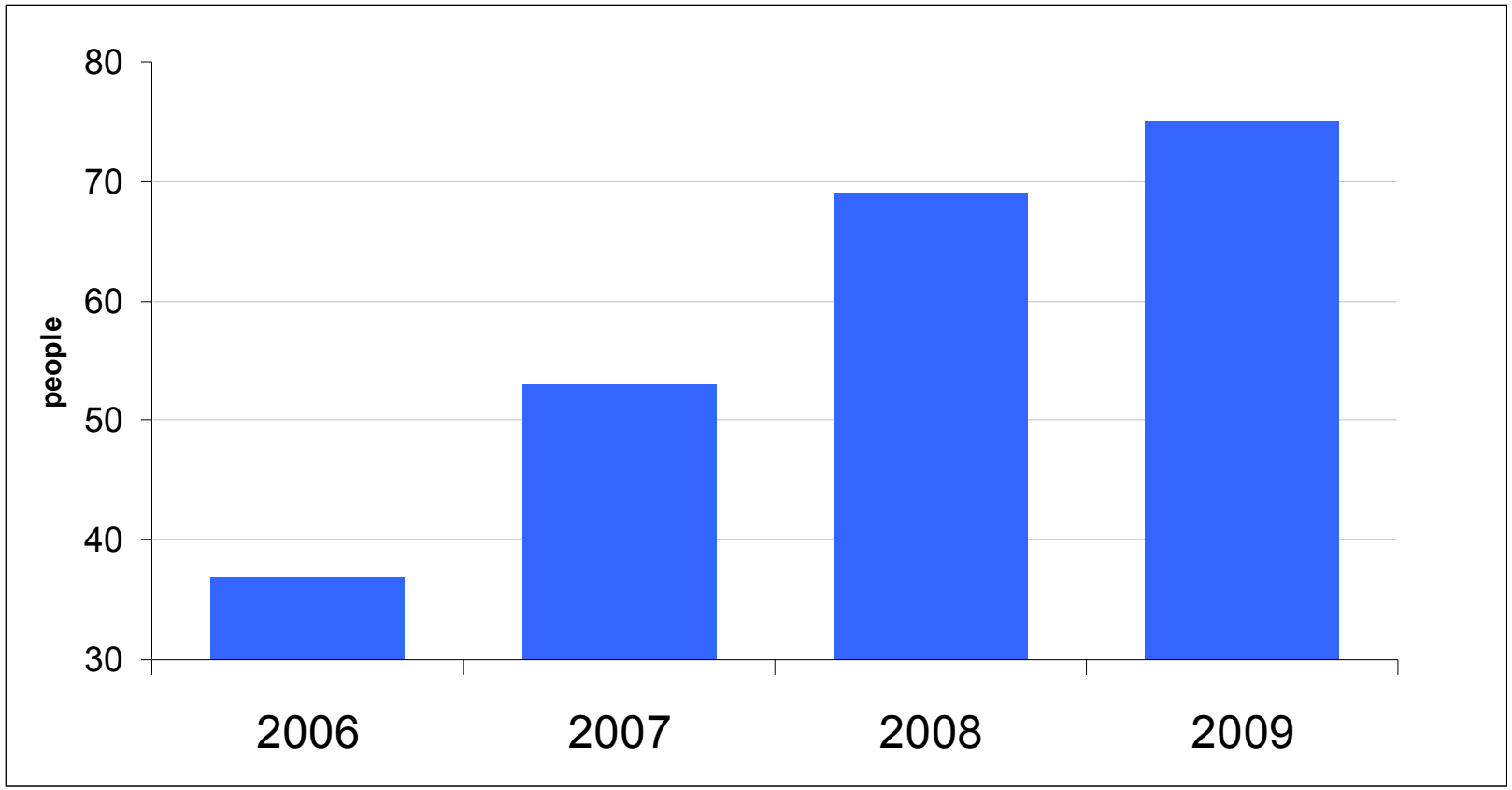
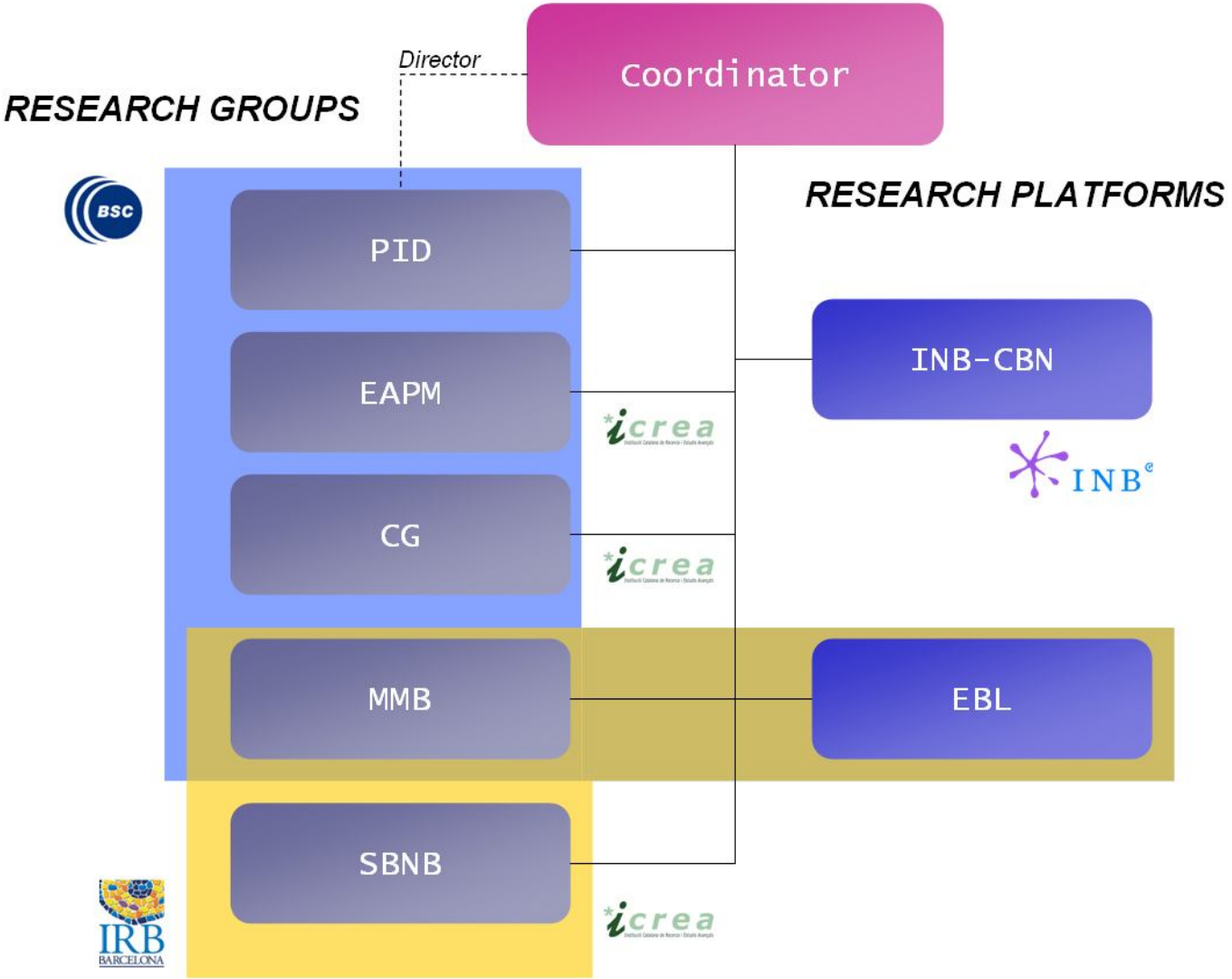
The Power of Computation in Life Sciences

