Introduction à la Bioinformatique 2023

Ezechiel Bionimian TIBIRI, Ph.D
Bioinformatique – Biologie Moléculaire

Plan

I. ~~Introduction~~

II. ~~Bases de la biologie moléculaire~~

III. Bases de la programmation

IV. ~~Recherche de séquences~~

V. Théorie et applications d'alignement de séquences

VI. Alignement de séquences multiples

VII. Évolution moléculaire et phylogénétique

VIII. ~~Analyse de génomes~~

IX. ~~Analyse de transcriptomes~~

X. ~~Analyses de protéines~~

Bioinformatique - définitions

- On trouve un grand nombre de définitions selon l'acception du terme et selon la prépondérance de "*bio*" sur "*informatique*" ou l'inverse.

Bioinformatique - définitions

La bioinformatique, c'est quoi ?

L'utilisation de l'informatique
pour l'analyse de données biologiques.

Bioinformatique - définitions

- *Surtout:*
  - *Biologie* + *Informatique*
  - *Biochimie* + *Informatique*
- *Mais aussi…*
  - *Médecine* + *Informatique*
  - *Pharmacie* + *Informatique*
  - *Chimie* + *Informatique*
  - *Mathématique* + *Informatique*
  - *Statistique* + *Informatique*
- *C'est un domaine pluridisciplinaire!*

Bioinformatique - définitions

- 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 ...

Bioinformatique - définitions

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

Bioinformatique - définitions

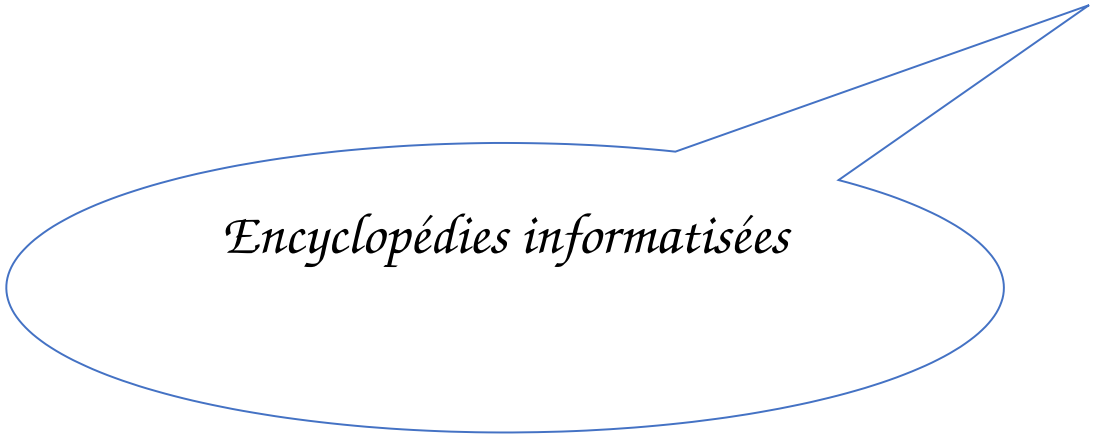
- Une définition de la bioinformatique : analyse de la bioinformation par des moyens informatiques.

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

Bioinformatique - définitions

Pourquoi faire ?

Acquérir puis stocker les informations biologiques sous la forme d'encyclopédies appelées bases de données;



Bioinformatique - définitions

Exemples de données 'biologiques' qui ne peuvent plus être gérées sans l'aide de l'informatique:

- *Séquences: ADN (génomes), ARN, protéines*
- *Structures 3D: ADN, ARN, protéines, sucres…*
- *Classification des espèces*
- *Voies métaboliques*
- *Expression des gènes (microarrays)*
- *Spectrométrie de masse*
- *Publications scientifiques*

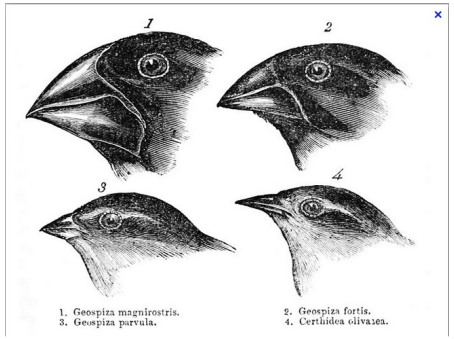
…

Beaucoup de 'omics',
mais… !

Bioinformatique - définitions

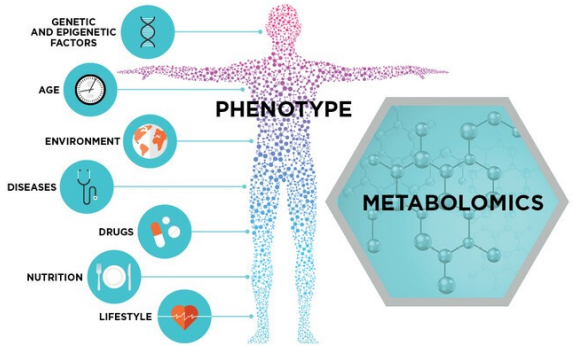
https://www.marketsandmarkets.com/pdfdownloadNew.asp?

Rappels historiques



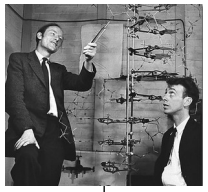
Https://fr.wikipedia.org/wiki/Pinsons_de_Darwin

- De la théorie darwinienne (1859) à la métabolomique (exploration cellulaire en temps réel)



