La bioinformatique

1- **Bioinformatique : définitions, descriptions, démarche et principales étapes.**

La bioinformatique est une discipline à part entière qui utilise un ensemble de techniques et d'outils issus des mathématiques, de la physique et de l'informatique dans le but d'analyser des données biologiques.

Ces données biologiques concernent essentiellement deux types de macromolécules : les acides nucléiques et les protéines.

Des données d'autres types sont maintenant générées, en particulier les données textuelles (ontologie - annotation) ou des mélange de très grands jeux de données (métabolomique).

Les énormes volumes de données biologiques ne peuvent évidemment pas être traités manuellement. Ils nécessitent l'utilisation de l'ordinateur pour le traitement automatique de l'information. Les données biologiques sont regroupées dans un grand nombre de bases de données (généralistes ou spécialisées), utilisables via Internet et le Web.

Les récents progrès des programmes de séquençage des génomes ont ouvert des voies dans les domaines de la biologie, de la santé et de l'agronomie.

La génomique s'intéresse à l'étude exhaustive des génomes : elle en analyse la structure afin d'identifier les gènes et les régions qui régulent l'expression de ces gènes. Cette régulation est assurée par la fixation de facteurs de transcription sur des régions particulières des gènes. Il en résulte un très grand nombre d'interactions protéine/ADN et protéine/protéine.

La transcriptomique s'intéresse à l'étude exhaustive des transcrits (ARN messagers et ARN non codanst). Elle s'appuie très largement sur les nouvelles technologies de séquençage à trés haut débit, en particulier la technique RNAseq, les méthodes SAGE, les puces à ADN

La protéomique étudie l'ensemble des protéines contenues dans une cellule (le protéome). Elle s'articule autour de trois thèmes majeurs : la prédiction de structures, la relation structure fonction et la phylogénie.

L'étude des réseaux d'interactions entre molécules (protéine/ADN, protéine/protéine, protéines/substrats, protéines/effecteurs). Le but est de décrire le fonctionnement global d'une cellule dans un environnement donné.

La modélisation moléculaire pour la conception de médicaments et l'étude de l'interaction entre macromolécules.

En conclusion :

La bioinformatique permet au biochimiste d'exploiter le potentiel de connaissances contenues dans les banques de données et de les analyser.

Elle permet à l'informaticien de mettre en œuvre ses compétences en algorithmique et en programmation, en développement d'interface et de bases de données.

Enfin, si l'informatique permet d'accéder aux données, les statistiques permettent d'en analyser le contenu

Définition:

La bioinformation est l'information liée aux molécules biologiques : leur séquence, leur nombre, leur(s) structure(s), leur(s) fonction(s), leurs liens de "parenté", leurs interactions et leur intégration dans la cellule ...

Cette bioinformation émane de diverses disciplines : la biochimie, la génétique, la génomique structurale, la génomique fonctionnelle, la transcriptomique, la protéomique, la biologie structurale (structure spatiale des molécules biologiques, modélisation moléculaire ...), ...

- **La bioinformatique** est l'analyse de la bioinformation par des moyens informatiques.

Definition du NCBI (2001): "Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline."

Description générale

Discipline récente (quelques dizaines d'années).

Discipline hybride : elle est fondée sur des concepts et des formalismes issus de la biologie, de l'informatique, des mathématiques et de la physique, de la chimie (techniques de séquençage, ...).

Discipline qui utilise tout le potentiel de traitement de l'informatique : modèles théoriques, algorithmes et programmes, bases de données, ordinateurs, réseau Internet, protocoles de communication, ...

Démarche

1. Compilation et organisation des données biologiques dans des bases de données: ces bases de données sont soit généralistes (elles contiennent le plus d'information possible sans expertise particulière de l'information déposée), soit spécialisées autour de thèmes précis.

2. Traitements systématiques des données : l'un des objectifs est de repérer et de caractériser une fonction et/ou une structure biologique importante. Les résultats de ces traitements constituent de nouvelles données biologiques obtenues "in silico".

3. Elaboration de stratégies : apporter des connaissances biologiques supplémentaires en combinant les données biologiques initiales et les données biologiques obtenues "in silico".

ces connaissances permettent, à leur tour, de développer de nouveaux concepts en biologie. Concepts qui, pour être validés, peuvent nécessiter le développement de nouvelles théories et outils en mathématiques et en informatique.

Principales étapes en biologie moléculaire, en informatique et en bioinformatique

1965	Première compilation de protéines ("Atlas of Protein Sequences"): *Margaret Dayhoff et al.*
1967	Article : "Construction of Phylogenetic Trees" - Fitch & Margoliash
1970	Algorithme pour l'alignement global de séquences : Saul Needleman & Christian Wunsch
1971	Premier microprocesseur Intel 4004
1972	Clonage de fragments d'ADN dans un virus, l'ADN recombiné : Paul Berg, David Jackson, Robert Symons
1973	Découverte des enzymes de restriction qui coupe spécifiquement l'ADN. Méthode de transfection (introduction d'un ADN étranger) des cellules eucaryotes grâce à un virus (vecteur).
1974	Programme de prédiction de structures secondaires des protéines : "Prediction of Protein Conformation" - Chou & Fasman. Vint Cerf et Robert Khan développent le concept des réseaux reliant des ordinateurs au sein d'un « internet » et développent deux protocoles fondamentaux "Transmission Control Protocol" (TCP) et "Internet Protocol" (IP).
1977	Micro-ordinateurs Techniques de séquençage d'ADN : Frederick Sanger / Maxam & Gilbert
1978-1980	**Mutagénèse dirigée**: Michael Smith Séquençage du 1er génome à ADN, le bactériophage phiX174 : Frederick Sanger **Premières bases de données** : EMBL, GenBank, PIR Accès téléphonique à la base de données PIR
1981:370.000 nucléotides GenBank:270 séquences	Micro-ordinateur IBM-PC 8088 Programme d'alignement local de séquences : Temple Smith & Michael Waterman
1983	IBM-XT disque dur (10 Mb)
1984	Amplification de l'ADN : réaction de polymérisation en chaîne (PCR) - Karry Mullis MacIntosh : interface graphique & souris
1985	"FASTA" : Programme d'alignement local de séquences - David Lipman & William Pearson
1987	Nouveau vecteur permettant de cloner des fragments d'ADN 20 fois plus grands : le YAC (Yeast Artificial Chromosome) qui rend possible l'étude de grands génomes.
1988	Taq polymérase, enzyme thermostable pour la PCR. Création du "National Centre for Biotechnology Information" (NCBI).
1989	INTERNET succède à ARPANET
1990	Clonage positionnel et premier essai de thérapie génique. "BLAST" : Programme d'alignement local de séquences - Altschul et al.
1991	"Expressed Sequences Tags" (EST) : méthode rapide d'identification des gènes (C. Venter).
1992	Séquençage complet du chromosome III de levure
1993	"European Bioinformatics Institute" (EMBL). Création à terme du "European

