The biological problem

- Extremely important
- Complex
- Diffuse and stochastic nature
- Bad defined problems
- Cover a wide scale (time and space)
- Typically defined with poor data
- Exponential growth (↑↑ Moore law)
- Cost of obtaining data ↓↓
- Cost of processing data ↑↑

The dual nature of computers

- Ordenador: machine to manage data.

- Computador: machine to do maths.

The biological problem is diverse in nature. Input of biologists should be important when selecting architectures

The dual nature of computers

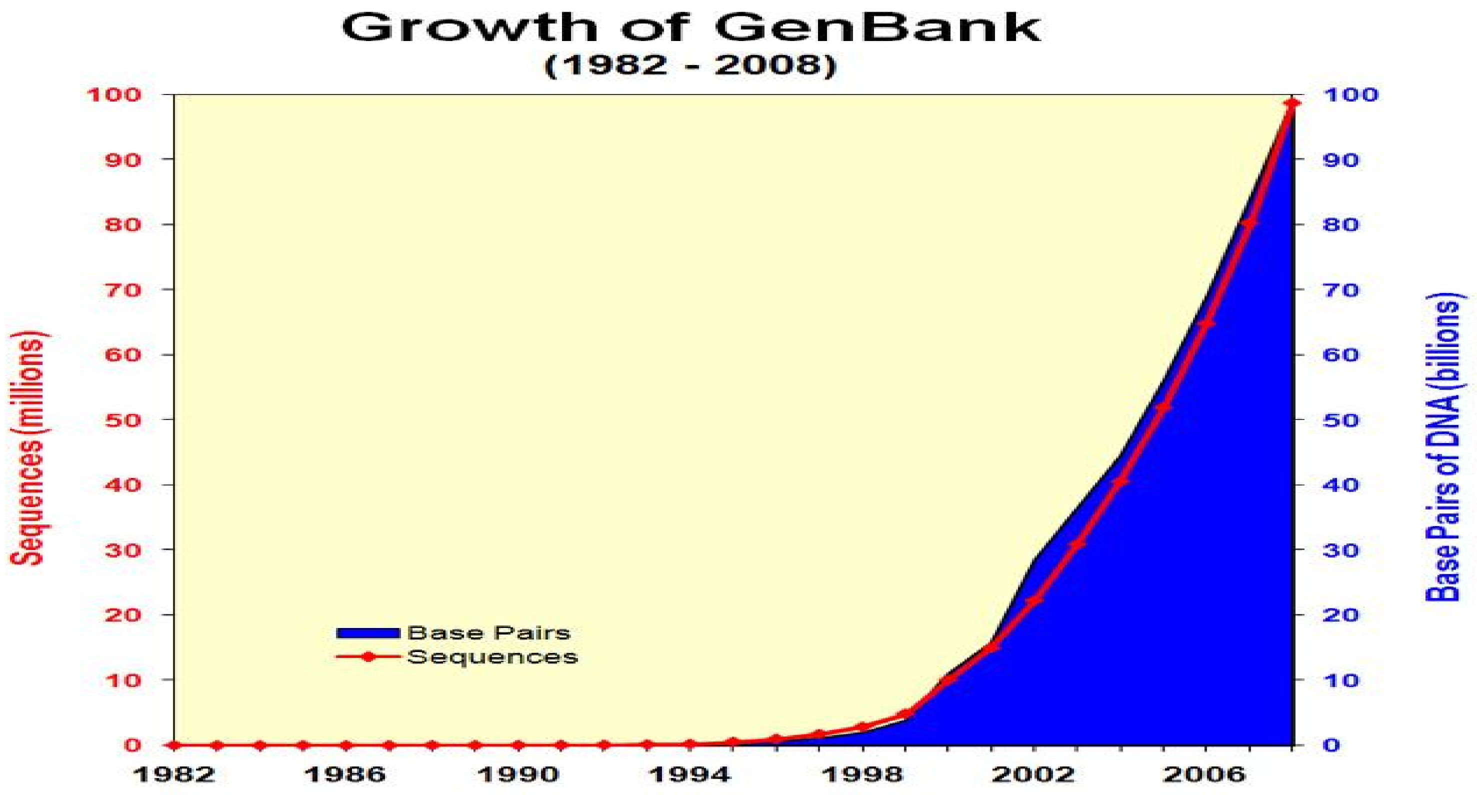
- Ordenador: machine to manage data.
- Computador: machine to do maths.

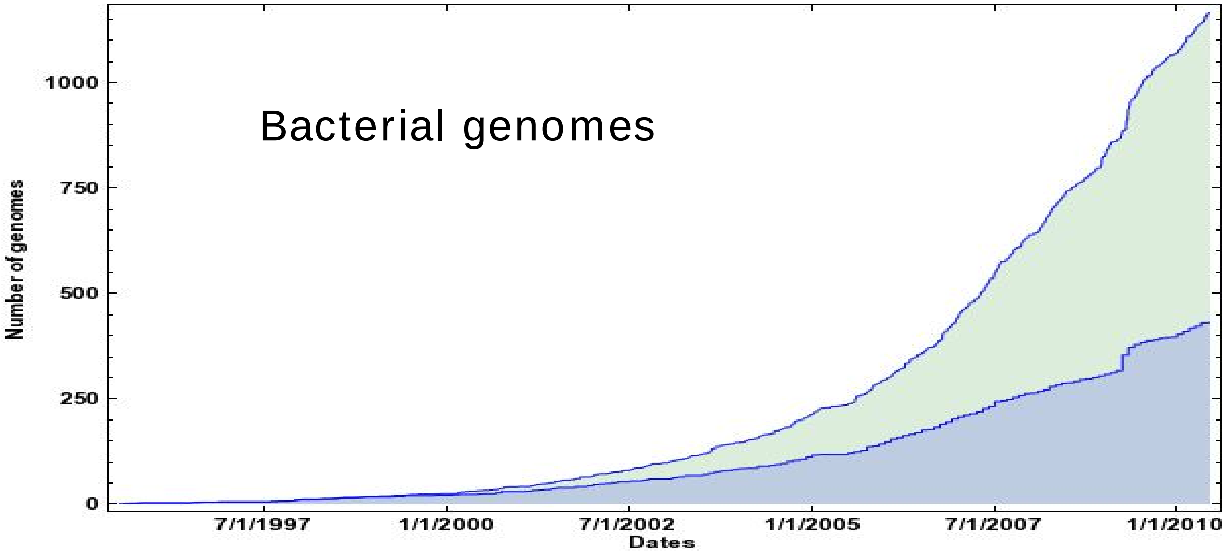
In biology capability does not always translate into FLOPs

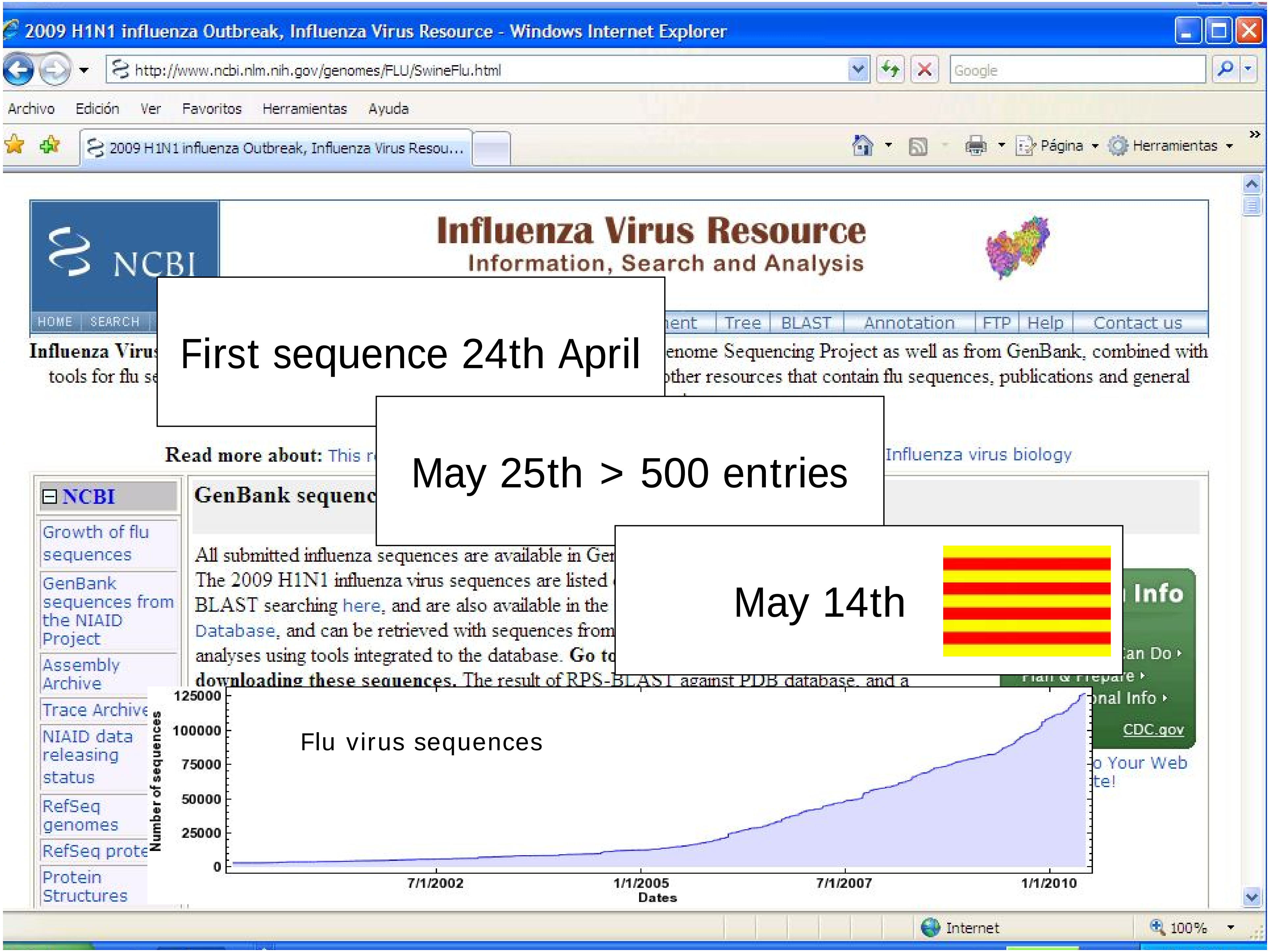
The computer as data manager

- Data growths exponentially
- Data management more complex
- Data integration a crucial issue
- Processing of data very costly