https://www.mtidx.com/our-technology/metabolomics

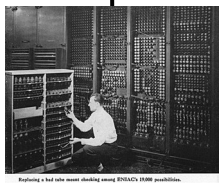
Structure De l'ADN 1953



1958 Dogme central

1957 FORTRAN

IBM 650



ENIAC 1945

Détermination d'une séquence protéique par ordinateur **1964**

Premiers arbres phylogénétiques **1967**

Margaret Oakley Dayhoff

Dayhoff the "mother and

1965 Atlas of Protein Sequence and Structure

Algorithme pour l'alignement global de séquences **1970**

1972 Clonage

1971 Premier microprocesseur Intel 4004

1969 UNIX et ARPANET

prédiction de structures secondaires 1974

TPC/IP

Séquençages ADN 1977

Algo. alignement Local 1981

1978 - 1980 EMBL, GenBank, PIR

1981 IBM-PC 8088

1983 IBM-XT (10 Mb)

PCR 1984

FASTA 1985

1988 NCBI

1989 INTERNET Python

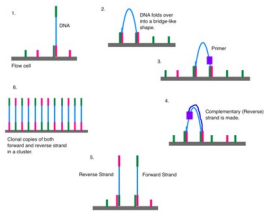
1990

EMBL-EBI 1993

11 génomes bactéries séq. 1997

2001 1er WGS humain

NGS 2007



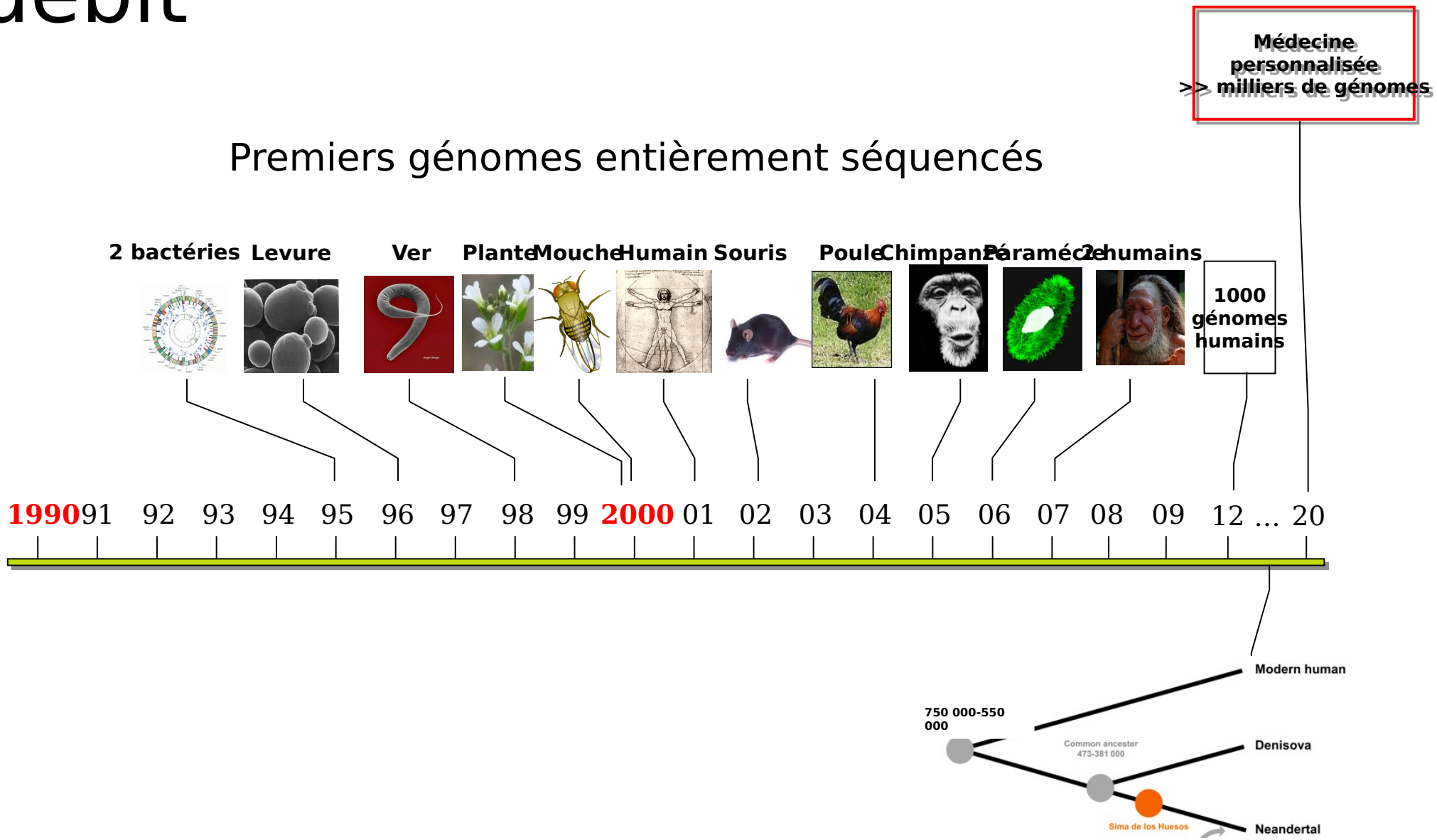
Mise q-Hise q 2011

2015 MinION

Médecine personnalisée >> milliers de génomes

2017

Production des données à haut débit



UNIX PEOPLE ARE HAPPY

UNIX @ BIOINFO

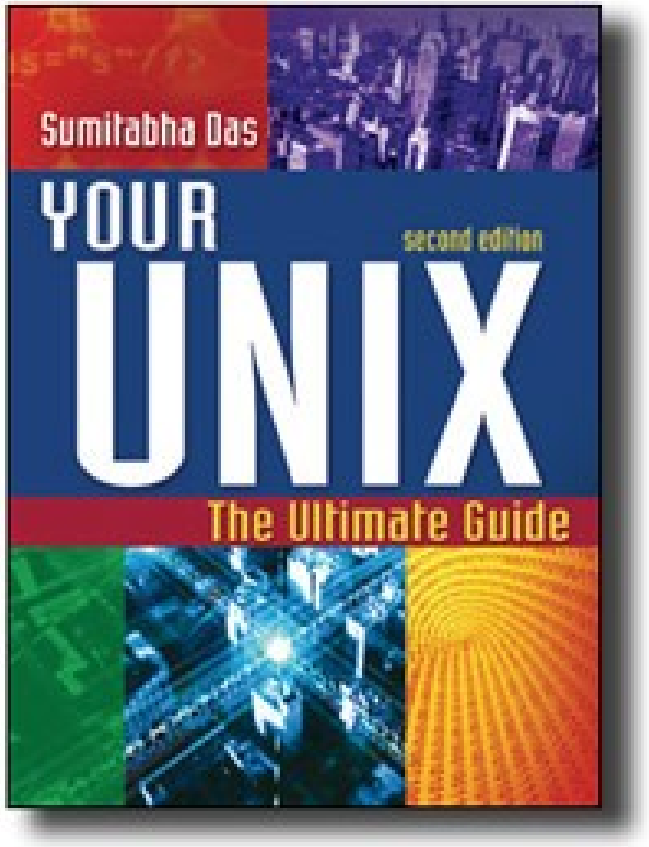
Ezechiel B. TIBIRI

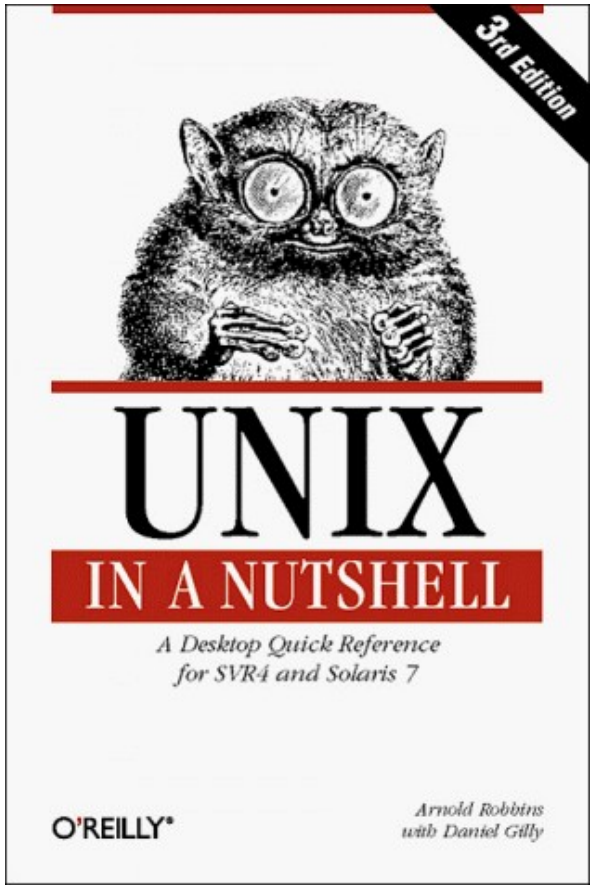
What will we cover?

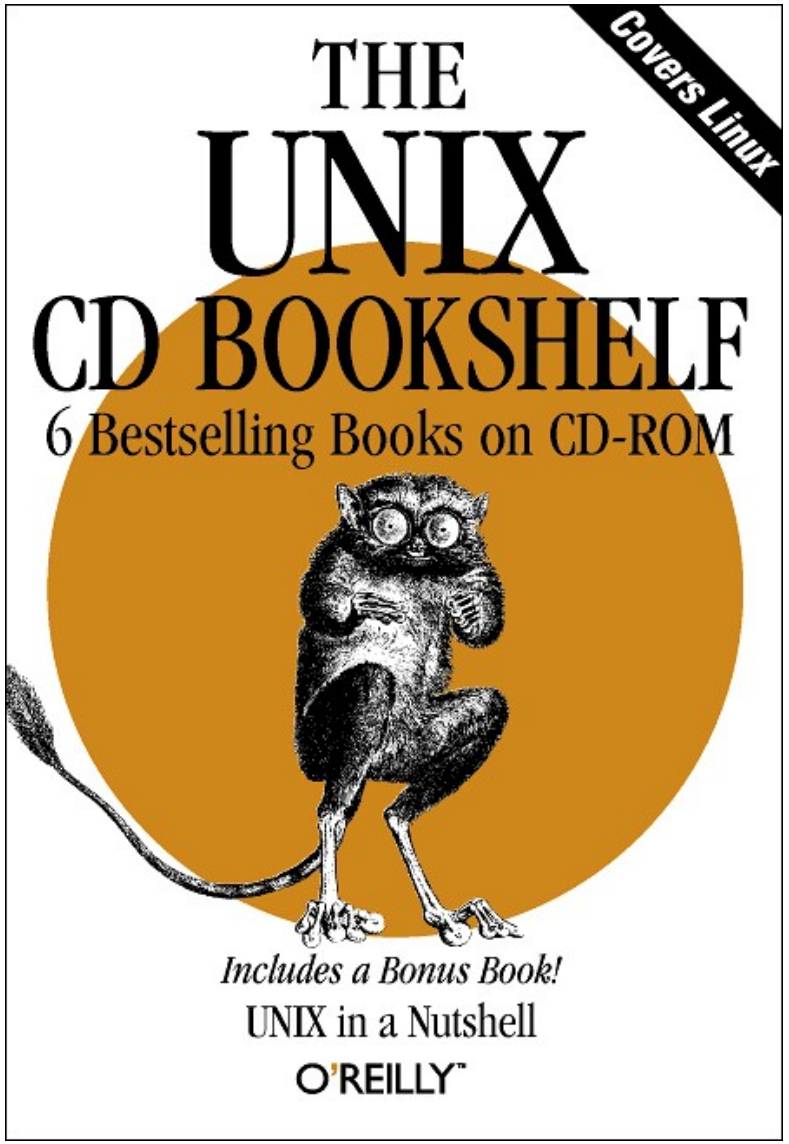
- Operating system overview
- Basics of the Unix command line interface
- Manipulating files and directories
- Manipulating data