	Bioinformatics Institute" (EMBL - EBI).
1995	Analyse du transcriptome : début des puces à ADN
1996	Séquençage complet de la levure (consortium européen)
1997	11 génomes bactériens séquencés Evolutions de BLAST : "Gapped BLAST" et "PSI-BLAST"
1998	Séquençage de 2 millions de nucléotides par jour. Interférence ARN.
2000	Séquençage du 1er génome de plante : Arabidopsis thaliana
2001	Séquence "premier jet" complète du génome humain
Année 2000	Epigénétique: développement de technologies d'analyse des modifications de l'ADN et des histones. Accès aux revues et journaux scientifiques : développement de l'open access". Montée en puissance de la biologie synthétique. Détermination de structures de systèmes biologiques de plus en plus complexes (ribosomes, spliceosome, virus, ...) - cryo-microscopie électronique et autres techniques ("femtosecond pulses / X-ray free-electron laser")
2007-2008	Avènement des nouvelles technologies de séquençage à très haut débit, dites de "seconde génération". Prise de conscience du phénomène "big data" (pas seulement en biologie) qui devient peu à peu une discipline scientifique.
Février 2015 : plus de 1.143.000.000.000 nucléotides	Plus de 18.900 génomes eucaryotes et procaryotes séquencés et des milliers en projet (Genomes OnLine).
	développement de la banque de données EMBL (banque européenne créée en 1980). développement de la banque de données Genbank (créée en 1982 et diffusée par le National Center for Biotechnology Information).

La bioinformation : molécules support, types et obtention

1. Deux types de molécules support de la bioinformation : les acides nucléiques et les protéines
2. Deux types de bioinformation : la séquence des nucléotides et la séquence des acides aminés
3. L'obtention des séquences

1. Deux types de molécules support de la bioinformation : les acides nucléiques et les protéines

ADN: Acide DésoxyriboNucléique	**ARN : Acide RiboNucléique**	**Protéine**
- macromolécule : chaîne nucléotidique - constituée par un enchaînement d'unités élémentaires : les désoxyribonucléotides - forme de stockage de l'information génétique. Cette information est représentée par une suite linéaire de gènes - formée de deux brins complémentaires enroulés en double hélice ce qui lui permet de se dupliquer en deux molécules identiques entre elles et identiques à la molécule mère On distingue : • l'ADN du génome du noyau • l'ADN du génome mitochondrial • l'ADN du génome chloroplastique	- macromolécule: chaîne nucléotidique - constitué par un Enchaînement d'unités élémentaires : les ribonucléotides -forme qui permet de transférer et de traiter l'information dans la cellule - le plus souvent formé d'un simple brin On distingue : - les ARN messagers ou ARNm : ils sont transcrits à partir d'un gène (ADN). Ils sont ensuite traduits en protéines. -les ARN de transfert - les ARN ribosomaux - les ARN nucléaires - les ARN cytoplasmiques	- macromolécule : chaîne polypeptidique - constitué par un Enchaînement d'unités élémentaires : les acides aminés - l'ensemble des protéines assurent les principales fonctions cellulaires - se replie sur elle même et adopte une conformation ou structure particulière dans l'espace. Cette structure tridimensionnelle est à l'origine de la fonction de la protéine et de la spécificité de cette fonction.

2. Deux types de bioinformation : la séquence des nucléotides et la séquence des acides aminés

Les chaînes nucléotidiques (ADN, ARN) et les chaînes polypeptidiques (protéines) sont des polymères d'unités élémentaires :

ADN : 4 désoxyribonucléotides = dCMP, dGMP, dAMP, dTMP

ARN : 4 ribonucléotides = CMP, GMP, AMP, UMP

l protéines : (essentiellement) 20 acides aminés = Ala (A), Cys (C), Asp (D), Glu (E), Phe (F), Gly (G), His (H), Ile (I), Lys (K), Leu (L), Met (M), Asn (N), Pro (P), Gln (Q), Arg (R), Ser (S), Thr (T), Val (V), Trp (W), Tyr (Y)

Elles possèdent 2 extrémités distinctes et sont donc orientées :

--- de l'extrémité dite 5' vers l'extrémité dite 3' pour les chaînes nucléotidiques

- de l'extrémité dite N-terminale vers l'extrémité dite C-terminale pour les chaînes-polypeptidiques

En conséquence :

- les chaînes nucléotidiques et polypeptidiques sont une succession ordonnée et orientée d'unités élémentaires

- les séquences sont leur transcription sous forme d'une succession ordonnée et orientée de lettres qui correspondent à ces unités élémentaires

Exemple de séquence nucléotidique	Exemple de séquence polypeptidique
AATTCCGGCA TAGAAACTCA AATCAAAGAG GAAGAAACAC CGATTCTCCT TTTCTCTCTC TAAACAACTA GATCAGATCT CTGAGTTTAA GGAAGCTTTC AGCCTATTCG ATAAGGATGG CGATGGTTGC ATCACAACCA AGGAGCTTGG AACTGTTATG CGATCATTGG GACAAAACCC AACTGAAGCA	MADQLTDDQI SEFKEAFSLF DKDGDGCITT KELGTVMRSL GQNPTEAELQ DMINEVDADG NGTIDFPEFL NLMARKMKDT DSEEELKEAF RVFDKDQNGF ISAAELRHVM TNLGEKLTDE EVDEMIREAD VDGDGQINYE EFVKVMMAK

Les séquences constituent l'un des principaux types de bioinformation qu'analyse la bioinformatique.

