Introduction à la bioinformatique structurale sous Linux

Karim Mezhoud
Ir. agronome
PhD. Toxicologie, Protéomique, Bioinformatique

Centre national des Sciences et Technologies Nucléaires
Sidi Thabet – Tunis

Master professionnel "Développement industriel des produits de Santé"

Objectifs du module 25h

Découvrir et acquérir des outils qui permettent de travailler sur la biologie structurale.

La biologie structurale est la science qui étudie la structure des bio-molécules.
Les bio-molécules sont les molécules d'origine vivant: protéines, substances bio-actives (Drugs).

Les outils sont un système d'exploitation, logiciels, bases de données, sites web et langage informatique (bash).

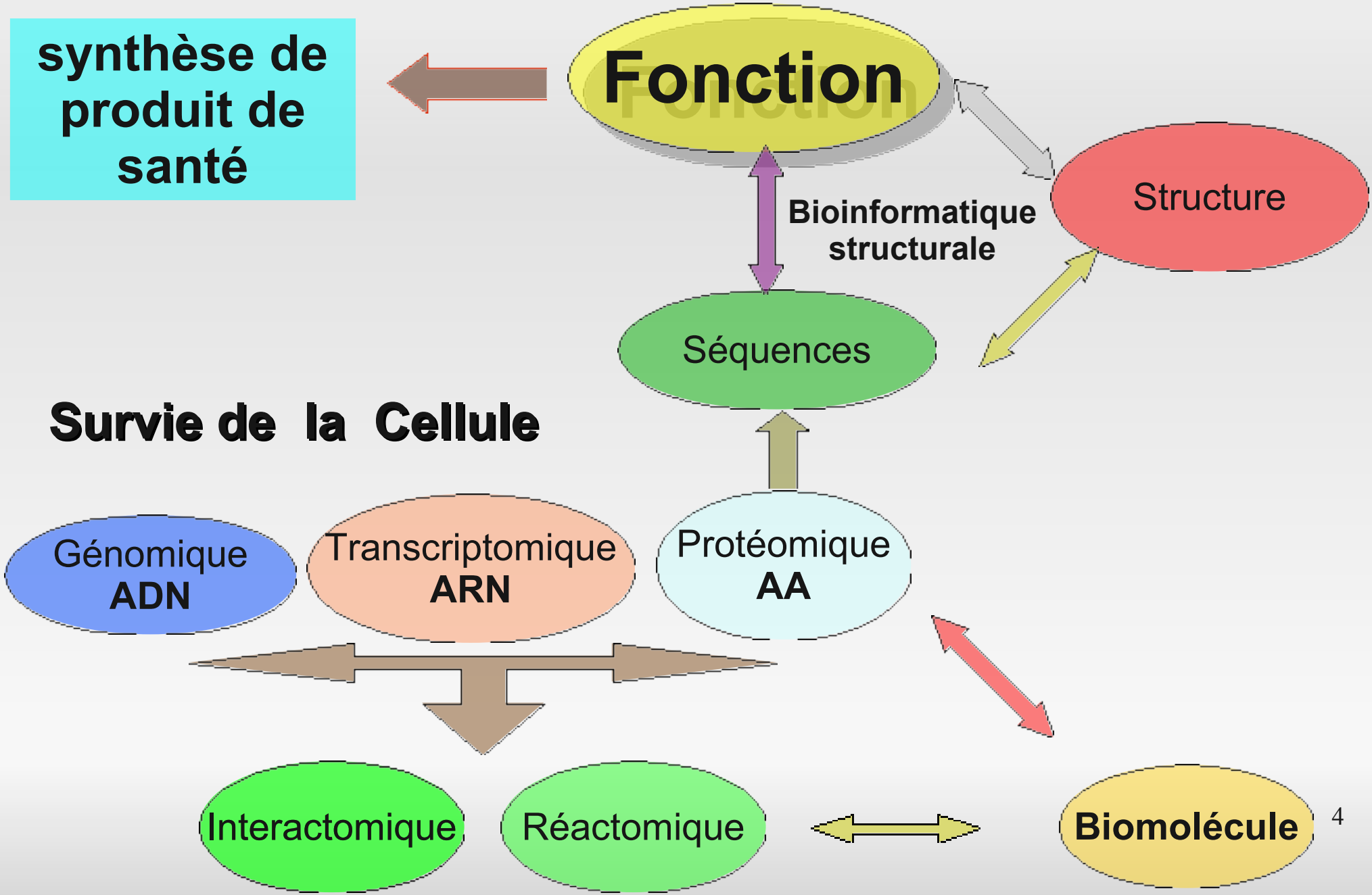
Semaine du 5 Avril 25h = 3 x 8 + 1

HORAIRE / JOURNEE	8h30 à 9h30	9h30 à 10h30	10h30 à 11h30	11h30 à 12h30		13h30 à 14h30	14h30 à 15h30	15h30 à 16h30
LUNDI								
MARDI	* Technique de coservation G2 TD Dr Boudhrioua (1h30) (SM) 10h		* Technique de coservation G2 TD Dr Boudhrioua (1h30) (SM) 12h					
MERCREDI	* An Text sens Cours Dr Boudhrioua (1h30) (SM) 10h		* An Text senso TD Dr Boudhrioua (1h30) 12h			Biotechnol Mol cours Dr Ben Amor (Ecole Vét)		
JEUDI	Bioinformatique (G1) TD Dr Mezhoud (CNSTN)					Bioinformatique (G2) TD Dr Mezhoud (CNSTN)		
VENDREDI	Anglais TD Mme Mallakh (Ecole Vet)							
SAMEDI								

2H + 2H

1 D. Surveillance (30%) + 1 D. synthèse (60%)

Prédiction de fonction



Programme des Travaux Dirigés

1. Installation de système d'exploitation Linux
   Exemple: Ubuntu
2. Recherche et manipulation des données biologiques (séquences de protéines) sous l'environnement shell: exemples
3. Principes d'alignement de séquences: exemples
4. De la séquence à la structure 3D: manipulation des données structurales: exemples
5. Étude des voies de signalisations et des interactions entre protéines: exemples
6. Docking et analyses des cibles moléculaires des médicaments: exemples