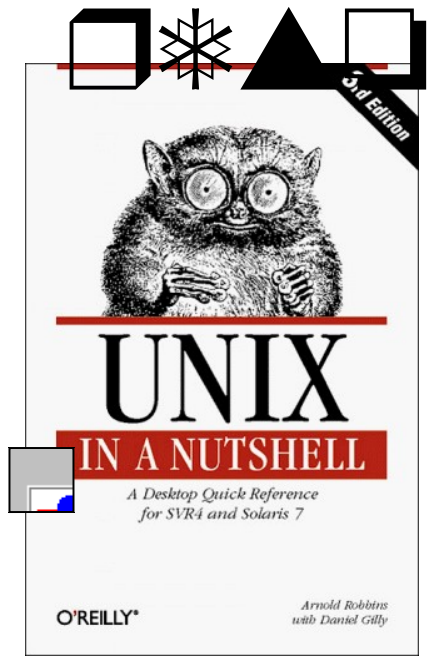
Who cares, how do I get an A?

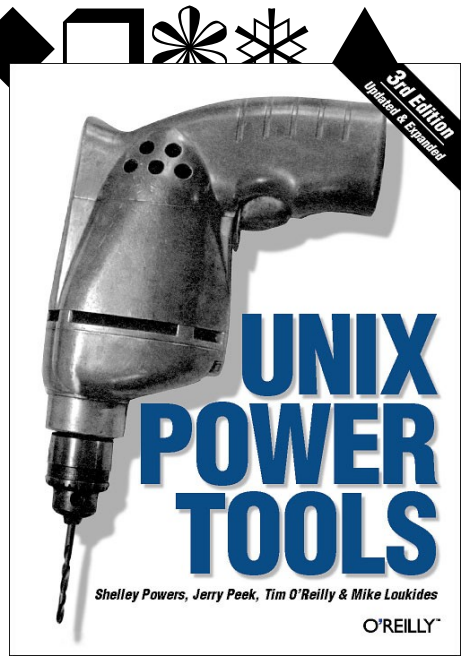
- Assignments: 25%
- Presence 10 %
- Project: 65%

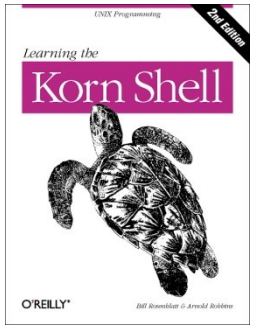

Sumitabha Das
YOUR
second edition
UNIX
The Ultimate Guide

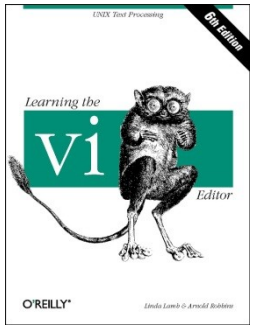

3rd Edition
UNIX
IN A NUTSHELL
A Desktop Quick Reference
for SVR4 and Solaris 7
O'REILLY®
Arnold Robbins
with Daniel Gilly

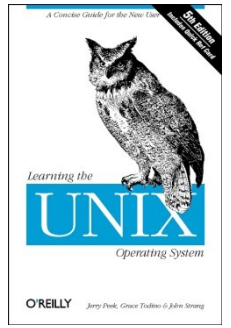


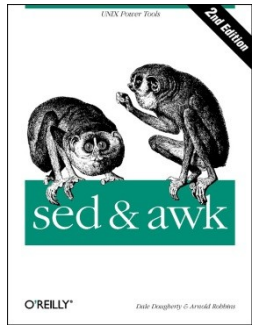












http://safari.oreilly.com

Administrivia

- Make sure you have a laptop (iCore5 & 8GB of RAM)
- Install Window subsystem Linux
- Check the github/slack regularly:

Cheating

- Don’t

Cheating

- Don’t
- Seriously, don’t

Individual Effort

- Assignments and project are open book, open notes, open computer/internet!
- This is a hands on course designed to familiarize YOU with the unix/linux environment.
- You will need these skills in future classes.
- Cheat and pay the price later.
- Why not learn this stuff now?

Lou Gerstner gives his keynote address at the eBusiness Conference and Expo in New York.

IBM to spend $1 billion on Linux in 2001

By Joe Wilcox
Staff Writer, CNET News.com
December 12, 2000, 8:50 a.m. PT

update IBM chief executive Louis Gerstner said Tuesday that his company will spend $1 billion on Linux next year.

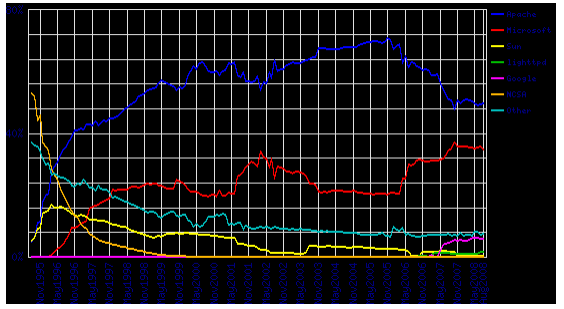
Gerstner made the announcement at the eBusiness Conference and Expo in New York, where IBM also revealed a Linux supercomputer win with Shell Oil.

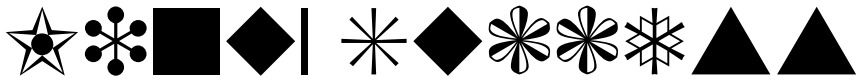
Info appliance makers adopt Linux

Just buzz or actual benefits? More info appliance makers are choosing Linux.

Intel	To use Linux for Intel-branded Web appliances
TiVo	Runs personal video recorder services on Linux
National Semiconductor	Offers Linux choice for its Web Pad platform
Sony	PlayStation 2 development system based on Linux
Transmeta	Bundling Linux for mobile applications with new chip
Lineo	Offers Linux development system for embedded info devices

SOURCE: Company announcements





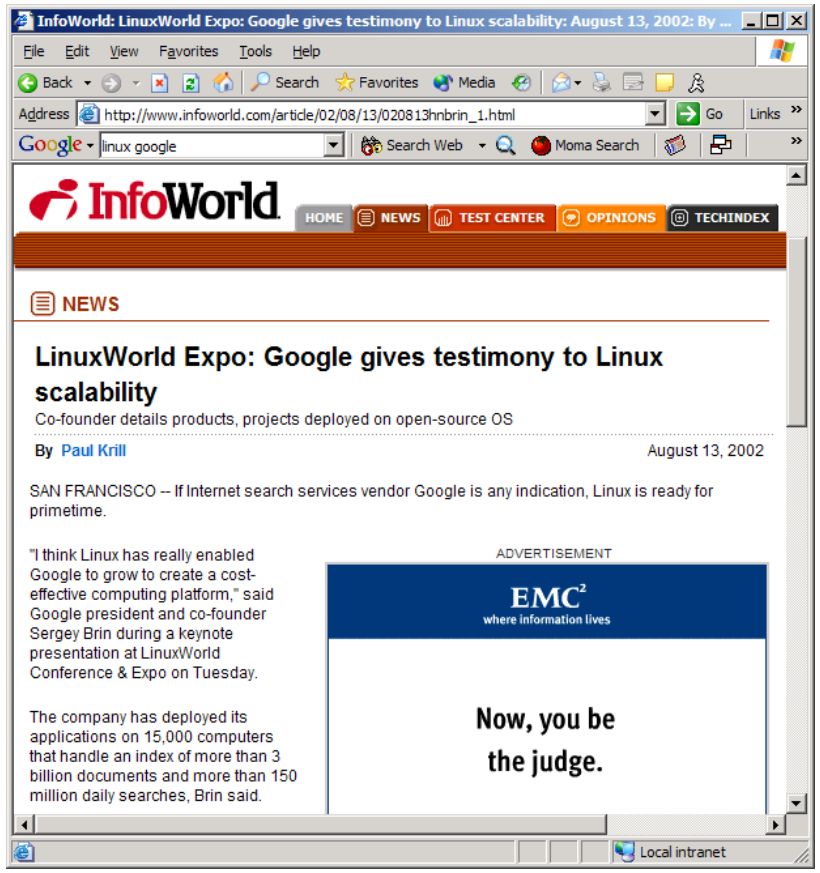
NYSE undertakes IBM mainframe migration to Unix and Linux