- Computers are now the limiting step in many biological problems
- Time is a major issue in Bio-research


Growth of GenBank
(1982 - 2008)
Sequences (millions)
Base Pairs of DNA (billions)
0
10
20
30
40
50
60
70
80
90
100
1982
1986
1990
1994
1998
2002
2006
Base Pairs
Sequences


Bacterial genomes
Number of genomes
0
250
500
750
1000
7/1/1997
1/1/2000
7/1/2002
1/1/2005
7/1/2007
1/1/2010
Dates


2009 H1N1 influenza Outbreak, Influenza Virus Resource - Windows Internet Explorer
http://www.ncbi.nlm.nih.gov/genomes/FLU/SwineFlu.html
Google
Archivo
Edición
Ver
Favoritos
Herramientas
Ayuda
2009 H1N1 influenza Outbreak, Influenza Virus Resou...
Página
Herramientas
NCBI
HOME
SEARCH
ent
Tree
BLAST
Annotation
FTP
Help
Contact us
Internet
100%

Influenza Virus Resource

Information, Search and Analysis

First sequence 24th April

Influenza Virus … enome Sequencing Project as well as from GenBank, combined with tools for flu se… other resources that contain flu sequences, publications and general

May 25th > 500 entries

Read more about: This r… Influenza virus biology

NCBI
- Growth of flu sequences
- GenBank sequences from the NIAID Project
- Assembly Archive
- Trace Archive
- NIAID data releasing status
- RefSeq genomes
- RefSeq prote…
- Protein Structures

GenBank sequenc…

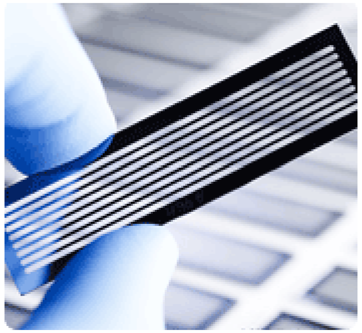
May 14th

🟨🟥🟨🟥🟨🟥🟨🟥🟨

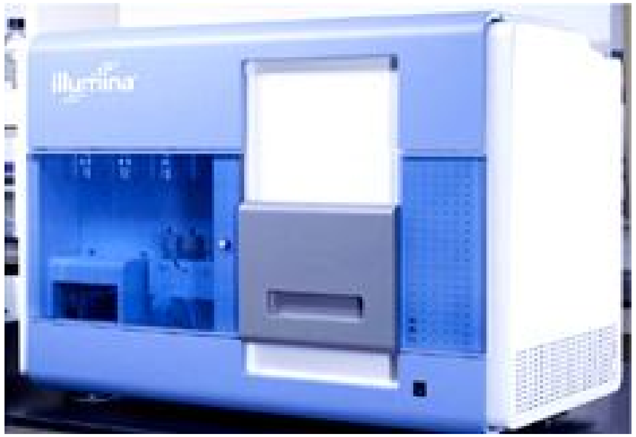
All submitted influenza sequences are available in Ge… The 2009 H1N1 influenza virus sequences are listed … BLAST searching here, and are also available in the … Database, and can be retrieved with sequences from … analyses using tools integrated to the database. Go to … downloading these sequences. The result of RPS-BLAST against PDB database, and a

…Info
…an Do
Plan & Prepare
…onal Info
CDC.gov
…o Your Web …te!

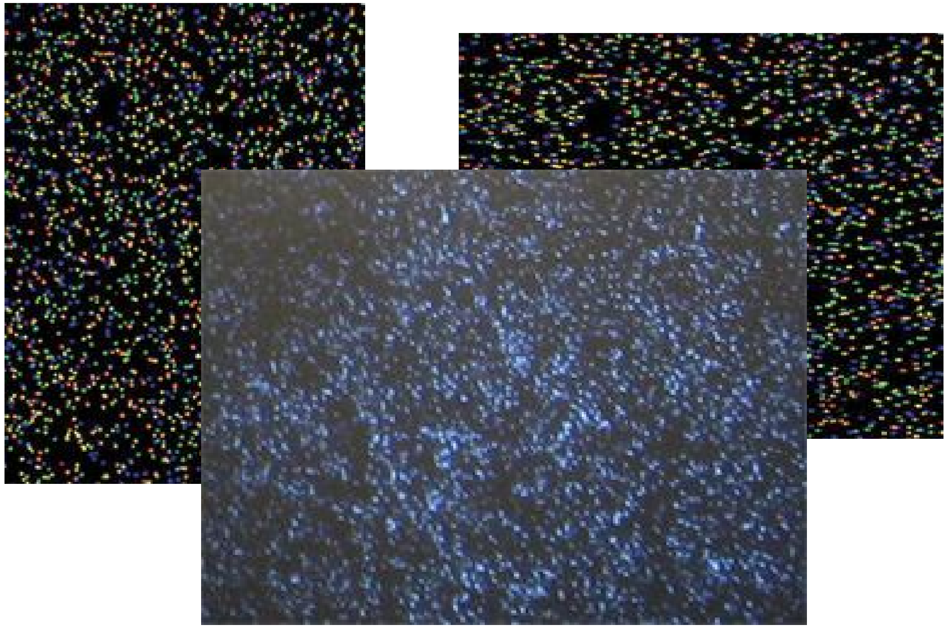
Flu virus sequences



Next generation sequencing

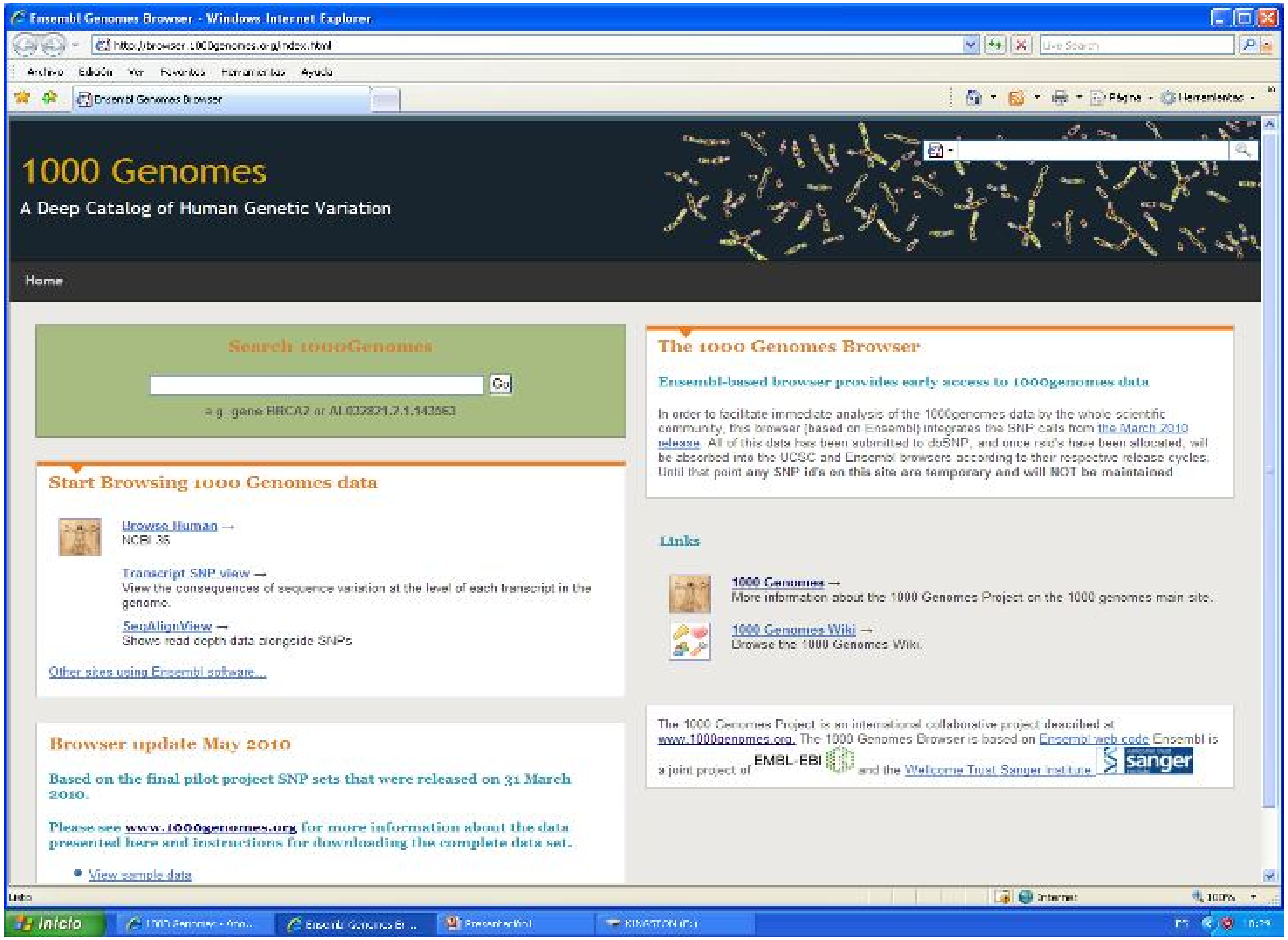


- Every experiment 2 Tb of data
- Every machine 2 experiments a week
- A medium sized center 10 machines



A sequencing center generates around 2 Petabytes data a year!

Massive (human) genomic projects


Ensembl Genomes Browser - Windows Internet Explorer
http://browser.1000genomes.org/index.html
Archivo Edición Ver Favoritos Herramientas Ayuda
Ensembl Genomes Browser
Página Herramientas
Home
Go
Listo
Internet
100%
Inicio

1000 Genomes

A Deep Catalog of Human Genetic Variation

Search 1000Genomes

e.g. gene BRCA2 or AL032821.2.1.143563

Start Browsing 1000 Genomes data

Browse Human →
NCBI 36

Transcript SNP view →
View the consequences of sequence variation at the level of each transcript in the genome.

SeqAlignView →
Shows read depth data alongside SNPs

Other sites using Ensembl software...

Browser update May 2010

Based on the final pilot project SNP sets that were released on 31 March 2010.

Please see www.1000genomes.org for more information about the data presented here and instructions for downloading the complete data set.