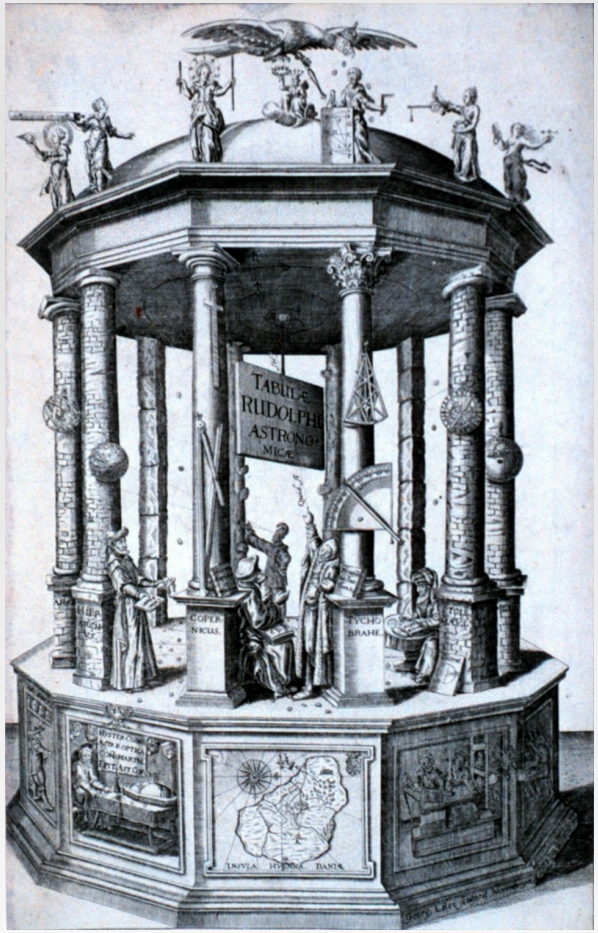
Empirisme européen

Renaissance européen grâce à l'empirisme

Avant:
Observation --> interprétation

Empirisme:
Expérimentation --> observation --> interprétation

L'empirisme a révolutionné la biologie



Histoire de la science

La génétique et l'héridité

Au cours du 20 ème siècle, Mendel a démontré le principe de l'hérédité

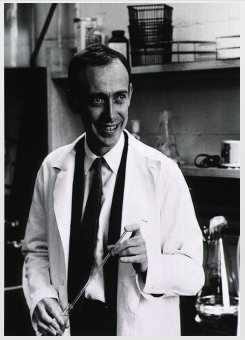
En **1930** la génétique commençait a prendre de l'ampleur et expliquer certaines hypothèses établies auparavant

Gregor Mendel

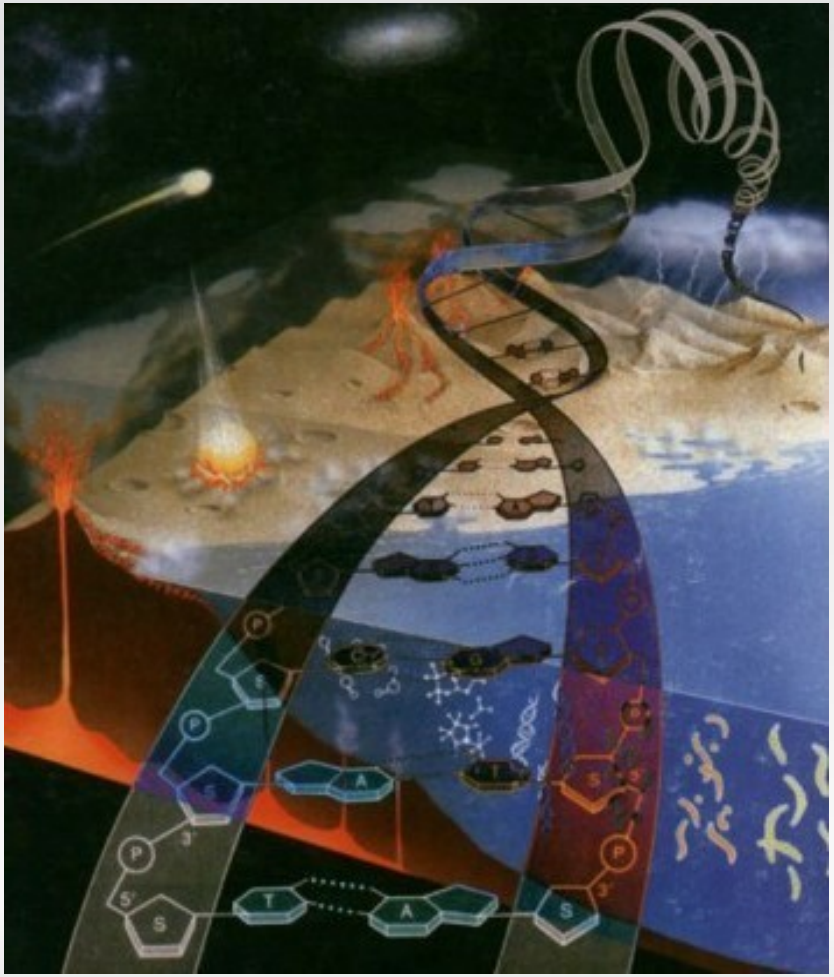
ubuntu

L'ADN

1961



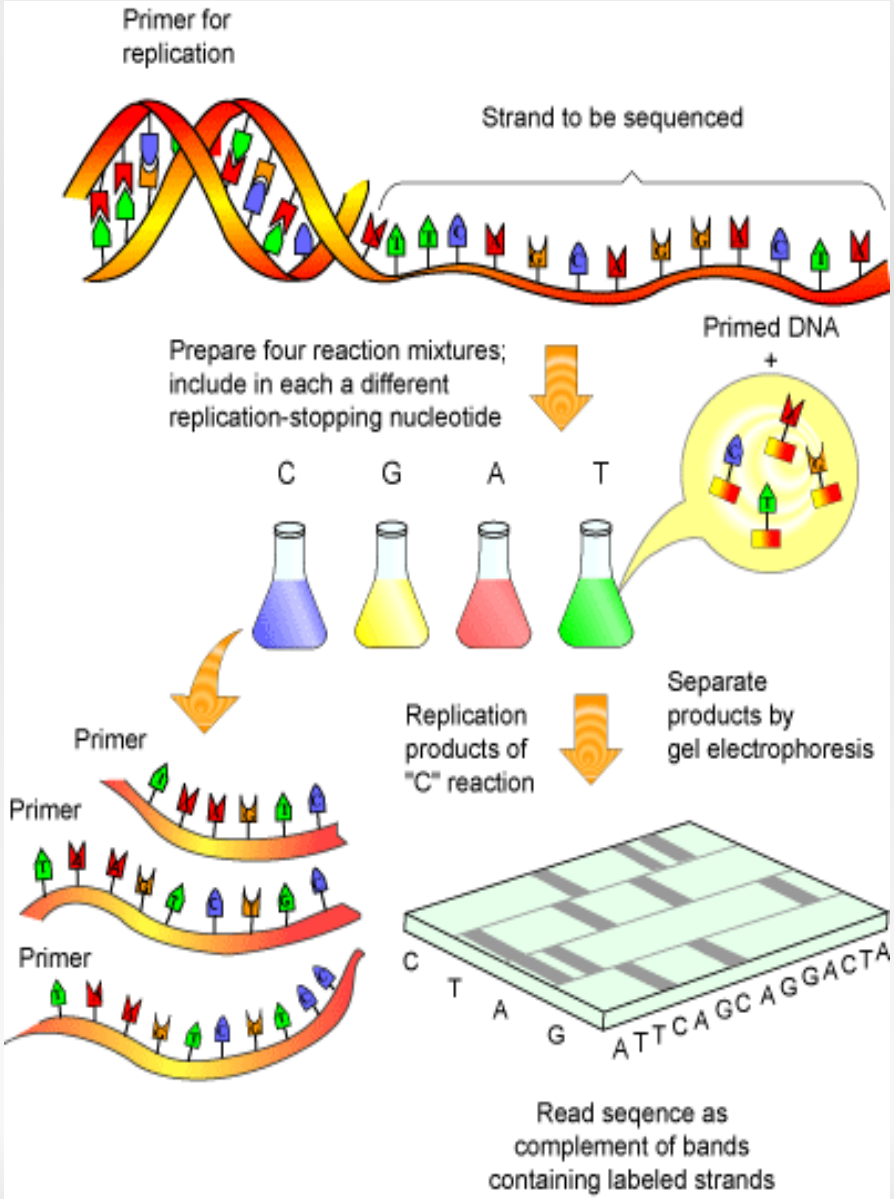
Watson et Francis Crick ont proposé pour la première fois la structure de l'ADN ensuite ils établissaient de codage génétique des 4 bases nucléotides.



