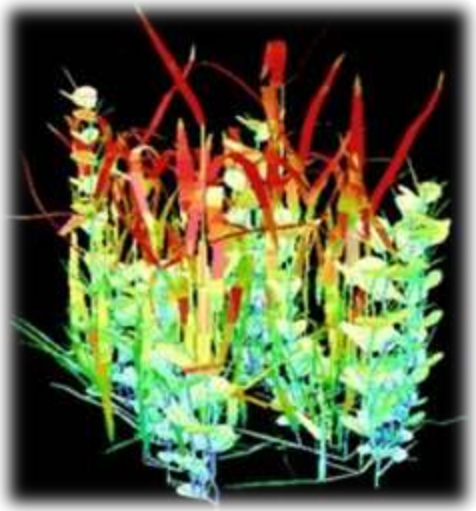
Université Bachir El Ibrahimi de Bordj Bou-Arréridj

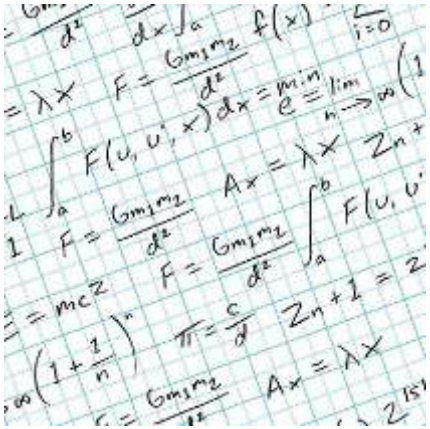
Cours de Modélisation et Bio-informatique

Partie 1: Modélisation

Dr ALILI Dahmane

Cours pour M1 Amélioration des pantes





Explication:

Classification en fonction de la forme

a) **Modèle conceptuel**

Les modèles conceptuels ou verbales sont des modèles basés sur l'utilisation d'un langage naturel pour la description et le diagnostic d'un système et/ou sa dynamique.

Par exemple, la description d'un plan expérimental lors de la réalisation d'une expérience en plein champs est considérée comme un modèle conceptuel.

Dans ce cas, la planification des facteurs expérimentaux, les niveaux de chaque facteur, ainsi que l'ensemble des témoins constituent les différentes parties de ce modèle.

•**Modèle physique**

C'est une représentation physique d'un système étudié, qui projette le fonctionnement réel de l'objet étudié. Les modèles physiques sont caractérisés par leurs structures dimensionnelles définissant les relations entre les différentes parties du modèle.

L'ensemble des structure moléculaires biochimiques (ADN, acides aminées, protéine, sucres…etc.) sont considérées comme des modèles physiques.

•**Modèle diagrammatique**

Le modèle diagrammatique est un modèle basé sur des représentations graphiques, schéma (Sketch)….qui illustre des relations et des tendances entre plusieurs objets du système. Il se répartie en plusieurs modules. Chacun de ces derniers représente une composante du système étudié.

•**Modèle mécaniste**

Dans ce cas, le fonctionnement du système biologique est contrôlé par des processus mécanistes caractérisés par une variation spacio-tomporelle (Ex. la croissance de la plante, la photosynthèse, l'évapotranspiration…etc.).
De plus, ces modèles décrivent les mécanismes physiques, chimiques et biologiques qui déterminent le fonctionnement du système.

•**Modèle descriptif**

•On parle des modèles descriptifs lorsque les processus du fonctionnement du **système biologique sont contrôlés par un paramètre unique** (Ex. variation de la densité d'une population de bactéries dans un milieu contrôlé sous l'effet de température).

•**Modèle dynamique et modèle statique** :

•**Le modèle dynamique** représente le changement *continu* du système au cours du temps.

•Cependant, **le modèle statique** décrit les variations *temporaires* qui affectent le fonctionnement du système biologique.

•**Modèle déterministe et modèle stochastique:**

Quand certain paramètres varient de forme aléatoire le modèle est stochastique et à chaque tournée peut donner des résultats différents.

Quand tous les paramètres sont constants le résultat est toujours le même. Un modèle stochastique est un modèle qui contient une (ou plusieurs) variable aléatoire. L'équation de base des modèles stochastiques est la suivante:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + a_3 X_3 + a_4 X_4 + \varepsilon$$

Mise en œuvre d’un modèle

2.1.1. Formulation qualitative : Diagramme de Forrester

Les diagrammes de Forrester ont été inventés par Jay Forrester et sont un outil de base pour la formulation qualitative de modèles dynamiques. Ce diagramme ne contient pas des équations mais montrent les objets et ses interrelations entre les différents compartiments du système étudié, selon les hypothèses posées. Ils servent comme base pour la formulation quantitative.

Tous les transferts d’énergie et de matière entre les différents compartiments du système doivent être représentés lors de la formulation qualitative.

Explication:

Formulation qualitative : Diagramme de Forrester

Les diagrammes de Forrester ont été inventés par Jay Forrester et sont un outil de base pour la formulation qualitative de modèles dynamiques. Ce diagramme ne contient pas des équations mais montrent les objets et ses interrelations entre les différents compartiments du système étudié, selon les hypothèses posées. Ils servent comme base pour la formulation quantitative.

Tous les transferts d'énergie et de matière entre les différents compartiments du système doivent être représentés lors de la formulation qualitative.

Les diagrammes de Forrester permettent de représenter graphiquement des modèles à base d'équations différentielles,

Il est possible par exemple de modéliser dans le cadre d'un diagramme de Forrester, un modèle représentant **les naissances et les décès de prédateurs et de proies**.