Exemples d'autres types de bioinformation (directe ou obtenue "in silico")

Les structures tridimensionnelles des protéines et aussi, malgré leur nombre plus restreint, des acides nucléiques (en particulier les ARN de transfert).	Protein Data Bank
Les données obtenues en protéomique (gels d'électrophorèse bidimensionnel).	SWISS-2DPAGE
Le changement d'un nucléotide dans un gène quelconque ("Single Nucleotide Polymorphism").	SNP
La taxonomie (classification) des organismes.	Taxonomy
Les réseaux d'interactions qu'établissent les molécules biologiques.	BioCarta
Les voies métaboliques	KEGG

L'ontologie : l'organisation hiérarchique de la connaissance sur un ensemble d'objets par leur regroupement en sous-catégories suivant leurs caractéristiques essentielles.	GO
Les données bibliographiques (diffusion des résultats de la recherche par les articles).	PubMed

3. L'obtention des séquences

Séquence des nucléotides

- par la méthode de F. Sanger (1977) au départ
- puis par des techniques de plus en plus sophistiquées, automatisées et de masse

Séquence des acides aminés

- par la méthode de P. Edman (1950) au départ
- puis par traduction "in silico" des séquences nucléotidiques

Le stockage de la bioinformation : les bases de données et les banques de données

Les fichiers contenant l'information biologique sous la forme de séquences est l'élément central autour duquel les banques de données se sont constituées à l'origine.

On peut distinguer :

- les bases de données généralistes : elles correspondent à une collecte des données la plus exhaustive possible et qui offrent un ensemble plutôt hétérogène d'informations

- **les bases de données spécialisées** : elles correspondent à des données plus homogènes établies autour d'une thématique et qui offrent une valeur ajoutée

Il existe un grand nombre de bases de données d'intérêt biologique : voir une liste quasi exhaustive avec les liens vers les bases de données.

1. Les banques généralistes Les banques généralistes sont indispensables à la communauté scientifique car elles regroupent des données et des résultats essentiels dont certains ne sont plus reproduits dans la littérature scientifique.

- **NCBI** ("National Center for Biotechnology Information")
- **EBI** ("European Bioinformatics Institute")
- **Uniprot**
- **PDB** ("Protein Data Bank")

Dans le cadre de l'analyse des séquences, par exemple, le fait que la majorité des séquences connues soit réunie en un seul ensemble est un élément fondamental pour la

recherche de similitudes avec une nouvelle séquence. D'autre part, la grande diversité d'organismes qui y est représentée permet d'aborder des analyses de type évolutif.

La principale mission des banques généralistes est de **rendre publiques les séquences et tout autre type d'information.** Cette notion de mise à la disposition du public a été capitale dans le cas par exemple de la diffusion des résultats du séquençage du génome humain.

On y trouve également de l'information qui accompagne les séquences (annotations, bibliographie, ...) et une expertise biologique directement liées aux séquences traitées.

La présence de références à d'autres bases permet d'avoir accès à d'autres informations. Les multiples liens entre les groupes de données dans les banques généralistes sont d'une complexité étonnante.

La qualité des données contenues dans ces bases présente un certain nombre de lacunes. Les organismes responsables de la maintenance de ces banques ont pris conscience de la nécessité de vérifications des données soumises ou saisies (surtout pour les séquences anciennes).

Maintenant, de nombreuses vérifications sont faites systématiquement dès la soumission de la séquence : c'est la "curation".

Il existe désormais un recueil de séquences référencées, annotées et contrôlées : The Reference Sequence (RefSeq) collection

Exemples de grandes bases de données généralistes

EMBL - EBI : Banque européenne créée en 1980 et financée par l'EMBO (European Moleculary Biology Organisation). Elle est aujourd'hui diffusée par l'EBI ("European Bioinformatics Institute", Cambridge - UK).

Genbank - NCBI : Créée en 1982 par la société IntelliGenetics et diffusée maintenant par le NCBI ("National Center for Biotechnology Information", Bethesda - Marylandet).

DDBJ ("DNA Data Bank of Japan") : Créée en 1986 et diffusée par le NIG ("National Institute of Genetics", Japon).

Swissprot & TrEMBL : Elle a été constituée à l'Université de Genève à partir de 1986. Elle est maintenant développée par le SIB (Swiss Institute of Bioinformatics) et l'EBI. Elle regroupe (entre autres) des séquences annotées de la PIR-NBRF ainsi que les séquences codantes traduites de l'EMBL (TrEMBL).

Ces banques s'échangent systématiquement leur contenu depuis 1987 et adoptent un système de conventions communes (The DDBJ/EMBL/GenBank Feature Table Definition).

PIR-NBRF ("Protein Information Ressource") : banque de protéines créée sous l'influence du NBRF ("National Biomedical Research Foundation") à Washington. Elle diffuse maintenant des données issues du MIPS ("Martinsried Institute for Protein Sequences"), de la base Japonnaise JIPID ("Japan International Protein Information Database") et des données propres de la NBRF.

UniProt ("Universal Protein Resource") : c'est la base de données des protéines : ExPASy Proteomics Server. Consortium [EBI - SIB - PIR]

GOLD ("Genomes OnLine Database") : base de données qui recense les milliers de génomes séquencés ou en voie de séquençage.

"The Quick Guide" : autre base de données qui recense des génomes séquencés (descriptions des organismes, liens vers les centres de séquençage et vers la bibliographie).

2. Les banques spécialisées

Pour des besoins spécifiques liés à l'activité d'un groupe de personnes, ou encore par compilations bibliographiques, de nombreuses bases de données spécifiques ont été créées au sein des laboratoires. Certaines sont inconnues ou mal connues et attendent qu'on les exploite davantage.

Les bases de données spécialisées sont d'intérêt divers et la masse des données qu'elles contiennent peut varier d'une base à une autre. Ces bases correspondent à des améliorations ou à des regroupements par rapport aux données issues des bases généralistes.

Exemples de banques spécialisées

Late Embryogenesis Abundant Proteins database (LEAPdb - G. Hunault & E. Jaspard): cette base de données contient un grand nombre d'informations sur les proéines LEA impliqués dans la tolérance à de nombreux stress, notament la déshydratation et le froid.

Pour l'instant, on les a mises en évidence principalement chez les plantes.

Small Heat Shock Proteins database (sHSPdb - G. Hunault & E. Jaspard) : cette base de données contient un grand nombre d'informations structurales sur les cystéines de plus de 400 protéines cristallisées. Elle a aussi pour but de servir à la mise au point d'un logiciel de prédiction des cystéines impliquées dans la formation de pont disulfure.