By Mark Fontecchio, News Writer
14 May 2007 | SearchDataCenter.com

RSS FEEDS: IT infrastructure news
Add to Google

The New York Stock Exchange (NYSE) is migrating off a 1,600 millions of instructions per second (MIPS) mainframe to IBM System p servers running AIX and x86 Hewlett-Packard Co. (HP) servers running Linux, with the first part of the move going live today.

Linux at Google



The UNIX Philosophy

- Small is beautiful
  - Easy to understand
  - Easy to maintain
  - More efficient
  - Better for reuse
- Make each program do one thing well
  - More complex functionality by combining programs
  - Make every program a filter

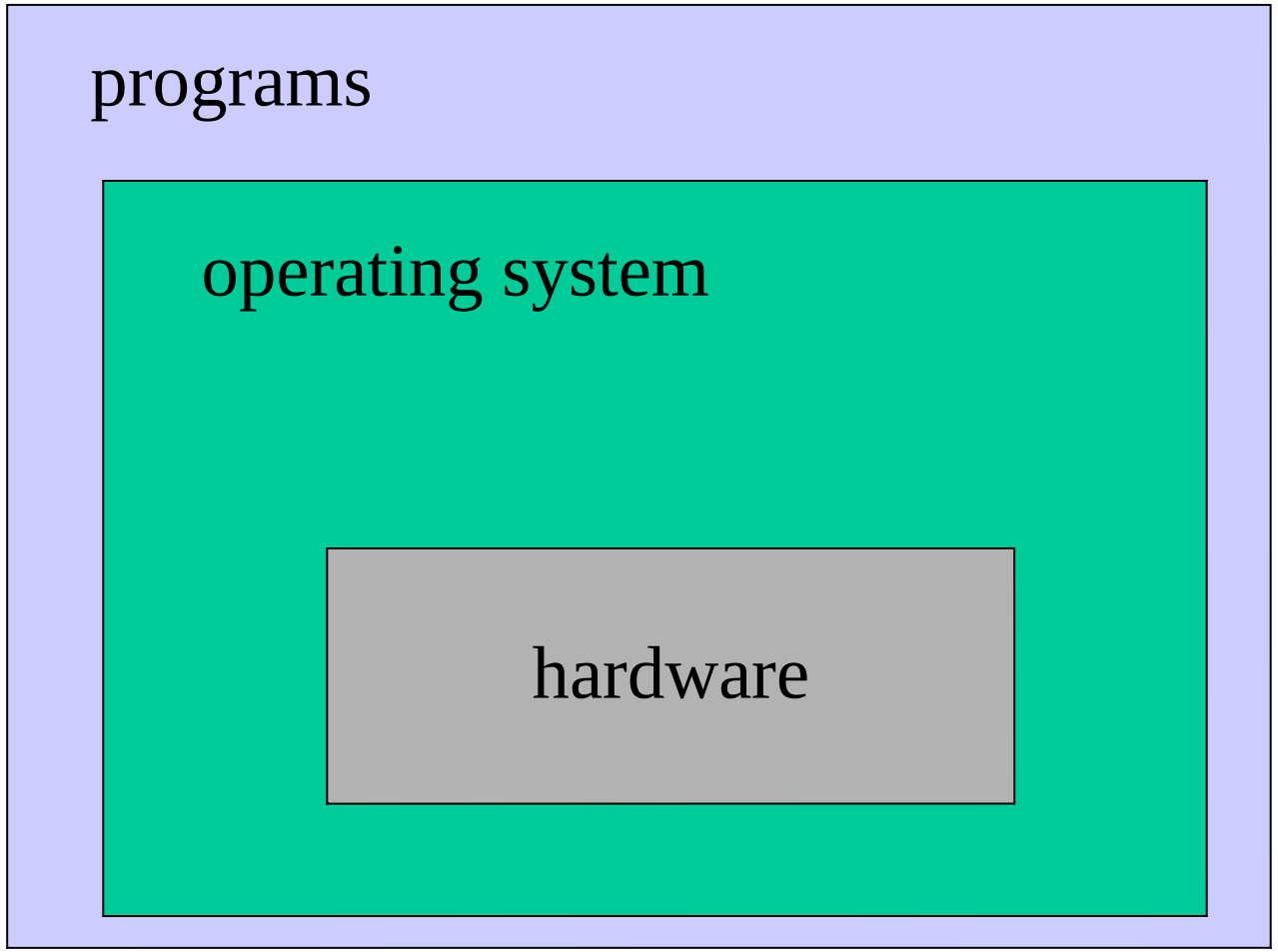
The UNIX Philosophy *..continued*

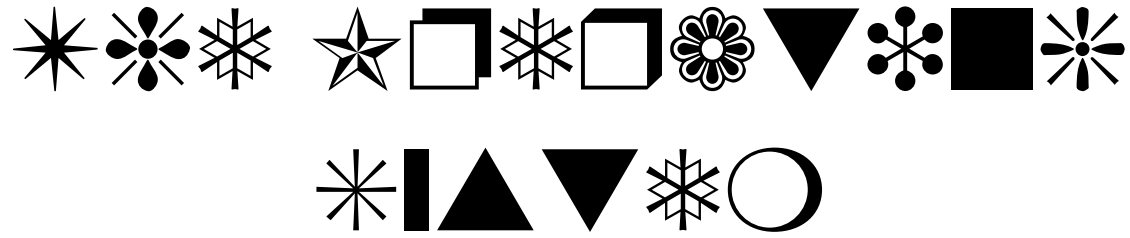
- Portability over efficiency
  - Most efficient implementation is rarely portable
  - Portability better for rapidly changing hardware
- Use flat ASCII files
  - Common, simple file format (yesterday's XML)
  - Example of portability over efficiency
- Reusable code
  - Good programmers write good code; great programmers borrow good code

UNIX Highlights / Contributions

- Portability (variety of hardware; C implementation)
- Hierarchical file system; the file abstraction
- Multitasking and multiuser capability for minicomputer
- Inter-process communication
  - Pipes: output of one programmed fed into input of another
- Software tools
- Development tools
- Scripting languages

