Clouds in biosciences

A journey to High Throughput Computing in life sciences

Vincent Breton

July 28th 2014

Enrico Fermi school of physics

A journey to High Throughput Computing in life sciences…

- Part I
  - Who am I?
  - Introduction to the countries we will explore
- Part II: Grid usage in life sciences
- Part III: Clouds in life sciences
- Part IV: Entering a new world

Concepts – acronyms used

- Grid computing is Cloud computing
  - **Platform as a service** (**PaaS**) is a category of cloud computing services that provides a computing platform and a solution stack as a service
- High Throughput Computing
  - Analyzing large volumes of data
  - Cluster, Grid and Cloud computing best fitted for embarrassingly parallel calculations
- High Performance Computing
  - Supercomputers best fitted to run complex models
  - Out of the scope of this talk

More than 60 life sciences !

- 1.1 Affective neuroscience 1.2 Anatomy 1.3 Astrobiology 1.4 Biochemistry 1.5 Biocomputers 1.6 Biocontrol 1.7 Biodynamics 1.8 Bioinformatics 1.9 Biology 1.10 Biomaterials 1.11 Biomechanics 1.12 Biomedical science 1.13 Biomedicine 1.14 Biomonitoring 1.15 Biophysics 1.16 Biopolymers 1.17 Biotechnology 1.18 Botany 1.19 Cell biology 1.20 Cognitive neuroscience 1.21 Computational neuroscience 1.22 Conservation biology 1.23 Developmental biology 1.24 *Ecology* 1.25 *Environmental science* 1.26 Ethology 1.27 Evolutionary biology 1.28 Evolutionary genetics 1.29 Food science 1.30 Genetics 1.31 Genomics 1.32 Health Sciences 1.33 Immunogenetics 1.34 Immunology 1.35 Immunotherapy 1.36 Kinesiology 1.37 Marine biology 1.38 Medical devices 1.39 Medical imaging 1.40 Medical Social Work 1.41 Microbiology 1.42 Molecular biology 1.43 Neuroethology 1.44 Neuroscience 1.45 Oncology 1.46 Optogenetics 1.47 Optometry 1.48 Parasitology 1.49 Pathology 1.50 Pharmacogenomics 1.51 Pharmaceutical sciences 1.52 Pharmacology 1.53 Physiology 1.54 Population dynamics 1.55 Proteomics 1.56 Psychiatric social work 1.57 Psychology 1.58 Sports science 1.59 Structural biology 1.60 Systems biology 1.61 Zoology

Life Sciences communities requiring High Throughput data analysis are currently using grid and/or cloud infrastructures

Table of contents – part I

- Who am I?
- A journey to High Throughput Computing in life sciences

A short biography (I/II)

- Background
  - Physicist by training
  - Interest for life sciences by education
- CV
  - 1990: PhD in Nuclear Physics at CEA Saclay
  - 1990-1998: hadronic physics (SLAC – TJNAF)
  - 1998-2002: LHCb@CERN
  - 2000-2014: interface between physics and life sciences

A short biography (II/II)

- The Grid and I…
  - 2000-2010: deployment of biomedical applications on grid infrastructures (DataGrid, EGEE)
  - 2010-2014: France-Grilles
- Today, my professional life is shared between:
  - Leading the France National Grid Initiative
  - Exploring the impact of radiation on evolution
- Mediator between grid technologists and researchers in life sciences and healthcare

A journey to High Throughput Computing in life sciences

- Lands visited
  - Molecular biology
  - Structural biology
  - Drug discovery
  - Medical imaging

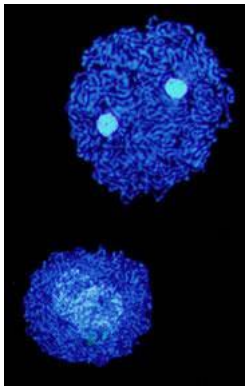
Welcome to the land of molecular biology