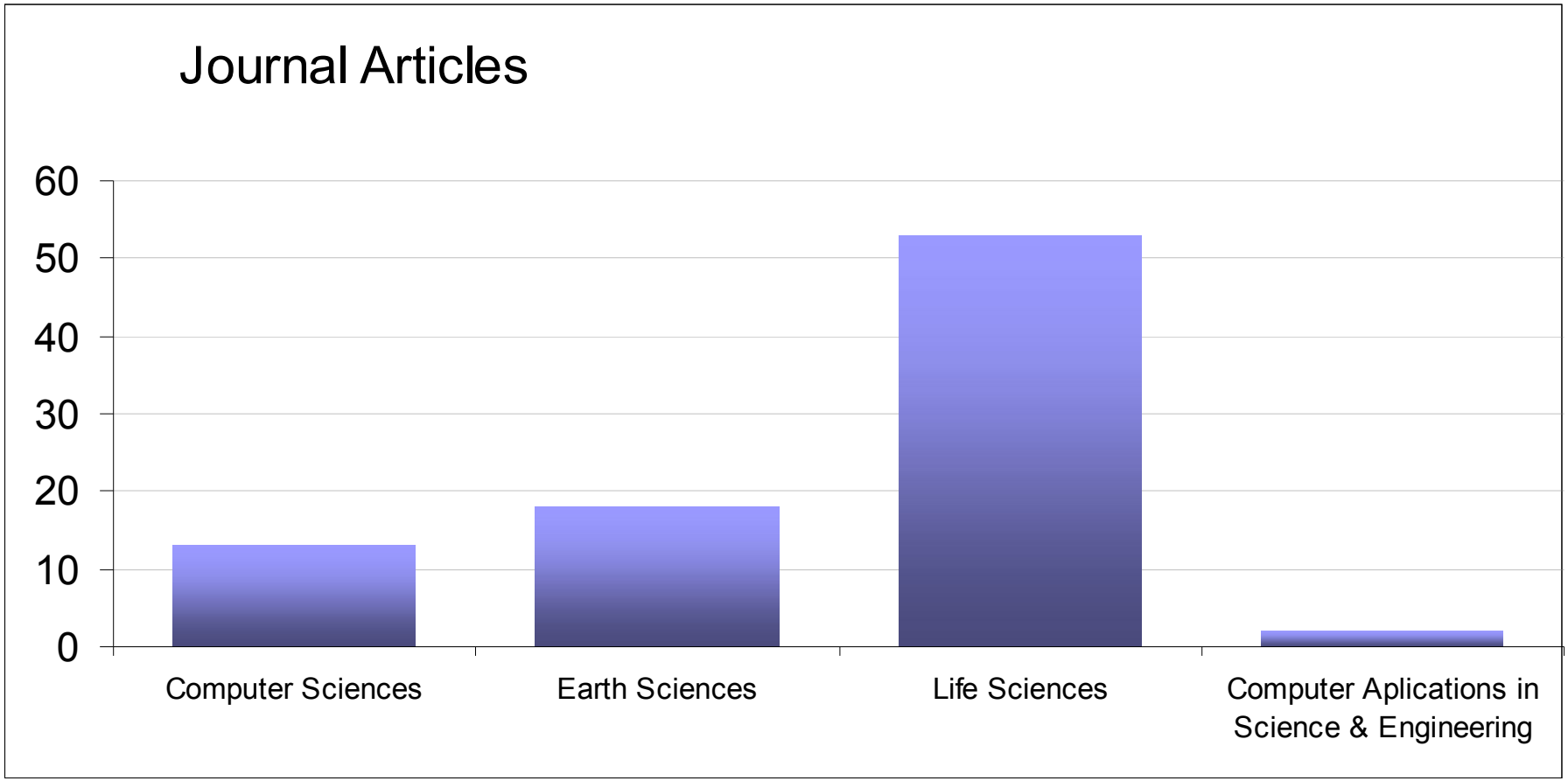
Ramon Goñi

Life Sciences Department





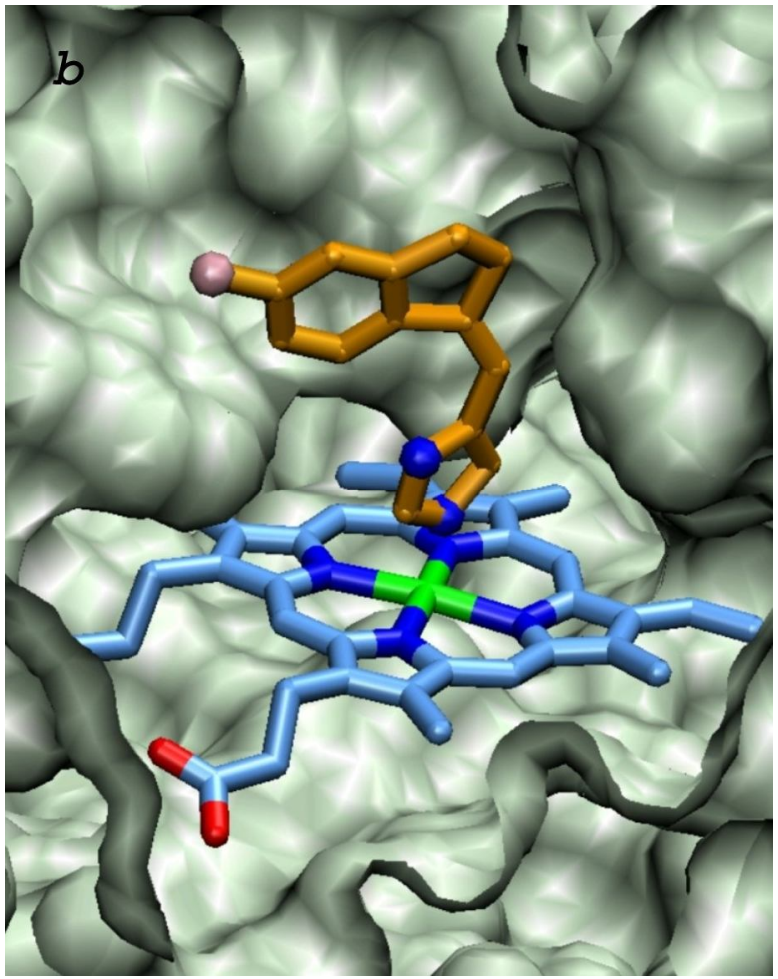
Life Science Research



Research Lines

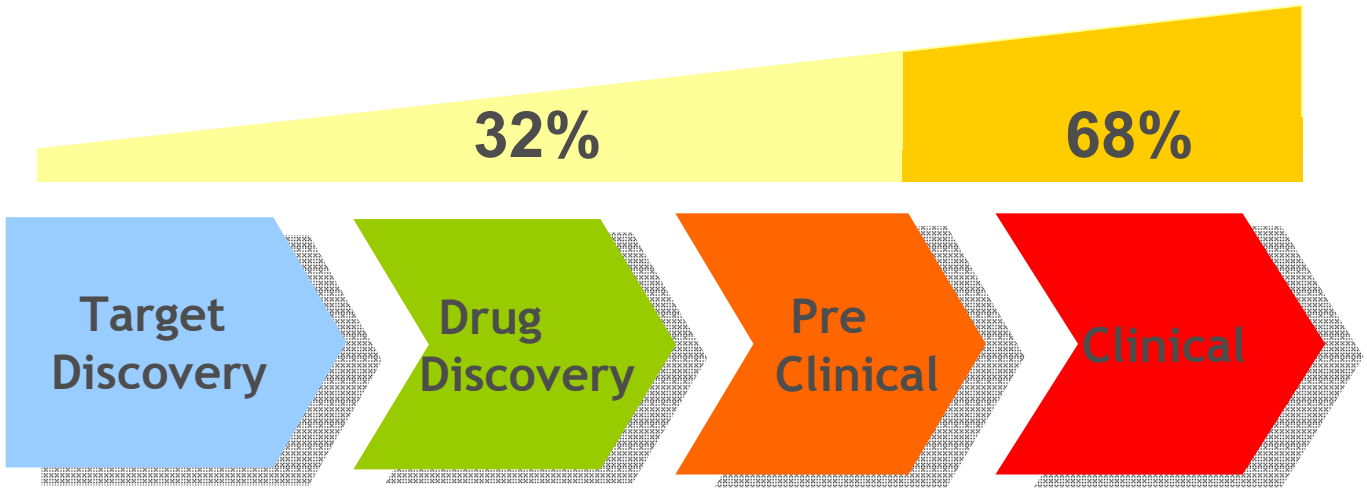
- Computer simulation (Drug Discovery)
  - **Molecular Dynamics**
  - **Protein-Protein Interaction**
  - **Protein-Ligand Docking**
- BioSupercomputing
  - **Computational Biology under GPU & CELL**
  - **User-friendly computing access Web-Services**
- Data analysis and Data Management (Target Discovery)
  - **Next Generation Sequencing**
  - **Genomics & DNA structure**

Computer Simulation & Drug Discovery

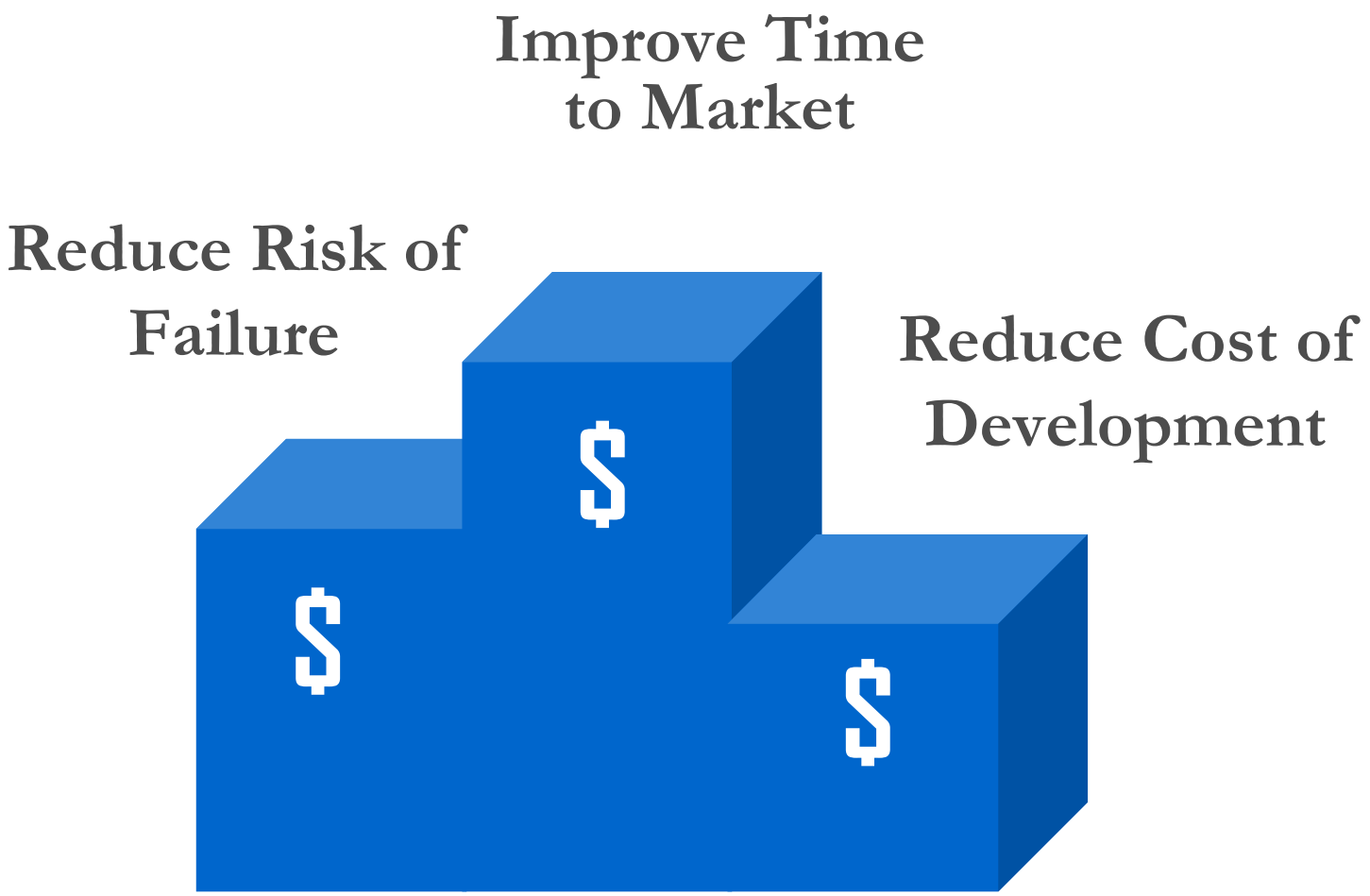


Drug Development

Cost: $1.2B / drug



Computer-Based Drug Design



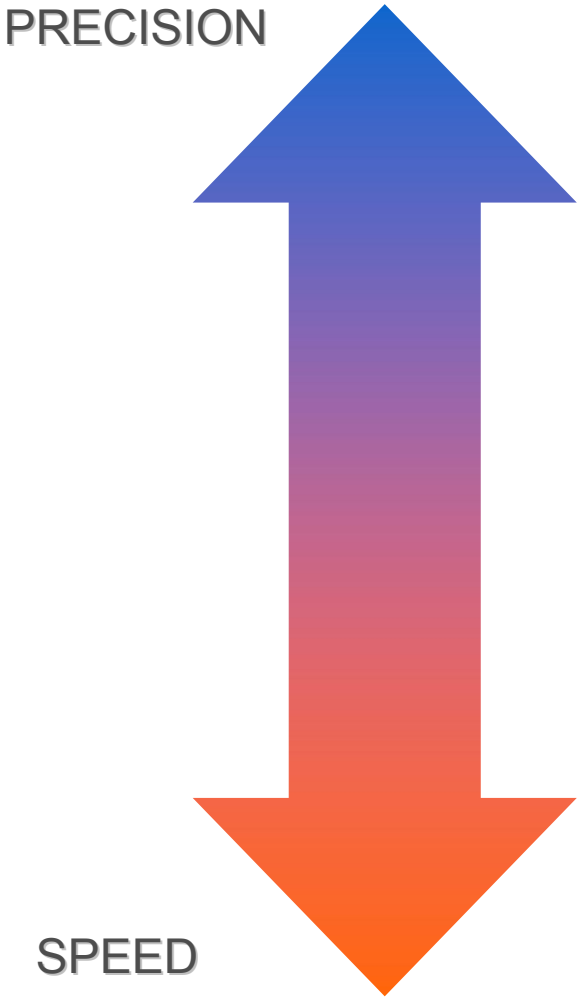
Computer Simulation

- Why and when we use it
  - **To validate a known model**
  - **As a cost-effective alternative**
  - **As the only realistic approach to solve a problem**

- The structure of bio-molecules are hardly modeled. The dynamics through experiments are only available for small molecules.

- There are different methods with different levels of complexity and realism

Molecular Simulation



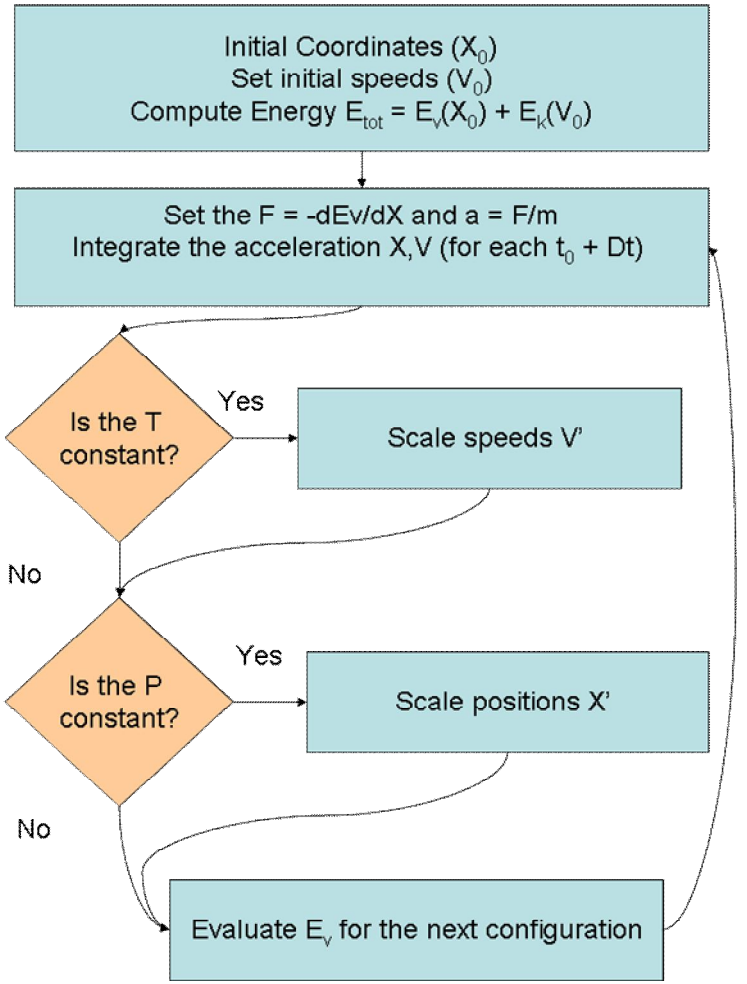
Quantum Mechanics

Molecular Dynamics (Newton motions)

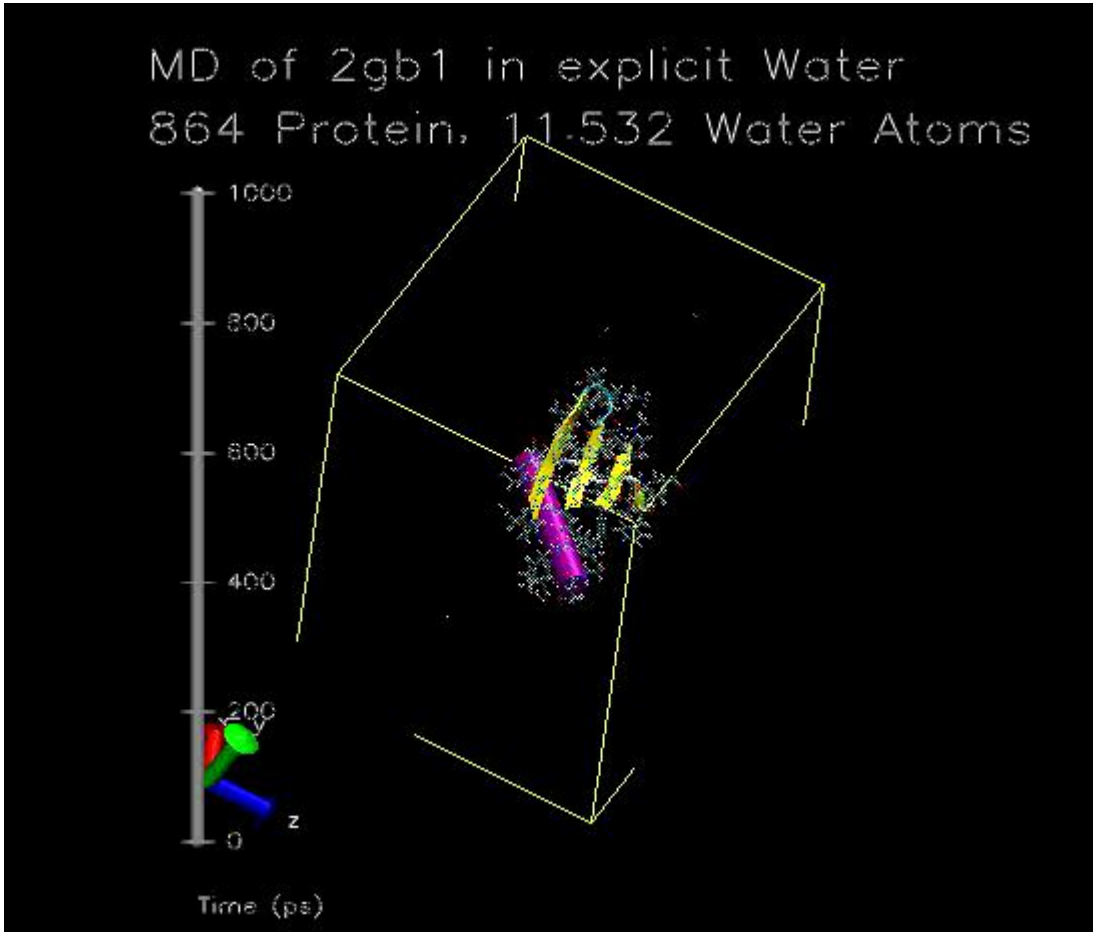
Corse Grained – MD (Pseudo-atoms)

Molecular Dynamics

- Atoms and molecules are allowed to interact for a period of time by approximations of known physics.