- View sample data

The 1000 Genomes Browser

Ensembl-based browser provides early access to 1000genomes data

In order to facilitate immediate analysis of the 1000genomes data by the whole scientific community, this browser (based on Ensembl) integrates the SNP calls from the March 2010 release. All of this data has been submitted to dbSNP, and once rsid's have been allocated, will be absorbed into the UCSC and Ensembl browsers according to their respective release cycles. Until that point any SNP id's on this site are temporary and will NOT be maintained

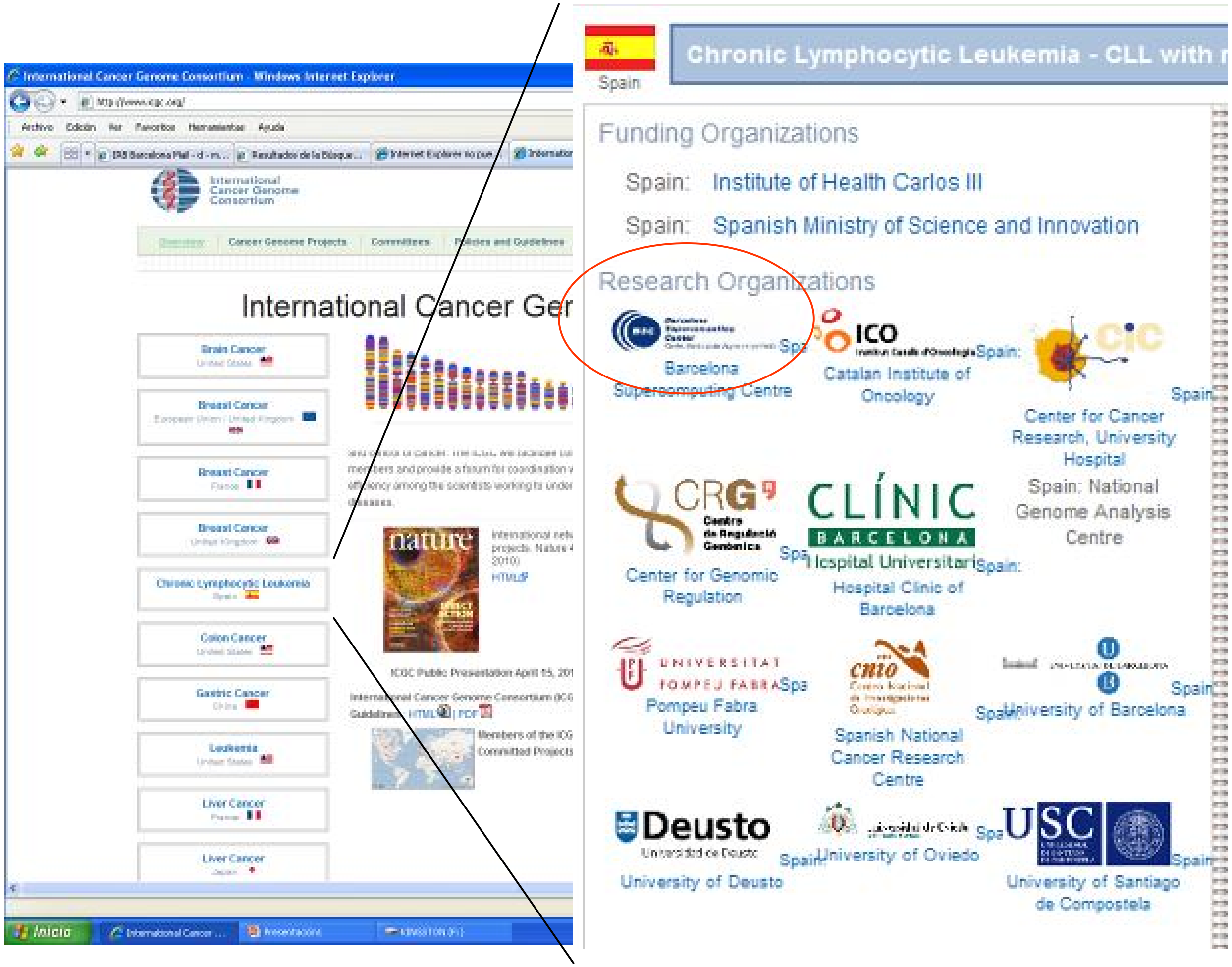
Links

1000 Genomes →
More information about the 1000 Genomes Project on the 1000 genomes main site.

1000 Genomes Wiki →
Browse the 1000 Genomes Wiki.

The 1000 Genomes Project is an international collaborative project described at www.1000genomes.org. The 1000 Genomes Browser is based on Ensembl web code Ensembl is a joint project of EMBL-EBI and the Wellcome Trust Sanger Institute sanger

Massive (human) genomic projects



International Cancer Genome Consortium - Windows Internet Explorer

http://www.icgc.org/

Archivo Edición Ver Favoritos Herramientas Ayuda

International Cancer Genome Consortium

Overview | Cancer Genome Projects | Committees | Policies and Guidelines

International Cancer Ger

Brain Cancer
United States

Breast Cancer
European Union / United Kingdom

Breast Cancer
France

Breast Cancer
United Kingdom

Chronic Lymphocytic Leukemia
Spain

Colon Cancer
United States

Gastric Cancer
China

Leukemia
United States

Liver Cancer
France

Liver Cancer
Japan

members and provide a forum for coordination
efficiency among the scientists working to under
diseases.

International netw
projects. Nature 4
2010)
HTML

ICGC Public Presentation April 15, 201

International Cancer Genome Consortium (ICG
Guidelines HTML | PDF

Members of the ICG
Committed Projects

Spain

Chronic Lymphocytic Leukemia - CLL with r

Funding Organizations

Spain: Institute of Health Carlos III

Spain: Spanish Ministry of Science and Innovation

Research Organizations

Spain: Barcelona Supercomputing Centre

Spain: Catalan Institute of Oncology

Spain: Center for Cancer Research, University Hospital

Spain: National Genome Analysis Centre

Spain: Center for Genomic Regulation

Spain: Hospital Clinic of Barcelona

Spain: Pompeu Fabra University

Spanish National Cancer Research Centre

Spain: University of Barcelona

Spain: University of Deusto

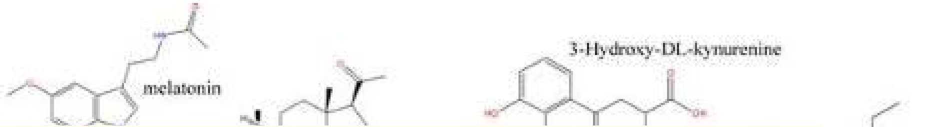
Spain: University of Oviedo

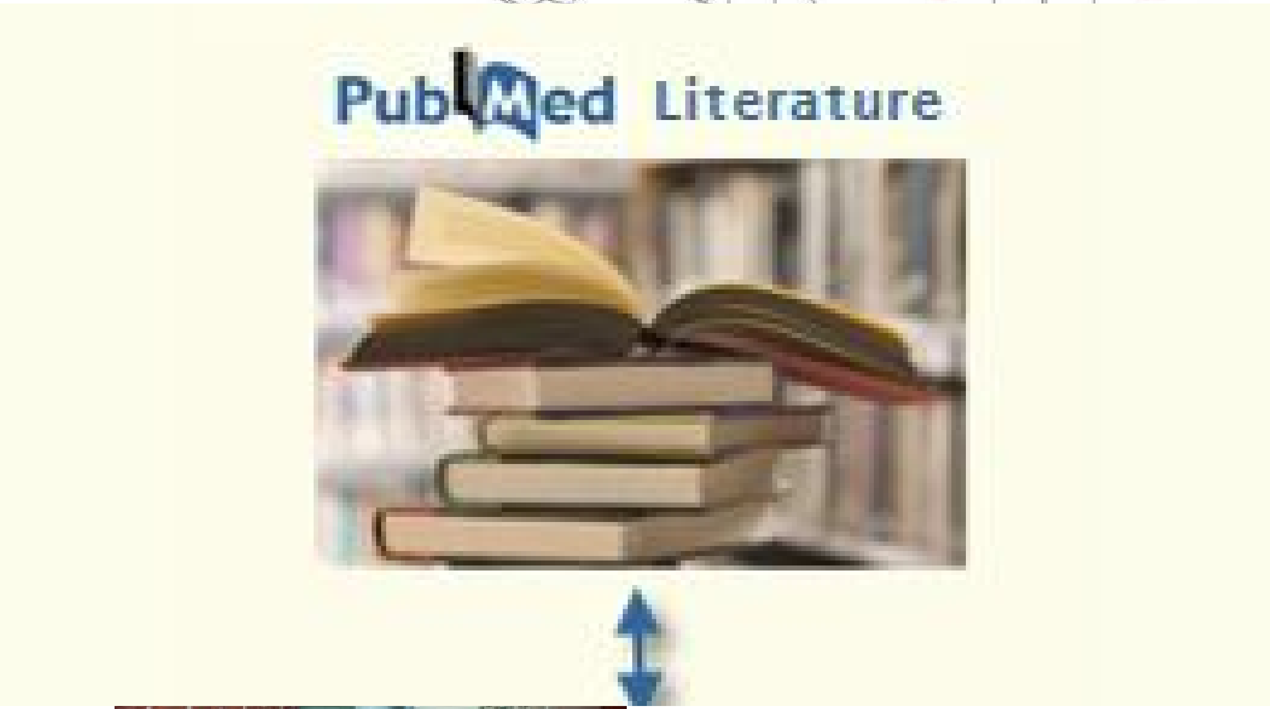
Spain: University of Santiago de Compostela