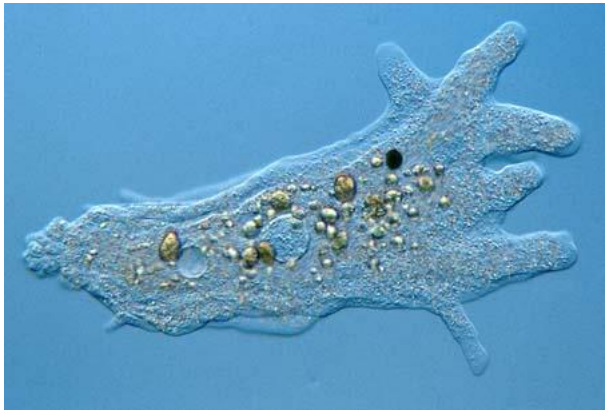
- Change in scale in the last 10 years
- Technological revolution: high throughput sequencing
- Encyclopedic approach: all genes, all proteins, all interactions, …
- New perspective: from the genome to the organism biological properties
- Biologists are flooded by an avalanche of heterogeneous data
- 25% of the time to collect data, 75% to analyze the data

Credit: J-F Gibrat

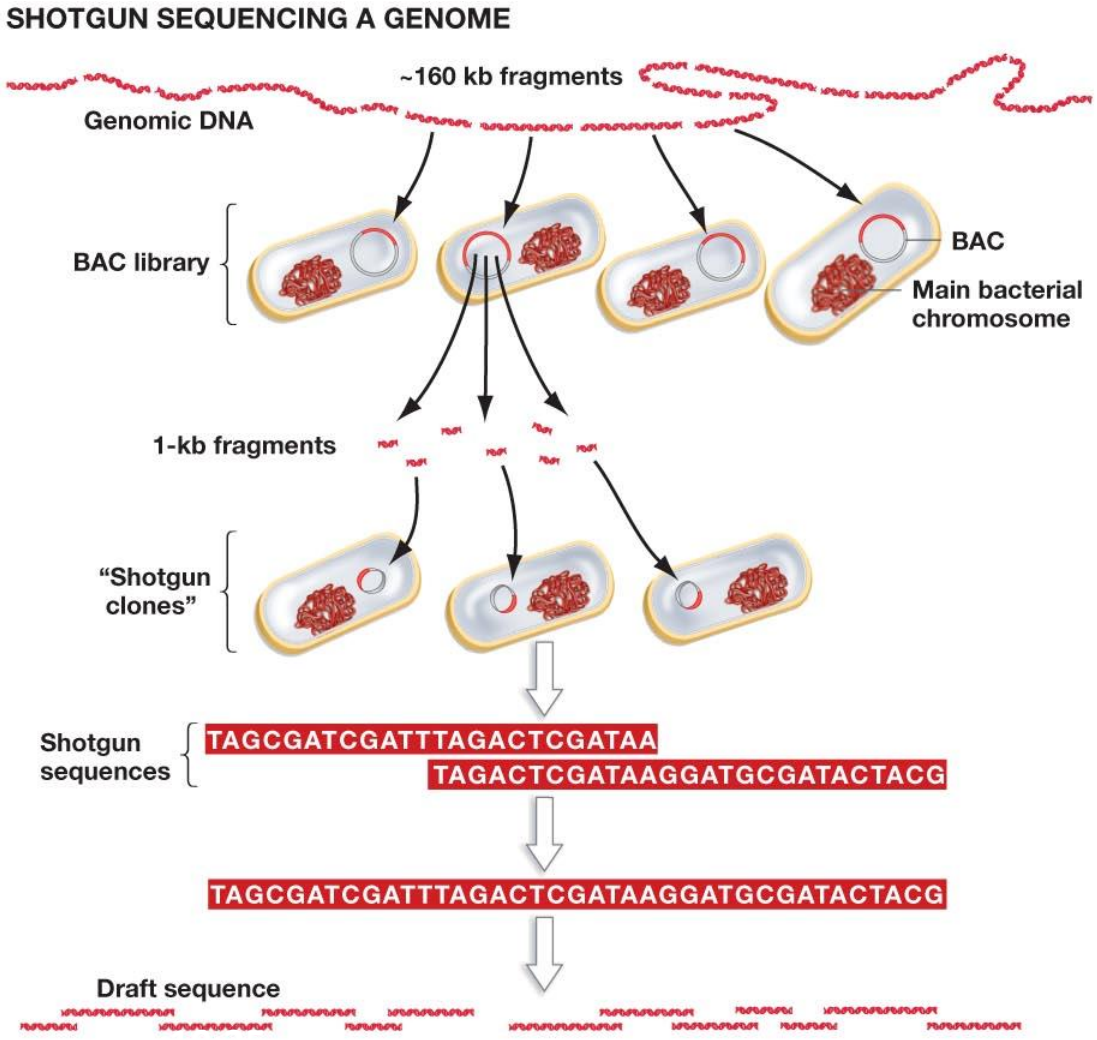
Sequencing genomes

- Genome = DNA sequence (4 nucleotids: A, C, G, T)
  - Smallest non viral genome: *Carsonella ruddii* (0,16M base pairs)
  - Largest genome: *Polychaos dubium* (670G base pairs)
- Human genome sequencing (3G base pairs)