RESID Database : Base de données sur les acides aminés peu fréquents (sous-partie de la base de données PIR)

Les bases de motifs

L'utilisation de bases spécialisées comme les bases de motifs est devenue un outil essentiel dans l'analyse des séquences pour tenter de déterminer la fonction de protéines inconnues ou savoir à quelle famille appartient une séquence non encore caractérisée.

a. Les bases de motifs nucléiques

La plupart de ces bases consiste à recenser dans des catalogues les séquences des différents motifs pour lesquels une activité biologique a été identifiée. Certains motifs sont simples et non ambigus, d'autres correspondent à des activités biologiques plus complexes et engendrent donc des séquences moins précises. Pour ces derniers types de motifs, des compilations ont été établies pour donner des listes annotées de motifs qui peuvent être communs à plusieurs séquences.

Il existe principalement deux bases de motifs nucléiques qui sont régulièrement actualisées et qui correspondent à un travail de synthèse bibliographique : il s'agit des bases de facteurs de transcription TFD (Ghosh, 1993) et TRANSFAC (Knüppel et al., 1994).

b. Les bases spécialisées de motifs protéiques

La base PROSITE (ExPASy Proteomics Server) peut être considérée comme un dictionnaire qui recense des motifs protéiques ayant une signification biologique.

Elle est établie en regroupant, quand cela est possible, les protéines contenues dans Swissprot par famille comme par exemple les kinases ou les protéases. On recherche ensuite, au sein de ces groupes, des motifs consensus susceptibles de les caractériser spécifiquement.

La conception de la base PROSITE repose sur quatre critères essentiels :

- collecter le plus possible de motifs significatifs
- avoir des motifs hautement spécifiques pour caractériser au mieux une famille de protéines
- donner une documentation complète sur chacun des motifs répertoriés
- faire une révision périodique des motifs pour s'assurer de leur validité par rapport aux dernières expérimentations

Voir un exemple : motif "EF-hand" des protéines fixant le calcium comme la calmoduline par exemple.

3. Liens Internet et références bibliographiques

Base de données sur les acides aminés peu fréquents (sous-partie de la base de données "Protein Information Resource" - PIR): RESID Database

Bases de données sur les propriétés physico-chimiques des acides aminés (sous-partie de la base de données "Expasy - Swiss-Prot"): ProtScale ; Swiss-Prot

Base de données PROWL : propriétés physico - chimiques des acides aminés, peptides, protéines. PROWL

Algorithmes et programmes de comparaison de séquences

Interprétation des résultats : E-value, P-value

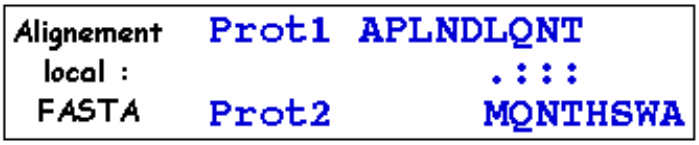
1. Définitions: Il existe 3 grandes classes d'algorithmes de comparaison de séquences :

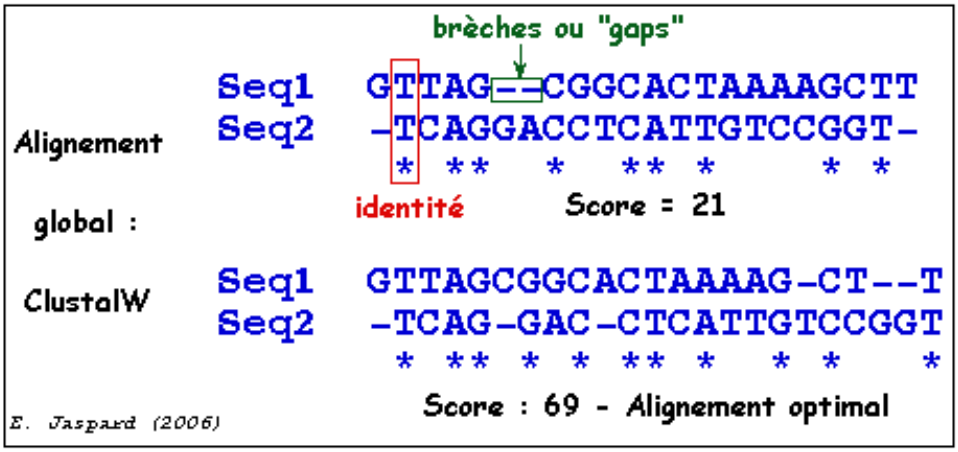
- méthode de programmation dynamique
- méthode heuristique
- méthode d'apprentissage machine

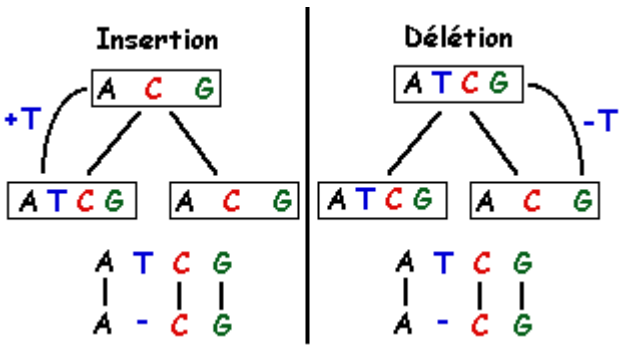
Alignement : processus par lequel deux (ou n) séquences sont comparées afin d'obtenir le plus de correspondances (identités ou substitutions conservatives) possibles entre les lettres qui les composent.

- alignement local : alignement des séquences sur une partie de leur longueur

- alignement global : alignement des séquences sur toute leur longueur
- alignement optimal : alignement des séquences qui produit le plus haut score possible
- alignement multiple : alignement global de trois séquences ou plus
- brèches ou "gap" : espace artificiel introduit dans une séquence pour contre-balancer et matérialiser une insertion dans une autre séquence. Il permet d'optimiser l'alignement entre les séquences.







indel :

- "in" = insertion
- "del" = délétion

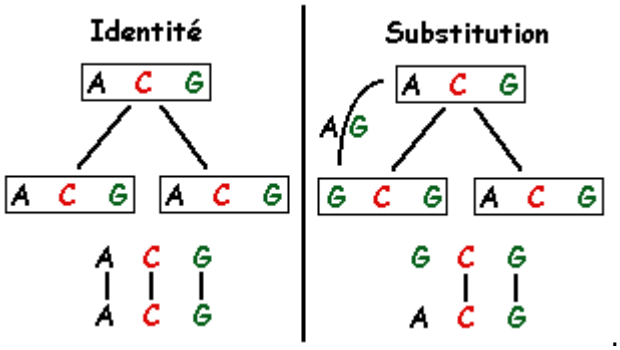
Similarité: c'est le pourcentage d'identités et/ou de substitutions conservatives entre des séquences. Le degré de similarité est quantifié par un score. Le résultat de la recherche d'une similarité peut être utilisé pour inférer l'homologie de séquences.