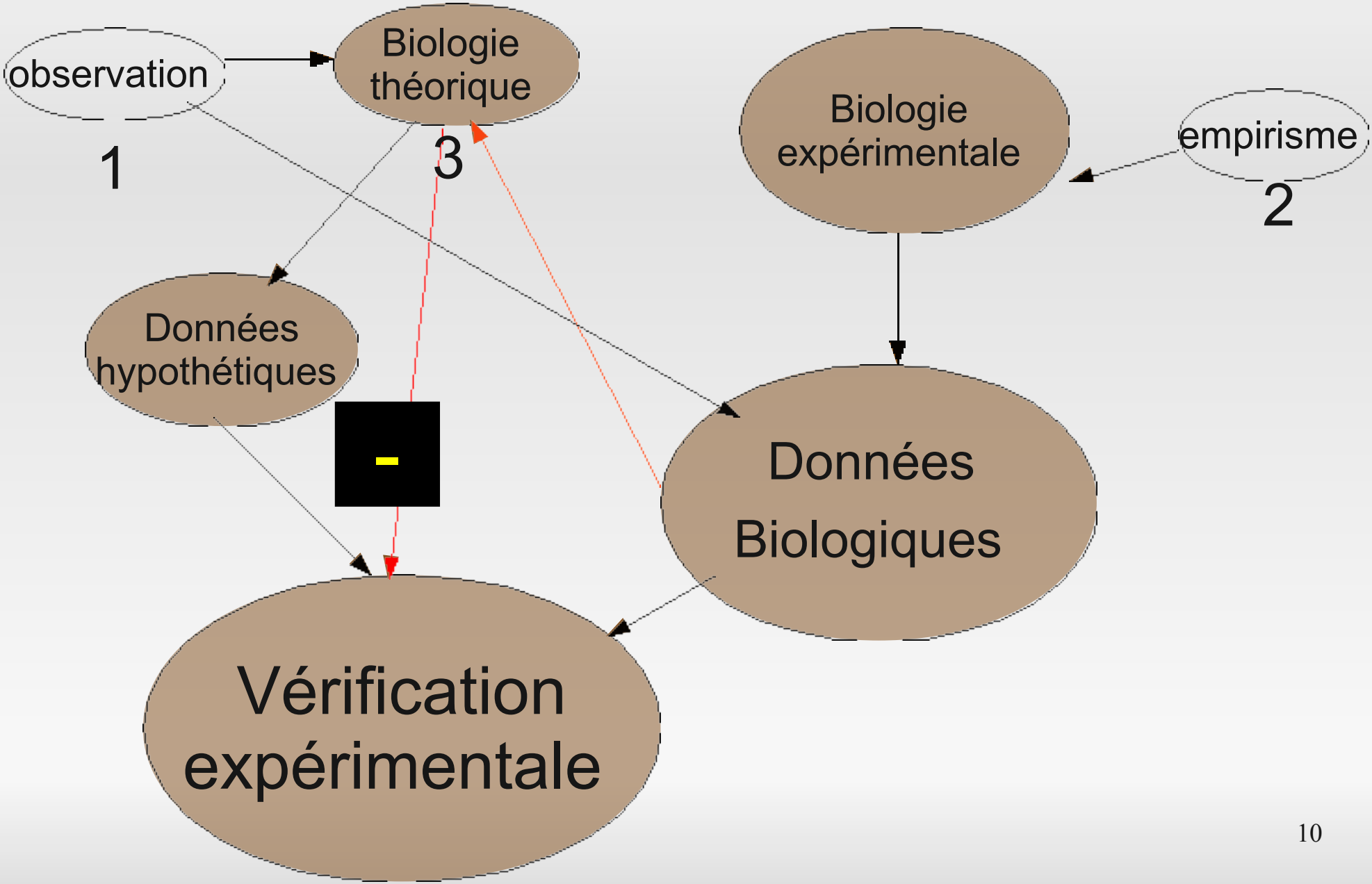
ubuntu

Séquençage des génomes

1984



Évolution de la recherche



Introduction à la bioinformatique
Cas de la drosophile sans yeux

- Modèle animal: embryologie, génétique
- Gène ***eyeless:***
  - Inactivé (knock-out) → mouche sans yeux
  - → eyeless: rôle dans le devlp de l'œil

Chez les Humains:

- Gène: ***aniridie***
  - Inactivé → œil sans iris
  - Suffisamment Muté → œil sans iris

Drosophila melanogaster

Chez la Drosophile knock-out:

- Insertion du gène ***aniridie***
  - → drosophile avec yeux normaux
  - → Ressemblance entre eyeless et aniridie

1ère Partie: introduction générale
Cas de la drosophile sans yeux

Il y a 40 ans:
Recherche des similarités entre séquences d'ADN eyeless et aniridie se fait manuellement avec un logiciel de traitement de texte

- Temps considérable
- Risque d'erreur
- Effet néfaste pour la vue

Fin 1989:
Développement des programme informatique conçus pour comparer des séquences ADN

Aujourd'hui:
Un biologiste peut trouver en quelques secondes des dizaines de correspondances entre des séquences: alignement, analyse phylogénique, identification des motifs...

Etiquettes dans les séquences des gènes

- ADN ou protéine = codage unique par une séquence de caractères = code-barre
- Information codée: produit, fonction, rôle métabolique

```
>gi|225639772|gb|FJ803732.1| Drosophila americana strain HI99.58 twin
of eyeless (Toy) gene, partial cds
GATACAGGCAATGCCCATACCAACTCCGAGAGCCCACCTTTGCAGGCCGTAGCACCCCGCCTACCCCTAA
ACTCTGGTTTCAACACAATGTACTCATCCATCCCGCAACCGATCGCAACCATGGCCGAAAGTTACAAGTA
AGTCTTCTAAGGTCTTCGGCGGCTAGCTAGGAGGCTAAAGAGAAAATTGTATCTAGGAAATTTCACTACC
CTAACGTTCTCAAACTCATTTCCAAAGATCTTTGCATATATATTTAAATTCCTACATAAGAACCTATTGA
TGACGTGATTTAACGTTCAACGATTAAAGAAAAGTTTCAAAATGGGGACGTGCTTCACACTTTTCGGTGG
AAAATGCGTACCCAAAGGTTGAATGGGCGTAGAGTTTGGGTGTGAGATCATTGGCAAAGATCTCTAAGAT
AATAAATTTGAAACGTGTTTCACACTTTTCGACAAACCCAACCCTTCTGGCGTGTAGTGGAGCCACAGGT
AGTTTCAAACTTCGTTTCAAAACCCGACATAAGAACTTAGTGAAAACGTTTTTCGCAACTGTGTAAATTT
GGCGGAAAATTAAAAACGAATCGGCTTCATACTGATTTTAAGAGGTCTTTGTGAAAATGTATTTGAGACA
TGTTTCACACTTTTCGGCAAATCCACCCCTCTTGGGCGATTTAATTAATTTATTTAATTAACAAAGATCT
CAAAAAATGTGACAAGTGTTAATTTTGTCTTTATTTTATTT
```