De manière générale, les diagrammes de Forrester sont utilisés pour étudier comment **des écosystèmes fonctionnent.**

« Modélisation du cycle du carbone et de l'azote au moyen du modèle MO MO S »

(*Micro-Organismes* et *Matière Organique* du *Sol*)

Explication:

Les transformations de C et de N dans les sols sont en grande partie assurées par des microorganismes. Ce sont donc ces microorganismes qui déterminent la cinétique de décomposition et qui sont responsables du couplage entre les dynamiques de C et N.

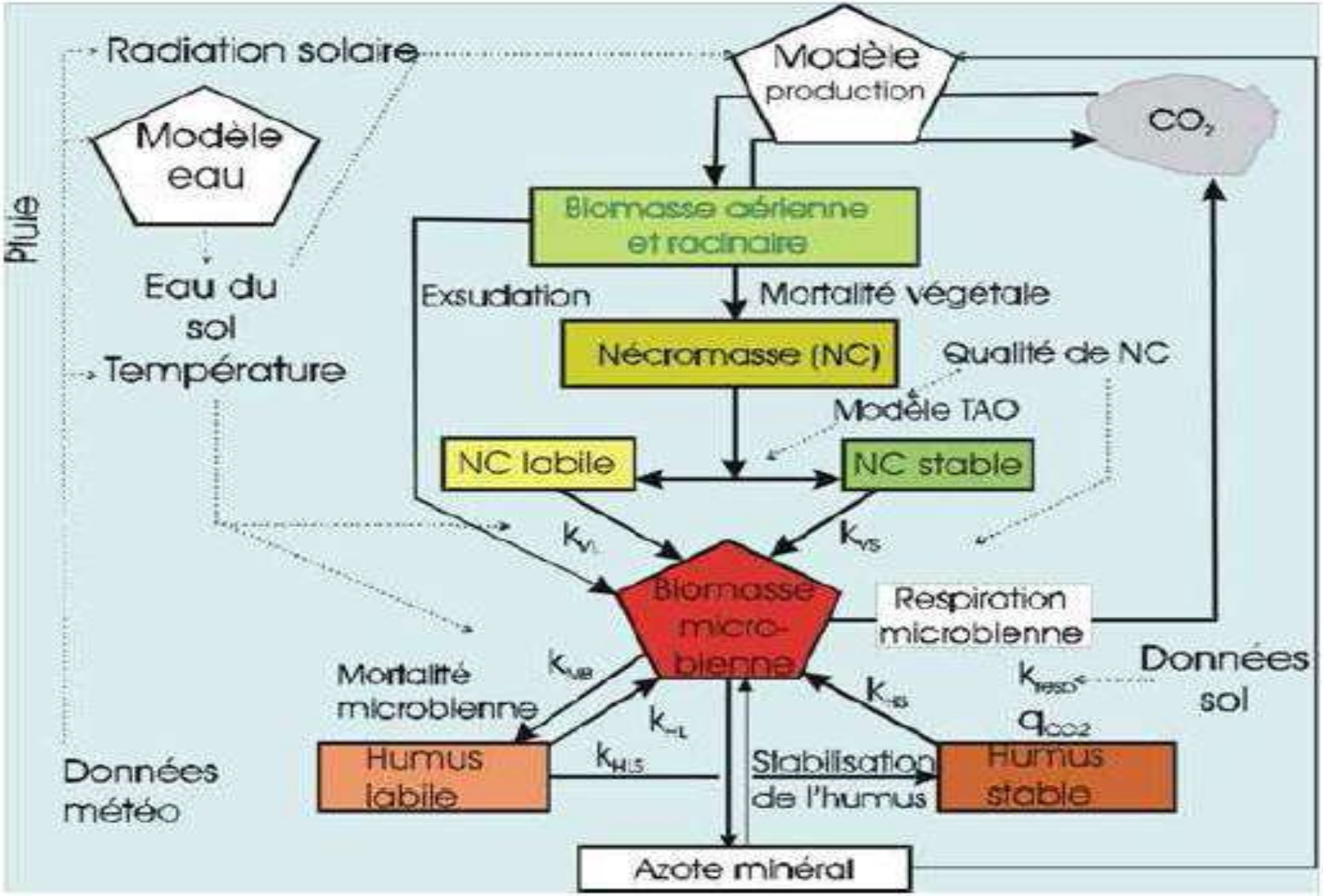


Diagramme relationnel de la Modélisation des Transformations Organiques par les Microorganismes du Sol: MOMOS, (Pansu *et al*., 2010).

Explication:

Les matières organiques du sol (MOS) désignent l'ensemble des substances et des composés carbonés
d'origine végétale et animale. Elles se localisent surtout dans l'horizon superficiel (0-20 cm).

La nature des MOS se répartit en 3 catégories :

-La matière organique vivante (vers de terre, insectes, champignons, bactéries,…).

-La matière organique fraîche principalement d'origine végétale.

-**La matière organique labile** (glucides simples, acides aminés) et la matière organique stable (l'humus)
issues de la décomposition de la matière organique végétale.

Explication:

Le modèle couplé visera à prédire simultanément **les flux du C et de N de l'atmosphère**, respectivement par les stomates des plantes et les rhizobiums, et leur translocation vers les tiges et les racines, les flux de C et N restitués par les plantes aux microorganismes du sol, le flux de C vers atmosphère par la respiration microbienne, les flux de N entre les microorganismes et la phase minérale (minéralisation et immobilisation microbienne), l'absorption de l'azote minéral par les racines et sa translocation vers les parties aériennes pour chaque plante.