  - 10 year effort
  - 3 billion USD
- Time has changed…





Shotgun sequencing



1. Cut DNA into fragments of ~160 kb, using sonication.

2. Insert fragments into bacterial artificial chromosomes; grow in *E. coli* cells to obtain large numbers of each fragment.

3. Purify each 160-kb fragment, then cut each into a set of 1-kb fragments, using sonication, so that 1-kb fragments overlap.

4. Insert 1-kb fragments into plasmids; grow in *E. coli* cells. Obtain many copies of each fragment.

5. Sequence each fragment. Find regions where different fragments overlap.

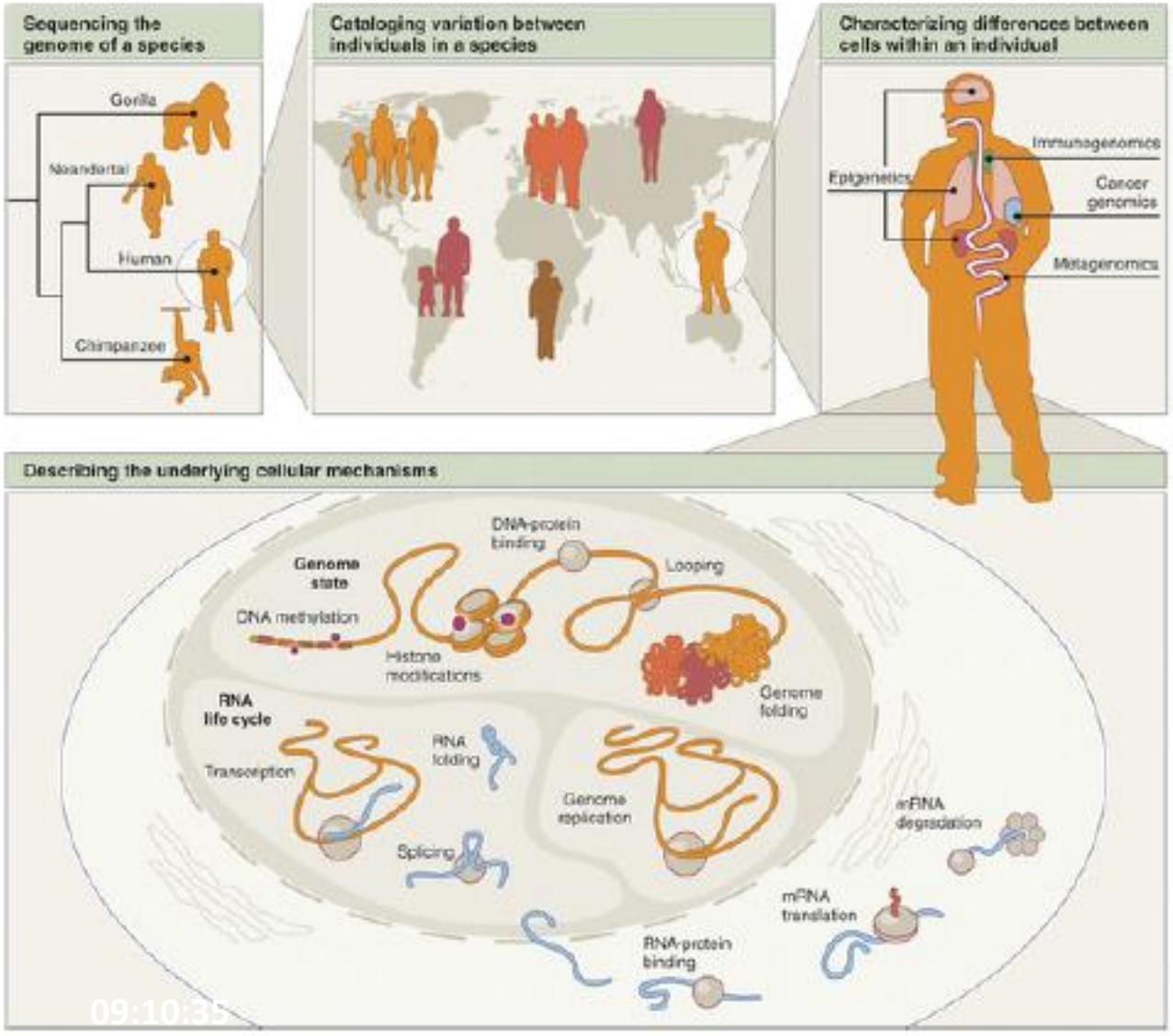
6. Assemble all the 1-kb fragments from each original 160-kb fragment by matching overlapping ends.