Compraison des séquences protéiques correspondant aux gènes eyless et aniridie par CLUSTAL

```
CLUSTAL 2.0.12 multiple sequence alignment


sp|P26367|PAX6_HUMAN        ---------MQNSHSGVNQLGGVFVNGRPLPDSTRQKIVELAHSGARPCD 41
tr|Q07DR6|Q07DR6_DAPPU      MPHKADFCFHHSGHSGVNQLGGVYVNGRPLPDSTRQKIVELAHSGARPCD 50
                                     :..**********:**************************

sp|P26367|PAX6_HUMAN        ISRILQVSNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSKIAQYK 91
tr|Q07DR6|Q07DR6_DAPPU      ISRILQVSNGCVSKILGRYHETGSIKPRAIGGSKPRVATSCVVTKIADYK 100
                            *******************:*****:*************. **:***:**

sp|P26367|PAX6_HUMAN        RECPSIFAWEIRDRLLSEGVCTNDNIPSVSSINRVLRNLASEKQQ----M 137
tr|Q07DR6|Q07DR6_DAPPU      RECPSIFAWEIRDRLLSEGVCNNDTIPSVSSINRVLRNLAAQKEQQSAAA 150
                            *********************.**.***************::*:*

sp|P26367|PAX6_HUMAN        GADGMYDKLRMLNGQTGSWGTRPGWYPGTS---------VPGQPTQDG-- 176
tr|Q07DR6|Q07DR6_DAPPU      ASESVYDKLRMFNGQAGGWAWYPGAPPPTHPALSLPHGHVPATPLPSGGG 200
                            .::.:******:***:*.*.  **  * *          **. *  .*

sp|P26367|PAX6_HUMAN        -----CQQQEGGGENTNSIS------SNGEDSDEAQMRLQLKRKLQRNRT 215
tr|Q07DR6|Q07DR6_DAPPU      GHHLIRDDPHKRDCNDGNISXGNSDLHNSSGDEDPQMRLRLKRKLQRNRT 250
                                  :: .  . * ..**       *....::.****:**********

sp|P26367|PAX6_HUMAN        SFTQEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKW 265
tr|Q07DR6|Q07DR6_DAPPU      SFTNEQXESLEKEFERTHYPDVXARERLAEKIGLPEAR------------ 288
                            ***:** *:************* ****** **.*****

sp|P26367|PAX6_HUMAN        RREEKLRNQRRQASNTPSHIPISSSFSTSVYQPIPQPTTPVSSFTSGSML 315
tr|Q07DR6|Q07DR6_DAPPU      --------------------------------------------------


sp|P26367|PAX6_HUMAN        GRTDTALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPTSPSVN 365
tr|Q07DR6|Q07DR6_DAPPU      --------------------------------------------------


sp|P26367|PAX6_HUMAN        GRSYDTYTPPHMQTHMNSQPMGTSGTTSTGLISPGVSVPVQVPGSEPDMS 415
tr|Q07DR6|Q07DR6_DAPPU      --------------------------------------------------


sp|P26367|PAX6_HUMAN        QYWPRLQ 422
tr|Q07DR6|Q07DR6_DAPPU      -------
```

(*) Identique
(:) propriété conservé (acide, basique, aromatique, hydroxyl)
(.) semi conservé

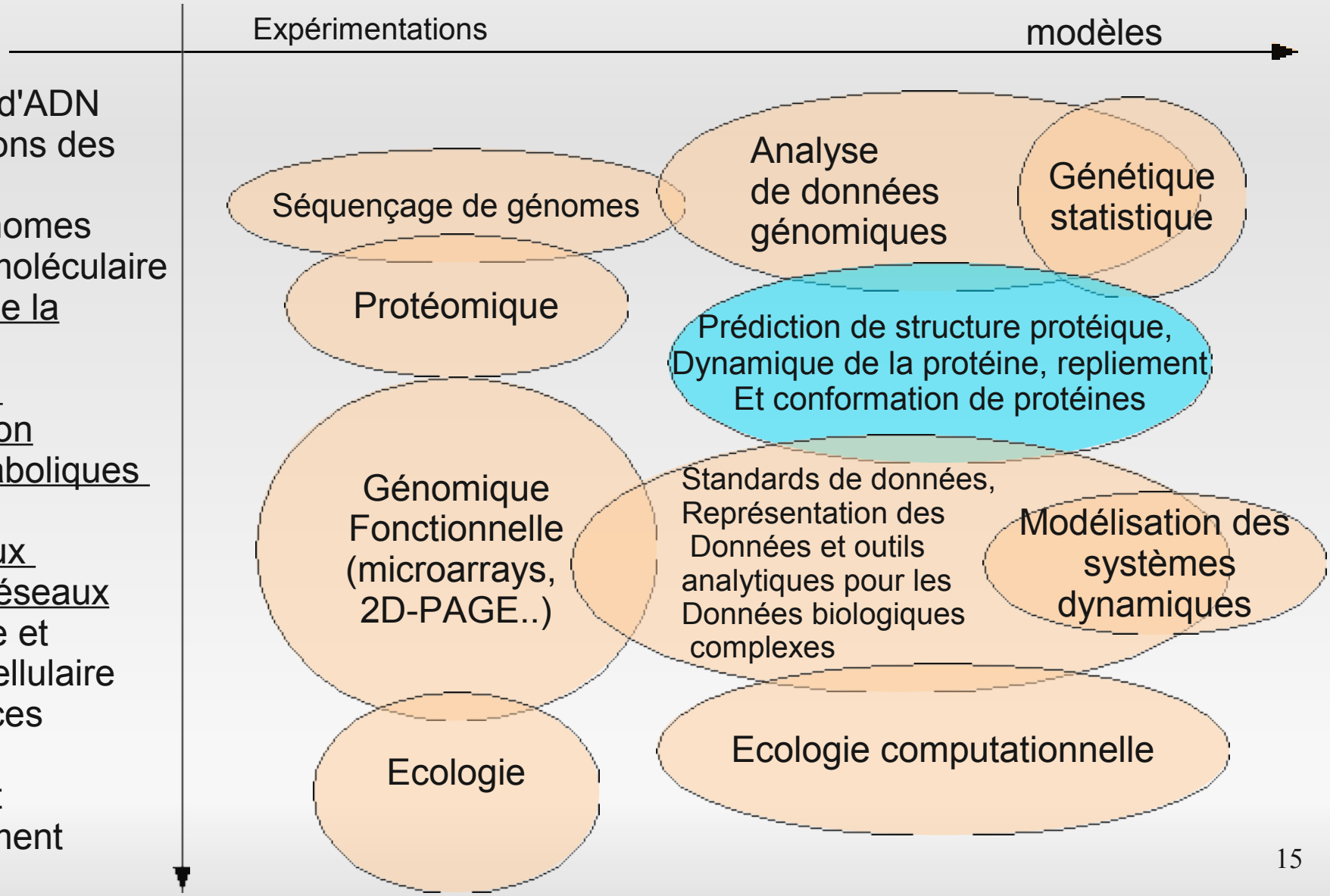
Étudier le gène eyeless = étudier le gène aniridie

http://www.ebi.ac.uk/Tools/clustalw2/

But de la Bioinformatique?

