-Villanueva et al, J Immunol, 2003).
La description détaillée du projet GRIV se trouve sur le site Web www.GRIV.org

5. Logiciel de design de nouveaux antigènes vaccinaux. (2001-2005)

a. Problématique biologique :

Les cytokines sont des protéines, messagers essentiels du système immunitaire, et leur surproduction peut être la cause de nombreuses maladies comme les maladies auto-immunes (sclérose en plaques, polyarthrite rhumatoïde, etc….) ou certains cancers.
Une façon de lutter contre ces maladies est de bloquer cette surproduction de cytokines, et cela peut être fait par l'administration passive d'anticorps ciblant ces cytokines. Cette approche passive a été essayée avec un grand succès dans la polyarthrite rhumatoïde et la maladie de Crohn.

Les concepteurs du logiciel ont développé une nouvelle stratégie, qui est basée non plus sur l'administration passive d'anticorps, mais sur une immunisation active contre la cytokine (en d'autres termes un vaccin « anti-cytokine »), qui fait que les anticorps anti-cytokine seront générés par l'organisme lui-même. Pour définir les morceaux de cytokine capables d'induire la production d'anticorps neutralisants (ces morceaux constituent le vaccin), ils ont utilisé une approche de modélisation moléculaire.

b. Traduction bioinformatique :

Les cytokines sont des protéines très importantes de la biologie, et beaucoup ont été cristallisées. Les concepteurs du logiciel ont utilisé les données structurelles tri-dimensionnelles des cytokines présentes dans la PDB (Protein Data Bank) pour définir des épitopes utilisables pour les vaccins anti-cytokine. Leur approche a notamment impliqué l'utilisation de logiciels développés en interne pour le calcul de distances spatiales entre les atomes des différents acides aminés.

6. Logiciel UMD-central

Les progrès réalisés dans le clonage de gènes impliqués dans les maladies, tant monogéniques que polygéniques, couplés à l'apparition de nouvelles méthodes d'identification des mutations, ont abouti à l'émergence d'un nouveau champ de la génétique : l'analyse des mutations. La nécessité d'une annotation précise des caractères génétiques, biochimiques et phénotypiques associés à chaque mutation ne pouvant être réalisés que par des experts (curators), a amené les concepteurs du logiciel à se spécialiser dans la création de banques de données spécifiques d'un locus (Locus Specific Databases ou LSDB).

Afin de fournir à la communauté scientifique un outil générique permettant, d'une part, de créer ces LSDBs et, d'autre part, de disposer d'un large éventail de fonctions d'analyse de ces données, ils ont créé le logiciel UMD® (Universal Mutation Database). Ce logiciel est aujourd'hui reconnu au plan international comme l'un des outils de référence par la société HGVS (Human Genome Variation Society) ainsi que l'organisation HUGO (Human Genome Organization), de plus Ce logiciel est distribué gratuitement à la communauté scientifique depuis 1998.

UMD est en constante évolution et les concepteurs du logiciel ont intégré ainsi de nouvelles fonctionnalités permettant notamment d'aborder la question de l'hétérogénéité allélique (un gène – plusieurs maladies) via des études de corrélations génotype-phénotype, d'intégrer l'ensemble des séquences non codantes du locus (introns, séquences régulatrices) mais également d'aborder des problèmes plus complexes comme l'aide à l'interprétation des mutations introniques. De plus ils ont intégré la gestion des images provenant par exemple d'hybridation sur coupes de tissus, de résultats de Western Blots… qui sont d'un grand apport pour les utilisateurs tant chercheurs que généticiens impliqués dans le diagnostic. Enfin, le développement du projet génome ainsi que celui du consortium international des SNPs (« Single Nucleotide Polymorphism ») ont fourni une grande quantité d'informations sur la structure du génome que nous extrayons et associons à chaque locus des LSDBs (séquences génomiques exhaustives, SNPs…).

Parallèlement, nous avons entrepris le développement d'un nouvel outil (UMD central) permettant d'effectuer des requêtes croisées entre ces LSDBs afin d'élargir leur domaine d'utilisation à celui de l'hétérogénéité génétique (une maladie – plusieurs gènes).