Il intégrera les entrées N par restitution et fertilisation ainsi que les pertes éventuelles par les agrosystèmes vers l'atmosphère et les eaux souterraines.

Application de modèle MOMOS dans le système légumineuse-céréales

La validation du modèle MOMOS a pour but de répondre à plusieurs questions de recherche sur le système légumineuses-céréales, à savoir:

- Les légumineuses et les céréales en monoculture ou en association, sont-ils des sources des puits de carbone ?
- Les légumineuses fixatrices peuvent-elles constituer des sources suffisantes d'azote pour les céréales en association ou rotation? Dans quelle mesure le phosphore limite– t-il la fixation d'azote ?
- A quelle échelle de temps, les sources en N et C sont-elles disponibles ? Comment assurent-elles la fertilité et la durabilité des systèmes ?

Développement d'un modèle mécaniste de succession végétale (FAPROM)

Le modèle FAPROM
Fallow Production Model (Martineau & Saugier 2004)

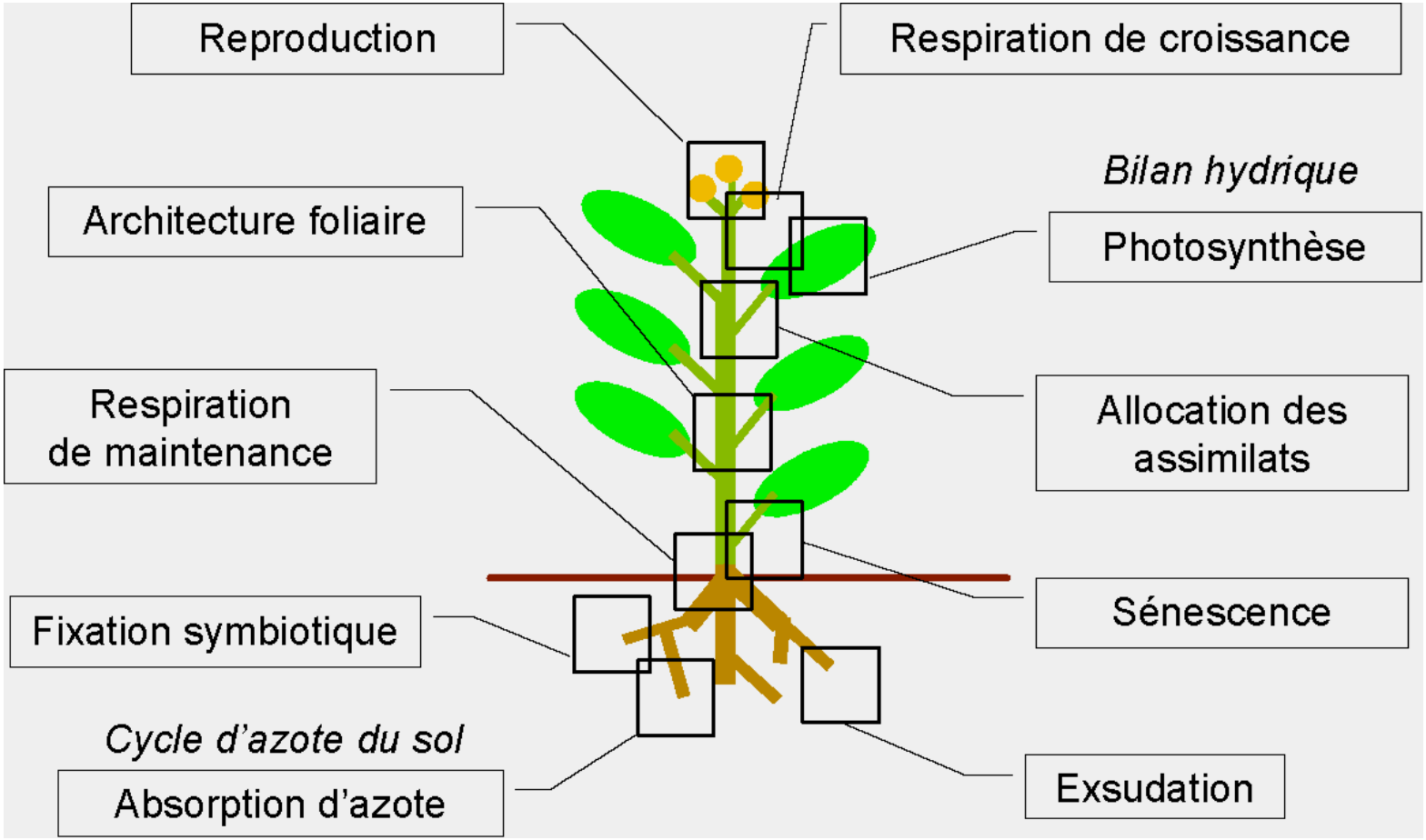


Fig. 1 - Processus simulés dans le modèle Faprom.

Explication de la figure 1:

La végétation est décrite comme **un mélange d'espèces en compétition pour la lumière et l'azote**. Les parties aériennes occupent plusieurs strates horizontales dans lesquelles les feuilles sont distribuées au hasard.