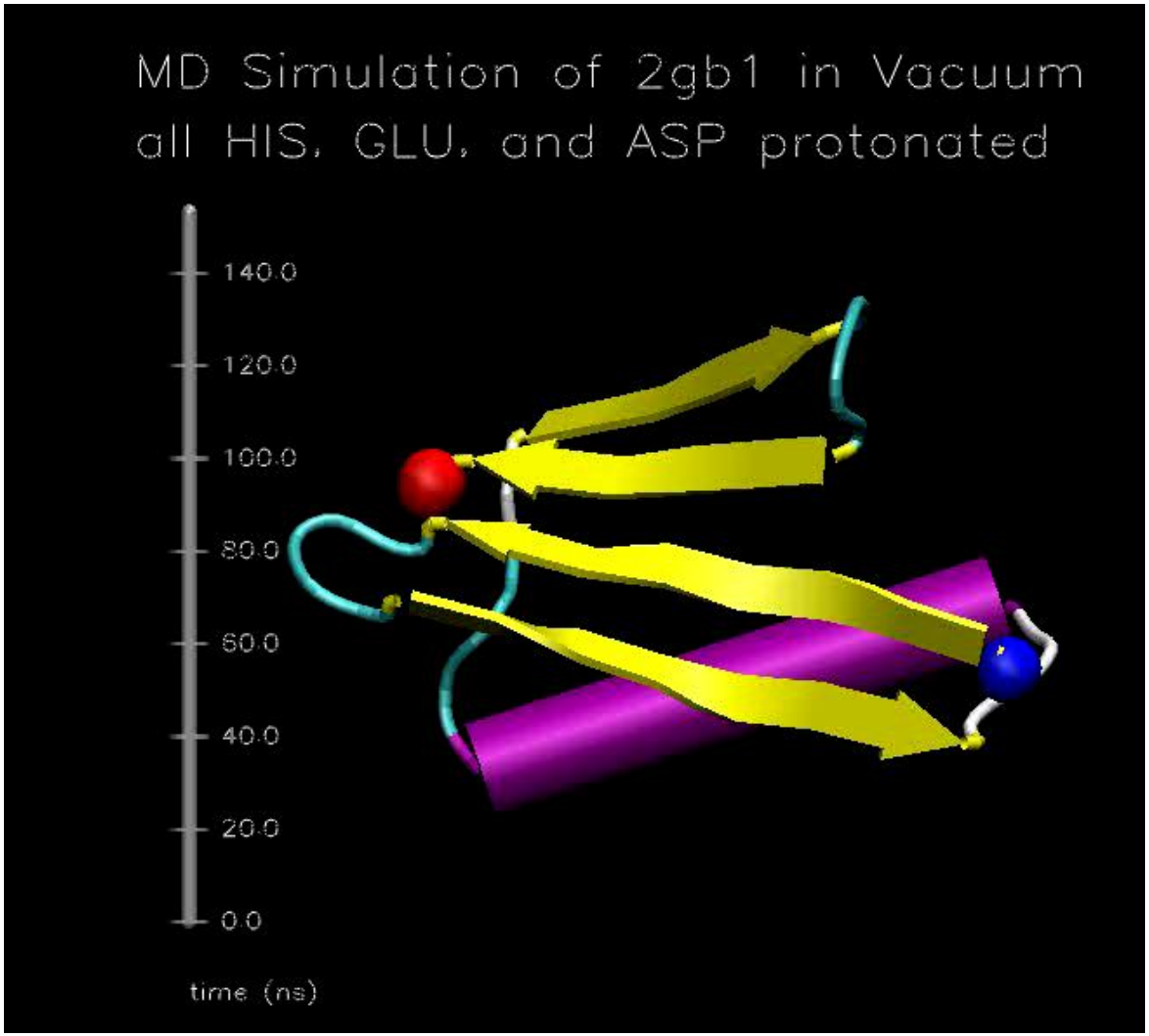


Molecular Dynamics of Solvated Protein



- Snapshot every femtosecond (10^{-15}s)
- System of 10^4 atoms
- 10 operations per atom pair-mate
- Using 16 processors we are able to simulate more than 10 nanoseconds (10^{-9}s) per day

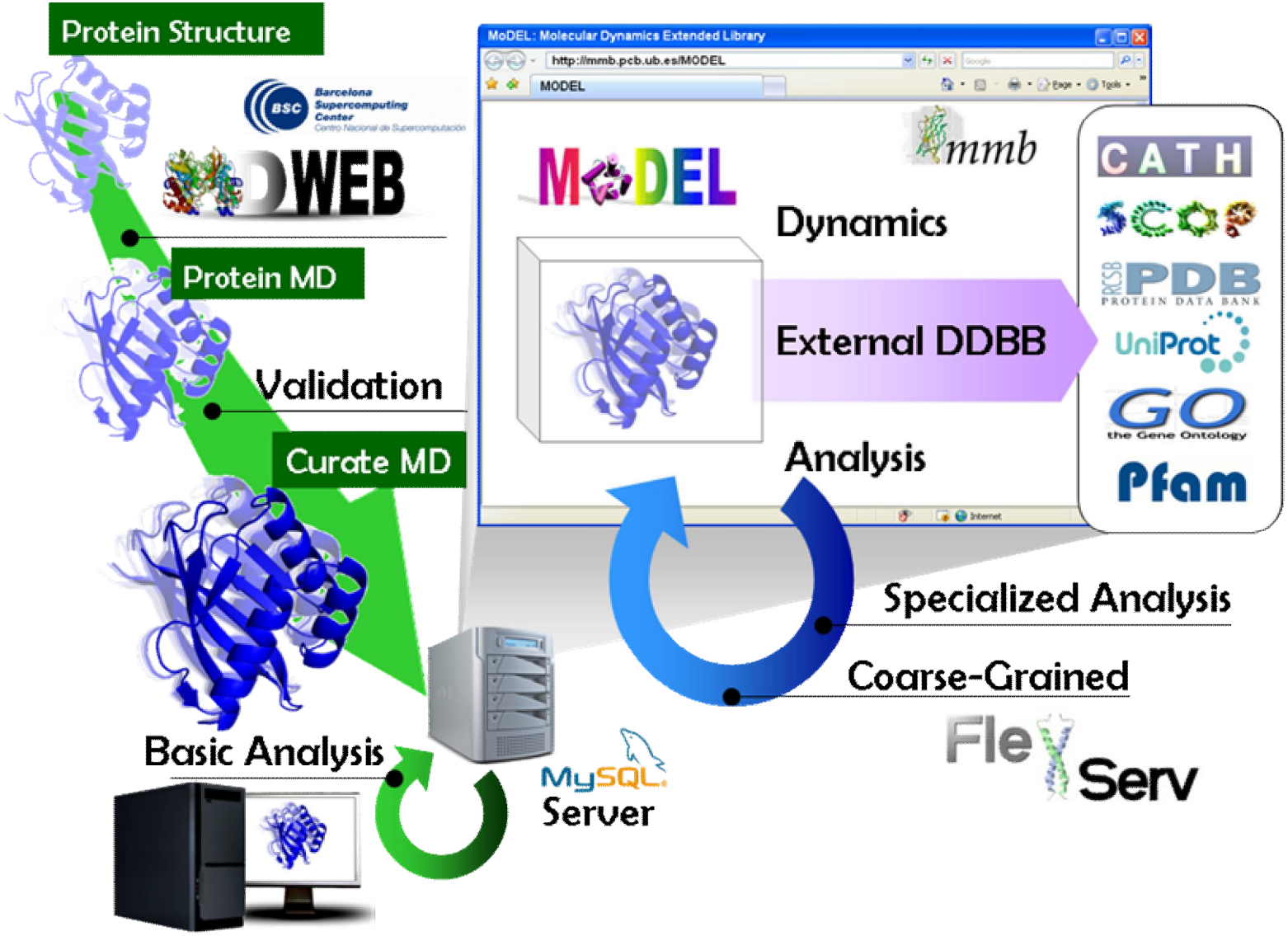
MD Scale

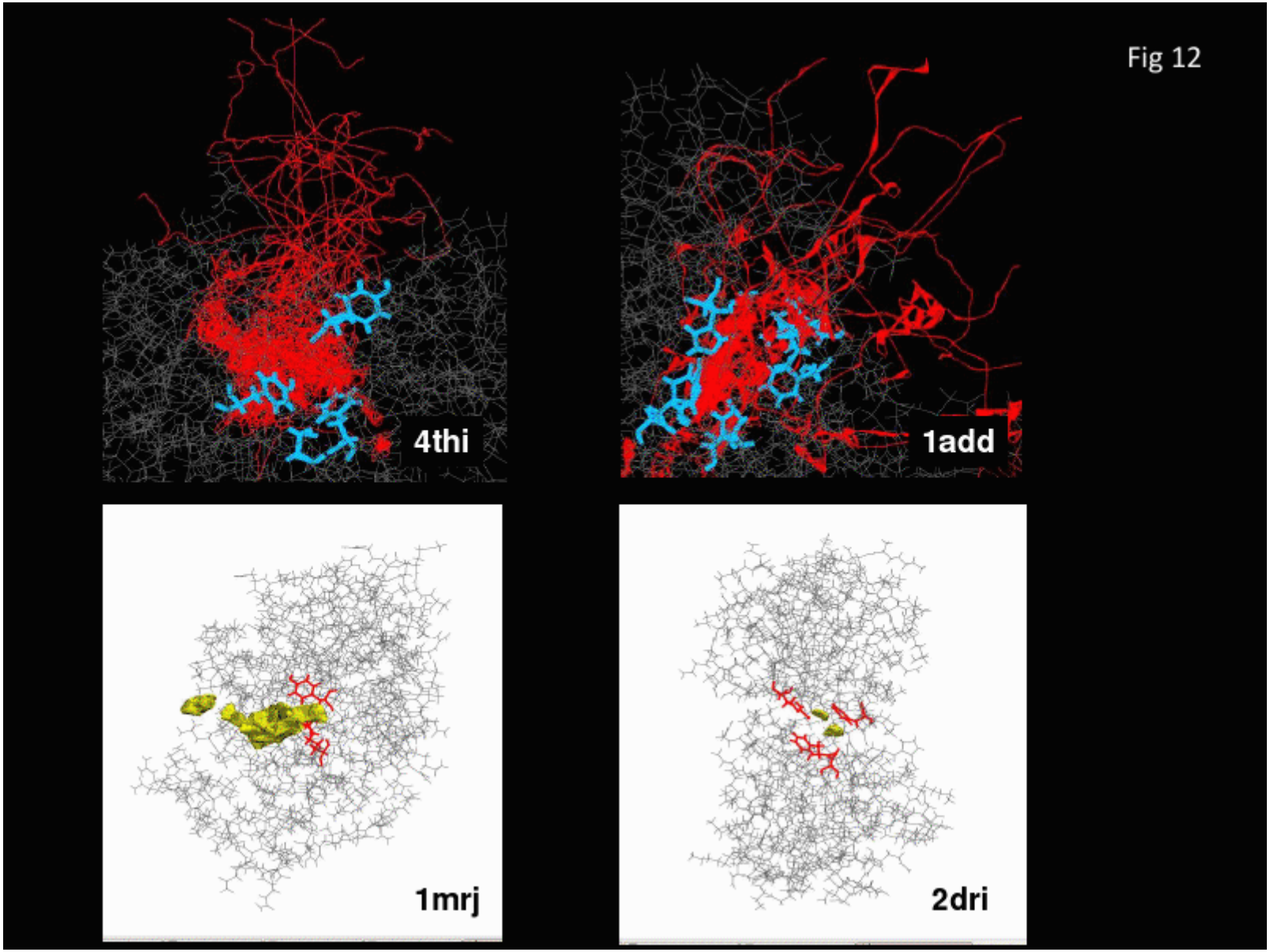


- In the last years the improvement of MD simulation was mainly through software-optimization