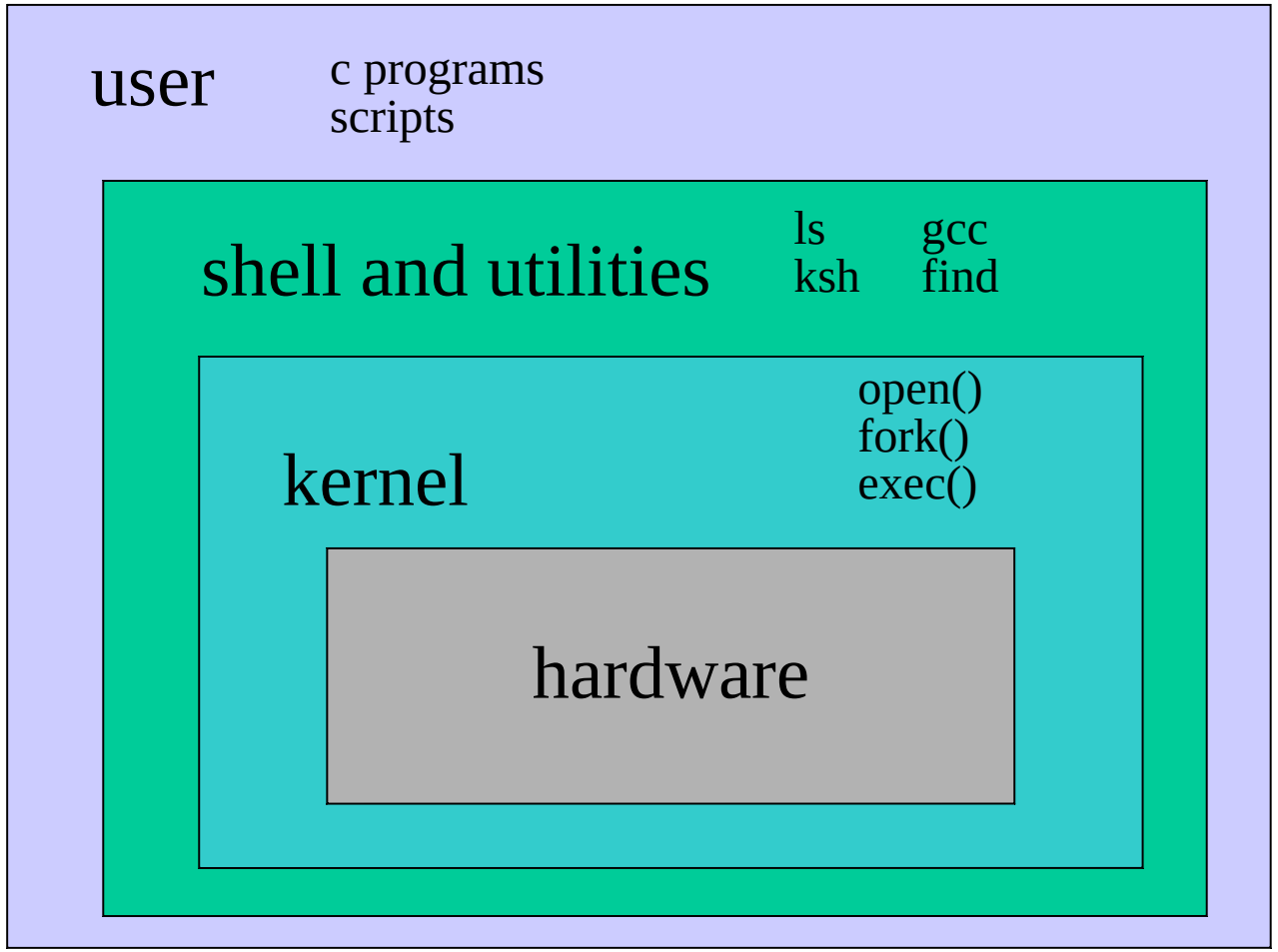
Operating System Structure

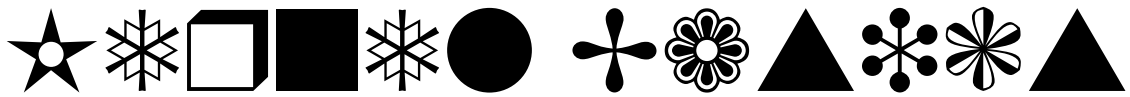




- The government of your computer
- Kernel: Performs critical system functions and interacts with the hardware
- Systems utilities: Programs and libraries that provide various functions through systems calls to the kernel

Unix System Structure

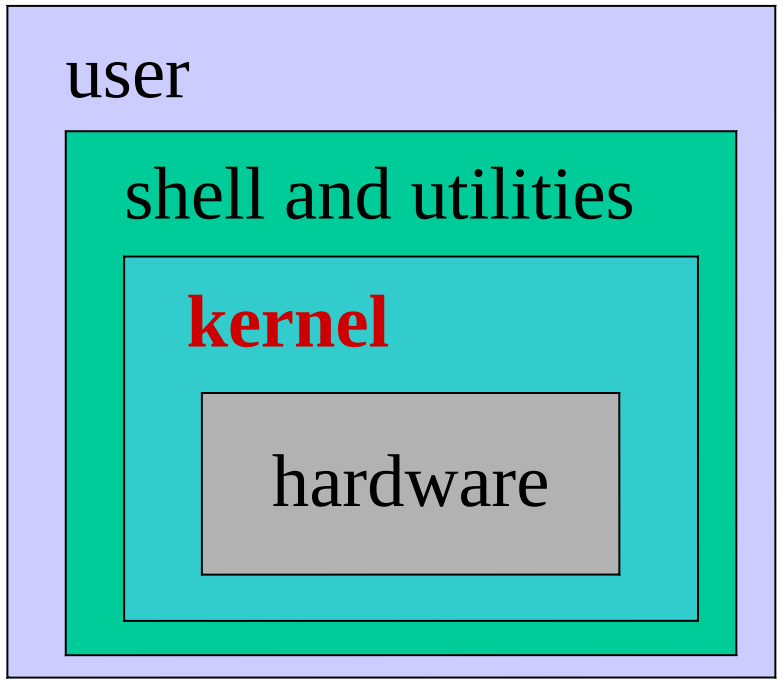




- The kernel is …
  - a program loaded into memory during the boot process, and always stays in physical memory.
  - responsible for managing CPU and memory for processes, managing file systems, and interacting with devices.

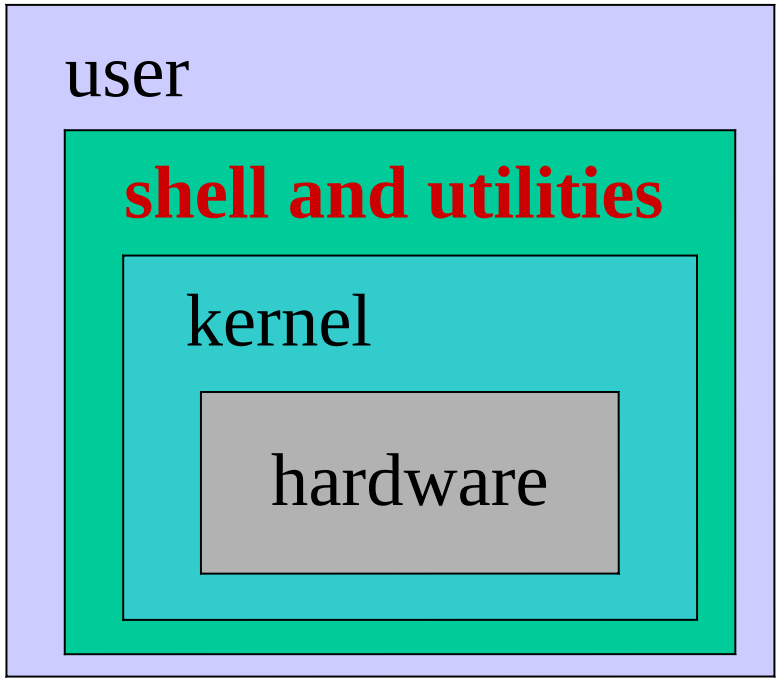
The Kernel

- Manage resources
  - Storage
  - Memory
  - CPU
  - Display
  - Network
- Sharing
  - Users
  - Tasks
- Communication



Shell & Utilities

- The rest of the operating system
- Focus of this course
- Cause of debate in Linux community



UNIX on Windows

Two recommended UNIX emulation environments:

Windows Subsystem for Linux (WSL)

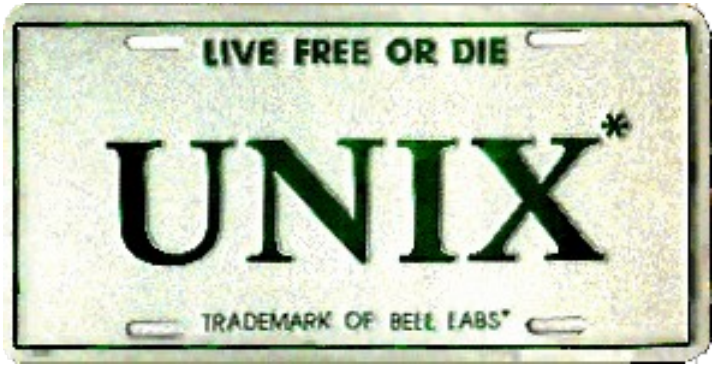
- **https://learn.microsoft.com/en-us/windows/wsl/install**

Cygwin (GPL)