Interactions entre la technologie et la biologie

Séquence d'ADN
Organisations des gènes
et des génomes
Évolution moléculaire
Structure de la protéine, repliement, conformation
Voies métaboliques
Régulation
Signaux
Réseaux
Physiologie et
Biologie cellulaire
Inter-espèces
Interaction
Écologie et environnement



La première ére de l'information en biologie

- La taxonomie fut le premier problème d'information traité en biologie
- De nos jours, les biologistes ont atteint un point de surcharge d'information sur les gènes.
- Que représente la bioinformatique pour les biologistes:
  - 1er volet stockage, distribution de données:
    - conception de formats de données
    - Schéma de bases de données
    - Développement d'outils d'interrogation des données
  - 2ème volet: analyse des données collectées
    - Comparaison des séquences d'ADN ou protéines
    - <u>Identification de fonction</u>
    - <u>Prédiction de structure de protéine</u>

Une ère nouvelle dans la collecte de données

- 1990 – Internet: structure centrale d'information
  - Développement des moyens de communication

Biologie → Bioinformatique ← Informatique

- Méthodes expérimentales progressent:
  - Séquençage du génome Humain....
  - La spectrométrie de masse
  - La cristallographie
  - La Résonance magnétique nucléaire

Comment configurer son PC pour faire de la bioinformatique?

Disponibilités sur le réseau (net)

1. Bases de données stockées dans des serveurs (ordinateurs à grande capacité (processeurs..)
2. Les serveurs fonctionnent sous Unix: système d'exploitation sans interface graphique

Moyens pour faire de la bioinfo

1. Utiliser un système d'exploitation équivalent
    1. Profiter des outils de pointes en (bio)informtique
        1. Installation des programmes

Quels sont les informations et les logiciels disponibles?

Bases de données des informations biologiques

Home | Feedback | Support

15 Mar 2010

BioMed Central
Databases

home | browse catalog | search catalog | BioMed Central home

- Databases home
- Catalog of databases on the web
- Search catalog of databases on the web
- Biology Image Library
- Contact us

Databases on the web

BioMed Central maintains this collection of links to biomedical databases. BioMed Central does not endorse these databases, but we hope you find the collection useful. Please contact us if you have a suggestion for a database we could add, or to tell us about changes affecting a database we already list.

Browse the links by subject area below, or search by name, description or other information.

Biochemistry [305]	Infectious diseases [39]
Bioinformatics [202]	Medical education [9]
Biotechnology [12]	Medical genetics [231]
Blood disorders [23]	Medical imaging [6]
Cancer [43]	Microbiology [143]
Cardiovascular disorders [5]	Molecular biology [287]
Cell biology [73]	Musculoskeletal disorders [4]
Chemical biology [30]	Nephrology [4]
Clinical pathology [13]	Neurology [13]
Clinical pharmacology [2]	Neuroscience [28]
Complementary and alternative medicine [2]	Ophthalmology [5]

Quick search

Go!

www.HIVSecrets.com/FreeReport
Ads by Google

http://databases.biomedcentral.com/browsecatalog

Quels sont les informations et les logiciels disponibles?

Liens utiles des programmes utiles en biologie

debian

Choisissez un serveur proche de vous

États-Unis | Lancer

À propos de Debian | Actualités | Obtenir Debian | Assistance | Le coin du développeur | Plan du site | Recherche

Note : le document original n'existe plus.

Debian-Med : Biologie moléculaire et génétique

La liste ci-dessous comprend divers logiciels qui présentent un intérêt pour le projet Debian-Med. Seulement certains d'entre eux sont actuellement disponibles sous forme de paquets Debian officiels. Cependant, notre but est d'inclure dans le projet Debian-Med tout logiciel qui peut sensiblement améliorer la qualité d'une distribution Debian personnalisée (*Custom Debian Distribution*).

Pour une meilleure vue d'ensemble de la disponibilité du projet sous forme de paquets Debian, les couleurs suivantes sont utilisées :

- Vert : le projet est disponible sous forme de paquet Debian officiel
- Jaune : le projet est disponible sous forme de paquet Debian non officiel
- Rouge : le projet n'est pas encore disponible sous forme de paquet Debian officiel

Si vous connaissez un projet qui vous semble être un bon candidat pour Debian-Med, ou si vous en avez fait un paquet Debian non officiel, n'hésitez pas à en envoyer une description sur la liste de diffusion de Debian-Med.

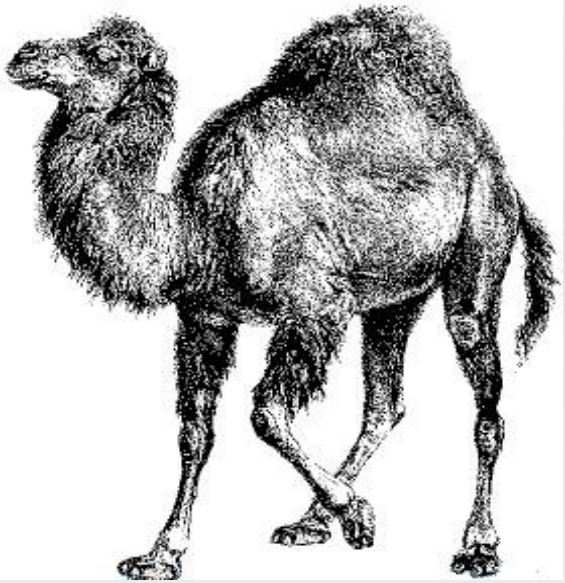
Paquets Debian officiels

Pourquoi installer un programme récupéré sur le web?

- Une page web pourrait résoudre : la consultation d'une base de données. Mais qu'en est-il de la recherche des correspondances à partir de 10000 requêtes, et suite à cette recherche, du tri de l'information afin de trouver des relations entre résultats??

  Objectif d'un logiciel :Traiter des volumes de données considérables et de nature complexe.