Chaque espèce comporte **quatre organes**: feuilles, tiges, racines et organes reproducteurs. Le modèle prédit la croissance et **la production de biomasse** de chaque espèce, et prend **en compte les cycles du carbone et de l'azote à travers le sol et les plantes**. Il utilise un pas de temps **journalier**, avec comme entrées des variables **météorologiques**. Il simule divers processus : photosynthèse,

respiration de croissance et d'entretien, allocation du carbone, absorption, fixation et remobilisation d'azote, exsudation racinaire, sénescence des tissus, et dispersion et germination des graines

Partie 2: Bioinformatique

"La bioinformatique fournit des **bases de données** centrales, **accessibles mondialement**, qui permettent aux scientifiques de présenter, rechercher et analyser de l'information.
Elle propose des **logiciels** d'analyse de données pour les études de données et les comparaisons et fournit des outils pour la modélisation, la visualisation, l'exploration et l'interprétation des données", selon une définition de l'Institut Suisse de Bioinformatique.

La bioinformatique. C'est quoi ?

- *Ensemble de méthodes, de logiciels et d'applications en ligne qui permettent de gérer, manipuler, et analyser des données biologiques.*
- *La bioinformatique met en jeu plusieurs champs disciplinaires :*

Informatique ***Mathématiques formelles*** ***Statistiques***

Biologie

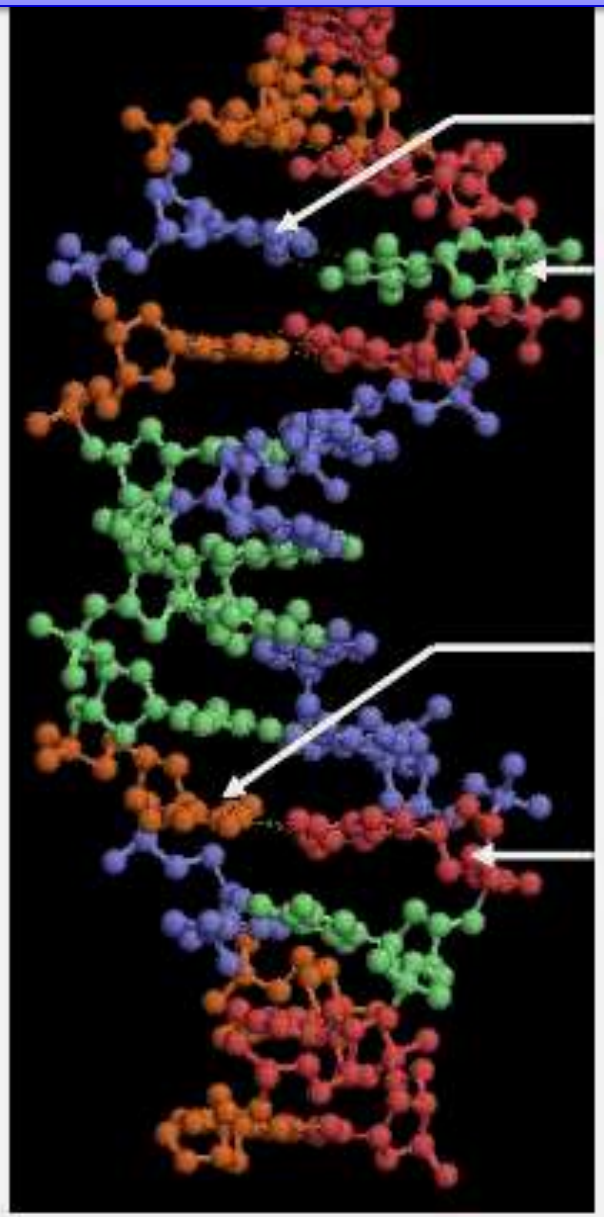
Biologie «in silico»

Traitement automatisé d'informations biologiques

«in vivo» (dans le vivant, cobaye; souris blanche, rat,…),
«in vitro» (dans le verre, tube à essai),
«in silico» (dans la silice, composant des «puces» des ordinateurs)

«in silico» Terme apparu dans «A. Danchin, C. Médigue, O. Gascuel, H Soldano, A Hénaut - From data banks to data bases - Research in Microbiology (1991) 142: 913-916»

Ressemblance ou similitude entre séquences



Nucléotide à **Adénine**

Nucléotide à **Thymine**

L'adénine est toujours associée à la thymine.

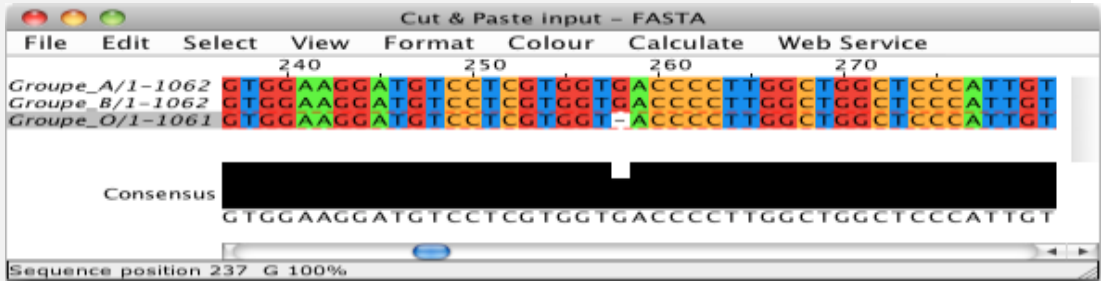
Les bases azotées sont complémentaires deux à deux: **les deux chaînes de l'ADN sont donc complémentaires l'une de l'autre.**

Nucléotide à **Cytosine**

Nucléotide à **Guanine**

La cytosine est toujours associée à la guanine.

(Image obtenue à partir du logiciel RasTop).