- **http://sources.redhat.com/cygwin**

Next Time

- Basic UNIX concepts
- Introduction to the shell
- Introduction to basic commands



LINUX Tutorials

- http://www.linux-tutorial.info/modules.php?name=Tutorial&pageid=224
- http://www.tldp.org/LDP/intro-linux/html/index.html
- http://www.slackbook.org/html/index.html

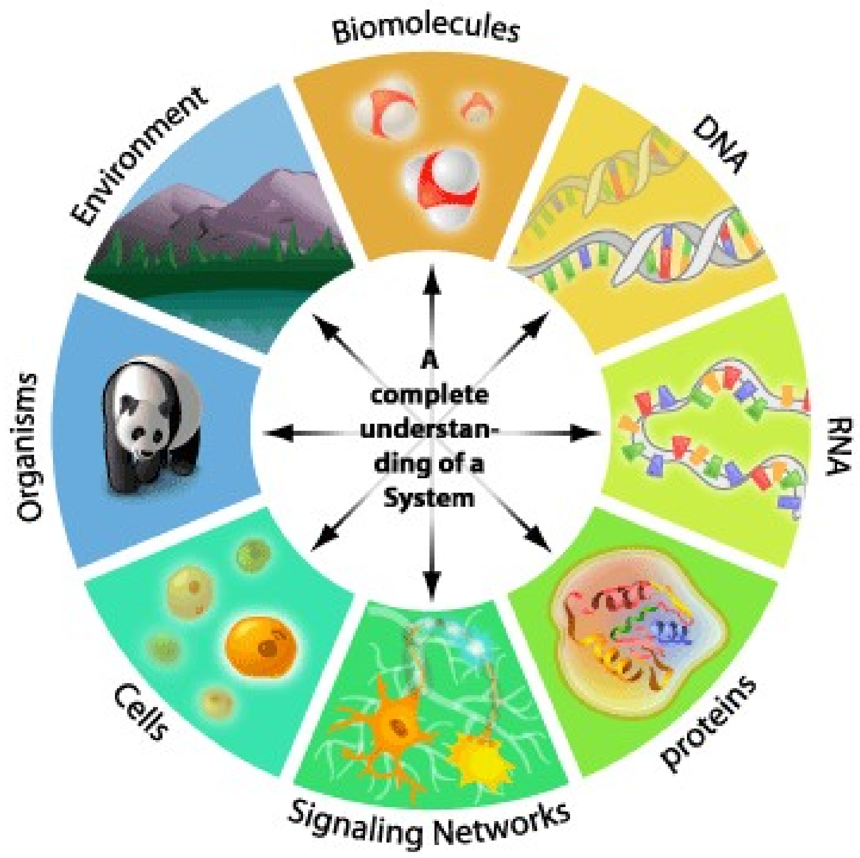
Introduction aux bases de données et aux ressources

Recherche de séquences

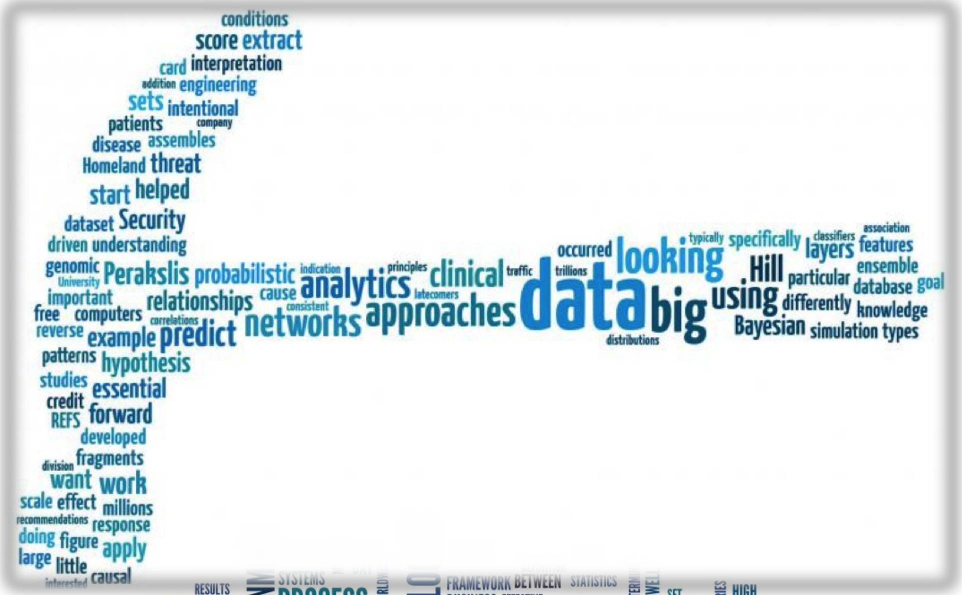
Objectifs du cours

- Comprendre la structure et la disposition des ressources de données du NCBI et de l’EBI
- Comprendre la différence entre les bases de données, les outils et les repositories.
- Recherche de données dans des bases de données spécifiques à l'aide de numéros d'accès, de noms de gènes, etc.
- Utiliser les ressources NCBI et EBI

Data



Biomolecules
DNA
RNA
proteins
Signaling Networks
Cells
Organisms
Environment
A complete understan-ding of a System



data
big
approaches
networks
analytics
clinical
looking
using
predict
probabilistic
Perakslis
relationships
hypothesis
essential
forward
Security
helped
threat
start
interpretation
score extract
engineering
intentional

BIG
DATA
INFORMATION
STORAGE
LARGE
SEARCH
PETABYTES
ANALYSIS
RESEARCH
MANAGEMENT
SOFTWARE
DEVELOPMENT
TECHNOLOGY
SETS
MILLION
SCIENCE
VOLUME
PROCESS
INTERNET
NEW
FUTURE

Introduction

- Plusieurs bases de données et ressources en ligne
- Besoin de savoir laquelle :
  - ✓ Quelles sont les bases de données et les ressources existantes
  - ✓ Quels sont les outils disponibles pour exploiter ces ressources ?
  - ✓ Quels sont les outils disponibles pour rechercher dans les ressources?

Bases de données biologiques

Bases de données biologiques

- Bases de données biologiques sont :

  - ✓ Publique ou privée
  - ✓ Protéine, nucléotide, structure, littérature, annotation…
  - ✓ Généralisée ou spécialisée
  - ✓ Centré sur la séquence (aa ou nt) ou le génome

Bases de données biologiques

Quelques noms de banques de données:

- **Séquences en acides nucléiques** (DNA et mRNA); EMBL, GenBank, DDBJ
- **Séquences en acides aminés (**protéines); Swiss-Prot, wwPDB

- **Références bibliographiques**; PubMed

- **Informations générales sur les gènes et/ou les maladies**; EntrezGene, OMIM, HMGD

- **informations sur la structure tridimensionnelle des protéines ou de l'ADN**; PDB

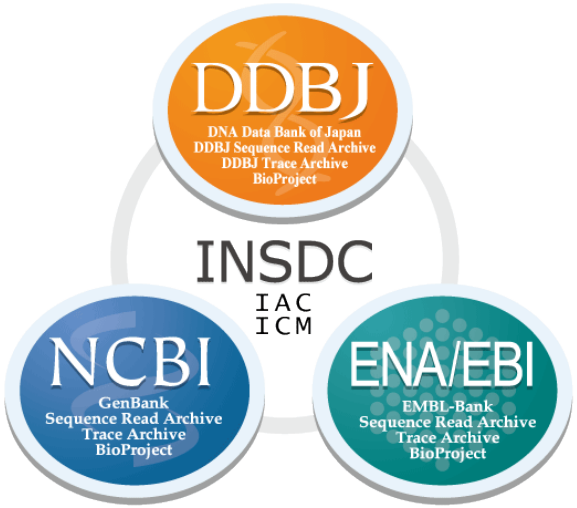
- Il existe aussi des banques spécialisées, comme Newt, qui

Bases de données primaires

- International Nucleotide Sequence Database Collaboration (INSDC)

- Données de séquences génomiques stockées dans 3 bases de données publiques

- Chacun a son propre numéro d'accès et ses propres outils



Bases de données secondaires

- Des bases de données spécialisées construites à partir de données de séquences primaires