7. Assemble sequences from different BACs (160-kb fragments) by matching overlapping ends.

Next generation sequencing

- Since 2007, new sequencing technologies
- One “run” (a few days) produces up to 3 billion “reads” = fragments of 2x100 base pairs
  - A few TOctets of raw data
  - individual sequence read has about 0.5% error rate
- Sequencing cost dropped from 10.000 $ to 0.03 $ per million of sequenced nucleotids

What is it interesting for?



- Whole genome re-sequencing
- Ancient genomes
- Metagenomics
- Cancer genomics
- Genomic epidemiology

Credit: Shendure & Aiden (2012) Nat. Biotech.

Sequencing scenarii

- Interest for a new genome requires assembly
  - process of taking a large number of short DNA sequences and putting them back together to create a representation of the original
  - Algorithms based on read overlapping benefit from large RAM (1 TO) -> HPC
- Working with a reference genome requires comparative analysis
  - Alignment algorithms (BLAST) find regions of local similarity between sequences
  - Phylogeny algorithms (PhyML) build evolutionary relationships between genomes
  - Comparative analyses are easily parallelized at data level -> HTC

Bioinformatics

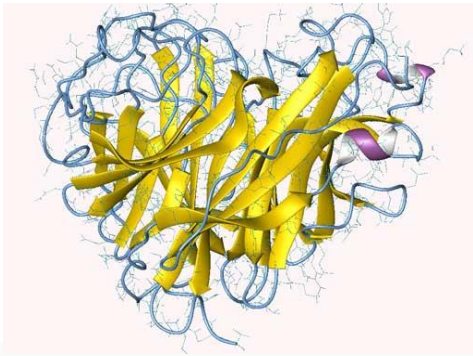
- Bioinformatics = computing methods to handle, organize and analyze biological data
  - Focused on the analysis of the sequences (DNA, RNA, proteins), their structure and interactions
  - No interest for image analysis
- The role of bioinformatics
  - Handle high throughput biological data
  - Organize the data
  - Extract biological information from raw data