Homologie : 2 séquences sont homologues si elles ont un ancêtre commun. L'homologie se mesure par la similarité : une similarité significative est signe d'homologie sauf si les séquences présentent une faible complexité.

Faible complexité ("low-complexity regions"): régions qui contiennent peu de caractères différents. Exemples : (a)

FFFPPPPPVVV, 3 acides aminés différents seulement (région riche en proline) - queue poly-A des ARN. Ces régions posent des problémes dans l'analyse des séquences car elles génèrent un score biaisé.

Exemple de programme qui analyse ce type de régions : "SEG"



Mésappariement : non correspondance entre deux lettres. Un mésappariement peut être :

- soit la substitution d'un caractère par un autre, c'est-à-dire une mutation
- soi l'introduction d'un "gap"

Score: un score global permet de quantifier l'homologie. Il résulte de la somme des scores élémentaires calculés sur chacune des positions en vis à vis des deux séquences dans leur appariement optimal. C'est le nombre total de "bons appariements" pénalisé par le nombre de mésappariements.

Score élémentaire:

- ADN: la valeur du score élémentaire est de 1 (les deux bases sont identiques, bon appariement) ou de 0 (les deux bases sont différentes, mauvais appariement).
- protéines : cette valeur est extraite d'une matrice de substitution.

2. Algorithme de Needleman & Wunsch et algorithme de Smith & Waterman

Tous deux sont des algorithmes de programmation **dynamique** utilisés pour obtenir l'alignement **global** ou **local** (respectivement) **optimal** de deux séquences protéiques ou d'acides nucléiques.

La programmation dynamique est une méthode développée par R. Bellman (1955) qui permet de résoudre de nombreux problèmes dont la solution directe n'est pas possible puisque de complexité exponentielle.

Exemple : calcul de la distance d'édition entre deux chaînes de caractères (séquences protéiques ou d'acides nucléiques).

La programmation dynamique une méthode de résolution ascendante qui détermine une solution optimale du problème à partir des solutions de tous les sous-problèmes.

L'algorithme de Needleman & Wunsch et l'algorithme de Smith & Waterman se déroulent globalement en deux étapes :

- la construction, ou descente, qui permet de calculer le meilleur score, c'est à dire le coût de la transformation de la première séquence en la seconde (étape de programmation dynamique)
- la construction de l'alignement lui-même, ou remontée

Ces algorithmes **n'utilisent pas d'heuristique** : ils sont donc **sensibles mais longs**.

Algorithme de Needleman & Wunsch	**Algorithme de Smith & Waterman**
alignement **global** optimal de 2 séquences $F(i,j)=\max\begin{cases}F(i-1,j-1)+s(x_i,y_j)\\F(i-1,j)+d,\\F(i,j-1)+d.\end{cases}$ La ligne i = 0 et la colonne j = 0 sont initialisées aux valeurs de pénalité des gaps. La fonction de récurrence ne réinitialise pas la valeur à 0 si aucune valeur positive n'est présente.	alignement **local** optimal de 2 séquences $F(i,j)=\max\begin{cases}0,\\F(i-1,j-1)+s(x_i,y_j),\\F(i-1,j)+d,\\F(i,j-1)+d.\end{cases}$ La ligne i = 0 et la colonne j = 0 sont initialisées à 0. N'importe quelle case de la matrice de comparaison peut être un point de départ pour le cacul des scores finaux. Si ce score devient inférieur à zéro, la fonction de récurrence **réinitialise la valeur à 0** et la case peut être utilisée comme un nouveau point de départ.

F(i,j) : valeur à la position (i,j) de la matrice.

s (xi,yj) : valeur obtenue à partir de la matrice de substitution pour les nucléotides ou les acides aminés (xi,yj) correspondant à la position (i,j) de la matrice. C'est donc le score correspondant à l'alignement des lettres xi et yj. Ce score prend, par exemple, les valeurs suivantes :

- identité : +3
- non identité : -1
- s (xi,-) et s (-,yj) est la fonction simple de pénalité de l'alignement d'un résidu avec un gap : -5

Remarque : si on opte pour d'autres valeurs, on obtient d'autres alignements optimaux, d'où le choix crucial de la meilleure matrice de substitution lors des alignements.

La fonction de pénalité d'un gap est définie par : f(n) = d + [e . (n-1)], où :

- n = longueur du gap
- d = pénalité d'ouverture d'un gap
- e = pénalité d'extension d'un gap

Exemple : un gap de longueur n = 3, avec une pénalité d'ouverture d = -10 et d'extension e = -2, aura un score de f(3) = -10 + (-2 x 2) = -14

Application : alignement de la séquence 1 = ACGCT avec la séquence 2 = ACT

On remplit la 1ère ligne et la 1ère colonne de la matrice qui correspondent à un gap à la 1ère position :

- l'alignement du A de la séquence 2 avec l'insertion d'un gap dans la séquence 1 coûte : -5
- celui du C de la séquence 2 avec l'insertion d'un second gap de la séquence 1 coûte : -5 + -5 = -10
- et ainsi de suite ...

F(1,1) aura pour valeur la valeur maximale de l'une des possibilités suivantes:

	j	0	1	2	3
i		- (gap)	A	C	T
0	- (gap)	0	-5	-10	-15
1	A	-5	3	-2	-7
2	C	-10	-2	6	1
3	G	-15	-7	1	5
4	C	-20	-12	-4	0
5	T	-25	-17	-9	-1

F(2,1) aura pour valeur la valeur maximale de l'une des possibilités suivantes :

l F(1,0) + s(C,A) = -5 + -1 = -6

l F(1,1) + s(C,-) = 3 + -5 = -2

l F(2,0) + s(-,A) = -10 + -5 = -15

Et ainsi de suite.