Aujourd'hui les seuls outils disponibles correspondent aux banques de données générales et aux banques de données spécifiques d'un gène. Dans le premier cas, l'hétérogénéité génétique ne peut être abordée par manque d'informations sur le phénotype clinique, dans le second par une trop grande spécificité (seules les informations sur un locus sont disponibles). Il était ainsi important de proposer un outil permettant de combler ce manque. L'approche des concepteurs du logiciel consiste à utiliser la richesse des informations phénotypiques disponibles dans les LSDBs et à les utiliser grâce à un outil capable de réaliser une interrogation croisée de l'ensemble des UMD-LSDBs. Pour cela, ils ont choisi de développer l'outil UMD-central à partir du langage 4D. Cet outil pourra communiquer avec l'ensemble des UMD-LSDB grâce à des process de communication installés dans chaque application. Un prototype a déjà démontré la faisabilité de cette approche. Afin d'optimiser les performances, nous avons choisi de maintenir un lexique de l'ensemble des termes cliniques employés dans les différentes LSDBs grâce à une procédure automatique de mise à jour des informations. Ainsi, lorsqu'un utilisateur se connecte à UMD-central via Internet, l'ensemble des termes de ce lexique lui est proposé afin de faciliter la saisie. Après avoir choisi sur quelles banques de données il souhaite

réaliser sa recherche, UMD-central limitera ses requêtes aux UMD-LSDBs intégrant les termes sélectionnés.

7. Logiciel MULTALIN : Alignement multiple de sequence

- Calcul de score pour l'alignement de séquences protéiques

D'un point de vue informatique, les problèmes de recherche de motifs et d'alignement de séquences de protéines semblent identiques aux problèmes correspondants pour les séquences nucléotidiques : seule change *a priori* la taille de l'alphabet. Pourtant au contact des biologistes on découvre qu'il n'en est rien et que les problématiques sont très différentes : les taux de similarité entre objets à comparer sont plus faibles et entrent en jeu les propriétés chimiques des acides aminés. Tout ceci conduit à des problèmes d'optimisation à critères multiples qui rendent les définitions de scores parfois assez arbitraires.

Dans ce cadre, les concepteurs du logiciel se sont concentrés sur des problèmes de score directement rencontrés dans les logiciels développés par l'équipe de l'IGBMC, en particulier sur l'amélioration du score utilisé par le logiciel Ballast . Il s'agit de passer pour ces applications précises d'une approche de type score, à une approche plus fondée statistiquement.

Ce logiciel traite un problème d'alignement local de séquences de protéines. Il retraite la sortie du logiciel standard BLAST qui recherche dans une base de données les séquences présentant une similarité avec une séquence *query*. Alors que BLAST utilise uniquement les qualités propres de chaque similarité pour calculer son score, Ballast se concentre sur des segments privilégiés (zones de la *query* rencontrant de nombreuses similarités dans la base) et néglige les similarités isolées, jugées peu importantes. Le score de Ballast est ainsi plus significatif que celui de BLAST mais, en l'absence de traitement statistique, n'est pas normalisé.

Ce groupe de chercheurs est en cours de la mise au point d'une nouvelle version de ce logiciel, au sein de laquelle la méthode de score est modifiée de façon à permettre une évaluation statistique approchée, sous forme de *p -values* associées aux scores. Ceci les a conduit, avec les concepteurs du logiciel IGBMC, à redéfinir plusieurs étapes du traitement et en particulier à traiter algorithmiquement et statistiquement le problème de la fragmentation des segments privilégiés et de leur redondance. Le nouvel algorithme ainsi défini est implanté et en cours de test à l'IGBMC. Un article décrivant les résultats obtenus est en préparation.

8. Logiciel MREPS

mreps est un logiciel de recherche de répétitions dites maximales dans les séquences d'ADN. Les répétitions maximales sont des répétitions successives, appelées parfois *périodicités* dans la littérature informatique et *répétitions en tandem* dans la littérature génomique. La naissance de *mreps*, il y a trois ans maintenant, a suivi les travaux théoriques dans lesquels nous avons proposé un algorithme très efficace (en temps linéaire) pour rechercher toutes les répétitions maximales exactes dans un texte.