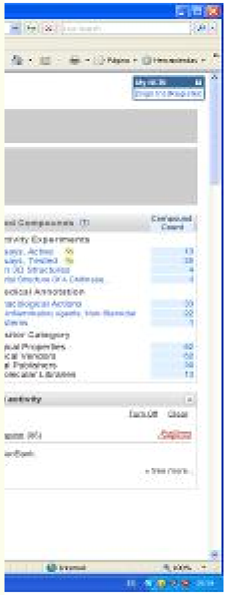
50 cancers

25000 cancer genomes

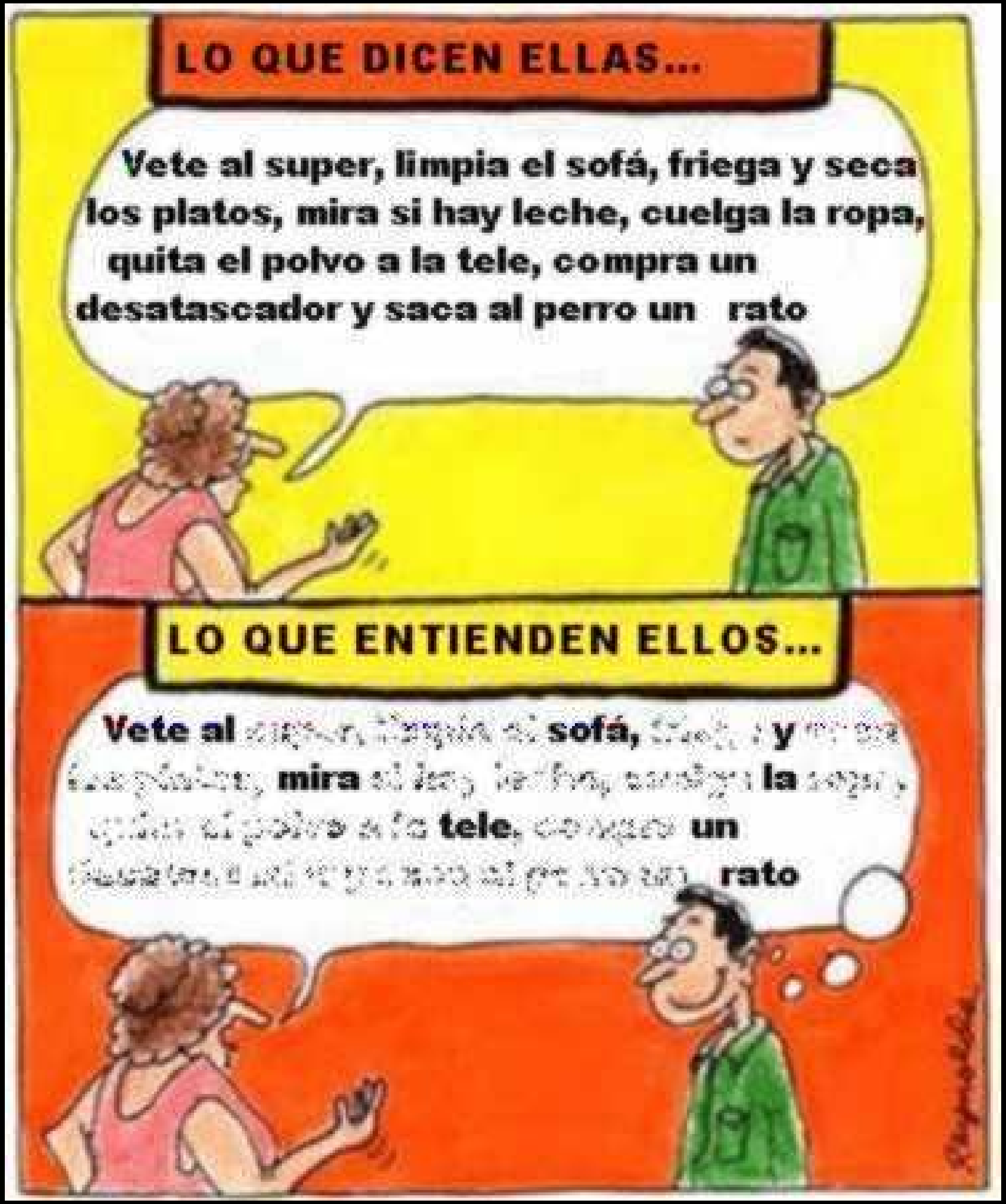
Structures, chemical, literature,...

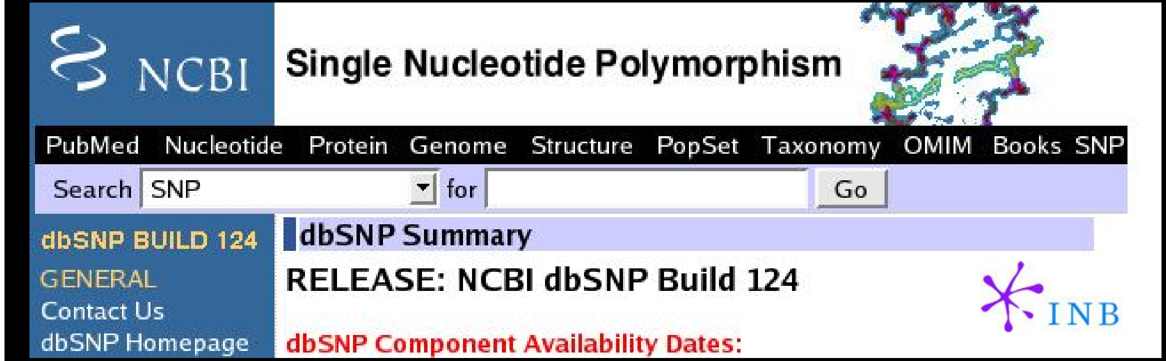






Chemical banks in the order of 10^7 compounds


LO QUE DICEN ELLAS...
Vete al super, limpia el sofá, friega y seca los platos, mira si hay leche, cuelga la ropa, quita el polvo a la tele, compra un desatascador y saca al perro un rato
LO QUE ENTIENDEN ELLOS...
Vete al sofá, y mira la tele, un rato



NCBI
PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books SNP
Search SNP for Go
INB

Single Nucleotide Polymorphism

dbSNP BUILD 124

GENERAL

Contact Us

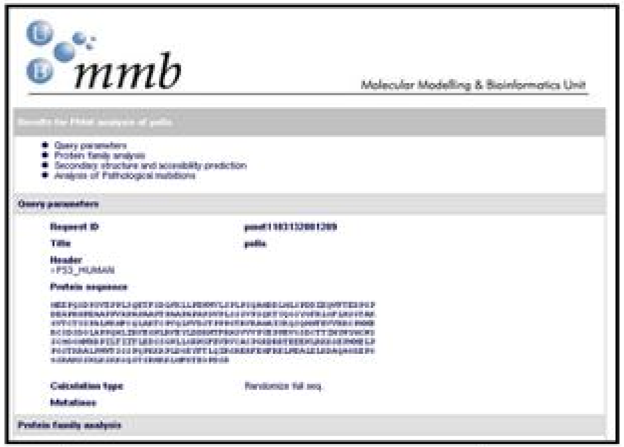
dbSNP Homepage

dbSNP Summary

RELEASE: NCBI dbSNP Build 124

dbSNP Component Availability Dates:

SNPs responsible
90% interspecie variability



- Query parameters
- Protein family analysis
- Secondary structure and accesibility prediction
- Analysis of Pathological mutations

Query parameters

Request ID

Title

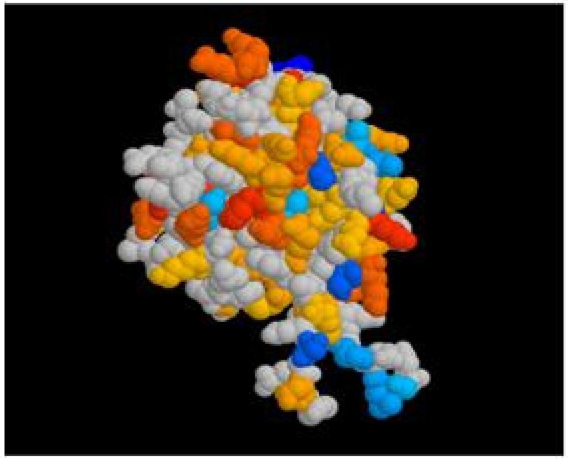
Header
>P53_HUMAN

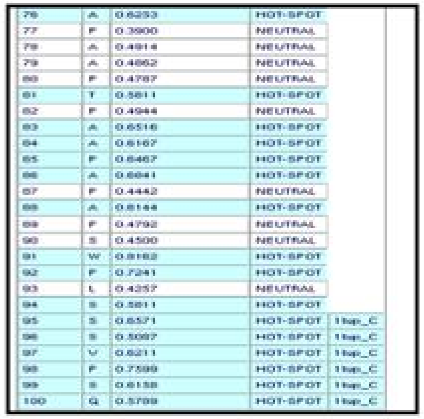
Protein sequence

Calculation type Randomize full seq.