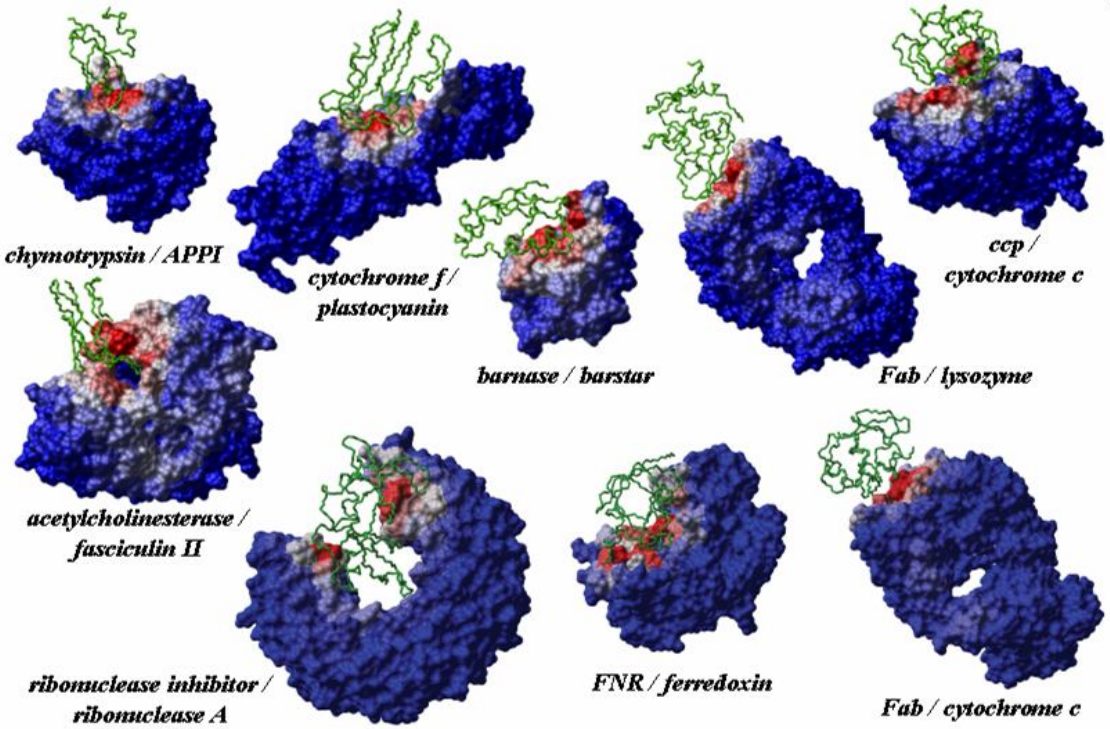
- Today the millisecond scale (10^{-3}s) is reached using specific hardware (512-processor)

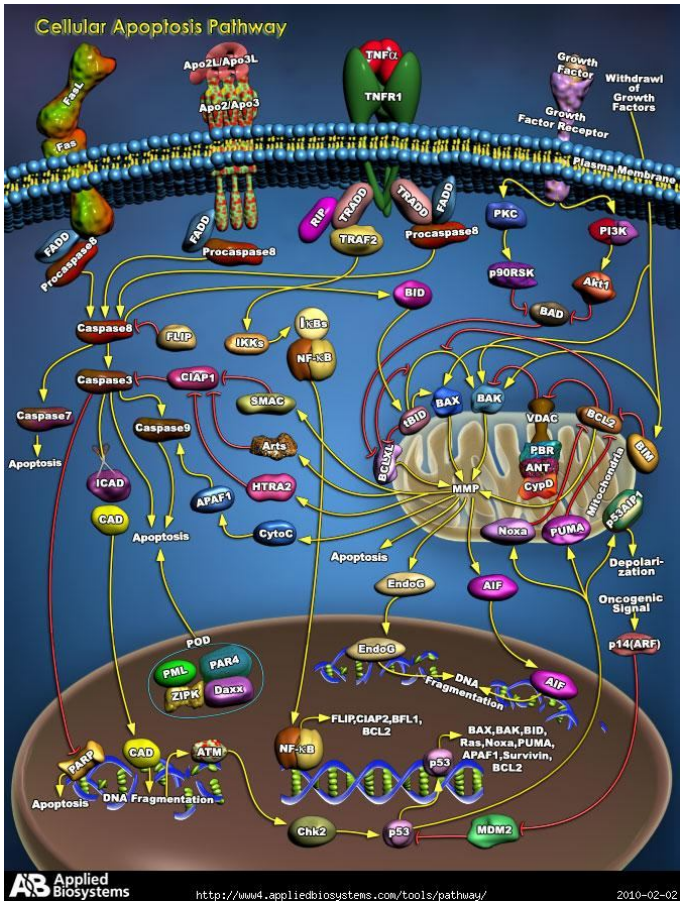
- Atoms and molecules are allowed to interact for a period of time by approximations of known physics





Protein Interaction

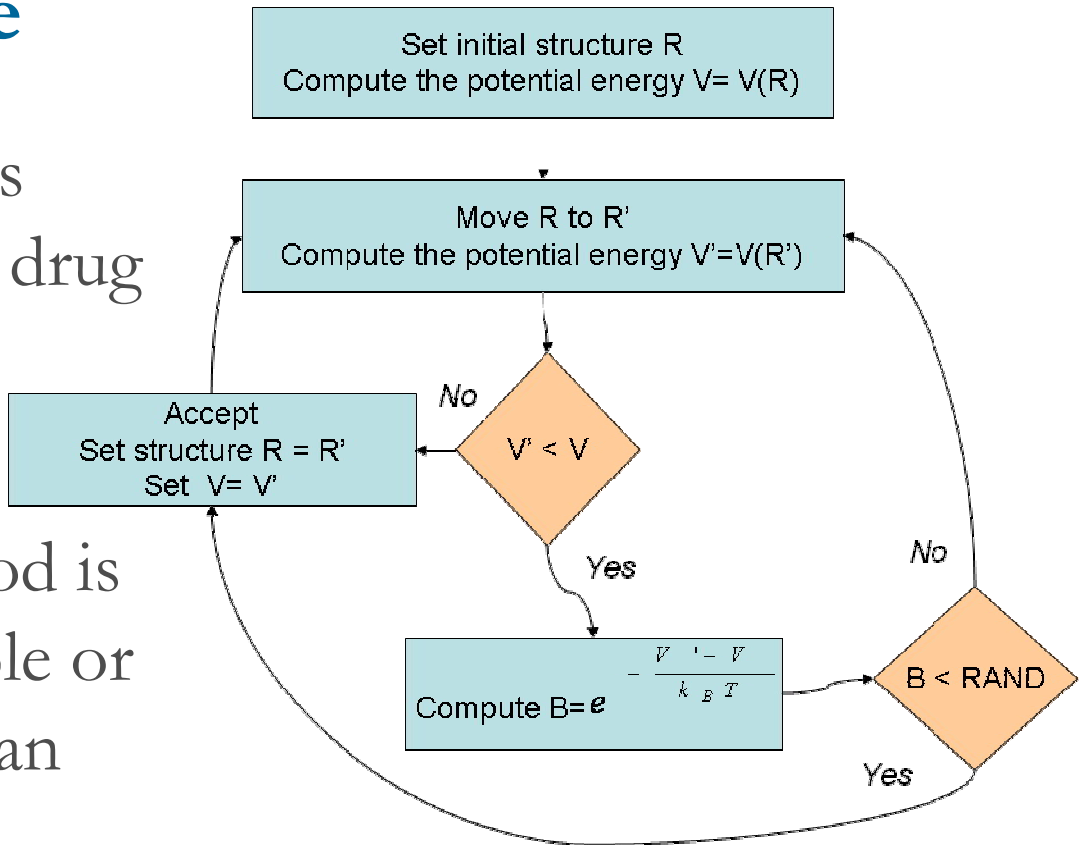




Simulate and Explore

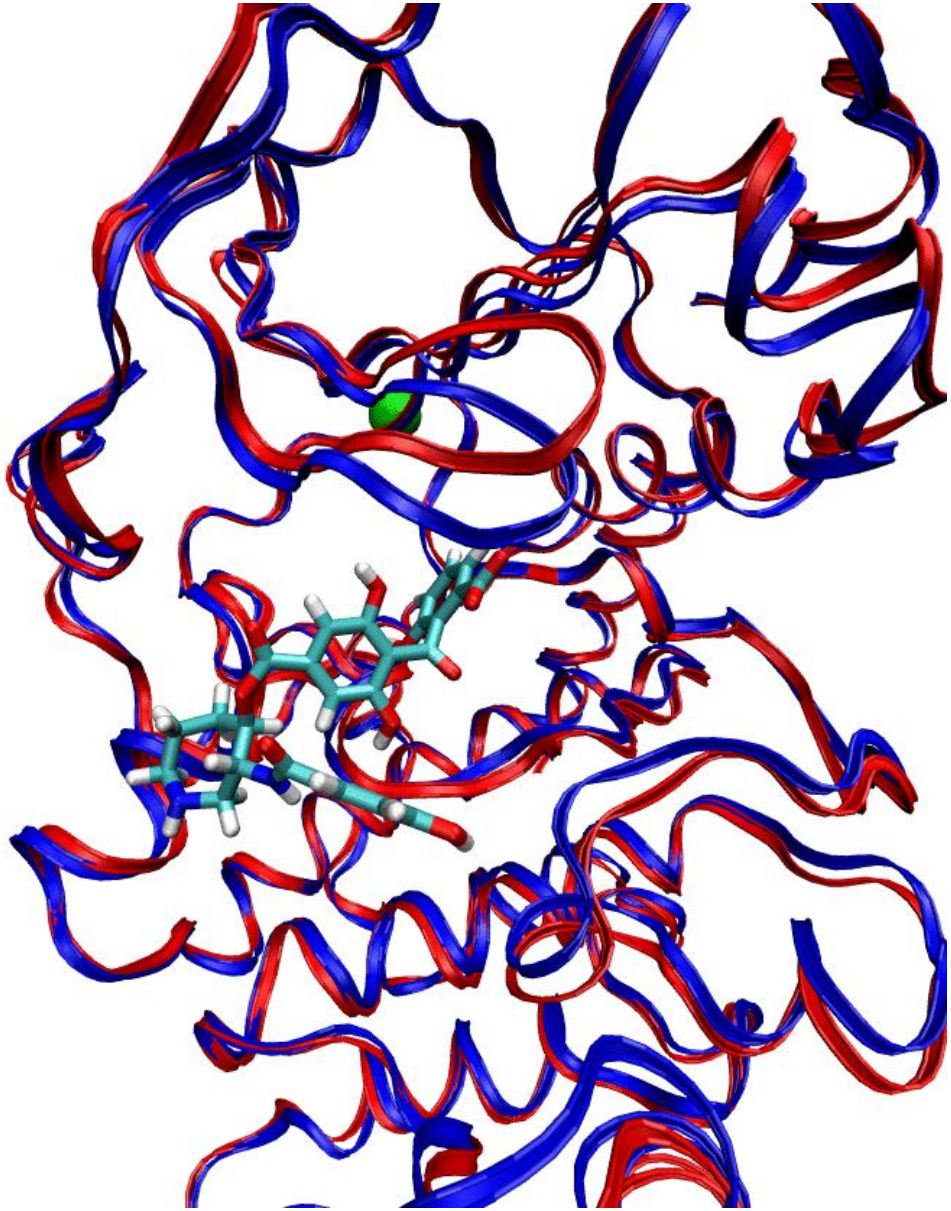
- Explore how to proteins interact. Explore how a drug binds its target.

- The Monte Carlo method is used when it is unfeasible or impossible to compute an exact result with a deterministic algorithm.