Depuis, les concepteurs poursuivent le développement de ce logiciel à la fois sur le plan théorique et appliqué. La version de *mreps* diffusé au début de l'année 2002 était la version 2.1. Elle a été présentée, sous forme de poster, à la conférence RECOMB'2002 . Elle implantait l'algorithme de recherche de répétitions approchées. Cette année, des améliorations

considérables ont été apportées à cette version. L'objectif général de ces améliorations a été d'augmenter la souplesse du logiciel, afin d'identifier des répétitions plus « floues » mais toujours biologiquement pertinentes.

Premièrement, la notion d'erreur, qui permet de rechercher des répétitions en tandem avec des mismatch a été revue et modifiée. Le logiciel n'autorise plus un nombre donné d'erreur par motif d'une même répétition, il fonctionne désormais avec un taux d'erreur (ou flexibilité) qui joue sur le degré maximal du « flou » des répétitions trouvées.

Parmi d'autres modifications, un paramètre de score a été introduit, dont l'objectif est double : d'une part il informe l'utilisateur sur la qualité d'une répétition trouvée et d'autre part, il est utilisé dans l'algorithme pour écarter les répétitions avec un score statistiquement attendu, de sorte que celles sorties par le logiciel soient « significatives ». Cela marque un changement dans la philosophie de l'approche, à savoir un passage d'une approche purement combinatoire vers une approche mixte. Cette dernière est basée sur une recherche exhaustive de tous les « éléments de base » (répétitions calculées par la version 2.1 de *mreps*) suivie par un traitement statistique de ces éléments afin de former des répétitions plus floues et biologiquement pertinentes.

Une autre modification consiste à rechercher des répétitions significatives dont l'exposant est inférieur à deux, ce qui n'était pas possible avec la version 2.1. Cette propriété est intéressante, elle permet de rechercher des fragments répétés séparés par une distance bornée.

A ce jour, la version 2.1 de *mreps* est diffusée sous la licence GPL par plusieurs voies : depuis le serveur Web du Loria : http://www.loria.fr/mreps/ et depuis la page des logiciels libres de l'INRIA(http://www.inria.fr/valorisation/logiciels/index.fr.html), ainsi que depuis le serveur *Collaborative Computational Project 11* (http://www.hgmp.mrc.ac.uk/CCP11/index.jsp) domicilié au *UK Human Genome Mapping Project Resource Centre* (http://www.hgmp.mrc.ac.uk/). mreps a également été déposé à l'APP. mreps est interrogeable via une interface Web depuis sa page de distribution ; il est également installé sur le serveur bioweb de l'Institut Pasteur (http://bioweb.pasteur.fr/seqanal/interfaces/mreps.html), serveur qui fournit une interface Web à plusieurs outils bioinformatiques existants. La dernière version stable de *mreps* est la 2.4.3 et la version 2.5 est en cours de finition.

9. Logiciel YASS

Le logiciel YASS *(Yet Another Similarity Searcher)* a été mis au point afin de rechercher les régions de similitude entre séquences génomiques. Une première version devrait être disponible courant décembre. Il se veut plus sensible que BLAST, et cela grâce à la recherche et au chaînage de mots plus petits qu'il regroupe à l'aide de deux critères statistiques issus du taux de mutations (substitutions et *indels*) de la séquence. Il travaille actuellement sur des données au format FASTA et indique la liste des similitudes trouvées classées par significativité, donne leur *e -value*, et éventuellement l'alignement observé.

Le programme a été développé en C Ansi sous Unix, a été testé sous Linux et Windows sur des chromosomes de levure (*Saccharomyces cerevisiae*). Les résultats ont été comparés à ceux d'autres logiciels tels que BLAST-NCBI, BLAT, et REPuter : le logiciel produit des alignements de qualité supérieure à REPuter sans donner de résultats redondants. Ses résultats

sont plus complets que BLAT ; il parvient même à trouver certaines répétitions significatives que BLAST ne peut distinguer.

- **Application : Recherche de régions de similitude dans les séquences d'ADN**

Les génomes de nombreuses espèces sont désormais disponibles (ou en cours de séquençage), et la comparaison de segments d'ADN ou de chromosomes complets est une des méthodes les plus fréquemment employées.