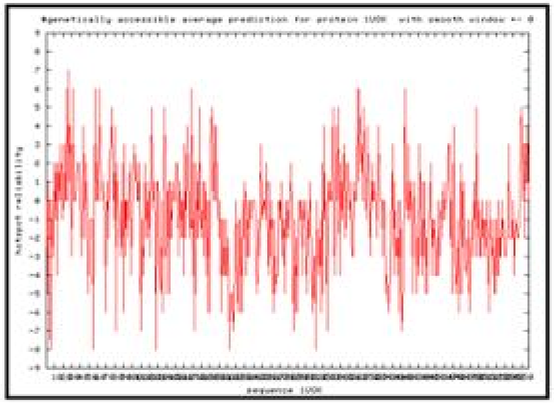
Mutations

Protein family analysis



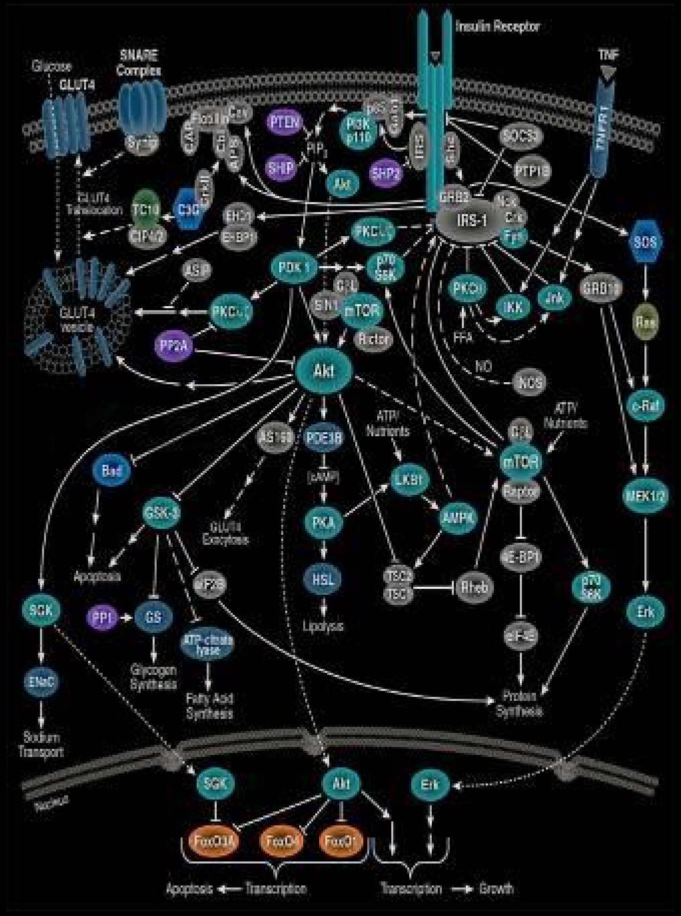


76	A	0.6253	HOT-SPOT	
77	P	0.3900	NEUTRAL	
78	A	0.4914	NEUTRAL	
79	A	0.4862	NEUTRAL	
80	P	0.4787	NEUTRAL	
81	T	0.5811	HOT-SPOT	
82	P	0.4944	NEUTRAL	
83	A	0.6516	HOT-SPOT	
84	A	0.6167	HOT-SPOT	
85	P	0.6467	HOT-SPOT	
86	A	0.6041	HOT-SPOT	
87	P	0.4442	NEUTRAL	
88	A	0.6144	HOT-SPOT	
89	P	0.4792	NEUTRAL	
90	S	0.4500	NEUTRAL	
91	W	0.8162	HOT-SPOT	
92	P	0.7241	HOT-SPOT	
93	L	0.4257	NEUTRAL	
94	S	0.5811	HOT-SPOT	
95	S	0.6571	HOT-SPOT	1tup_C
96	S	0.5087	HOT-SPOT	1tup_C
97	V	0.6211	HOT-SPOT	1tup_C
98	P	0.7599	HOT-SPOT	1tup_C
99	S	0.6158	HOT-SPOT	1tup_C
100	Q	0.5709	HOT-SPOT	1tup_C

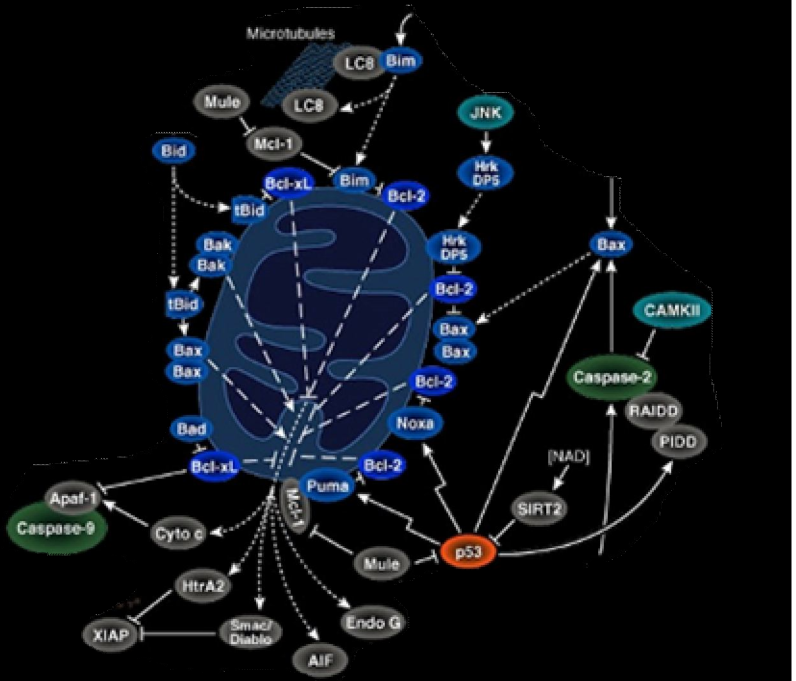


Responsible of
monogenic
pathologies

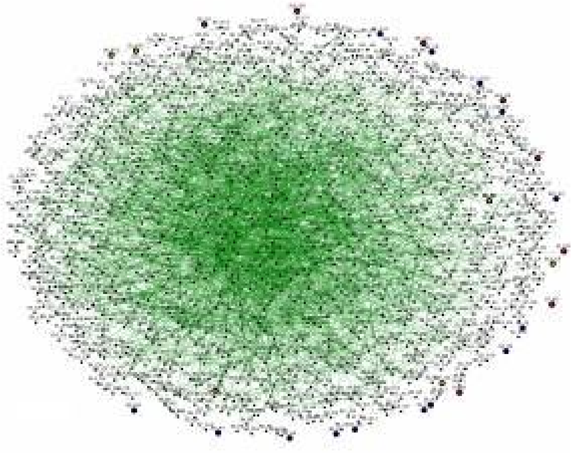
Modeling complex diseases using genomic data



?



Ex. the genomic approach to medicine



- Genome healthy vs pathological
- Trace changes
- Remove noisse
- Correlate changes with pathology

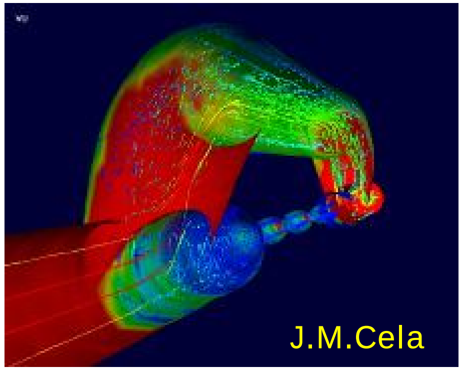
Every step is computationally demanding
the last simply impossible (for current computers)

The computer as a calculator

- Importance of simulation increases with:
  - Increase in data on biological systems
  - Better definition of the problem
- Different types of algorithms
- Must be robust to lack of information
- Often set-up conditions are unclear