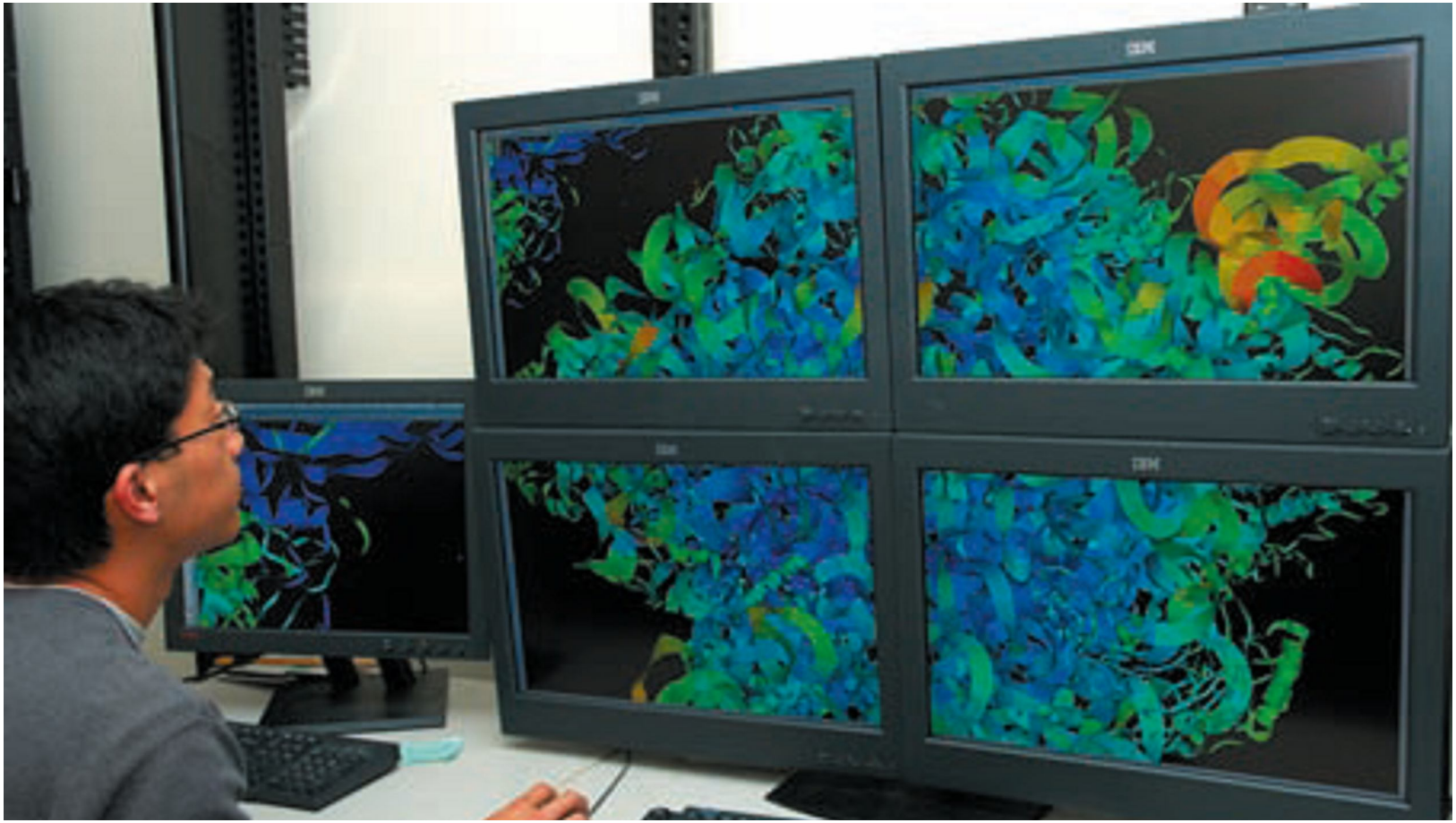


Drug Discovery

- Red: Cristal Protein
- Blue: Modeled Protein (Coarse Grained)
- Drug explores protein binding site using a Monte Carlo algorithm



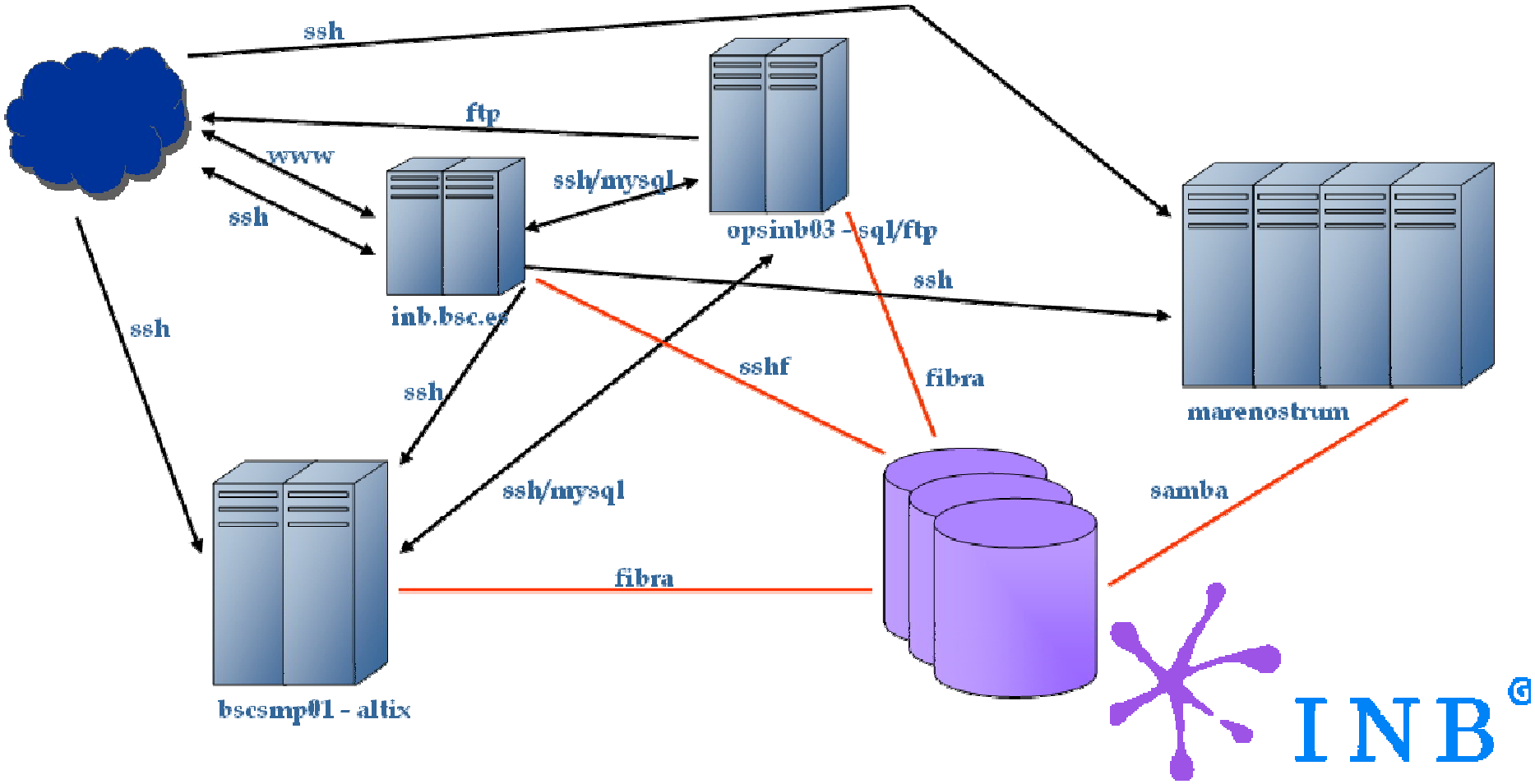
BioSupercomputing



Next Generation of Supercomputers

- We are currently porting our technology to new architecture:
  - **CELL processors: "Mare Incognito" machine, which aims at providing 10PF in the 2012 time frame . It will b a Supercomputer based on the Cell processor**
  - **GPUs: One of the most time consuming calculations in a typical molecular dynamics simulation is the evaluation of forces between atoms that do not share bonds. The high degree of parallelism and floating point arithmetic capability of GPUs can attain performance levels 20 times that of a single CPU core.**

BioSupercomputing open to "BioCommunity"

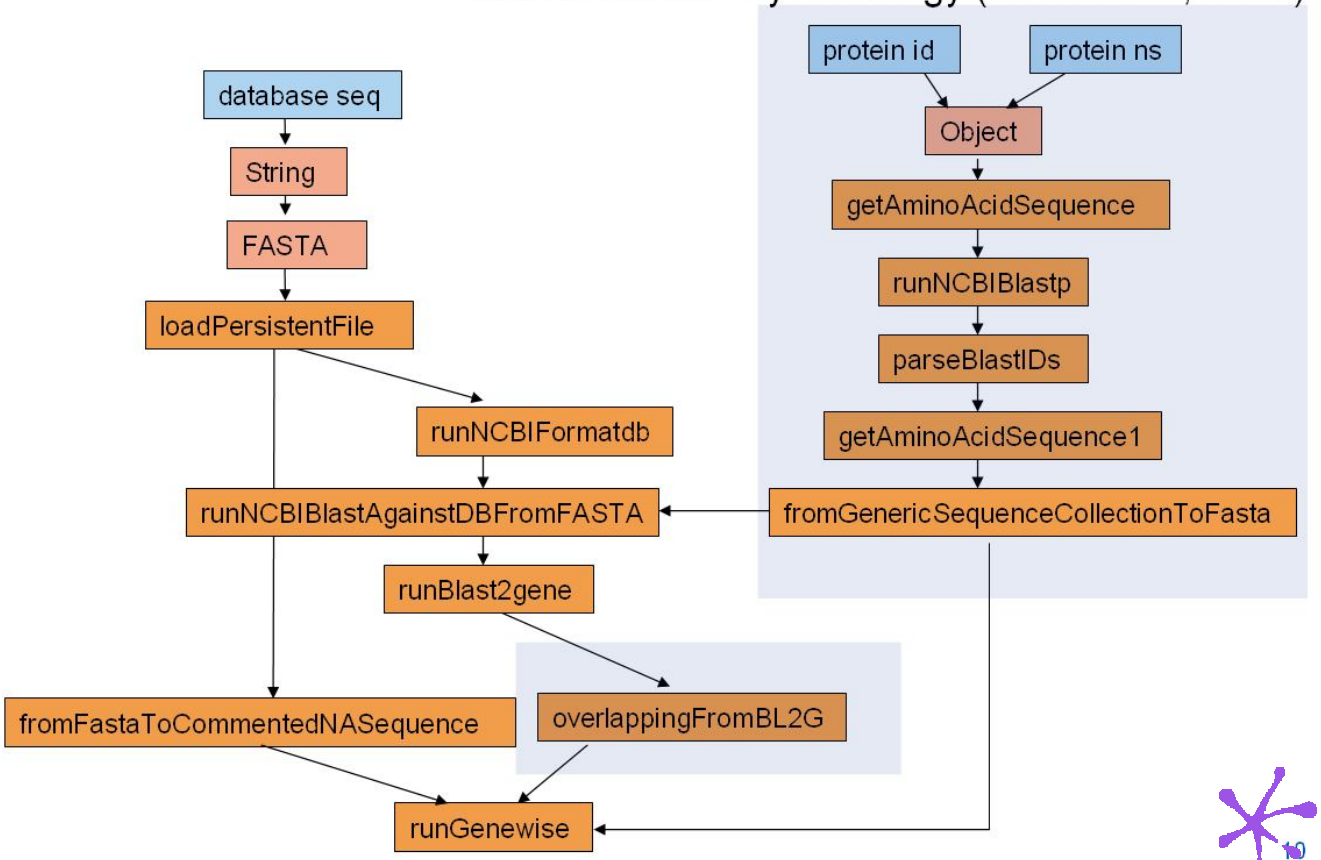


Pre-defined Workflows

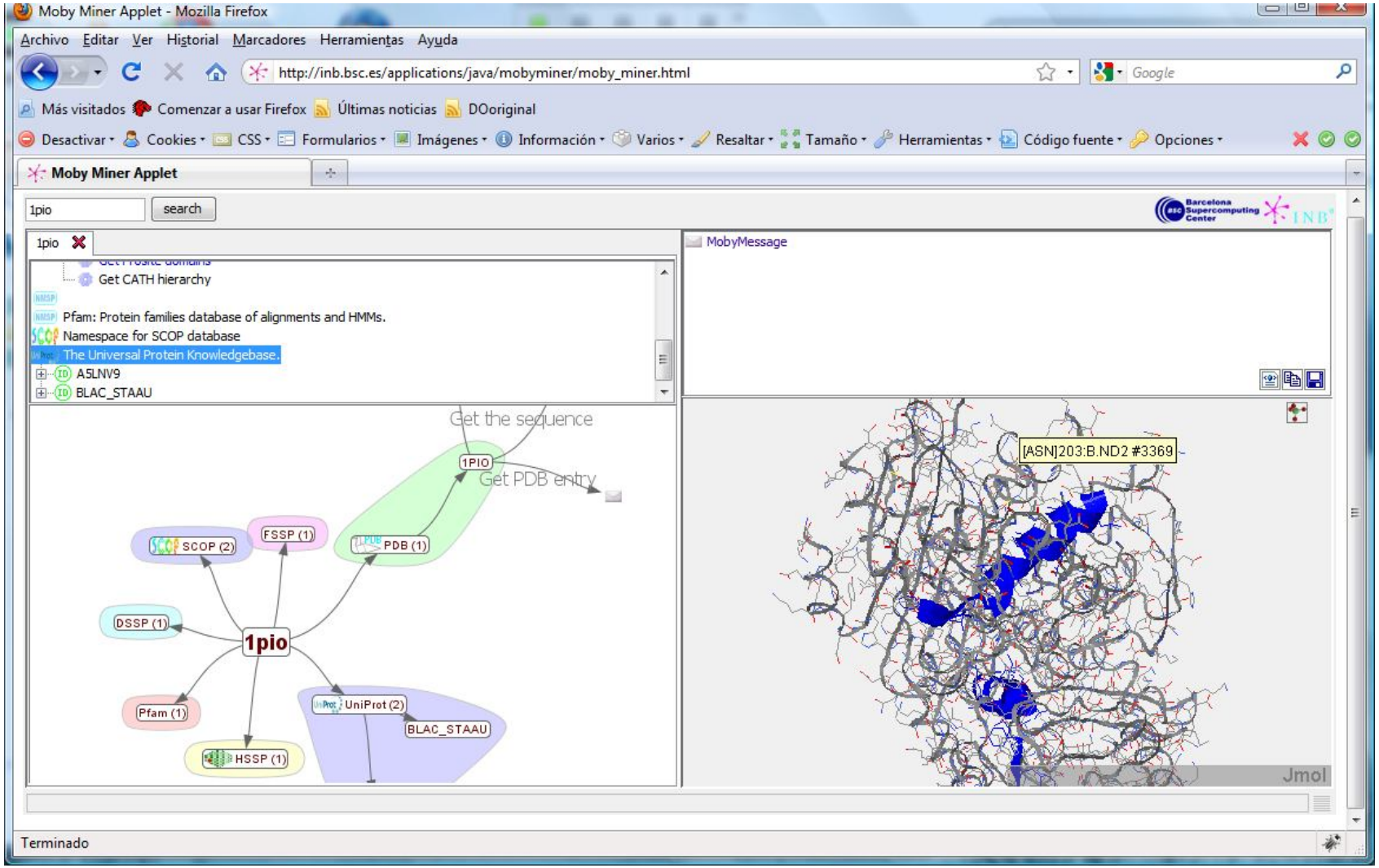
ISC 2008
A BioMoby-based workflow for gene detection using sequence homology. Royo, R., López, J., Torrents, D., Gelpi, J.L.