Pour reconstituer l'alignement, on démarre de la dernière case (5,3) et on détermine la case à partir de laquelle cette case a été atteinte :

a. la valeur -1 de la case (5,3) ne peut-être obtenue qu'en ajoutant +3 (soit une identité) à la valeur -4 [(case (4,2)].

Cela correspond à l'alignement du "T" de la séquence 1 avec le "T" de la séquence 2.

b. la valeur -4 de la case (4,2) peut être obtenue de 2 manières :

- en ajoutant +3 (soit une identité) à la valeur -7 [(case (3,1)]. Celà correspond à l'alignement du "C" de la séquence 1 avec le "C" de la séquence 2.
- en ajoutant -5 (soit un gap) à la valeur 1 [(case (3,2)]. Cela correspond à l'alignement du "C" de la séquence 1 avec un gap dans la séquence 2.

c. Et ainsi de suite.

Dés lors, on obtient 2 **alignements optimaux** qui ont **le même score** de +1.

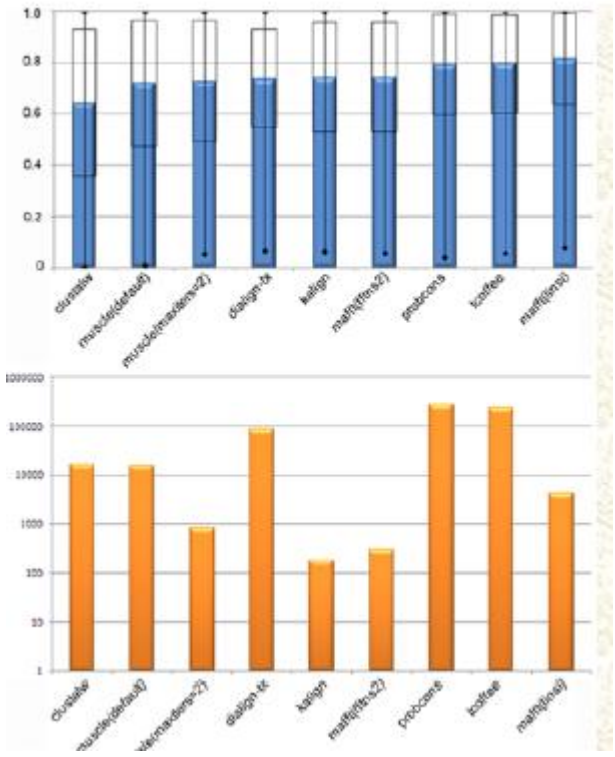
3. Diversité des programmes - spécificité selon le type de données analysées

Voir l'extrême diversité des programmes.	
Type de séquences	Protéines ou acides nucléiques (ADN et/ou ARN) ou les deux
Type d'alignement	Local ou global
Accessibilité	Serveur Web ou implémenté sur l'ordinateur (lignes de commandes)
Spécialisation de plus en plus prononcée du champs d'application des algorithmes / programmes	• recherche dans des bases de données • alignement de séquences 2 à 2 ("paiwise alignment") • alignement de séquences multiples • analyse de génome • recherche de motifs (sous-séquences spécifiques "signature") : ScanProsite • alignement de millions de courtes séquences (voir les nouvelles technologies de séquençage à très haut débit) • modélisation de structures homologues et superposition de structures 3D de protéines ("homology modeling"- "protein threading") • ...
Les "benchmarks" sont de vastes ensembles de données (homogènes, curées, testées) qui permettent de comparer les performances d'algorithmes / programmes.	Exemples de "benchmarks": • BAliBASE : le premier "benchmark" construit d'alignements de séquences protéiques • HOMSTRAD ("HOMologous STRucture Alignment Database") : curated database of structure-based alignments for homologous protein families. • PFAM ("Protein FAMilies") : contient toutes les familles de protéines identifiées (environ 14.000 en 2012). Chacune est représentée par un alignement multiple des séquences de la famille considérée auquel est adjoint un profil HMM ("Hidden Markov Model"). • Affycomp : pour l'analyse de l'expression de gènes - puces à ADN Affymetrix • "The Protein Classification Benchmark collection" : pour l'annotation fonctionnelle par apprentissage machine

Comparaison des performances de plusieurs programmes d'alignement de séquences

- T-Coffee ("Tree-based Consistency Objective Function For alignment Evaluation")

Programme	score d'efficacité	temps de calcul
Probcons	79.4%	2.7 jours
T-Coffee	79.4%	2.7 jours
Mafft (linsi)	81.6%	1.2 heures
Kalign	74.3%	3 minutes !



Source : Thompson et al. (2011) Bleu : efficacité / Orange : rapidité (échelle log)

Les programmes sont de plus en plus **spécifiques du type de données biologiques** traitées ou du type d'analyse effectuée :

- analyse de génomes ou assemblage d'EST en contigs
- construction d'arbres phylogénétiques
- détection de SNP ("Single Nucleotide Polymorphism")
- recherche dans des banques généralistes ou spécialisées
-analyse de paramètres physico-chimiques d'acides aminés de protéines
-séquences consensus conservées ("pattern")
- recherche de motifs structuraux
-l analyse d'expression des gènes
- annotations

La comparaison de structures et la modélisation par homologie

On a de plus en plus d'informations qui tendent à démontrer que le nombre de repliements des protéines dans la nature est limité (quelques milliers). On peut donc regrouper les protéines selon le type de repliement qu'elles adoptent. Voir les bases de données CATH et SCOP, par exemple.
Remarque : les protéines dites "intrinsèquement non structurées" sont à part.
Le préalable de la modélisation par homologie ("homology modeling"- "protein threading") est de disposer d'au moins une protéine dont la structure 3D a été déterminée. Elle sert de "modèle" pour modéliser la structure 3D potentielle d'une protéine pour laquelle on ne dispose que de la séquence. Cette séquence doit bien sûr être proche (homologue) de celle de la protéine modèle. Il faut donc d'abord effectuer des alignements de séquences.