Langage de programmation



The Perl Programming Language

http://www.perl.org/
http://www.bioperl.org/wiki/

Python Programming Language

http://www.python.org/

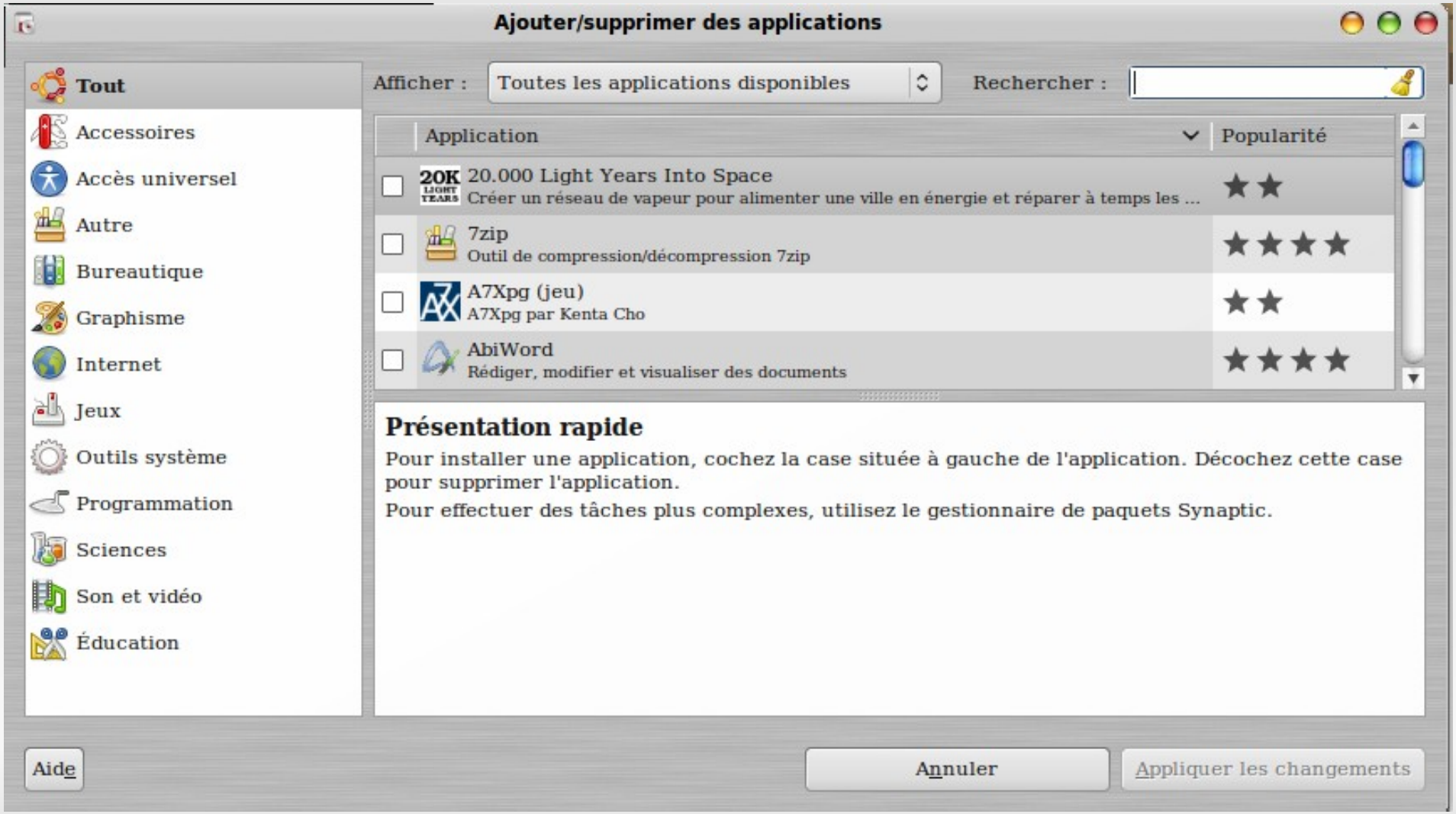
Biopython

2ème Partie: Installation d'un programme chez Ubuntu

- A partir de la Logithèque Ubuntu

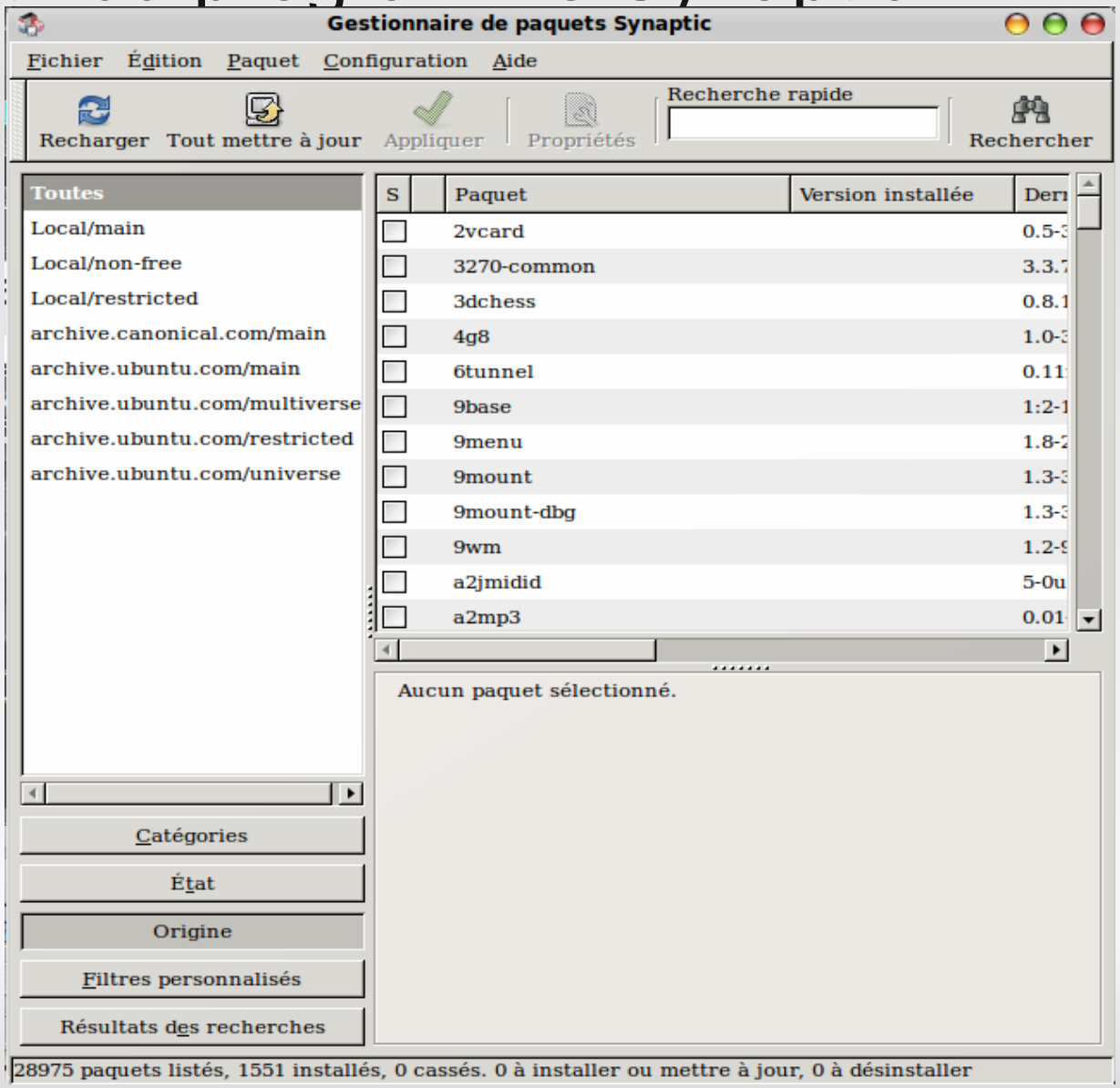
2ème Partie: Installation d'un programme chez Ubuntu