Chapitre 2.
Banques et bases de données biologiques

➢ *Harmonisation des fiches de données*

Exemple de la fiche GENBANK d'un plasmide d'E .faecalis

```
LOCUS       KC297657                673 bp    DNA     linear   BCT 18-MAR-2013
DEFINITION  Enterococcus faecalis strain 493/96 plasmid OrfC gene, partial cds;
            RNAI (rnaI) and RNAII (rnaII) genes, complete sequence; and OrfD
            gene, partial cds.
ACCESSION   KC297657
VERSION     KC297657.1  GI:460758767
KEYWORDS    .
SOURCE      Enterococcus faecalis
  ORGANISM  Enterococcus faecalis
            Bacteria; Firmicutes; Bacilli; Lactobacillales; Enterococcaceae;
            Enterococcus.
REFERENCE   1  (bases 1 to 673)
  AUTHORS   Wardal,E., Sadowy,E. and Hryniewicz,W.
  TITLE     Diversity of plasmid-associated genes among Enterococcus faecalis
            clinical isolates
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 673)
  AUTHORS   Wardal,E. and Sadowy,E.
  TITLE     Direct Submission
  JOURNAL   Submitted (10-DEC-2012) Molecular Microbiology, National Medicines
            Institute, Chelmska 30/34, Warsaw 00-725, Poland
COMMENT     ##Assembly-Data-START##
            Sequencing Technology :: Sanger dideoxy sequencing
            ##Assembly-Data-END##
```

➢ *Harmonisation des fiches de données*

En résumé, une fiche comporte de nombreuses informations :

Locus (lieu)	Identificateur (nom et taille de la séquence)
Definition	Description de la séquence
Accession / version	Numéro d'accès dans la base
Keyword / Source / Organism / Reference / Authors / Title / Journal	Informations diverses (taxonomie, publications…)
Features (caractéristiques)	Caractéristiques de la séquence / produits d'expression
Origin (origine)	Séquence (par blocs de caractères / par lignes)
//	Fin de l'entrée dans la base

Chaque séquence de GenBank possède deux identifiants : **le numéro Accession et le Gi** (prononcé ji-aïe !).

1. Le numéro Accession

C'est un identifiant unique à chaque séquence. Il est composé de lettres et de chiffres (parfois d'un underscore '_'). **Il n'est jamais modifié**, même si l'enregistrement est corrigé à la demande de l'auteur. Pour une combinaison 'accession.version' existe un numéro **Gi unique**.

2. Le Gi (GenInfo Identifier)

Le Gi est un nombre qui est spécifique à une séquence.

GenBank est une banque généraliste

GenBank contient l'ensemble des **séquences nucléiques**, quelle que soit leur nature (**ADN génomique, ARN messager**, …). Les séquences produites dans les laboratoires du monde entier à partir de plus de **100 000 organismes différents** sont régulièrement **soumises au NCBI.**

NCBI National Center for Biotechnology Information

GenBank croit ainsi de manière exponentielle, doublant de taille tous les dix mois. En **2006** elle contenait **plus de 65 milliards de nucléotides** dans plus de **61 millions de séquences**.

La banque se construit soit par **des dépôts directs** en provenance des **laboratoires**, soit par des dépôts en masse des **centres de séquençage** à grande échelle.

Le format FASTA

- ***Format commun de manipulation des données :***
  le format FASTA (Fast – alignment) (Alignement rapide)

<u>Objectif</u> : manipuler facilement des séquences dans les bases de données, à l'aide d'un format universel, compatibles avec les traitements de texte (sous forme de fichier texte), ou par copier – coller.

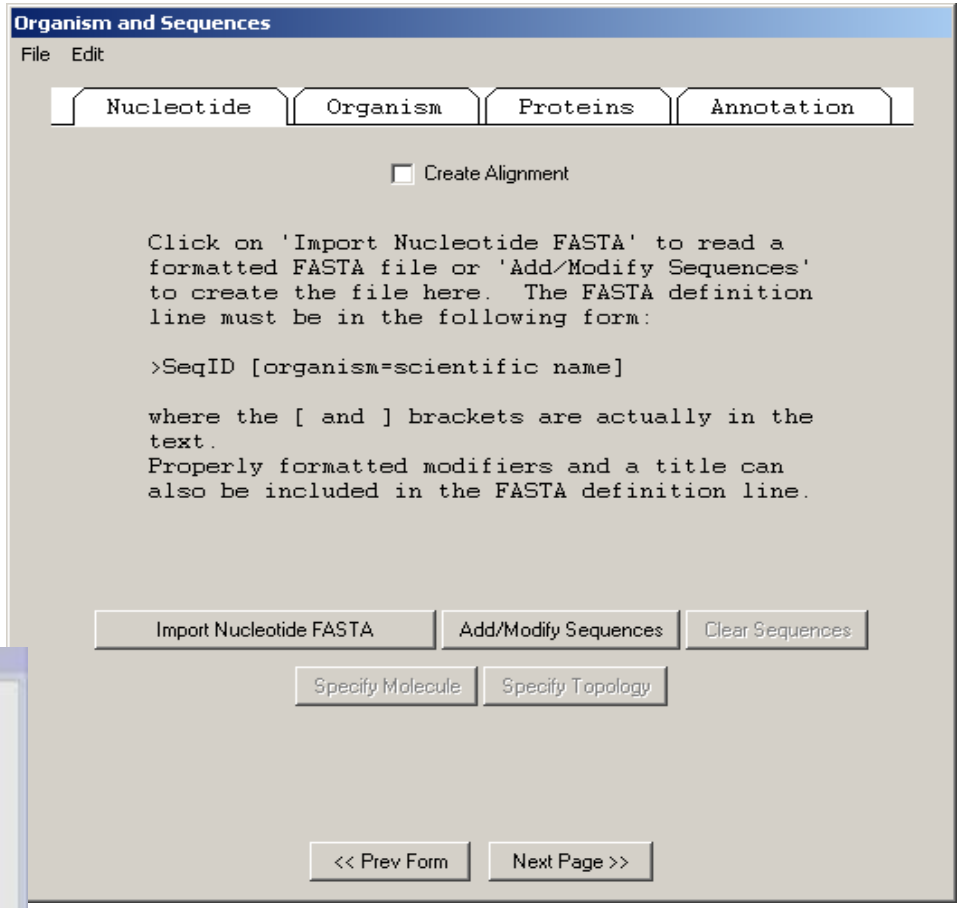
<u>Exemple de la fiche précédente du plasmide d'E.faecalis en format FASTA :</u>