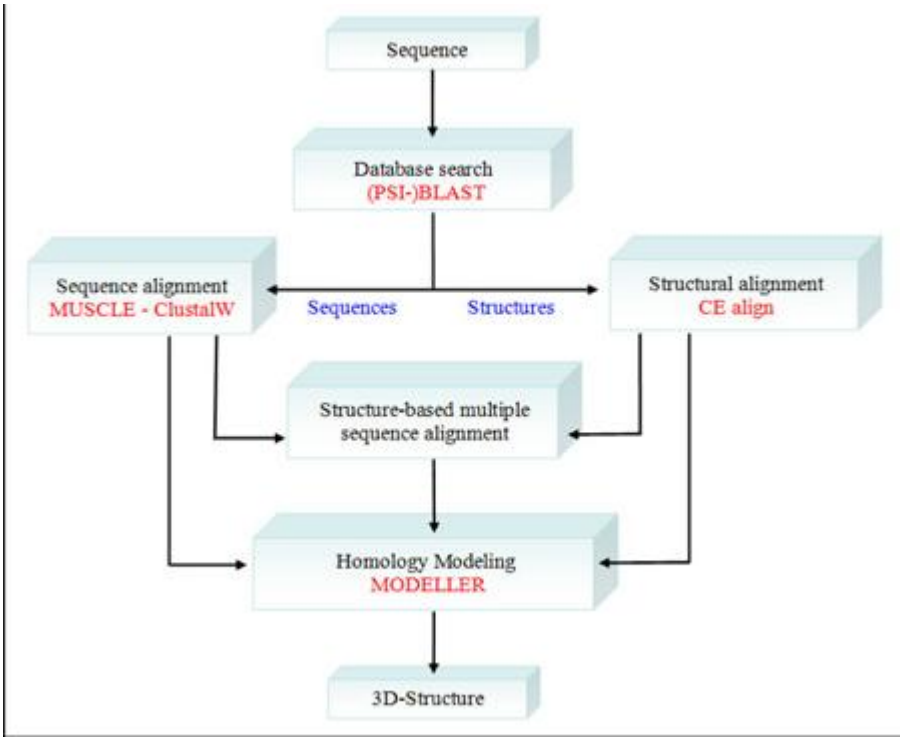
- Exemple de logiciel / interface Web qui renvoie un fichier au format PDB : ESyPred3D.
- Exemples d'autres programmes de modélisation structurale par homologie :

1. DeepView
2. Chimera
3. MolIDE

Figure ci-dessous : Procédure de "PyMod" qui intègre divers types de données et d'analyses :

- recherche dans une base de données de similarités avec la séquence requête
- alignement multiple de séquences
- modélisation de structures 3D par homologie avec le logiciel Modeller.

Chaque "bloc de procédure" est indépendant des autres : on peut donc, par exemple, effectuer un alignement multiple de séquences sans recherche préalable dans une base de données.



Source : Bramucci et al. (2012)

4. Programmes d'alignement local

Les méthodes de programmation dynamique permettent de calculer, sous un système de scores donné, l'alignement optimal, global ou local, entre deux séquences en un temps proportionnel au produit des longueurs des deux séquences. Appliquées à une banque de séquences, le temps de calculs de ces méthodes augmente linéairement avec la taille de la banque.

On définit 2 caractéristiques pour une méthode de comparaison de séquences :

- la sensibilité : c'est l'aptitude à détecter toutes les similarités considérées comme significatives et donc à générer le **minimum de faux-négatifs.**

- la sélectivité : c'est l'aptitude à ne sélectionner que des similarités considérées comme significatives et donc à générer le **minimum de faux-positifs**.

Les programmes des familles **Fasta et BLAST** sont des heuristiques qui réduisent le facteur temps en "sacrifiant" un peu de sensibilité. L'un et l'autre simplifient le problème :

- **en pré-sélectionnant** les séquences de la banque susceptibles de présenter une similarité significative avec la séquence requête

- et **en localisant** les régions potentiellement similaires dans les séquences

Ces étapes sélectives permettent :

- de n'appliquer les méthodes de comparaison, coûteuses en temps, qu'à un sous-ensemble des séquences de la banque

- et de restreindre le calcul de l'alignement optimal à des parties des séquences

Cette logique de recherche plus rapide dans son exécution, comporte donc le risque d'éliminer des séquences qui ont une similarité plus difficile à détecter ou d'aboutir à des alignements sub-optimaux.

La sensibilité et la sélectivité se réfèrent à une notion de résultat significatif ou non. Les programmes mesurent une signification statistique des résultats par rapport à un modèle aléatoire : un résultat est considéré comme significatif si la probabilité de l'obtenir par hasard est très faible.

Les systèmes de score partent du postulat que les résultats les plus significatifs du point de vue statistique sont aussi les plus pertinents du point de vue biologique. Or ce n'est pas toujours le cas car des résultats biologiquement intéressants peuvent être non significatifs sur un plan statistique.

En d'autres termes, la signification biologique d'une similarité entre des séquences n'est pas forcément estimable sur la seule valeur d'un score.

Programme FASTA - Pearson & Lipman (1988)

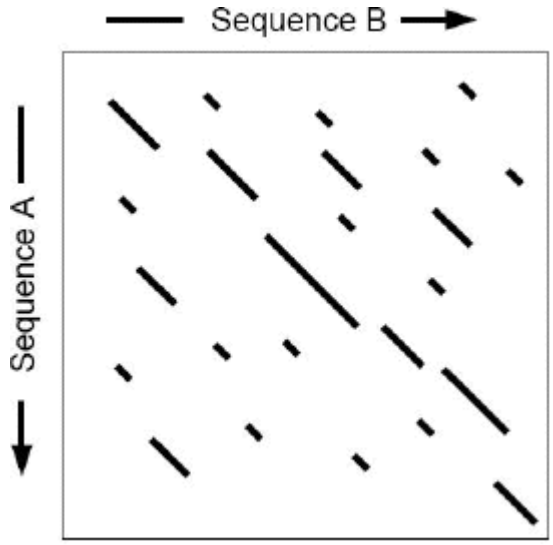
Le programme ne considère que les séquences présentant une région de forte similitude avec la séquence recherchée. Il applique ensuite localement à chacune de ces meilleures zones de ressemblance un algorithme d'alignement optimal.

La codification numérique des séquences, c'est-à-dire la décomposition de la séquence en courts motifs (nommés uplets) transcodés en entiers, confère à l'algorithme l'essentiel de sa rapidité.

Etape 1

- Les régions les plus denses en identités entre les deux séquences sont recherchées. Ces régions sont appelés points chauds ou "hot spots".
- C'est le paramètre "ktup" qui détermine le nombre minimum de résidus consécutifs identiques. Généralement : ktup = 2 pour les protéines - ktup = 6 pour l'ADN.
- Recherche des meilleures diagonales : plusieurs "hot spots" dans une même région génère des diagonales de similarité sans insertion ni délétions. Ces diagonales sont les régions ayant le plus de similarité. Elles sont représentées par un graphique de points ou "dotplot".