Cette comparaison est menée à bien, soit pour l'étude des homologies à des fins phylogénétiques, soit pour éventuellement localiser, sur un génome fraîchement séquencé, des loci susceptibles de jouer le même rôle que ceux d'espèces déjà plus largement étudiées. Elle peut avoir également pour but de trouver des éléments mobiles dans un génome, ou d'identifier des régions polymorphes en analysant les génomes de différents représentants d'une même espèce, ou dans bien d'autres situations.

De nombreux logiciels permettent la comparaison séquence à séquence (éventuellement contenues dans une base de données), afin de trouver des régions de similitude significatives, en ce sens qu'elles ont peu de chance de se produire par hasard. Il s'agit alors de trouver des alignements dits locaux dont les scores permettent de les distinguer nettement des alignements parasites dus au hasard.

L'algorithme classique pour cette tâche est celui de Smith et Waterman, qui présente l'avantage d'être exhaustif et l'inconvénient d'être gourmand en temps de calcul. En pratique, les algorithmes du type BLAST ou FASTA sont très largement utilisés. Ces programmes sont basés sur des heuristiques leur permettant d'éviter de considérer tout l'espace de recherche. Le principe général de ces heuristiques est de rechercher des sous-séquences répétées de manière exacte pour déterminer des zones susceptibles d'être des copies approchées d'un même fragment.

L'objectif des concepteurs de ce logiciel est ici, d'améliorer la sensibilité de tels algorithmes en s'intéressant d'une part aux propriétés statistiques des événements qui transforment les séquences (indels, mutations ponctuelles), et d'autre part à l'information conservée permettant de retrouver ces répétitions.

10. Logiciel GENETIX

Ceci est la version 4.05 du logiciel GENETIX. Cet ensemble de programmes, fonctionnant sous Windows95 et Windows NT, permet le calcul d'un ensemble de paramètres couramment utilisés en génétique des populations, et propose l'étude de leur signification statistique par l'emploi de tests de ré-échantillonnage de type permutations.

Pour chaque paramètre considéré, la distribution sous l'hypothèse nulle correspondante (par exemple l'équilibre de Hardy-Weinberg pour les F de Wright) est générée par une technique de ré-échantillonnage appropriée (par ex. permutation des allèles entre individus dans le cas du Fils). De ce fait, cette approche est préférable aux test paramétriques usuels parce que les lois des estimateurs sont en général mal connues, et parce que les niveaux de polymorphisme élevés qui sont rencontrés dans les études impliquant des locus hypervariables produisent des tableaux de contingence avec peu d'observations dans chaque case. Les procédures d'inférence statistique par permutation proposées par GENETIX constituent une alternative aux ré-

échantillonnages de type « bootstrap » ou « jackknife » ainsi qu'aux tests exacts quand ces derniers sont disponibles. Le bootstrap et le jackknife permettent une estimation de l'intervalle de confiance autour de la valeur observée, alors que les tests par permutations et les tests exacts estiment la probabilité d'écart à zéro sous l'hypothèse nulle.

Une adaptation du test de Mantel à l'étude des corrélations entre matrices de distances est également disponible. GENETIX contient également un programme d'analyse factorielle des correspondances adapté aux génotypes diploïdes et fournissant des sorties graphiques en 3D.

Conception du logiciel : GENETIX est organisé autour d'un éditeur de fichiers de données (tableur) analogue à une feuille de calcul Excel. Ainsi, il existe un seul fichier de base (NOMFICH.GTX) regroupant les données génotypiques individuelles multilocus pour l'ensemble des individus de toutes les populations à tous les locus. Un certain nombre de menus déroulants sont proposés, qui permettent d'effectuer les traitements dont les détails sont donnés ci-après. Pour la plupart des traitements (FSTATS, LINKDIS, Distances génétiques), GENETIX donne le choix entre un simple calcul de paramètres, ou bien la réalisation d'un test de ré-échantillonnage par permutations. Pour un traitement donné, il est possible de sélectionner des sous-ensembles de populations et / ou de locus ; choisir alors l'option de traitement affichée avant son lancement.

-
La Programmation : GENETIX 4.05 a été développé à l'aide du logiciel Delphi TM, version 5.0.