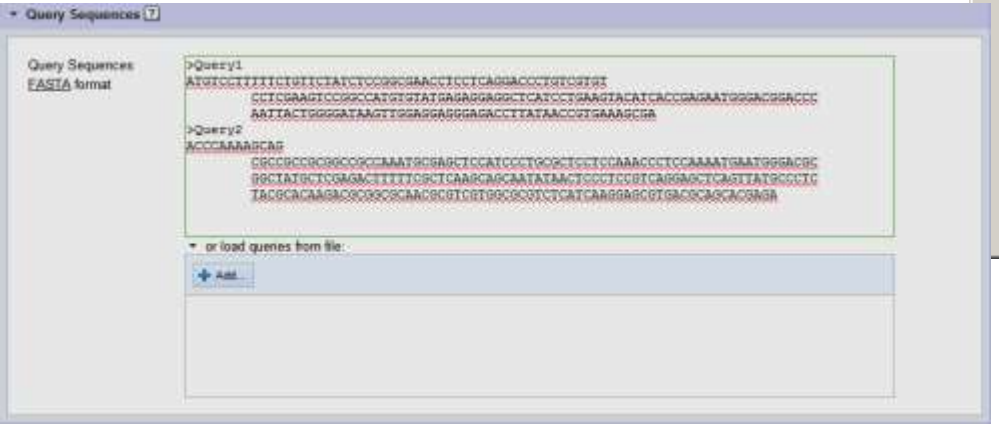
```
>gi|460758767|gb|KC297657.1| Enterococcus faecalis strain 493/96 plasmid OrfC
gene, partial cds; RNAI (rnaI) and RNAII (rnaII) genes, complete sequence; and
OrfD gene, partial cds
AGTATAAATGTTCCTGGTGTAAACGAGTTTACACGCTTAGAAAAGATCATAAAACAGCTAGATAAATTGT
TGAAGGTTTTATTATTGAATTGGCAGAATTTCAATCTATGCTATAATTAATACGGCAGCTCGCCTCGATT
GGAGGTGTGTTATTTGTGAAAGATTTAATGTCGTTGGTTATCGCACCAATCTTTGTAGGATTGGTTCTGG
AAATGATTTCTCGTGTGTTGGACGAGGAAGACGATAGCCGAAAGTAAGCTGCTATCAACACACACGCTAG
AAGTCGCAACTAGTGTAAAAAAAAGCAATCCTATTCGCCGTAGGATTGCTTTTTGTGTTATCTGTACGAT
TTAATGTCGTTTCGCACTTTTAGTATAGCATATTTTTATTTTGGGTCAAGTTTTGTGACTATGCAGGAAT
TGGTAAAGAATACAGTGGTAGCAATTTTCATCGATGCTATTTTATTAATAAAATAGTAGTAGAAAAATAT
ATTTATTGATAAACTTATAGTTATGAATCTGTATAGTTAGTTATAATAATTGGTATTTTTTTAGGAAAAT
TTGAGCTTTTGAATTGAATAAGAAGGAGTGATTTTATGGATTTAAAGTACAATGTTTTTGGTAATTCAAT
GTATTCTTTGAAAGAAATGGAGCTAATTCAACTAGCTTCACAA
```

FASTA est à l'origine un programme **d'alignement de séquences d'ADN et de protéine** développé par **William R. Pearson en 1988**. Un des héritages de ce programme est le format de fichier FASTA qui est devenu un format standard en bioinformatique.

Des programmes d'alignement **"FAST-P"** (pour **protéine**)

et **"FAST-N"** (pour **nucléotide**)