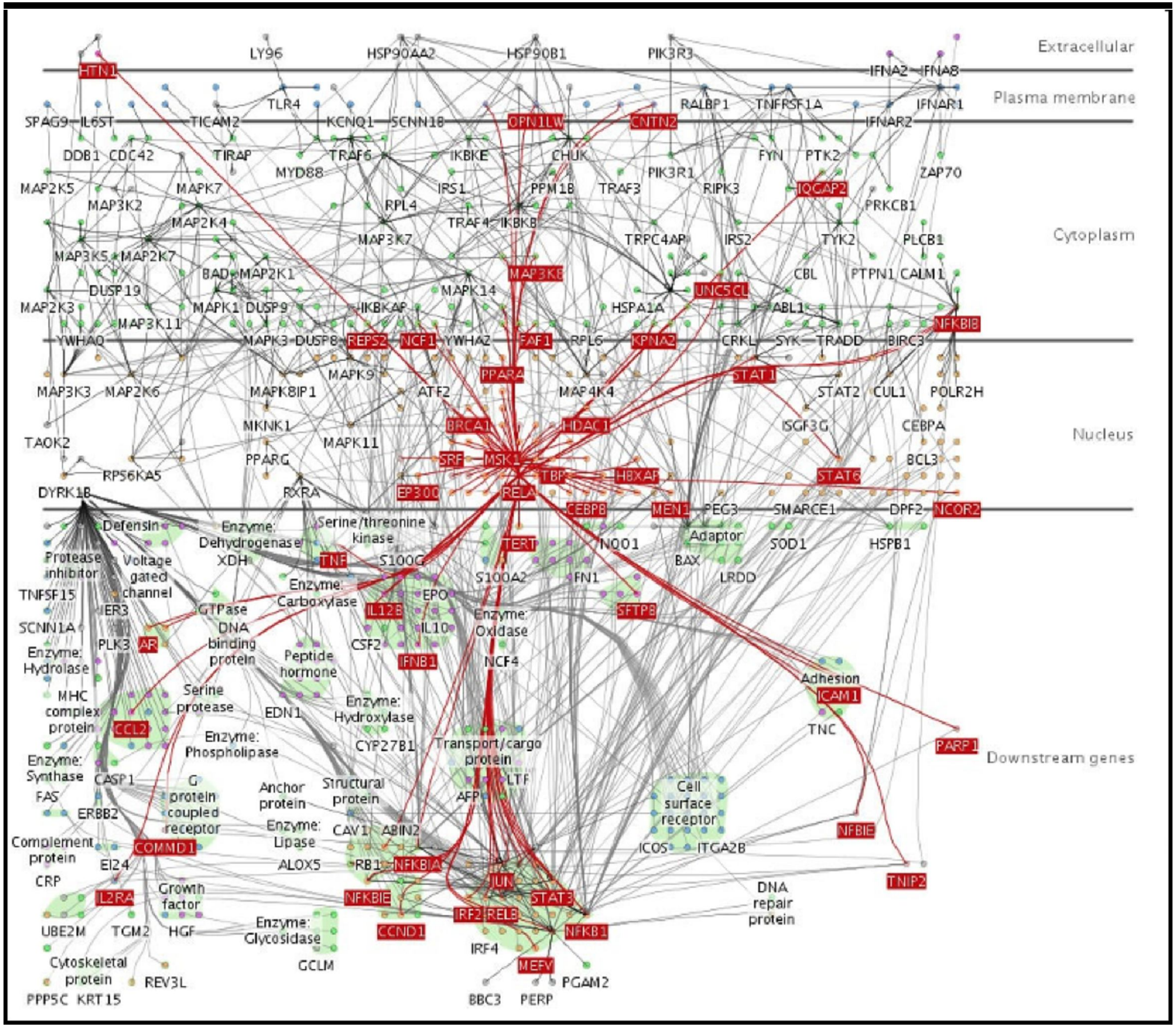
Simulation scenario in Life Sciences

- Ecosystem simulation
- Organ simulation

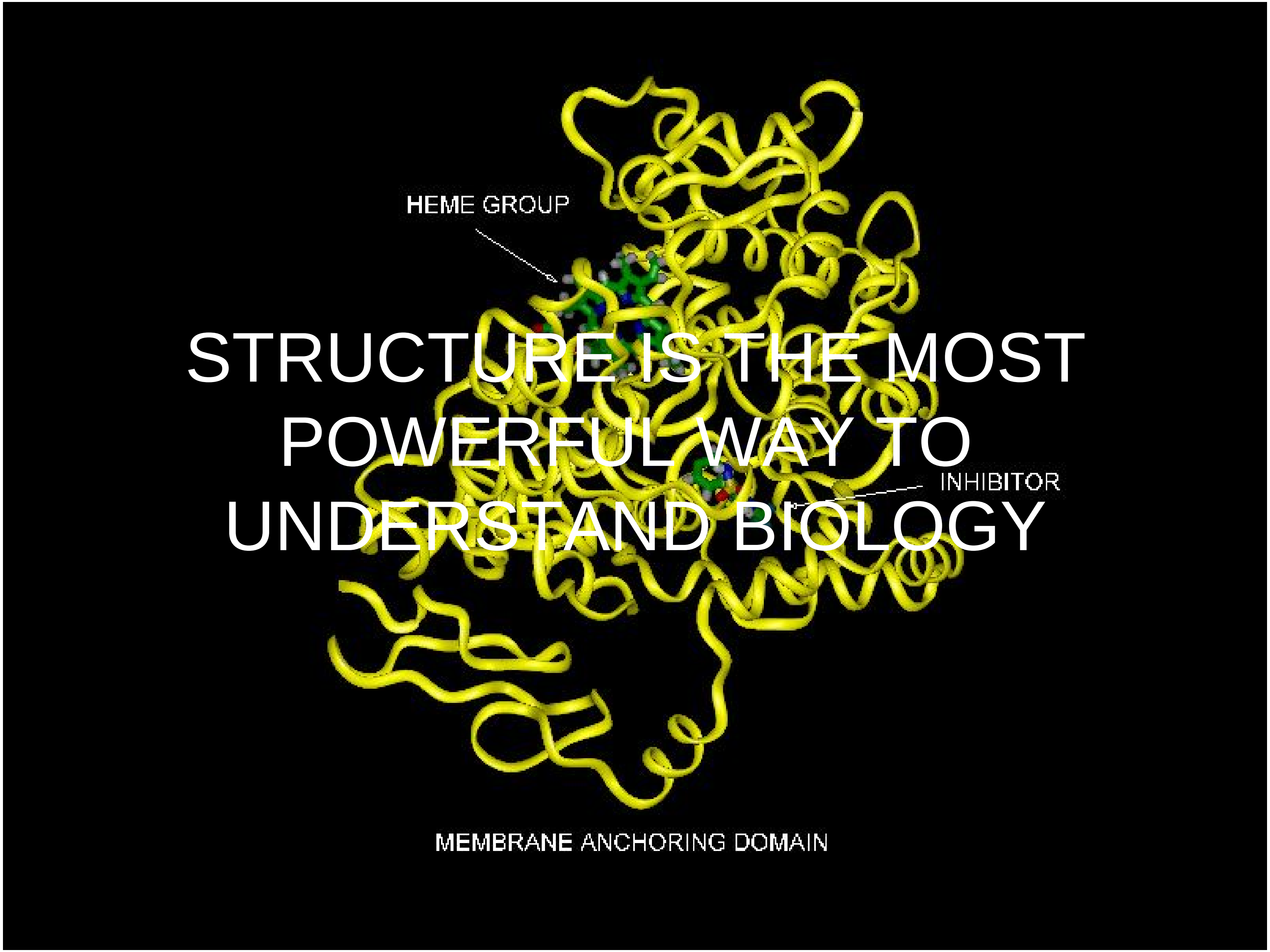


- Gene inter-relations (Bayesian logics)
- Cell simulation (systems biology)
- Molecular simulations
  - Structural prediction
  - Docking
  - Atomistic simulation
  - Cell-scale mesoscopic simulations

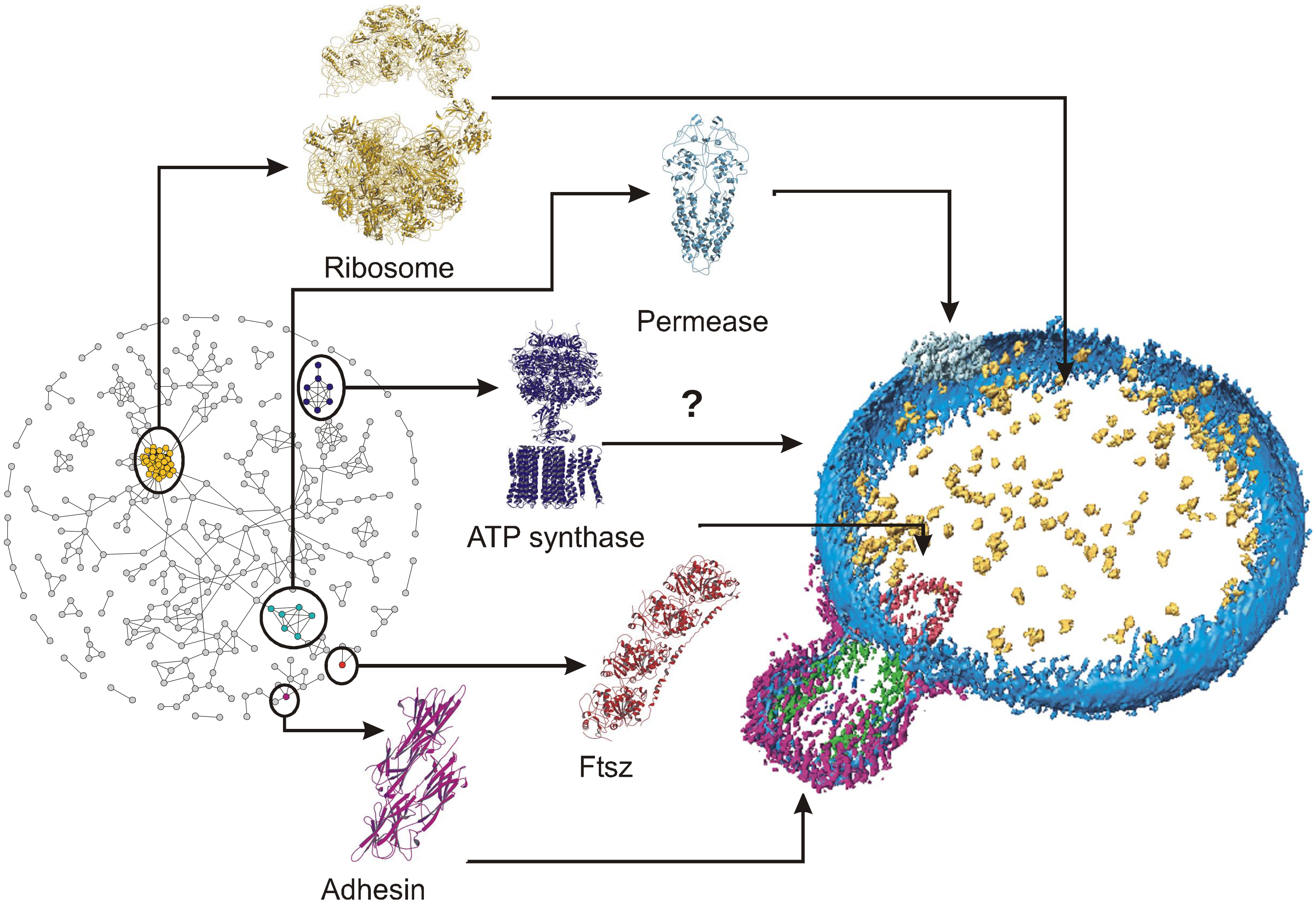
Breast cancer interactome



STRUCTURE IS THE MOST POWERFUL WAY TO UNDERSTAND BIOLOGY

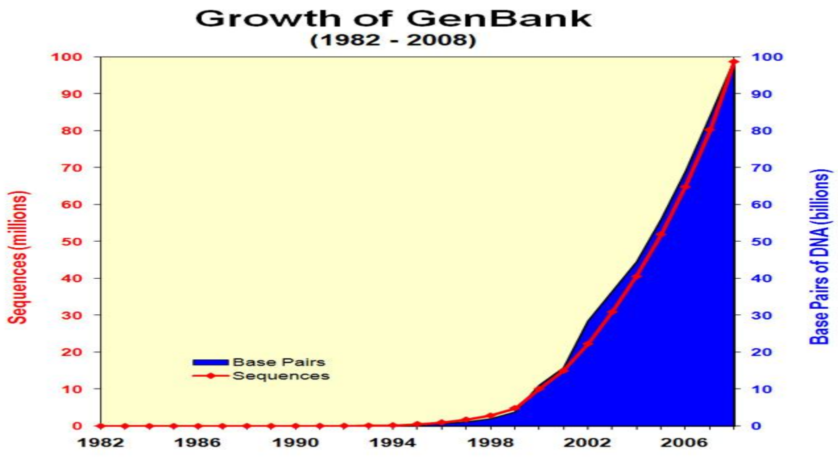


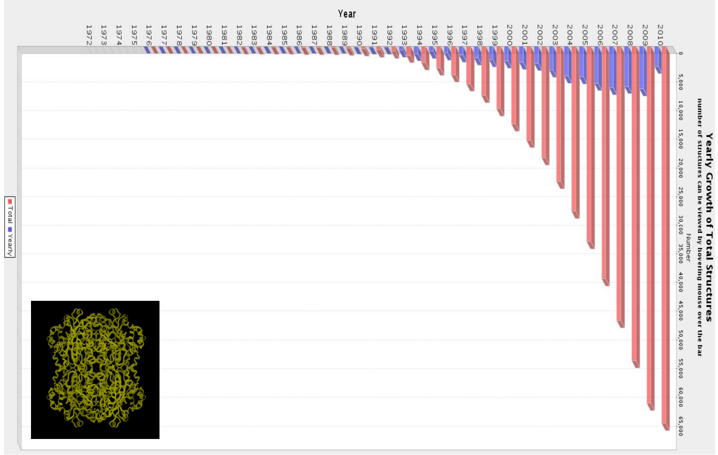
Moving from abstract networks and real cells