- A partir du programme ajouter/supprimer



2ème Partie: Installation d'un programme chez Ubuntu

- A partir du programme Synaptic



2ème Partie: Installation d'un programme chez Ubuntu

- A partir de la ligne de commande ou le terminal

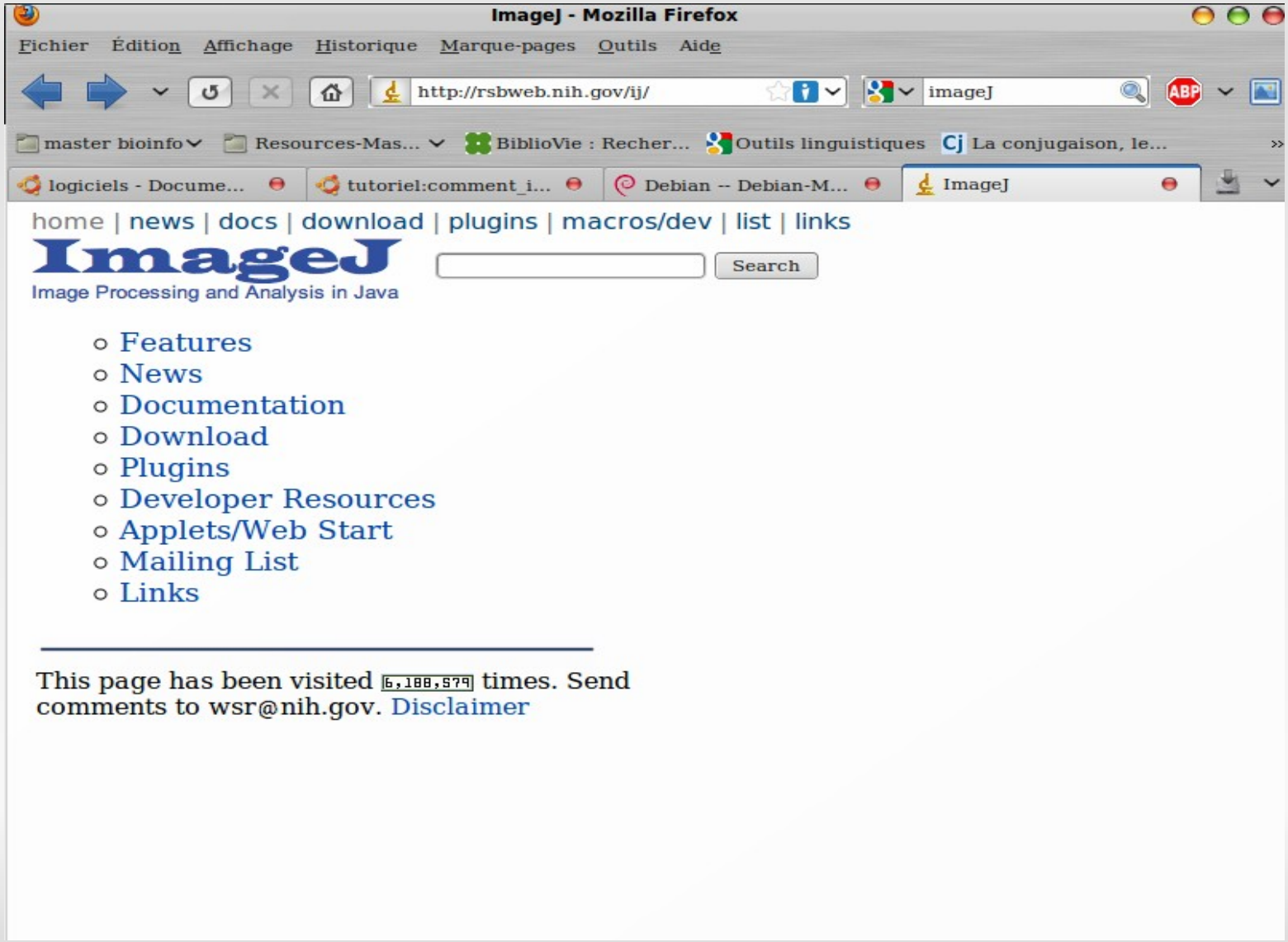
```
Sudo apt-get install nom-packet
```

- Sous le format .deb disponible sur internet

Télécharger le fichier.deb
Double clic sur le fichier

2ème Partie: Installation d'un programme chez Ubuntu

- Installation manuelle des packets



Choisir Linux, debian, 32 bit

logiciels - Docume... | tutoriel:comment_i... | Debian -- Debian-M... | Download

home | news | docs | download | plugins | macros/dev | list | links

Download

Platform Independent

To install ImageJ 1.42 on a computer with Java pre-installed, or to upgrade to the latest full distribution (including macros, plugins and LUTs), download ij142.zip (3MB) and extract the ImageJ directory. Use the *Help>Update ImageJ* command to upgrade to the latest pre-release version.

Mac OS X

Download ImageJ 1.43 (5.4MB) as a double-clickable Mac OS X application. Includes ImageJ64, which uses Java 1.6 in 64-bit mode on Intel Macs running OS X 10.5 or later. (Instructions)

Linux

Download ImageJ 1.42 bundled with 32-bit Java (43MB) or with 64-bit Java (36MB). Both versions include Java 1.6.0_10 from Sun and the ImageJ source code. (Instructions)

Windows

Download ImageJ 1.42 bundled with 32-bit Java 1.6.0_10 (26MB), with 64-bit Java 1.6.0_12 (22MB; requires XP x64 or Vista 64-bit) or without Java (3MB). On Vista, ImageJ must be installed in a directory that the user can write to (e.g., "Documents"). (Instructions)

Documentation

Download the ImageJ documentation (25MB), in HTML format, as a compressed zip archive. The JavaDoc API documentation (567K) is also available.

Source Code

Télécharger le tar.gz

Download - Mozilla Fi... | [bioinfo-intro - Navig... | bioinfo_intro_drosoph...

bioinfo_intro_drosophile.odp - Open

File Edit View Insert Format Tools Slide Show Window Hel

Slides

Normal | Outline | Notes | Handout | Slide Sort

Choisir Linux, debian, 32 bit

logiciels - Docume... | tutoriel:comment_i... | Debian -- Debian-M... | Download

home | news | docs | download | plugins | macros/dev | list | links

Download

Platform Independent

To install ImageJ 1.42 on a computer with Java pre-installed, or to upgrade to the latest full distribution (including macros, plugins and LUTs), download ij142.zip (3MB) and extract the ImageJ directory. Use the *Help>Update ImageJ* command to upgrade to the latest pre-release version.