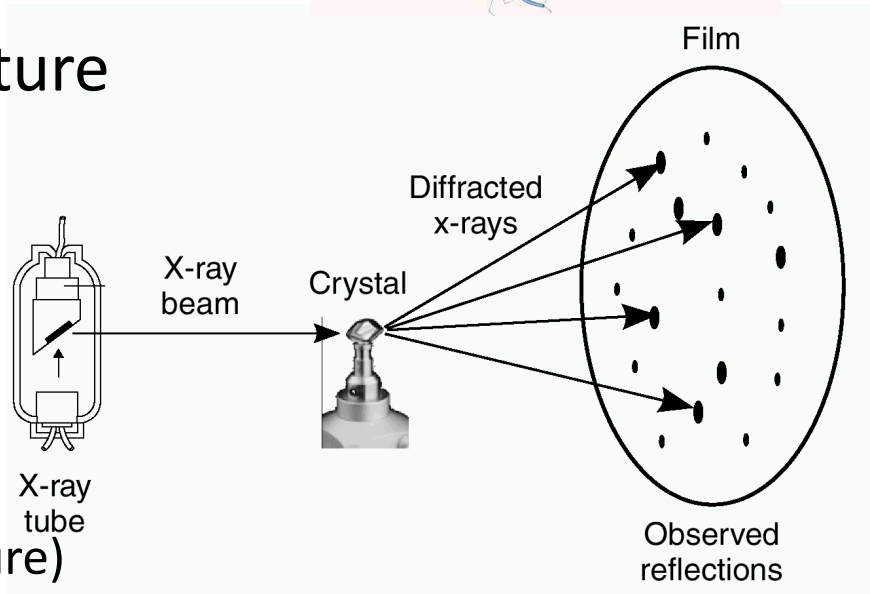
What characterize bioinformatics analysis?

- Many analyses can be parallelized at data level
  - Comparative analysis
- Analyses require treatment chains (pipelines, workflows) and integration of heterogeneous data
- Different programming languages (Perl, Python, Java, etc)
- Multiplication of programs and algorithms
  - 98 sequence alignment software tools
- A typical bioinformatics platform proposes hundreds of software tools

Welcome to the land of structural biology

- Structural biology studies the molecular structure of biological macromolecules
  - macromolecules carry out most of the functions of cells
- Techniques to measure the structure of macromolecules
  - Physical techniques
    - Mass spectrometry
    - Nuclear Magnetic Resonance
    - X-ray cristallography
  - Biological techniques
    - Bioinformatics (sequence ⇔ structure)