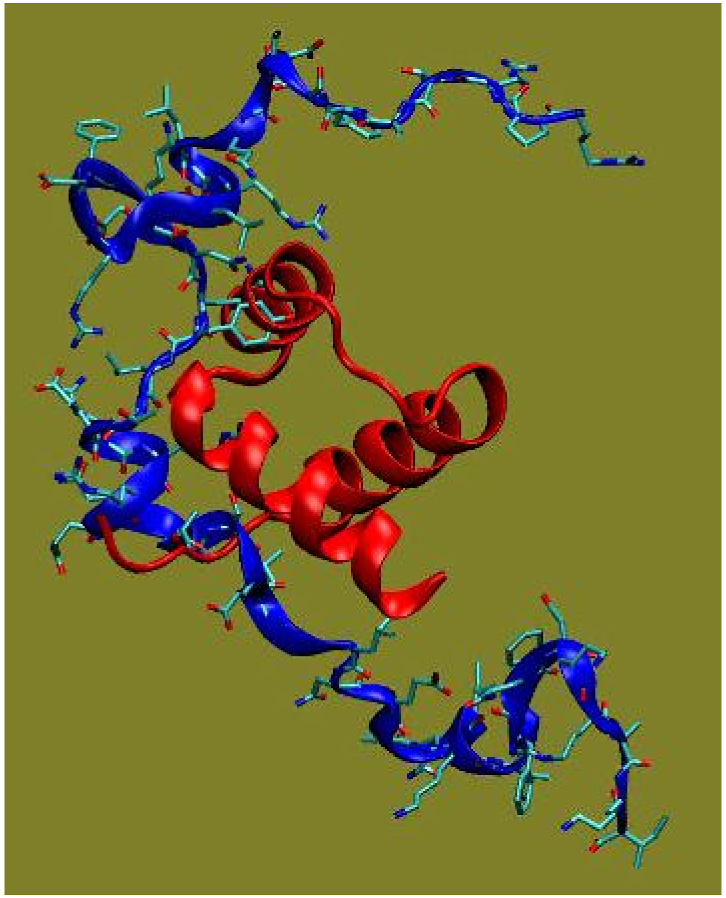


Predicting structure of complexes

Structure prediction







10^8 sequences vs 10^5 structures

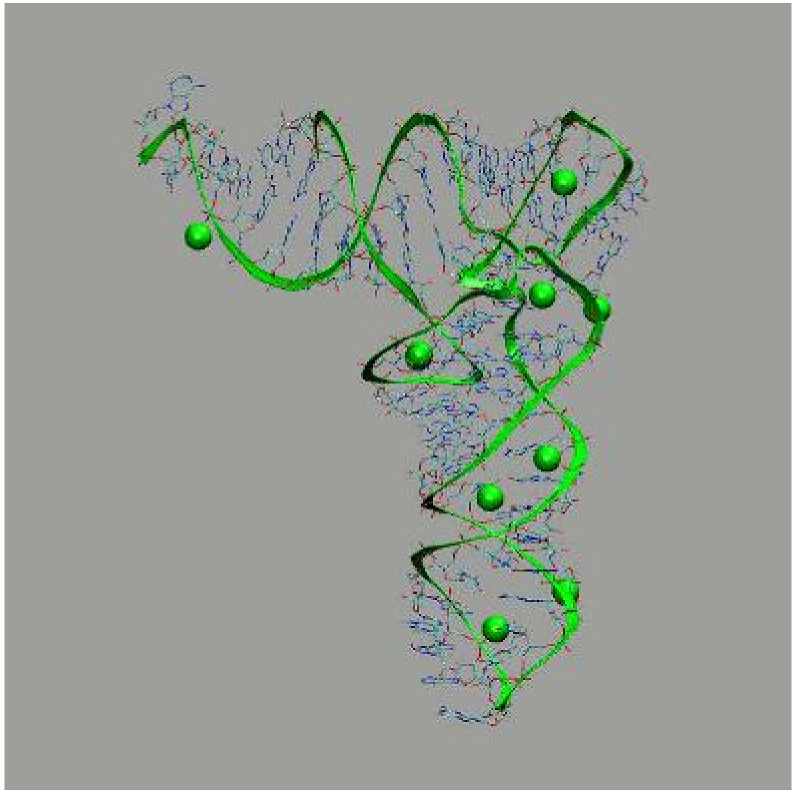
Molecular dynamics

$$\vec{f}_i = -\frac{\partial E_i}{\partial \vec{r}_i}$$

$$\vec{f}_i = m_i \vec{a}_i$$

$$\vec{v}_i = \int \vec{a}_i dt$$

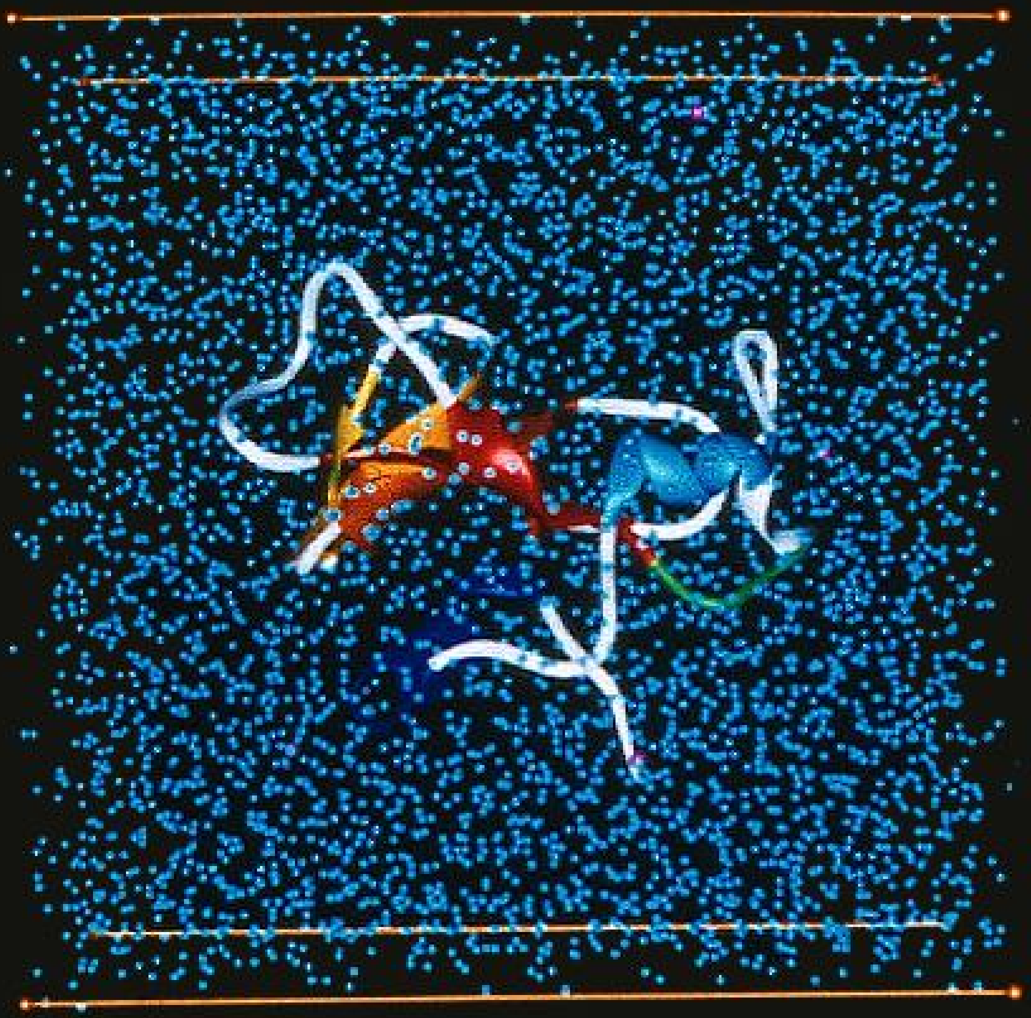
$$\vec{r}_i = \int \vec{v}_i dt$$



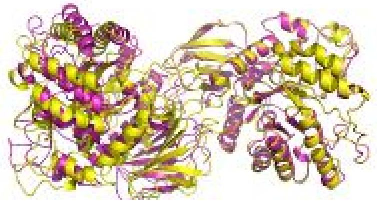
Integration step 1 fts (10^{-15} seg) → 1 mseg = 1 Eur Billion integration steps

Equivalent to follow evolution Nearthental→ H sapiens with photos every sec

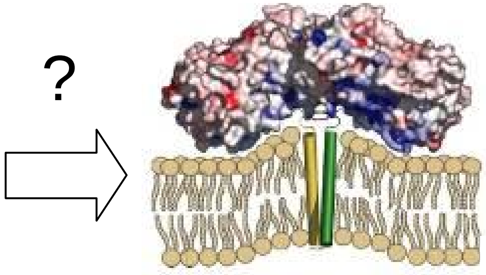
Less than 10000 atoms

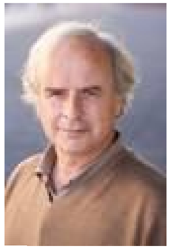


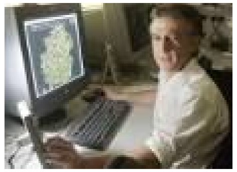
Tirado et al., Biochemistry, 36,7313, 1997