Workflow

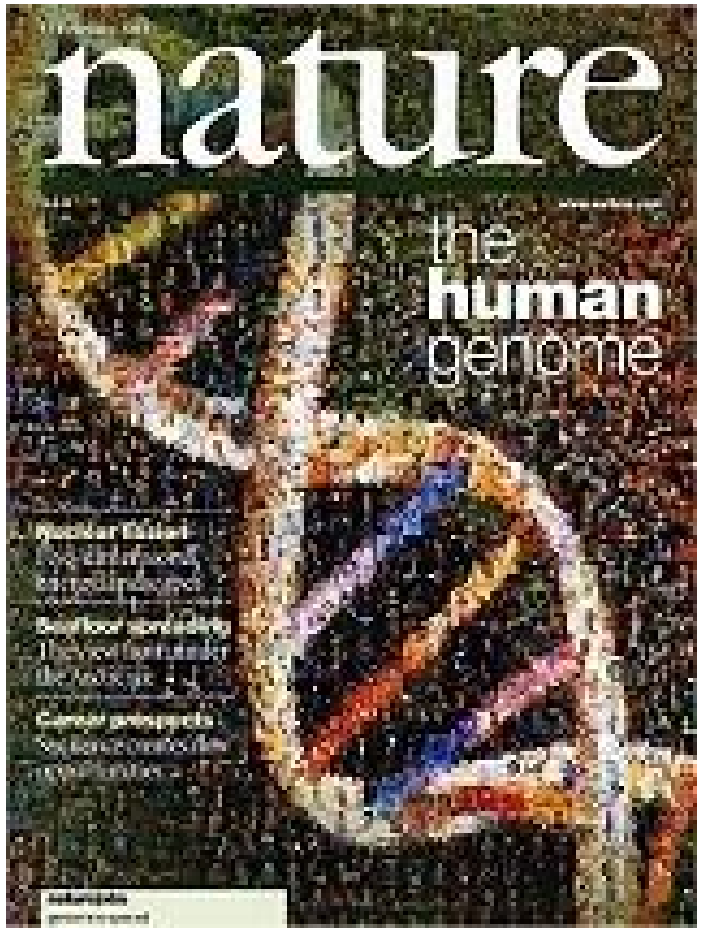
Gene detection by homology (D. Torrents, BSC.)

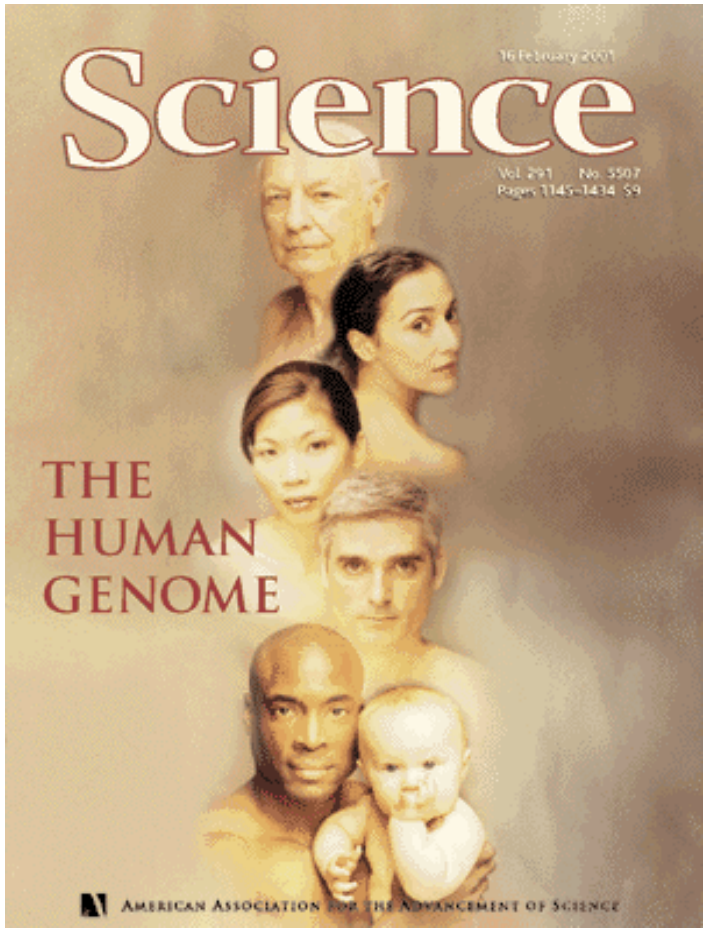


User-friendly for non-expert



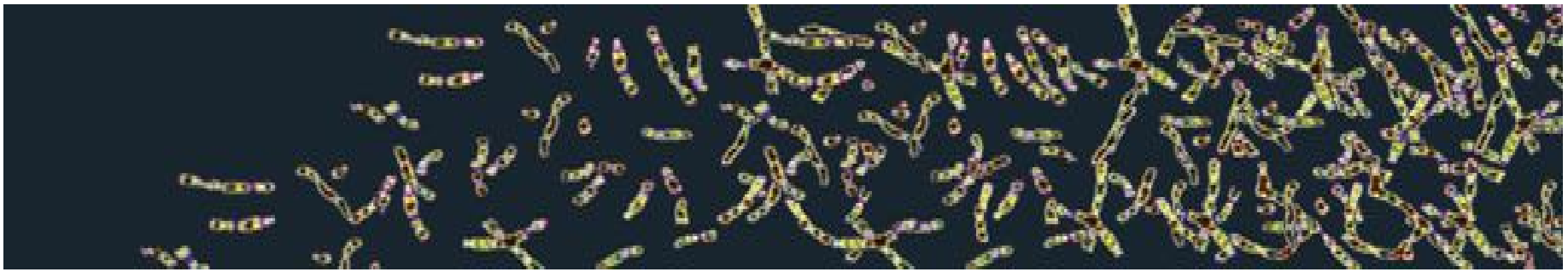
Data Analysis & Data Management





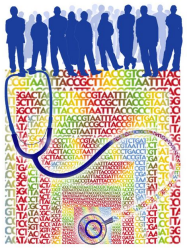
Genome Projects

http://www.ensembl.org



http://www.icgc.org

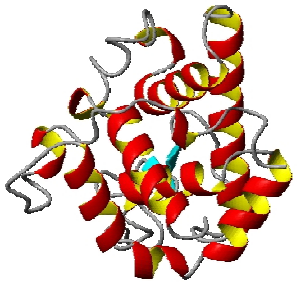
http://www.1000genomes.org



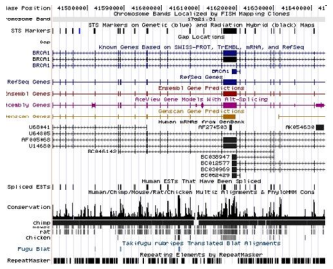
1000

International Cancer Genome Consortium

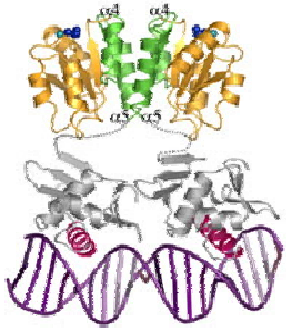
The Genome is the key for Target Discovery



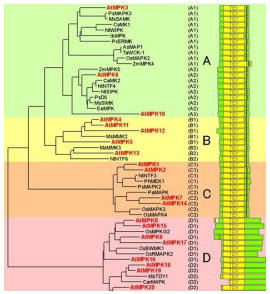
```
DEIGDAAKKLGDASYAFAKEVDWNNGIFLQ
APGKLQPLEALKAIDKMIVMGAAADPKLLK
AAAEAHHKAIGSISGPNGVTSRADWDNVNA
ALGRVIASVPENMVMDVYDSVSKITDPKVP
AYMKSLVNGADAEKAYEGFLAFKDVVKKSQ
VTSAA
```