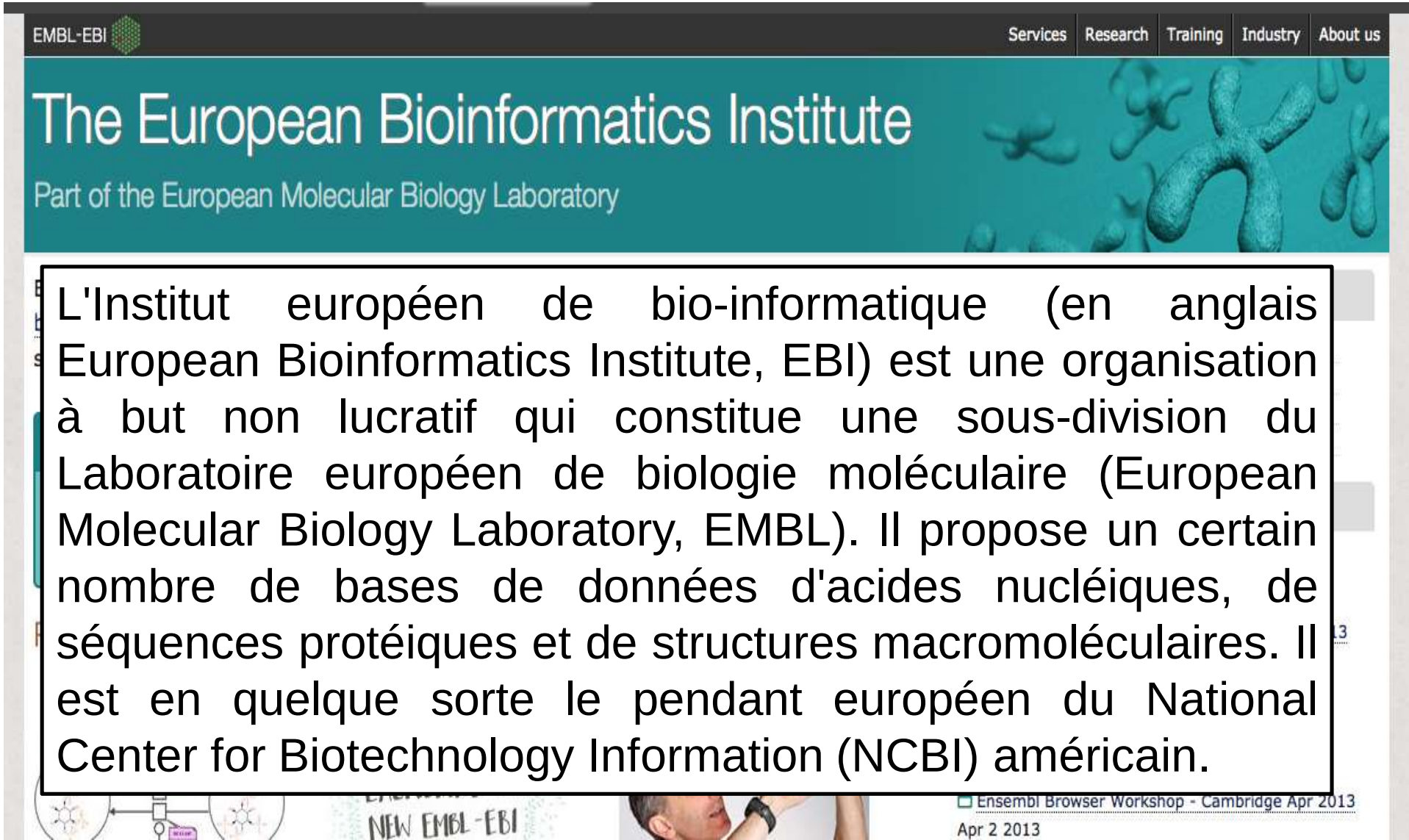
Une séquence au format FASTA commence par une ligne de titre (nom, définition ...), suivie par les lignes de la séquence.

La ligne de titre se distingue de la séquence par un symbole plus grand que (">") en début de ligne. La longueur de cette ligne ne doit pas excéder 200 caractères. Il est recommandé de mettre la séquence sous forme de lignes de 80 caractères maximum. Un exemple de séquence au format fasta

```
>gi|532319|pir|TVFV2E|TVFV2E envelope protein
ELRLRYCAPAGFALLKCNDADYDGFKTNCSNVSVVHCTNLMNTTVTTGLLLNGSYSENRTQIWQKHRTSNDSALIL
LNKHYNLTVTCKRPGNKTVLPVTIMAGLVFHSQKYNLRLRQAWCHFPSNWKGAWKEVKEEIVNLPKERYRGTNDP
KRIFFQRQWGDPETANLWFNCHGEFFYCKMDWFLNYLNNLTVDADHNECKNTSGTKSGNKRAPGPCVQRTYVAC
HIRSVIIWLETISKKTYAPPREGHLECTSTVTGMTVELNYIPKNRTNVTLSPQIESIWAAELDRYKLVEITPIGFAPTEVR
RYTGGHERQKRVPFVXXXXXXXXXXXXXXXXXXXXXXXVQSQHLLAGILQQQKNLLAAVEAQQQMLKLTIWGVK
```

Où trouver les outils de bioinformatique ?