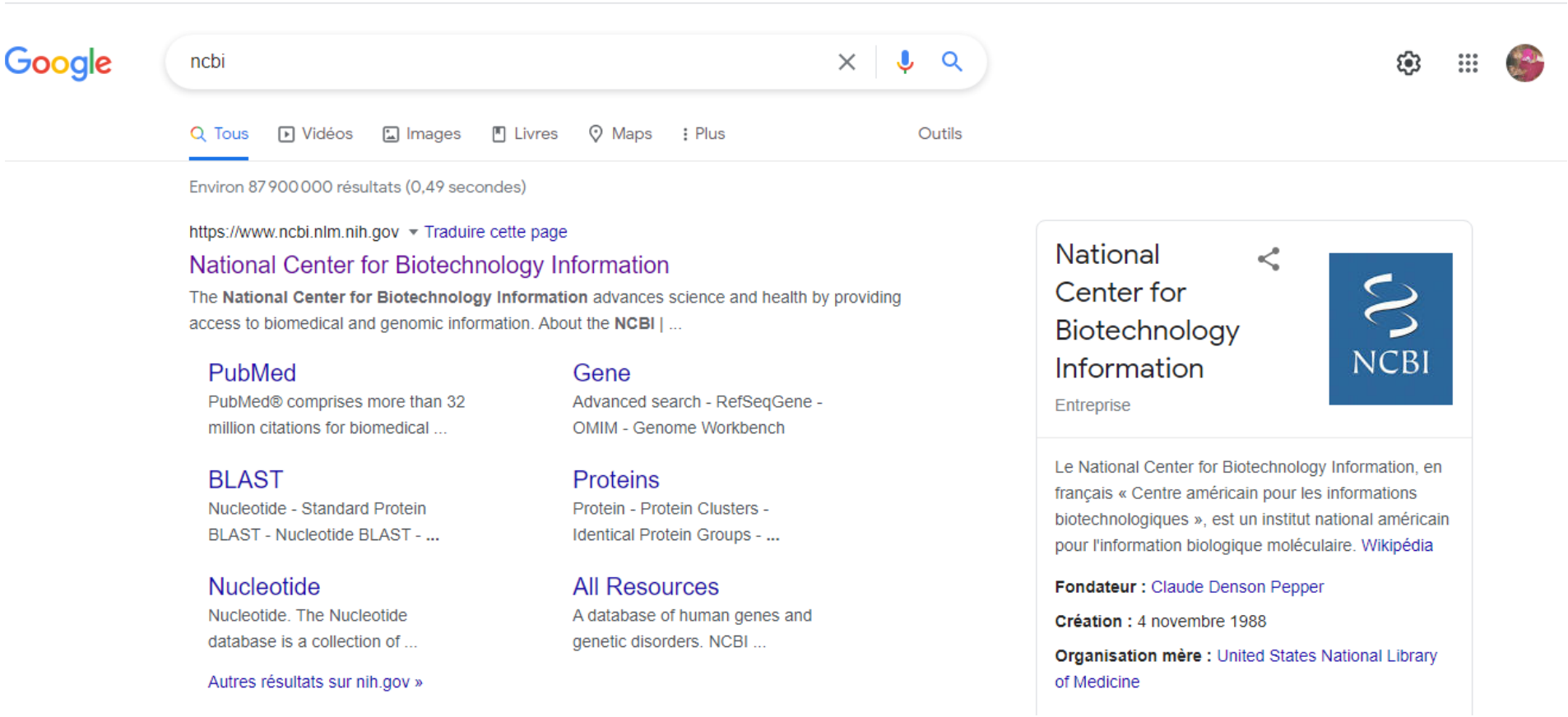
- Fournissent plusieurs ressources et annotations différentes

Ressources bioinformatiques les plus populaires

- National Centre for Biotechnology Information (NCBI)

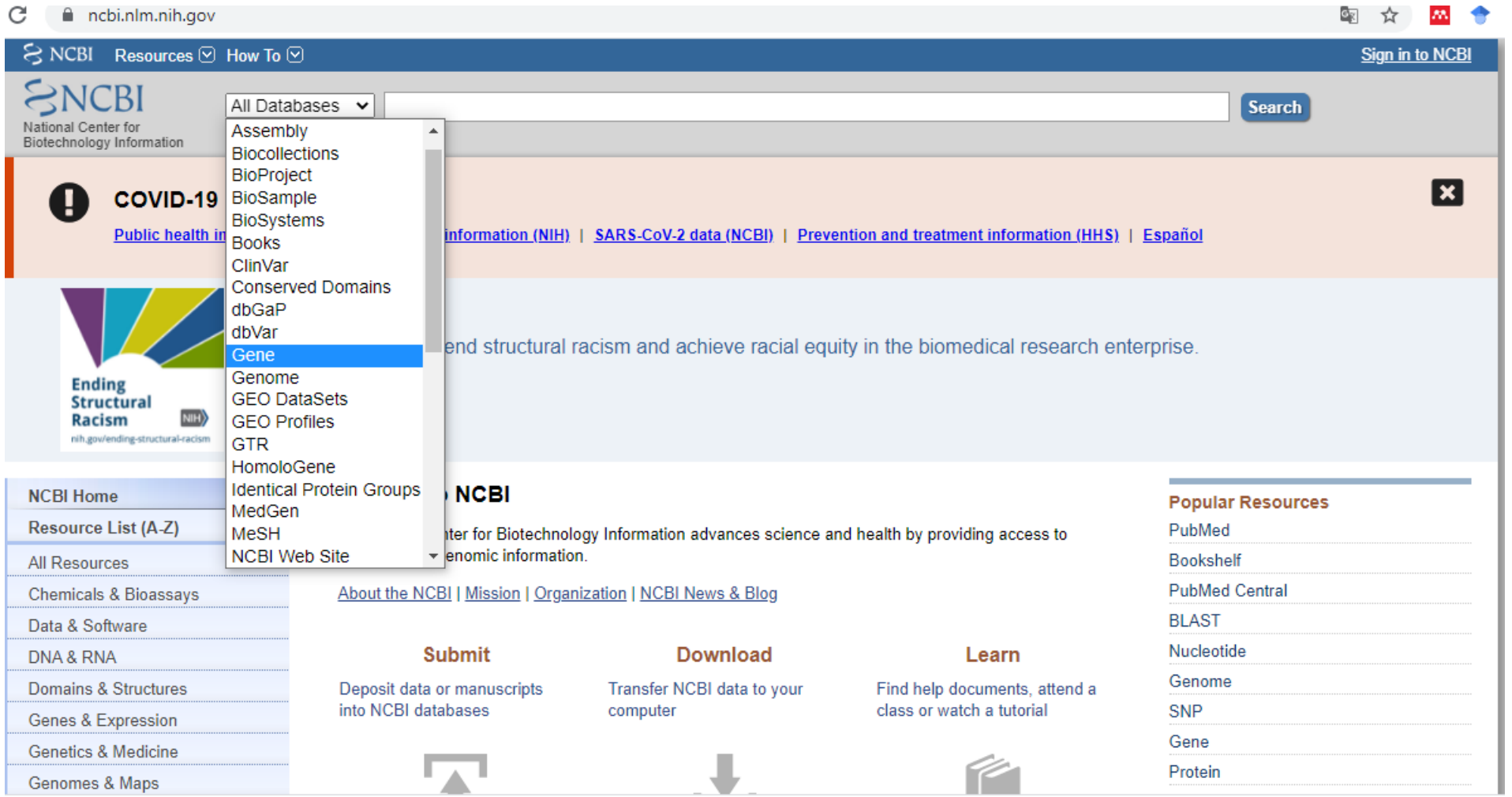
- European Bioinformatics Institute (EMBL-EBI)

Recherche dans les bases de données: NCBI



Recherche dans les bases de données: NCBI

Recherche dans les bases de données: NCBI



Bases de données de NCBI

- NCBI comprend plus de 30 bases de données
- la littérature : PubMed Central (PMC), Bookshelf et PubReader

- La santé: ClinVar, dbGaP, dbMHC, the Genetic Testing Registry, HIV-1/Human Protein Interaction Database et MedGen

- Les génomes: BioProject, Assembly, Genome, BioSample, dbSNP, dbVar, Nucleotide, Probe et RefSeq.

- Les génes: Gene, Gene Expression Omnibus (GEO), HomoloGene, PopSet, Refseq et UniGene.