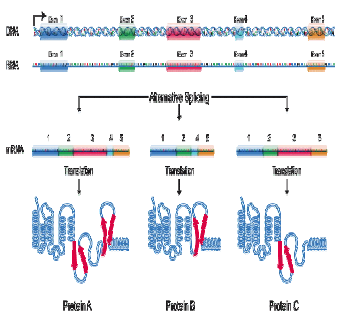
Origin



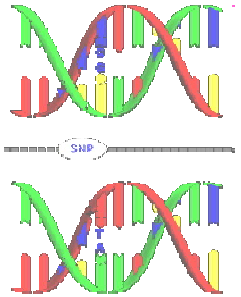
Regulation



Phylogeny



Isoform

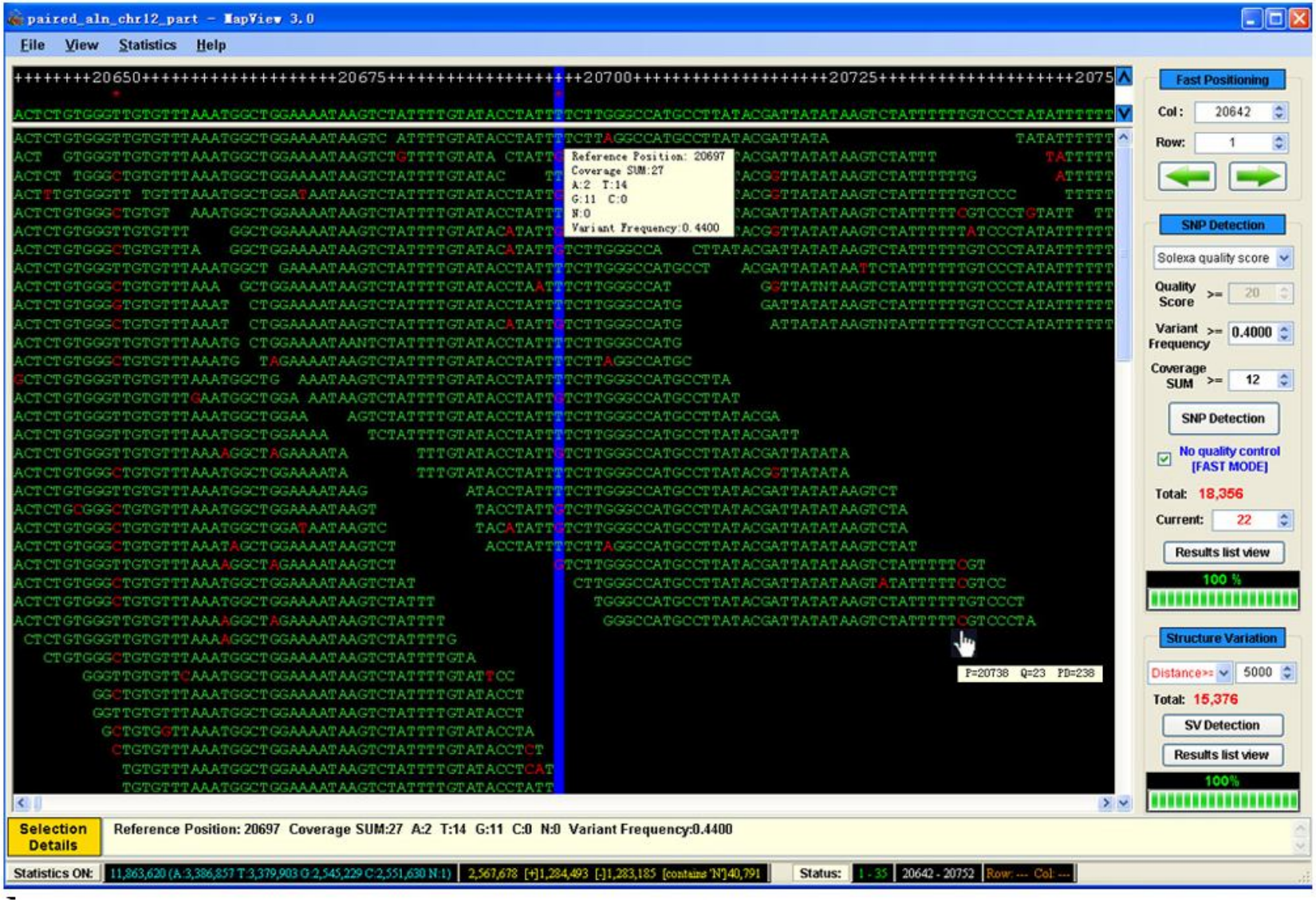


Mutations

Data Overload

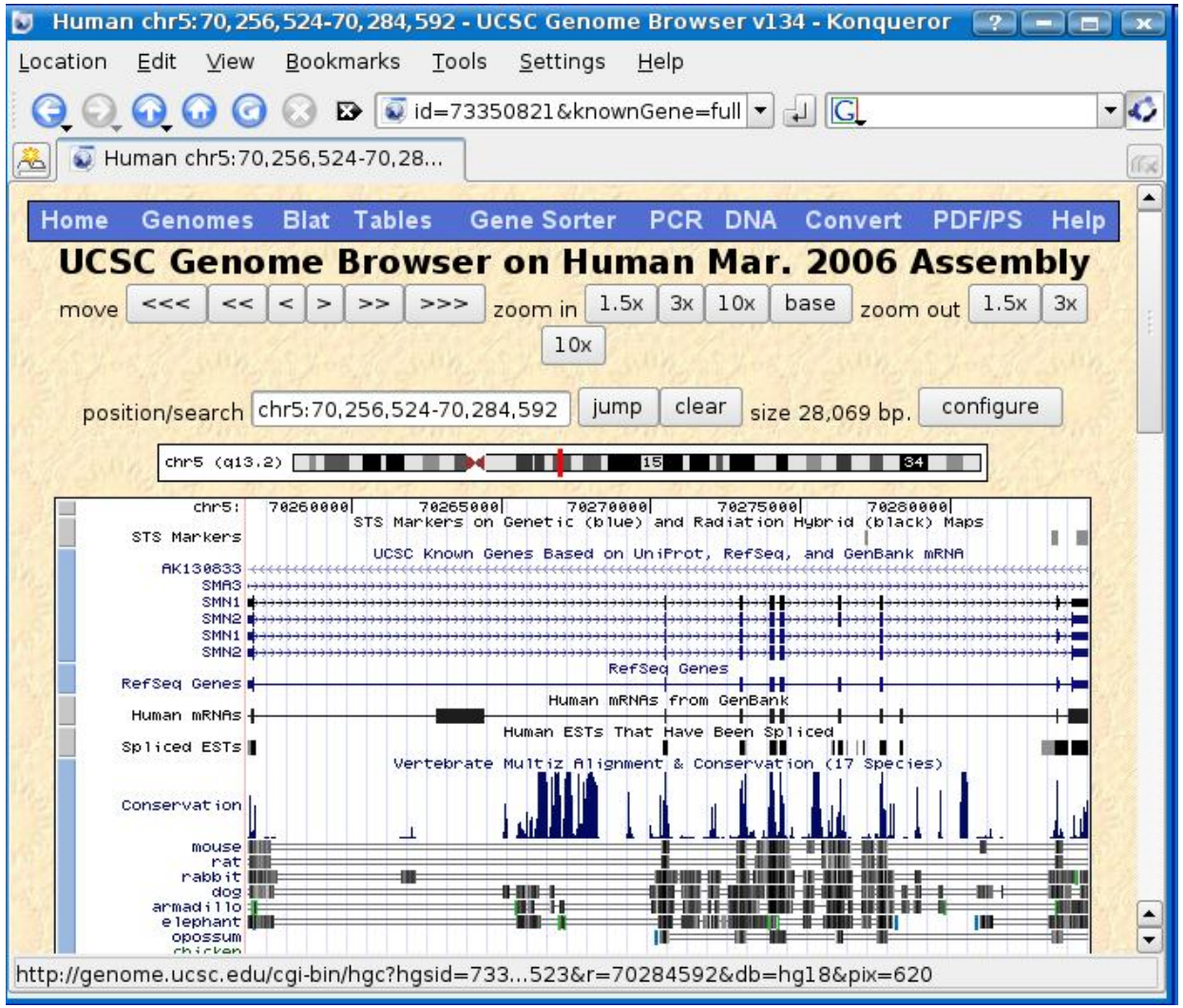
- Next Generation Sequencing platforms can sequence a complete genome fast and cheap. It is estimated that soon this service will cost $100.
- Approximately 115,200 Tiff formatted files are produced per run, each at about 8 megabytes (MB) in size. This is approximately 1 terabyte (TB) of data, which must be moved from the capture workstation to the analysis resource A mere 10–20 sequencing runs could overwhelm any storage and archiving system available to individual investigators
- A 1 GB network is essential within this environment, with 10 GB networks becoming more prevalent

The lack of expertise in data mining is the worst problem!



Our Challenge is …the Tomato

Genome Annotation



DNA sequence and DNA structure

Genomic Sequence

SEQUENCE1:

TGCACGTAGCTAAAAA[redacted]TGACGG[redacted]GCGTCGAGCAGCGAG

SEQUENCE2:

CGACTAGCACTACGTACATCGACATGCGACATCATCGACCTCGACAGTCGACGTCAACGACGAGTCAGCACGTGTAGTCGACAGTGAGCGGCAG

BLAST (sequence alignment): ID 49%

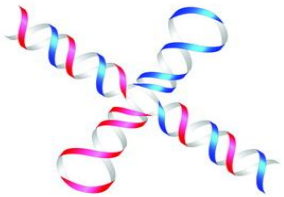
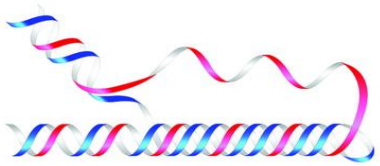
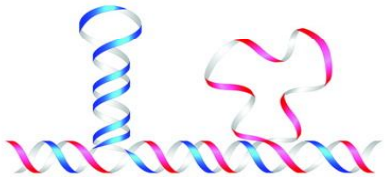
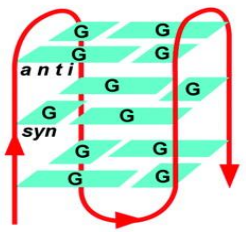
```
Sequence1      TGCACGTAGCTAAAAAGAGGAGGGAGGAGAG-GA----GAGGGATGACGGAAAAGAGGAGGGAGAGAGAG-AGGCGTC---GAGCAGCGAG
               .||||...| ||.|..||...|.||....|. ||    ||..|..||||..||.||.||||..||...|.| ||.||.|   ||||.|| ||
Sequence2 CGACTAGCACTACG-TACATCGACATGCGACATCATCGACCTCGACAGTCGACGTCAACGACGAGTCAGCACGTGTAGTCGACAGTGAGCGGC-AG
```

DNA Molecule

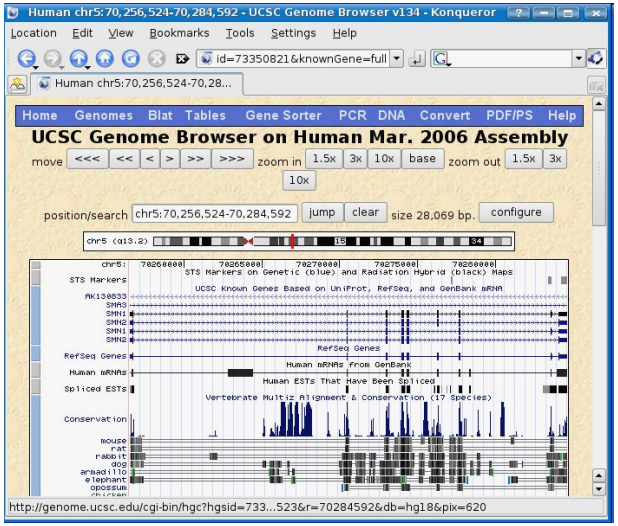
Sequence1

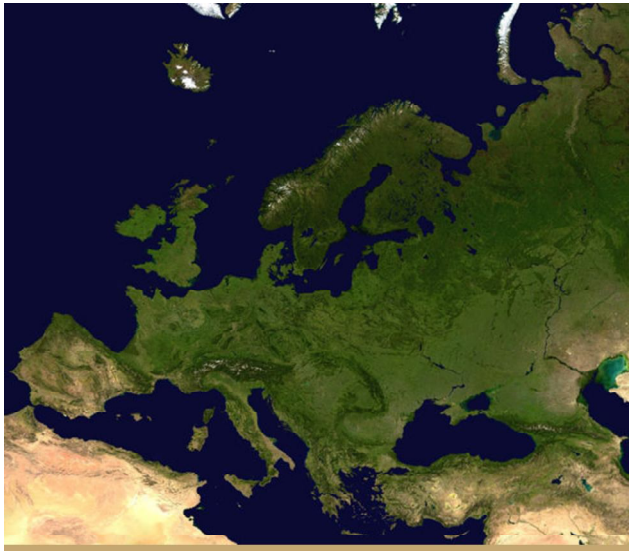
Sequence2

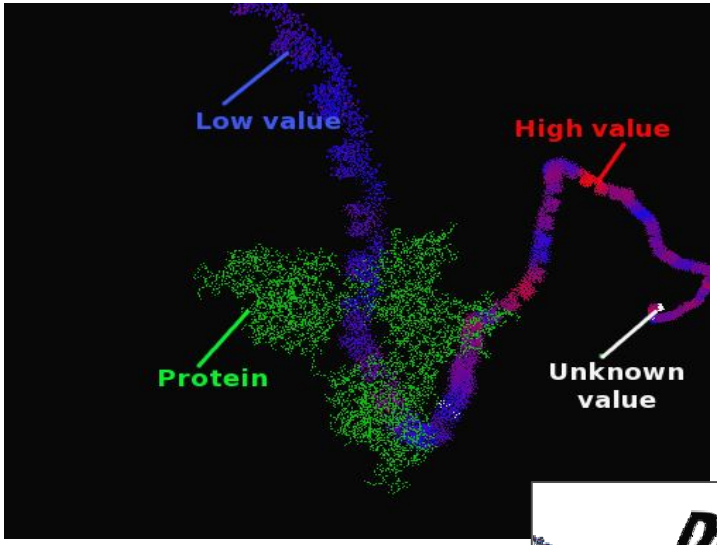
Unusual Conformations of DNA

Name	Conformation	General Seq. Requirements	Sequence
Cruciform		Inverted Repeats	TCGGTACCGA AGCCATGGCT
Triplex		$(R \bullet Y)_n$ Mirror Repeats	AAGAGGGGAGAA TTCTCCCCTCTT
Slipped (Hairpin) Structure		Direct Repeats	TCGGTTCGGT AGCCAAGCCA
Tetraplex		Oligo $(G)_n$ Tracts	$AG_3(T_2AG_3)_3$ single strand
Left-handed *Z* - DNA		$(YR \bullet YR)_n$	CGCGTGCGTGTG GCGCACGCACAC

Bacolla et al. (2004) J. Biol. Chem



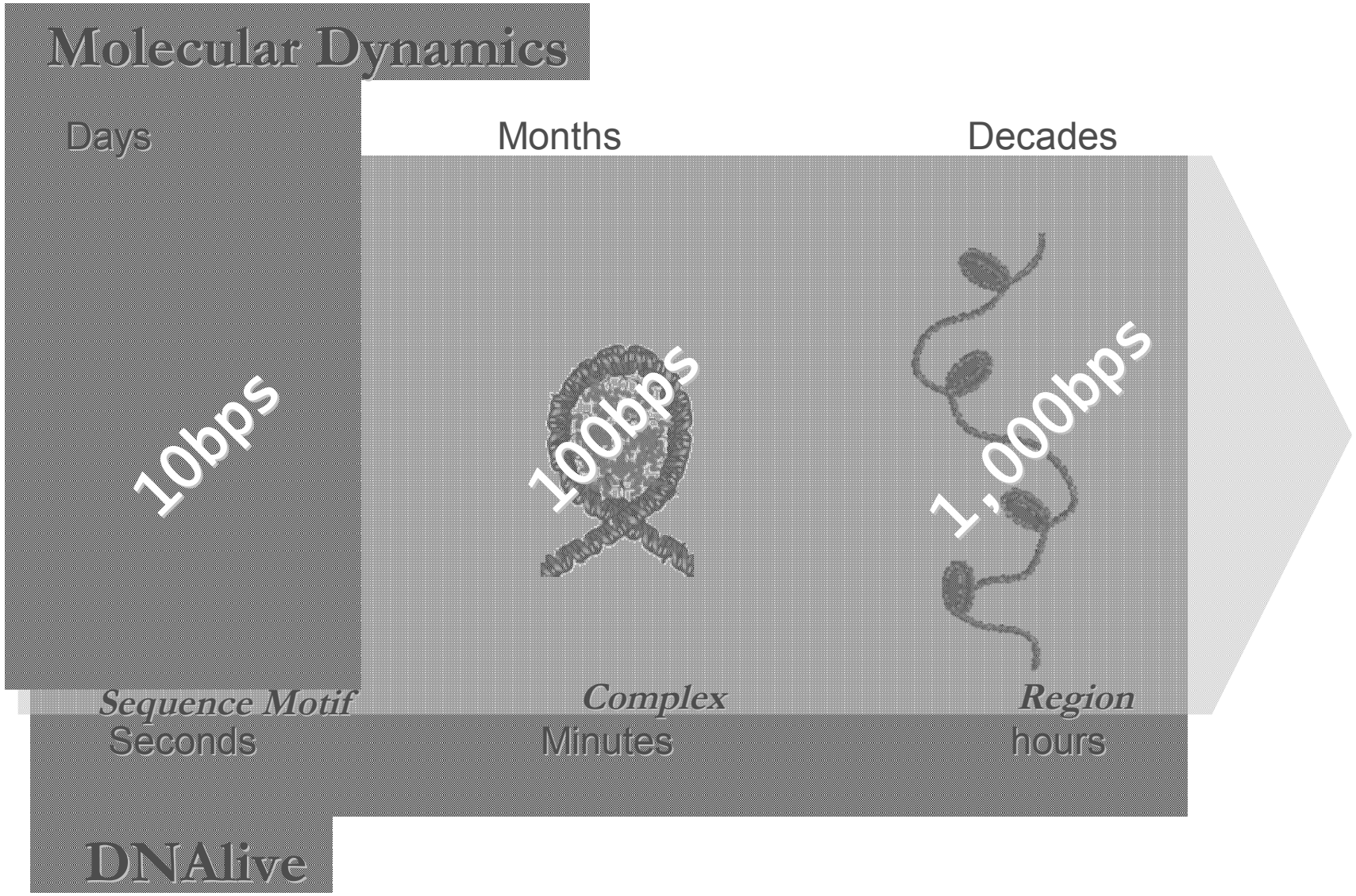




GENOME IN 3D !