-
La Configuration matérielle : Sont requis un ordinateur compatible IBM possédant un système d'exploitation Windows 95, Windows NT ou versions ultérieures, ainsi que les caractéristiques minimales suivantes : espace-disque disponible > 4 Mo ; RAM > 8 Mo ; processeur de type Intel 386 ou supérieur.

IV. Les logiciels de traitement des données biologiques

1. Le logiciel R

R est à la fois un logiciel de statistique et un langage de programmation.

R est un logiciel de traitement statistique des données. Il fonctionne sous la forme d'un interpréteur de commandes. Il dispose d'une bibliothèque très large de fonctions statistiques, d'autant plus large qu'il est possible d'en intégrer de nouvelles par le système des "packages", des modules externes compilés (sous forme de DLL sous Windows) que l'on peut télécharger gratuitement sur internet. R propose également une palette étendue de fonctionnalités graphiques. Il est possible d'utiliser R en mode intercatif sans jamais avoir à programmer.

R est un langage de programmation (de script) interprété dérivé de S (disponible dans le logiciel S-PLUS). A ce titre, il en intègre toutes les caractéristiques : données simples et structurées, opération d'entrée-sortie, branchements conditionnels, boucles indicées et conditionnelles, récursivité, etc. En particulier, il nous sera possible de créer de nouvelles fonctions de traitement de données avec le langage R.

R et RStudio sont deux logiciels distincts :

- R est un langage de programmation particulièrement puissant pour l'exploration, la visualisation et l'analyse statistique des données
- RStudio est un environnement de développement intégré (IDE) qui facilite l'utilisation de R.

La version de base de R est livrée avec une large gamme de fonctions à utiliser dans le domaine de l'environnement. Cette fonctionnalité est complétée par une pléthore de packages disponibles via CRAN, qui fournissent des méthodes spécialisées telles que les techniques d'ordination et d'analyse de cluster. Un bref aperçu des packages disponibles est fourni, regroupés par sujet ou type d'analyse.

a. Packages généraux

Ces packages sont généraux, ayant une large applicabilité au domaine de l'environnement. Le package **EnvStats** est le successeur du module S-PLUS EnvironmentalStats, tous les deux créés par Steven Millard. Un guide d'utilisation sous forme de livre vient de paraître : https://dx.doi.org/10.1007/978-1-4614-8456-1.

b. Modélisation des réponses des espèces et autres données

L'analyse des courbes de réponse des espèces ou la modélisation d'autres données implique souvent l'ajustement de modèles statistiques standard aux données écologiques et comprend une régression simple (multiple), des modèles linéaires généralisés (GLM), une régression étendue (par exemple, des moindres carrés généralisés [GLS]), des modèles additifs généralisés (GAM) et des modèles à effets mixtes, entre autres.

- L'installation de base de R fournit **lm()** et **glm()** pour l'ajustement des modèles linéaires et linéaires généralisés, respectivement.
- Les modèles des moindres carrés généralisés et les modèles à effets mixtes linéaires et non linéaires étendent le modèle de régression simple pour tenir compte du regroupement, de l'hétérogénéité et des corrélations au sein de l'échantillon d'observations. Le package **nlme** fournit des fonctions pour adapter ces modèles. Le package est pris en charge par Pinheiro & Bates (2000) Modèles à effets mixtes en S et S-PLUS, Springer, New York. Une approche mise à jour des modèles à effets mixtes, qui s'adapte également aux modèles mixtes linéaires généralisés (GLMM) et aux modèles mixtes non linéaires généralisés (GNLMM) est fournie par le package **lme4**, bien qu'il s'agisse actuellement d'un logiciel bêta et ne permet pas encore de corrélations au sein de l'erreur. structure.
- Le package recommandé **mgcv** s'adapte aux GAM et aux modèles mixtes additifs généralisés (GAMM) avec sélection automatique de lissage via une validation croisée généralisée. L'auteur de mgcv a également écrit une monographie d'accompagnement, Wood (2006) Generalized Additive Models ; Une introduction avec R Chapman Hall/CRC, qui est accompagné d'un package **gamair**.
- Alternativement, le package **gam** fournit une implémentation de la fonction S-PLUS **gam()** qui inclut les lissages LOESS.
- Les modèles de cotes proportionnelles pour les réponses ordinales peuvent être ajustés à l'aide de **polr()** dans le package MASS de Bill Venables et Brian Ripley.