Mac OS X

Download Ima
application. In
on Intel Macs

Linux

Download Ima
64-bit Java (36
and the Image

Windows

Download Ima
with 64-bit Jav
without Java (
that the user c

Ouverture de ij142-x86.tar.gz

Vous avez choisi d'ouvrir

ij142-x86.tar.gz

qui est un fichier de type : archive gzip

à partir de : http://rsbweb.nih.gov

Que doit faire Firefox avec ce fichier ?

Ouvrir avec | Gestionnaire d'archives (défaut)

Enregistrer le fichier

Toujours effectuer cette action pour ce type de fichier.

Annuler | OK

Documentation

Download the ImageJ documentation (25MB), in HTML format, as a compressed zip archive. The JavaDoc API documentation (567K) is also available.

Source Code

41.229.140.130 | Maintenant : Partiellement ensoleillé, 15 °C | Mer : 18 °C | Jeu

mar. 16 mars, 18:23

Décompréssser un dossier .tar

1- Double clic sur le dossier

2 avec la ligne de commande:

`tar -zxvf votre_archive.tar.gz` si le fichier est .tar.gz
`tar -xvf archive.tar` si le fichier est .tar

Chercher le fichier README pour lire les instructions

Fichier README de imageJ

README.html (Système de fichiers 200 GB) - gedit

Fichier Édition Affichage Rechercher Outils Documents Aide

Ouvrir Enregistrer Annuler

README.html

```
1 <html>
2 <head>
3 <title>ImageJ 1.42</title>
4 </head>
5 <body>
6 <h3>ImageJ 1.42</h3>
7 The ImageJ website is at
8 <a href="http://rsb.info.nih.gov/ij/">rsb.info.nih.gov/ij/</a><br>
9 and the Linux installation notes are at<br>
10 <a href="http://rsb.info.nih.gov/ij/docs/install/linux.html">rsb.info.nih.gov/ij/docs/install/linux/</a>.
11 </body>
12 </html>
```

```
mezhoud@sunny-deb:/media/8294-7434/cours/imageJ$ ls
ij142-x86.tar.gz  ImageJ
mezhoud@sunny-deb:/media/8294-7434/cours/imageJ$ cd ImageJ
mezhoud@sunny-deb:/media/8294-7434/cours/imageJ/ImageJ$ ls
ij142q-src.zip  ImageJ.png  luts    plugins     run
ij.jar          jre         macros  README.html  script
mezhoud@sunny-deb:/media/8294-7434/cours/imageJ/ImageJ$
```

HTML | Largeur des tabulations: 8 | Lig 1, Col 1 | INS

Présence de fichier run en couleur verte (excécutable)

- Dans le fichier il y a commande suivante:

```
./jre/bin/java -Xmx512m -jar ij.jar
```

Càd qu'on peut excécuter directement le programme en <u>étant dans le dossier ImageJ</u> par la commande suivante:

```
./run
```

2ème programme: PerlMol - Perl Modules for Molecular Chemistry

Introduction

Features

Quick Tutorial

Reference Manual

Examples

Download & Installation

Mailing List

Bugs and Requests

Support this project

SOURCEFORGE.NET

PerlMol - Perl Modules for Molecular Chemistry

PerlMol is a collection of Perl modules for chemoinformatics and computational chemistry with the philosophy that "simple things should be simple". It should be possible to write one-liners that use this toolkit to do meaningful "molecular munging". The PerlMol toolkit provides objects and methods for representing molecules, atoms, and bonds in Perl; doing substructure matching; and reading and writing files in various formats. For a more detailed feature list, see the Features section.

Recent News

- **2009-05-10** Updated distributions: Chemistry-Mol-0.37, Chemistry-Pattern-0.27, Chemistry-Ring-0.20, Chemistry-Canonicalize-0.11, Chemistry-Bond-Find-0.23 Chemistry-File-MDLMol-0.21, Chemistry-File-PDB-0.23, Chemistry-File-SMILES-0.46,

Project status

The current status could be called "alpha", because there are still many missing or untested features, but there is enough for being useful for some real work. The Feature list shows how complete the current implementation for a given feature is.

The following CPAN distributions have been released:

- Chemistry-Mol

Automated install of PerlMol

Introduction

Features

Quick Tutorial

Reference Manual

Examples

Download & Installation

Mailing List

Bugs and Requests

Support this project

SOURCEFORGE.NET

PerlMol - Download and Installation

The PerlMol modules are being developed for perl 5.8.x and are known to work under Windows, Linux, FreeBSD, Solaris, and OS X. They are backward-compatible with perl 5.6.x, but you are strongly encouraged to upgrade to 5.8.x because support for older versions may be removed in the future. PerlMol depends on other modules available on CPAN, such as Math::VectorReal. See the README files for details.

There are three main ways of installing the PerlMol modules:

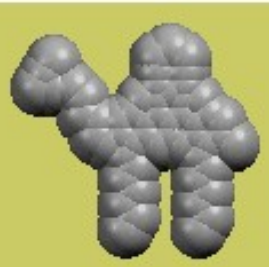
1. Automated installation using the CPAN.pm module - recommended way for UNIX-compatible systems.
2. Manual install of the PerlMol bundle - alternative way if CPAN.pm is not an option for some reason.
3. PPM package for Windows ActivePerl 5.8 - easiest way of installing under Windows.
4. CVS Repository - for developers that know what they are doing and are interested in trying the latest unreleased development versions.

Automated installation via CPAN.pm

If you have the CPAN module configured properly, you can install each module separately by using `perl -MCPAN -e 'install Chemistry::Mol'` (to give an example). To install the latest bundle do `perl -MCPAN -e 'install PerlMol'`. **But note that some of the most recent modules may not be included in the bundle yet.**

You can also download the tarballs for individual modules and install them yourself by following the instructions on the README files, either from CPAN or from SourceForge.net.

Exemple de programme



Introduction

Features

Quick Tutorial

Reference Manual

Examples

Download & Installation

Mailing List

Bugs and Requests

Support this project

PerlMol - Examples

Here are some sample scripts using PerlMol. Most of them have been submitted or suggested by PerlMol users; if you have something that you can share and could be educational, please send it (itubert@users.sourceforge.net)! The ideal example is relatively short (about 100 lines), self-contained, and does something potentially useful.

This scripts, along with some sample input and output files, are included in the PerlMol bundled distribution under the examples/ directory.

- combinatorial_enumeration
- file_conversion
- molgrep
- pdb_viewer
- peptide_builder
- polar_surface_area

Exemple: conversion de fichiers

Quick Tutorial

Reference Manual

Examples

Download & Installation

Mailing List

Bugs and Requests

Support this project

SOURCEFORGE.NET

Usage

```
./pbabel.pl [options] <input file> <output file>

Options:
    -i <input format>
    -o <output format>
    -3      build 3d coordinates
    -b      find bonds
    -a      aromatize
    -K      don't kekulize (only for reading SMILES)
    -u      unique (only for writing SMILES)
    -n      include name (only for writing SMILES)

Available file formats:
    dumper
    formula
    formula_pattern
    mdl
    midas
    mop
    pdb
    sdf
    sln
    smarts
    smiles
    xyz
```

Files

- pbabel.pl - the program
- amines.sdf - sample sdf file
- amines.smi - sample sdf file, generated with ./pbabel.pl -nu amines.sdf x.smi