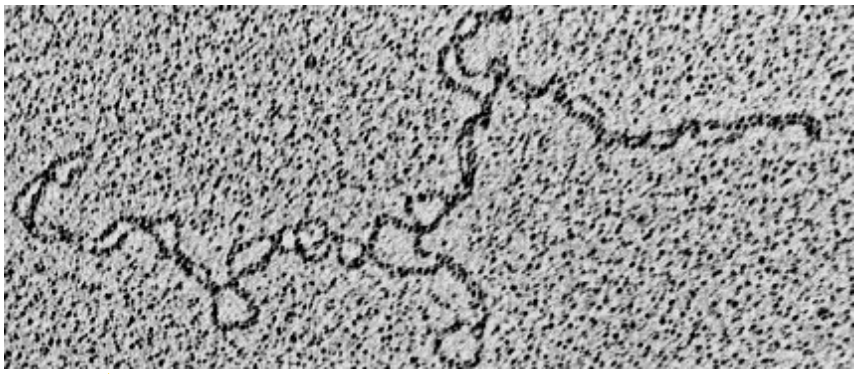
DNAlive

Large Scale DNA Dynamics

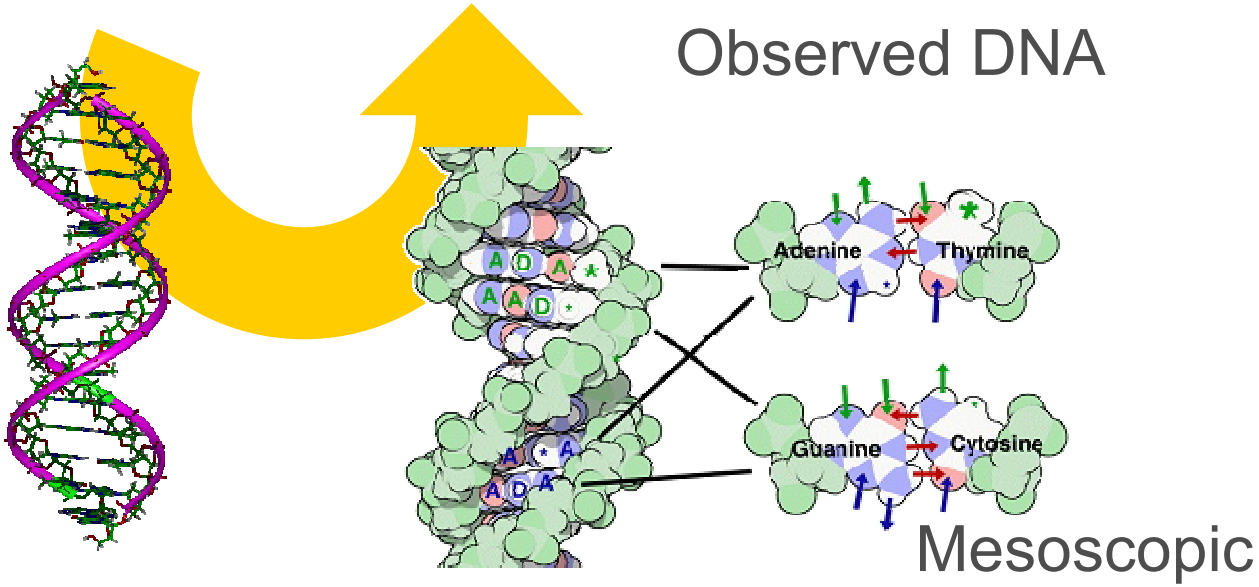


Representing the genomic DNA

Supercomputing Resources



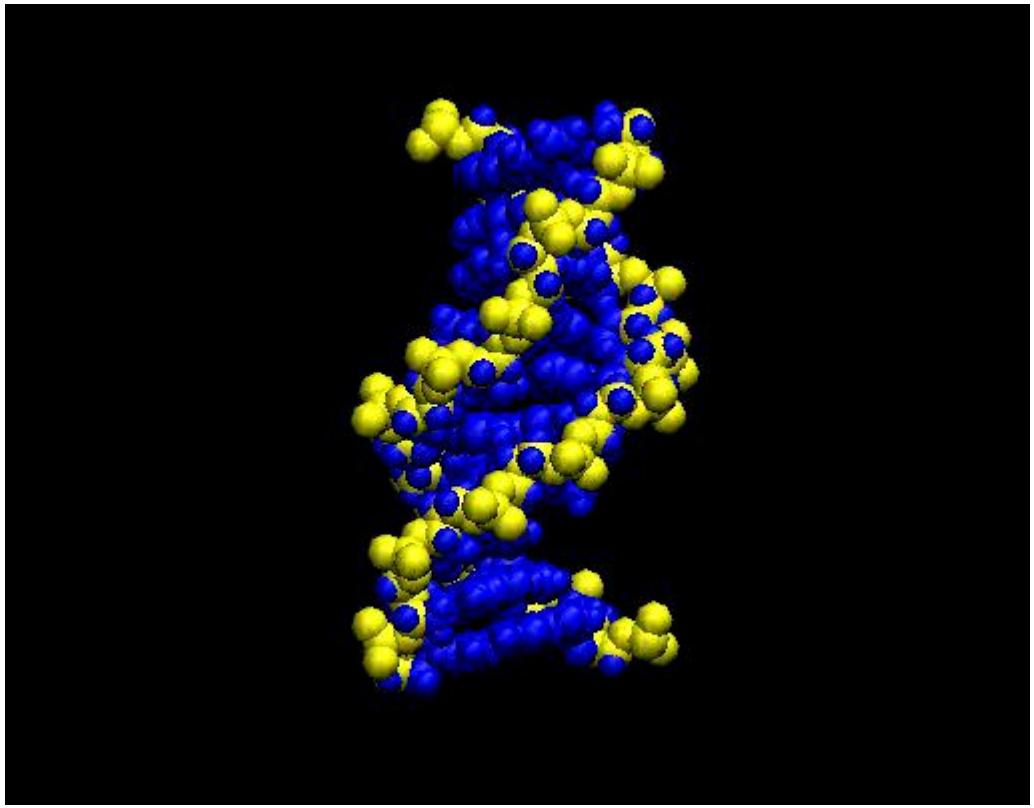
Observed DNA



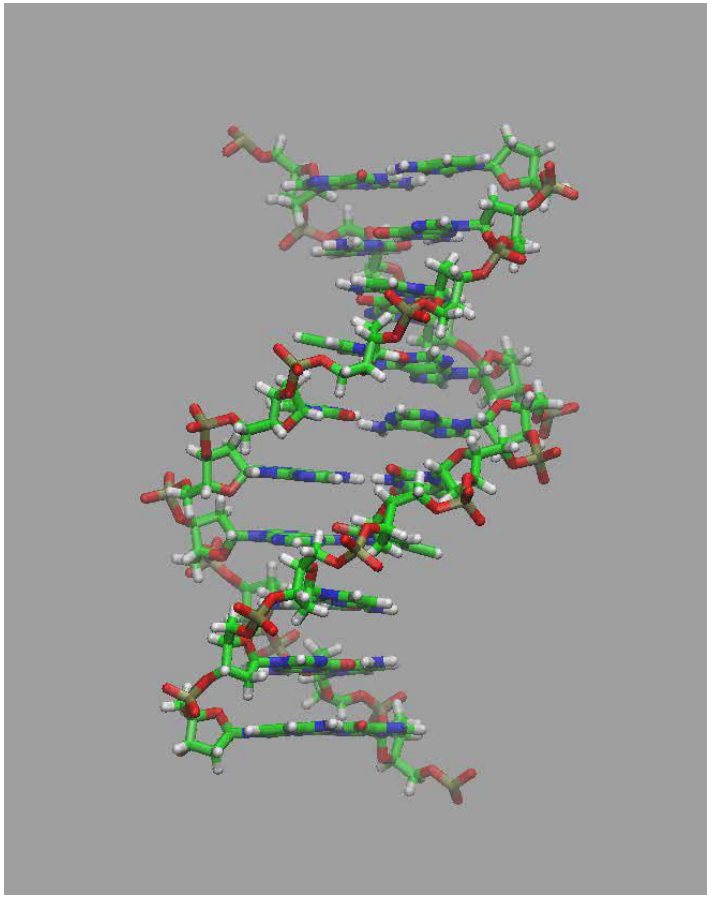
Molecular Dynamics

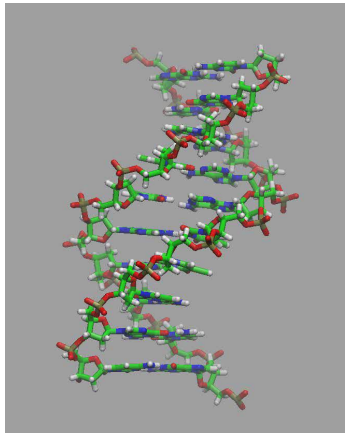
Mesoscopic

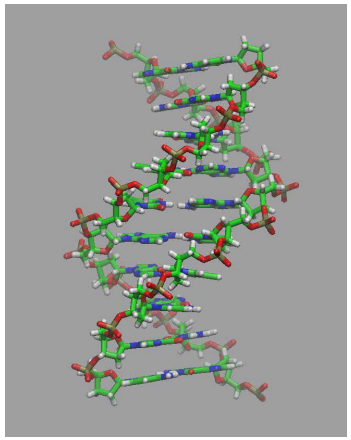
ParmBSC: a Force Field for long DNA Simulations



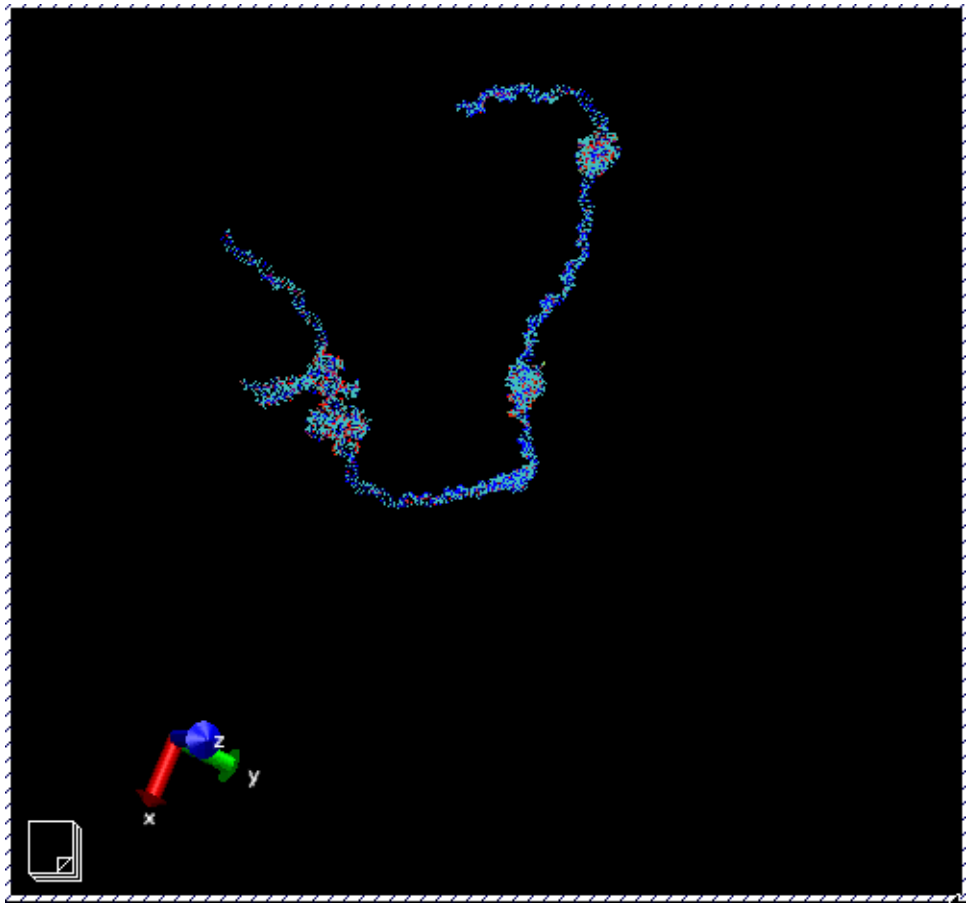
Study of Principal Component Analysis and Essential Dynamics







Genic DNA Dynamics



$$E = \sum_{i=1}^{M} \sum_{j=1}^{6} K_{ij} \left(\xi_{ij} - \xi_{ij}^{0} \right)^2$$

MonteCarlo Simulatio (n. steps = 100,000 x number of flexible dinucleotides)
Adjusted to accept 40% of the steps
Energy: Perez et al. (2007) JACS

Summary

- Computer simulation (Drug Discovery)
  - **Molecular Dynamics**
  - **Protein-Protein Interaction**
  - **Protein-Ligand Docking**
- BioSupercomputing
  - **Computational Biology under GPU & CELL**
  - **User-friendly computing access Web-Services**
- Data analysis and Data Management (Target Discovery)
  - **Next Generation Sequencing**
  - **Genomics & DNA structure**

Contact:

Ramon Goñi
Senior Researcher
BSC, Life Sciences
ramon.goni@bsc.es