Lorsqu'une séquence est comparée à une base de données, la première étape est effectuée pour chaque séquence présente dans cette base de données.



Etape 2

- Les dix meilleures diagonales sont réévaluées à l'aide d'une matrice de substitution et les extrémités de ces diagonales sont coupées afin de conserver les régions ayant les plus hauts scores seulement. Cette recherche de similitude est faite sans insertions ni délétions.

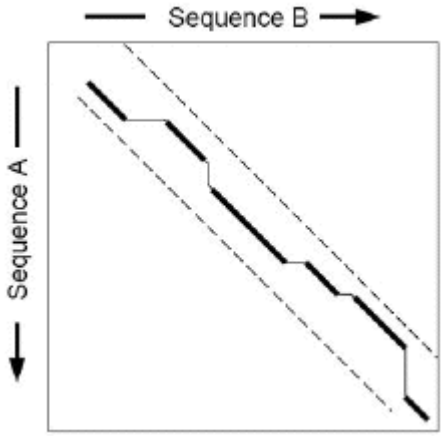
l Le score le plus élevé obtenu est appelé le score "init1". Il est attribué à la région ayant le plus fort score parmi les 10 analysées.

Etape 3

- Les diagonales trouvées à l'étape 1 dont le score dépasse un certain seuil ("cutoff"), sont reliées entre elles pour étendre la meilleure similarité.
- Ces nouvelles régions contiennent des insertions et/ou des délétions
- Le score des nouvelles régions est calculé en combinant le score des diagonales reliées diminué d'un score de pénalité de jonction des diagonales.
- Le score le plus élevé obtenu à cette étape s'appelle le score "initn".
- Cette étape permet d'éliminer les segments peu probables parmi ceux définis à l'étape précédente.

Etape 4

- La région initiale qui a généré le score"init1" est de nouveau évaluée avec un algorithme de programmation dynamique sur une fenêtre de résidus dont la largeur est déterminée par le paramètre "ktup". Le nouveau score est "opt".
- Les séquences de la base de données sont classées selon leurs scores "initn" ou "opt".
- Les séquences sont alignées avec la séquence cible à l'aide de l'algorithme de Smith & Waterman : le score final est le score Smith & Waterman.



Interprétation des résultats

La sortie de FASTA se décompose en trois parties :

- colonne 1 : échelle de valeurs
- colonne 2 : nombre de séquences dans la banque donnant un "z-score" = valeur
-colonne 3 : nombre de séquences dans la banque donnant une "E-value" = valeur
- "init1" = "initn" = "opt" : 100% de similarité
-"initn" > "init1" : plusieurs régions de similarité reliées par des gaps
- "initn" > "opt" : pas de similarité

Les programmes BLAST (Basic Local Alignment Search Tool) - Altschul et al. (1990)

Méthode heuristique qui utilise la méthode de Smith & Waterman.

C'est un programme qui effectue un alignement local entre deux séquences nucléiques ou protéiques.

La rapidité de BLAST permet la recherche des similarités entre une séquence requête et toutes les séquences d'une base de données.

Les différents programmes BLAST

Acides nucléiques

- 1. "MEGABLAST" est l'outil de choix pour identifier une séquence.
- 2. "Standard nucleotide BLAST" est mieux adapté à la recherche de séquences similaires mais pas identiques à la séquence requête.
- 3. L'option "Search for short and near exact matches" de "Nucleotide BLAST" est adapté à la recherche d'amorces ("primer") ou de courts motifs nucléotidiques

Protéines

l 1. Il n'y a pas d'équivalent de "MEGABLAST" pour les requêtes protéiques.

l 2. "Standard protein BLAST" est le mieux adapté à la recherche de séquences protéiques.

- 3. "PSI-BLAST (Position-Specific Iterated-BLAST)" est adapté à la recherche de similarité fine entre séquences protéiques. A utiliser quand une recherche BLAST a échoué ou renvoyé des résultats tels que : "hypothetical protein" or "similar to...".

- 4. "PHI-BLAST (Pattern-Hit Initiated-BLAST)" est adapté à la recherche de séquences protéiques qui contiennent un motif spécifié par l'utilisateur ET sont similaires à la séquence requête dans le voisinage proche du motif.

-5. "Search for short nearly exact matches" de "Protein BLAST" est adapté à la recherche de similarité dans le cas de courtes séquences peptidiques. Les valeurs des paramètres "Expect value cutoff" et "word size" sont modifiés la matrice PAM30 (plus stringente) remplace la matrice BLOSUM62. Une séquence requête inférieure à 5 acides aminés est déconseillée.

- 6. "Nucleotide query - Protein db [blastx]" est adapté pour trouver des séquences protéiques similaires à celles codées par une séquence requête nucléotidique. Trés utile pour l'analyse massive de séquence d'EST ("Expressed Sequence Tags").

- 7. "Protein query - Translated db [tblastn]" est adapté pour trouver des régions codantes des protéines homologues dans un ensemble de séquences nucléotidique nonannotées. Trés utile pour l'analyse de séquence d'EST et de brouillons de génomes (HTG).

- 8. "Conserved Domain Database (CDD)": ce service utilise le programme "Reverse Position Specific BLAST (RPS-BLAST)" pour identifier des domaines protéiques conservés en comparant la séquence requête contre des bases d'alignements de domaines conservés obtenues avec des matrices de scores de position spécifiques "Position specific scoring matrices (PSSMs)". Les bases de données sont : "SMART", "PFAM" et "LOAD" ("Library Of Ancient Domains").

- 9 "Conserved Domain Architecture Retrieval Tool (CDART)" permet d'examiner la structure en domaine de toutes les protéines de la base de données BLAST. Plus sensible qu'une recherche BLAST classique car CDART est lié au programme RPS-BLAST ("Reverse Position-Specific BLAST") qui est lui-même une "variation" du programme "PSIBLAST".

-10. "BLAST 2 Sequences" permet la comparaison de 2 séquences requête. Ne recquiert pas de format particiliers des séquences. La séquence entrée en second est considérée comme la "base de donnée" contre laquelle est effectuée la comparaison.