- Les proteines: Protein, the Conserved Domain Database (CDD), COBALT, Conserved Domain Architecture Retrieval Tool (CDART), the Molecular Modeling Database (MMDB), Refseqp et Protein Clusters.

- Les produits chimiques: Biosystems et PubChem

EMBL - EBI

- Maintenir la gamme la plus complète au monde de bases de données moléculaires librement accessibles et actualisées

- Proposer des formations en ligne et en direct pour l'utilisation de leurs ressources.

- https://www.ebi.ac.uk/training

EMBL - EBI

The EMBL-EBI website has been redesigned. Please send us feedback about this page.

EMBL's European Bioinformatics Institute

EMBL-EBI

Unleashing the potential of big data in biology

Find a gene, protein or chemical | All | Search

Example searches: blast keratin bfl1 | About EBI Search

Find data resources | Submit data | Explore our research | Train with us

Latest news

Organisations should embrace open science faster – interview with Prof. Dame Janet Thornton

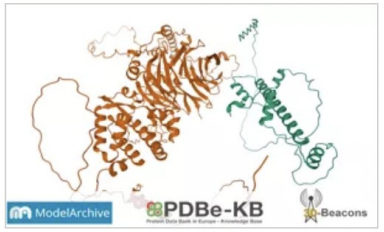
17 May 2022

Europe PMC: Harnessing the power of text mining to accelerate life sciences research

12 May 2022

2.4 billion sequences now available in the latest MGnify protein database release

11 May 2022



Predicted complexes from ModelArchive now on PDBe-KB pages

6 May 2022

EMBL - EBI

Services

Overview | A to Z | Data submission | Research infrastructure development programme | Support

The European Bioinformatics Institute (EMBL-EBI) maintains the world's most comprehensive range of freely available and up-to-date molecular data resources.

Developed in collaboration with our colleagues worldwide, our services let you share data, perform complex queries and analyse the results in different ways. You can work locally by downloading our data and software, or use our web services to access our resources programmatically.

— You can read more about our services in the journal Nucleic Acids Research

Tools & Data Resources

Search all tools & data resources

Tools

Clustal Omega

Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.

Web API | Multiple sequence alignment

InterProScan

InterProScan searches sequences against InterPro's predictive protein signatures.

Web API | Protein feature detection | Sequence motif recognition

BLAST [protein]

Fast local similarity search tool for protein sequence databases.

Web API | Sequence similarity search

BLAST [nucleotide]

Fast local similarity search tool for nucleotide sequence databases.

Data resources

Ensembl

Genome browser, API and database, providing access to reference genome annotation

Web API | EMBL-EBI Terms of use

UniProt

A comprehensive resource for protein sequence and functional annotation.

Web API | CC-BY

PDBe

The European resource for the collection, organisation and dissemination of 3D structural data (from PDB and EMDB) on biological macromolecules and their complexes.

Web API | CC0

Europe PMC

A database to search the worldwide life sciences literature

Web API | EMBL-EBI Terms of use

Browse by type

DNA & RNA | Gene Expression | Proteins
Structures | Systems | Chemical biology
Ontologies | Literature | Cross domain

Programmatic access

EMBL-EBI web services allow you to query our large biological data resources programmatically, so that you can develop data analysis pipelines or integrate public data with your own applications. The Web Services technology we use are built on open standards to ensure client and server software from various sources will work well together.

Browse EMBL-EBI web services

Principles of service provision

Open

Our data and tools are freely available, without restriction. The only exception is potentially identifiable human genetic information, for which access depends

Bases de données spécialisées

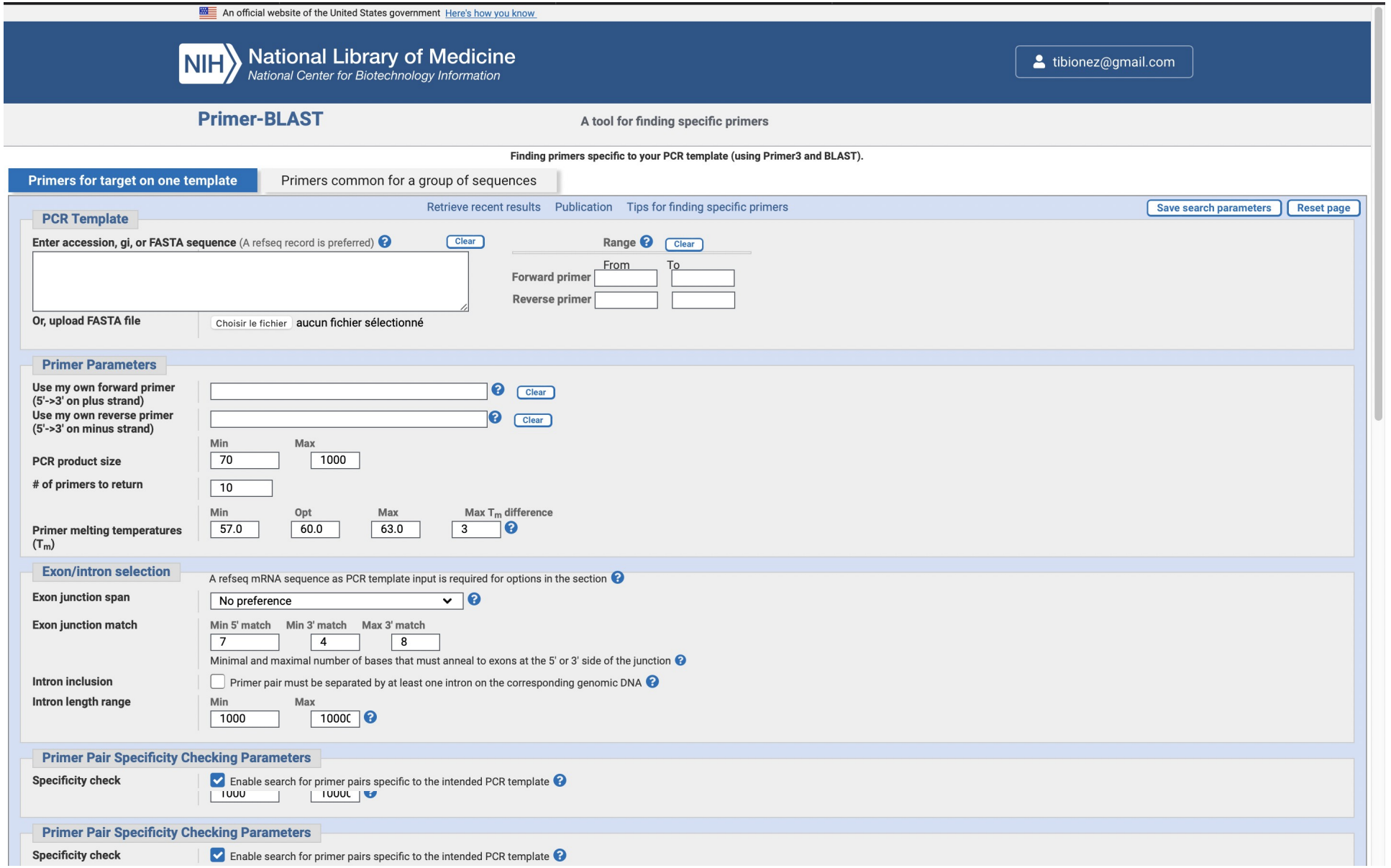
- Il existe un grand nombre de bases de données spécialisées

- La plupart des séquences sont également dans la banque GenBank/EMBL
- Peut contenir des génomes entiers
- Peut contenir des ressources spécialisées
- Contient des outils spécifiques pour l'exploitation des données

Bases de données spécialisées

- Plasmodium https://plasmodb.org/plasmo/app
- Les collections spécialisées de Sanger https://www.sanger.ac.uk
- Base de données sur les hépatites https://hcv.lanl.gov/content/sequence/HCV/news/old_news.html
- Base de données de recherche sur la grippe influenza https://www.fludb.org/brc/home.spg?decorator=influenza

Design d’amorce utilisant Primer Blast



Take home

- Une grande quantité de données existent
- Les bases de données primaires stockent les données brutes des séquences
- Les bases de données secondaires fournissent des informations sur l'annotation des données de séquence.
- Il est important de savoir comment et où les données sont stockées
- NCBI et EBI sont les deux ressources les plus populaires pour obtenir des données biologiques.