- Une famille binomiale négative pour les GLM pour modéliser la surdispersion dans les données de comptage est disponible dans MASS.
- Modèles pour les comptes et les proportions surdispersés
  - ✓ Le package **pscl** contient également plusieurs fonctions pour gérer les données de comptage trop dispersées. Des distributions binomiales de Poisson ou négatives sont fournies à la fois pour les modèles gonflés à zéro et les modèles de haies.
  - ✓ **aod** fournit une suite de fonctions pour analyser les comptes ou les proportions surdispersés, ainsi que des fonctions utilitaires pour calculer, par ex. Poids AIC, AICc, Akaike.
- La détection des points de changement et des changements structurels dans les modèles paramétriques est bien prise en charge dans le package segmenté et le package **strucchange** respectivement. segmenté est discuté dans un article de R News (R News, volume 8 numéro 1).

c. Autres analyses :

- Modèles arborescents
- Ordination
- Cluster analysis
- Théorie écologique
- Les dynamiques de population
  - ➢ Estimation de l'abondance des animaux et des paramètres associés
  - ➢ Modélisation des taux de croissance démographique
  - ➢ Séries temporelles environnementales
- Analyse de données spatiales
- La science du sol
- Hydrologie et océanographie
- Paléoécologie et données stratigraphiques

2. Le logiciel SPSS

La plate-forme logicielle IBM® SPSS® offre une analyse statistique avancée, une vaste bibliothèque d'algorithmes d'apprentissage automatique, une analyse de texte, une extensibilité open source, une intégration avec le Big Data et un déploiement transparent dans les applications.

Après l'exécution d'une expérience et la collecte des données, le biologiste, avoir besoin de convertir les nombres en information ; il doit donc trouver un moyen pour choisir l'hypothèses la plus proche à la vérité. Les tests statistiques sont la méthode préférée pour le faire, et des logiciels comme SPSS facilitent grandement la réalisation de ces tests.

SPSS est un programme puissant qui offre de nombreuses façons d'examiner rapidement des données scientifiques. SPSS peut offrir des statistiques descriptives de base, telles que des moyennes et des fréquences, ainsi que des tests avancés tels que l'analyse de séries chronologiques et l'analyse multivariée. Le programme est également capable de fournir des graphiques et des tableaux de haute qualité. Savoir comment faire fonctionner le programme rendra les travaux de recherche plus faciles et plus sophistiqués.

Ce programme va fournir au chercheur :

- Entrer le data
- Visionner les statistiques
- Ouvrir le data à partir d'un fichier Excel
- Exporter les résultats en fichier Excel

Parmi les analyses qu'ils peuvent être réalisé avec SPSS, on note :

- Analyses descriptives, ex : la fréquence
- Comparaison des moyennes, ex : test T pour échantillon apparié
- Analyse univariée de la variance (ANOVA)
- Mise en corrélation de variables, corrélations bivariés et partielles
- Analyse de la régression
- Tests non paramétriques, ex : Khi-deux

3. Le logiciel PAST

Un progiciel complet mais simple à utiliser pour exécuter une gamme d'analyses et d'opérations numériques standard utilisées en paléontologie quantitative. Le programme, appelé PAST (PAleontological STatistics), fonctionne sur des ordinateurs Windows standard et est disponible gratuitement. PAST intègre :

- La saisie de données de type tableur avec des statistiques univariées et multivariées,
- L'ajustement de courbes,
- L'analyse de séries chronologiques,
- Le traçage des données
- Une analyse phylogénétique simple.

De nombreuses fonctions sont spécifiques à la paléontologie et à l'écologie, et ces fonctions ne se trouvent pas dans les progiciels statistiques standard plus compliqués. PAST comprend également quatorze études de cas (fichiers de données et exercices) illustrant l'utilisation du programme pour des problèmes paléontologiques, ce qui en fait un ensemble pédagogique complet pour les cours de méthodes quantitatives.