Grid added value for structural biology

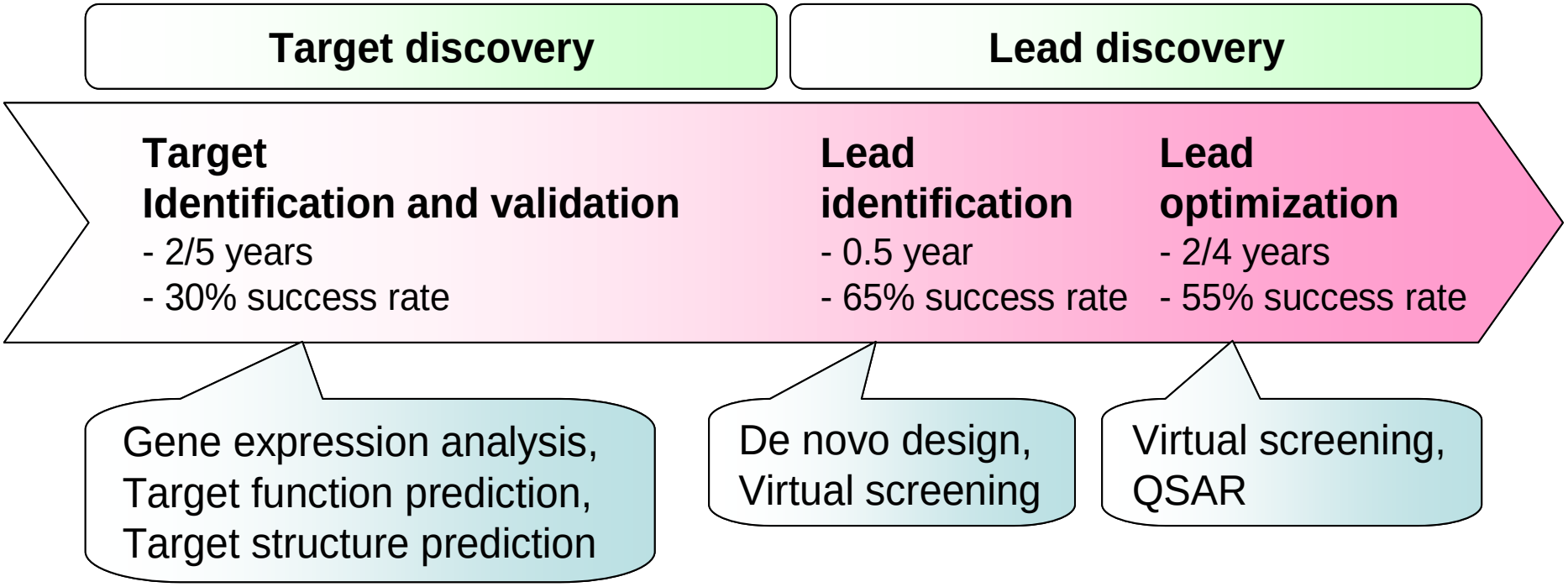
- Structural calculations from raw data are CPU demanding and easily parallelized by the data
  - Towards standardized pipeline analysis using reference software tools
- Example from mass spectrometry
  - Human cell contains 5 to 6000 different proteins
  - Goal: compare proteins expressed by healthy and cancerous cells
  - One mass spectrometer generates ≈ 50.000 fragmentation spectra in 5 hours ⇔ 15 GB of raw data

From structural biology to *in silico* drug discovery

- The *Protein Data Bank* (*PDB*) is a repository for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids
  - data typically obtained by X-ray crystallography or NMR spectroscopy
  - More than 100.000 structures in 2014
- Among them are biological targets for drugs
  - Biological target = biomolecule that changes its behaviour or function when a chemical compound binds to it

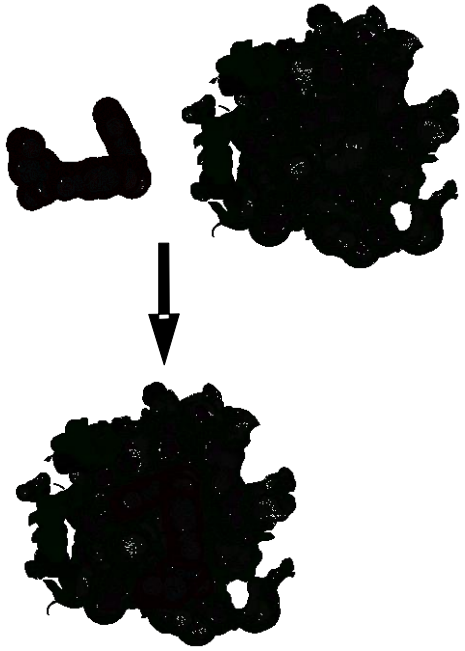
Searching for new drugs

- Drug development is a long (10-12 years) and expensive (~800 MDollars) process
- *In silico* drug discovery opens new perspectives to speed it up and reduce its cost

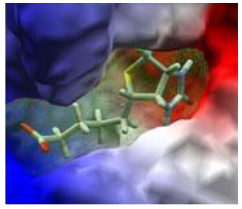


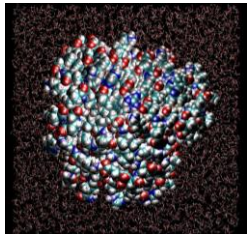
Screening

- Biologists identify a protein involved in the metabolism of the virus: the target
- The goal is to find molecules to prevent the protein from playing its role in the virus life cycle: the hits
  - Hits dock in the active site of the protein
- *in silico* vs *in vitro* screening
  - *In silico:* computational evaluation of binding energy
  - *In vitro:* optical measurement of chemical reaction constant