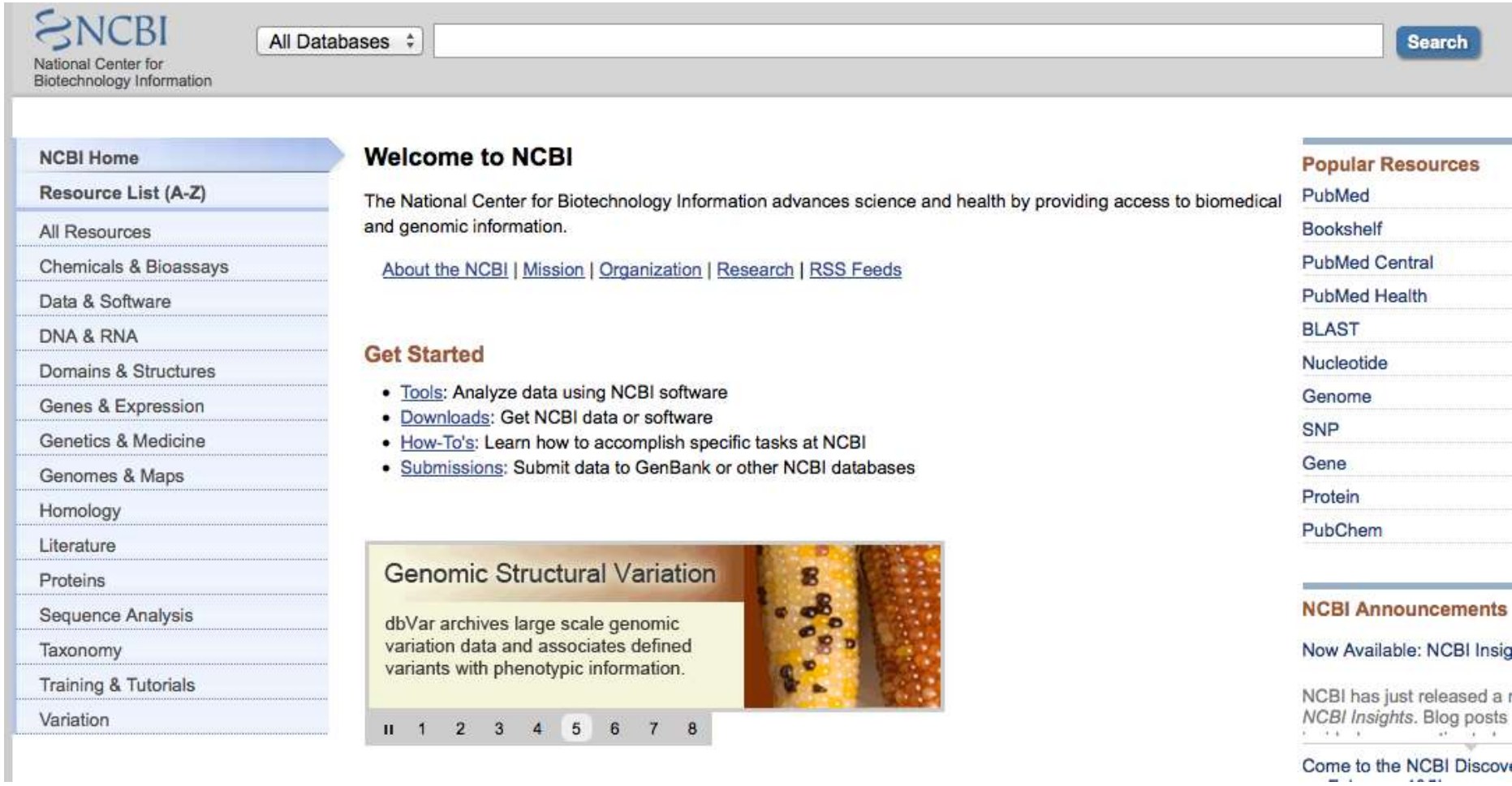
Le portail EBI



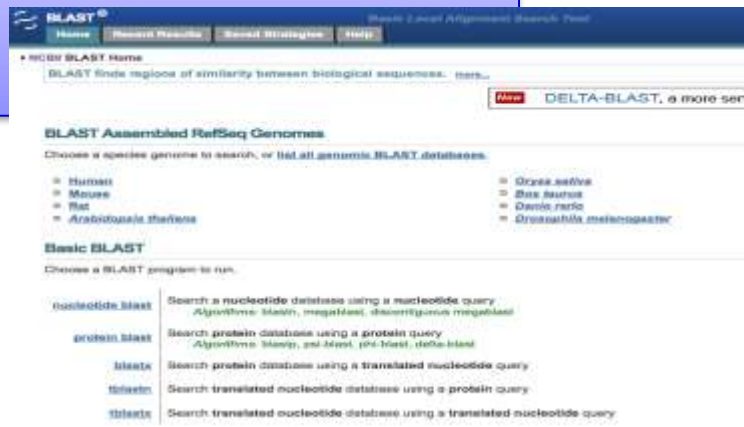
L'Institut européen de bio-informatique (en anglais European Bioinformatics Institute, EBI) est une organisation à but non lucratif qui constitue une sous-division du Laboratoire européen de biologie moléculaire (European Molecular Biology Laboratory, EMBL). Il propose un certain nombre de bases de données d'acides nucléiques, de séquences protéiques et de structures macromoléculaires. Il est en quelque sorte le pendant européen du National Center for Biotechnology Information (NCBI) américain.

Où trouver les outils de bioinformatique ?

Le portail NCBI



Le programme d'alignement BLAST



Le logiciel BLAST accessible depuis le portail

Le programme d'alignement **BLAST** (**Basic Local Alignment Search Tool**)

Les programmes BLAST effectuent une recherche rapide dans les banques de séquences nucléiques et protéiques combinée avec une estimation rigoureuse des statistiques pour apprécier la signification des similitudes.

Bases de séquences spécialisées

Bases de séquences	Adresse
• **Bases génériques (multi-organismes)**	
EMBL / trEMBL	http://www.ebi.ac.uk/embl/
Genbank / GenPept	http://www.ncbi.nlm.nih.gov/entrez
DDBJ (DNA Data Bank of Japan)	http://www.ddbj.nig.ac.jp/
SwissProt	http://www.expasy.org/sprot/
• **Bases spécialisées (organisme)**	
GenoList	http://genolist.pasteur.fr
Cyanobase	http://www.kazusa.or.jp/cyano/
TAIR (The Arabidopsis Information Resource)	http://www.arabidopsis.org
FlyBase (Database of the Drosophila Genome)	http://flybase.bio.indiana.edu/
MGD (Mouse Genome Database)	http://www.informatics.jax.org/
GDB (Human Genome data Base)	http://gdbwww.gdb.org/
• **Bases spécialisées (thématique)**	
PROSITE	http://www.expasy.org/prosite
eMOTIF	http://fold.stanford.edu/motif
EPD (Eukaryotic Promoter Database)	http://www.epd.isb-sib.ch/