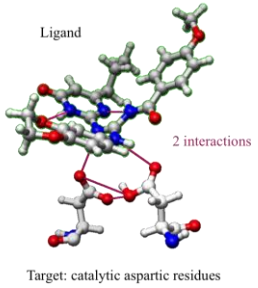


Virtual screening pipeline





AMBER



CHIMERA

WET LABORATORY

Millions of chemical compounds available in open source databases

In silico drug discovery very early
Identified as a potential “killer application” for the grid

Welcome to the land of medical imaging

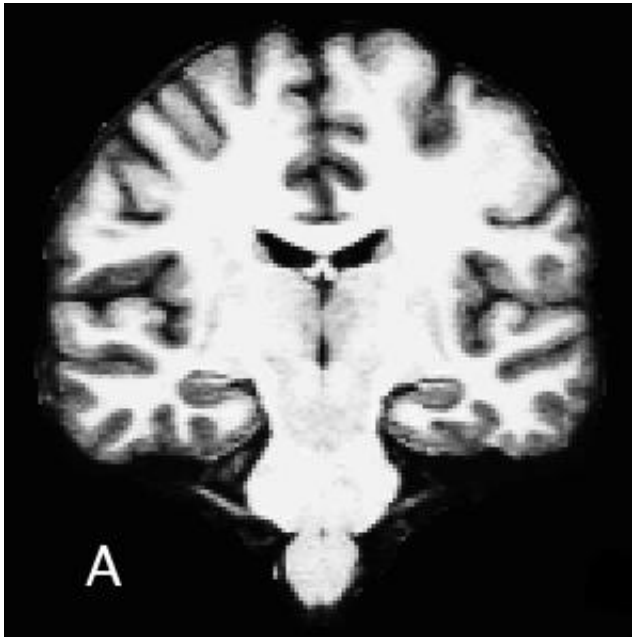
- *Medical imaging* is the technique, process and art of creating visual representations of the interior of a body for clinical analysis and medical intervention
- Medical imaging techniques are multiple
  - X-ray radiography, magnetic resonance imaging, medical ultrasonography or ultrasound, endoscopy, elastography, tactile imaging, thermography, medical photography and nuclear medicine functional imaging

Medical image simulation

- Variety of applications in research and industry
  - prototyping of new devices
  - evaluation of image analysis algorithms
- Commonly simulated image modalities
  - Magnetic Resonance Imaging
  - Ultrasound imaging
  - Positron Emission Tomography
  - Computed Tomography

Neurosciences, the need for high-throughput imaging research

- New imaging technologies significantly improve diagnostic and prognostic accuracy of neurodegenerative diseases
  - Especially true for Alzheimer's disease
- CPU-greedy tools for analysis and visualization of structural and functional brain imaging data
- Example : segmentation of cortical and subcortical anatomy and calculation of areas and thickness
  - About 24 hours to run for each scan



Life sciences need High Throughput computing

Scientific discipline	Data to be processed
Molecular Biology	High Throughput Computing of NGS data
Structural biology	High Throughtput analysis of Nuclear Magnetic Resonance and Mass Spectrometry data
Neurosciences	High Throughput analysis of brain images
Drug discovery	High Throughput computing of molecular structures

Additional features

- Need for comparative analysis in biology and medicine -> extensive use of databases
- Security is
  - Critical for medical data (privacy issues) and pharmaceutical data (intellectual property issues)
  - Much less for biological data, except for personalized medicine
- HPC is needed mostly at the interface with computational chemistry and for genome assembly
- Hundreds of bioinformatics algorithms and databases but a handful of structural biology software

Grid computing is part of the answer (security issues, flexibility)