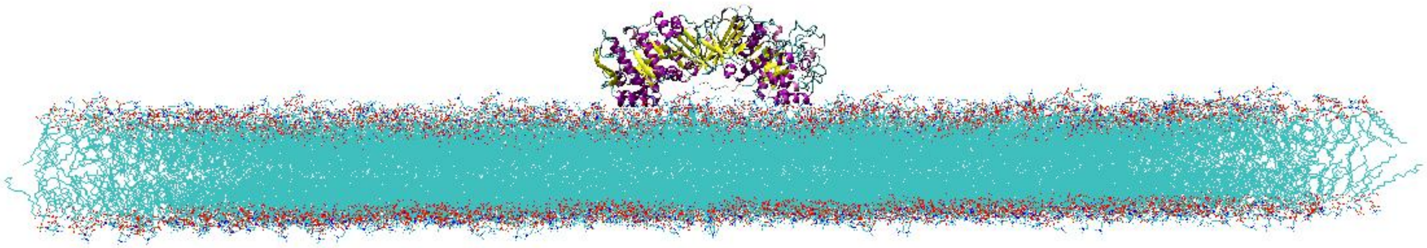


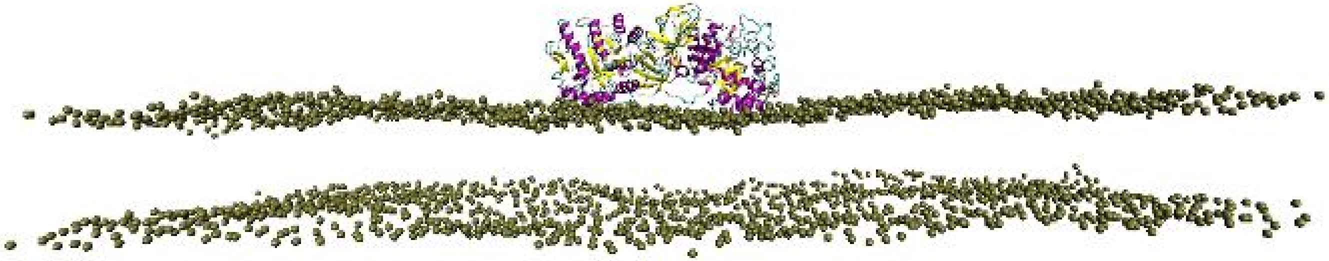
J.Fort, et al JBC 2007

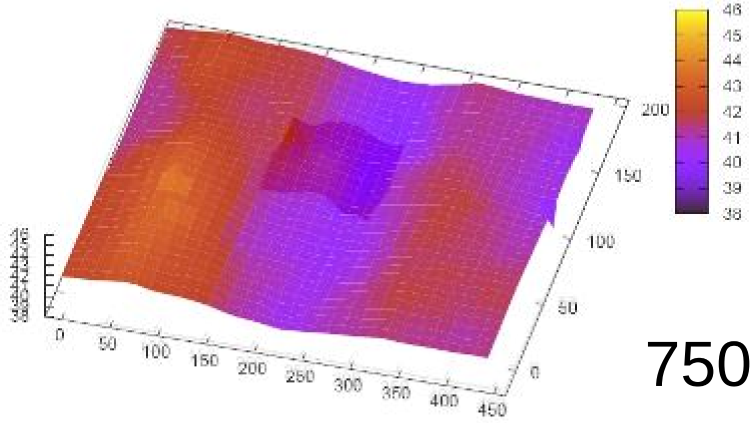




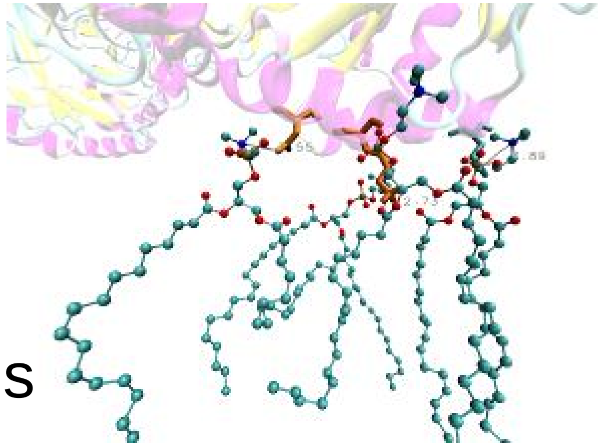


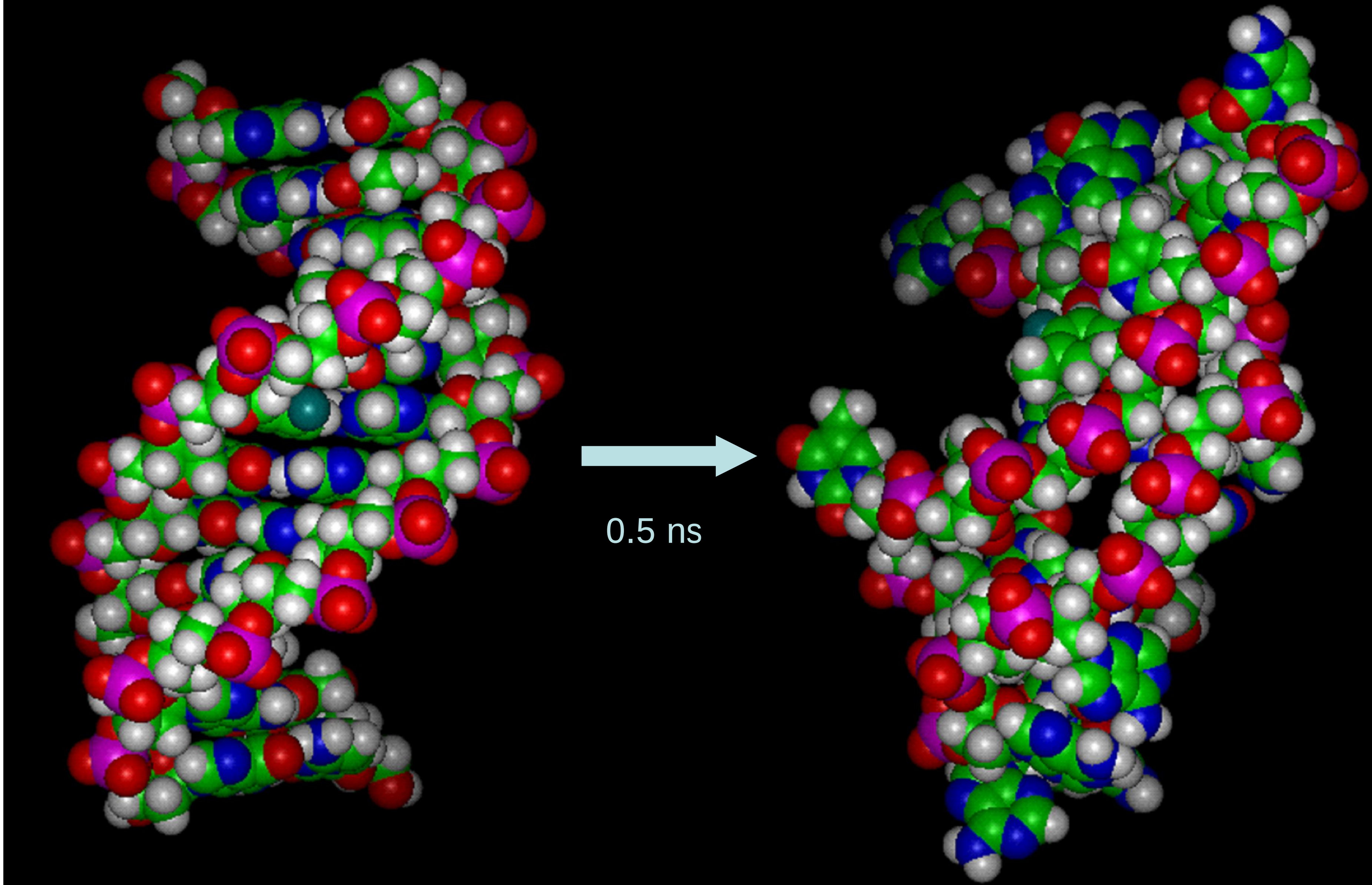




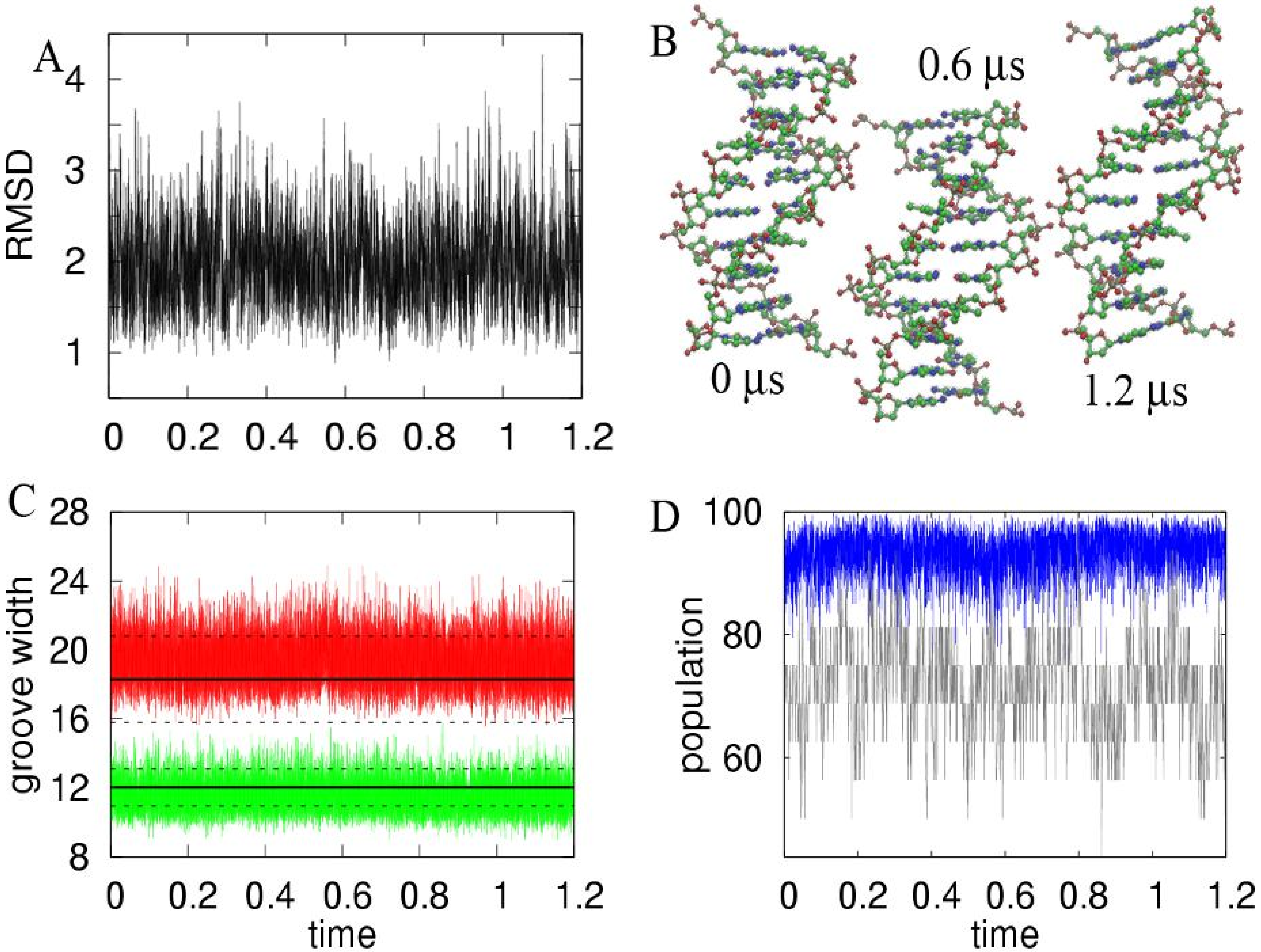


750000 atoms



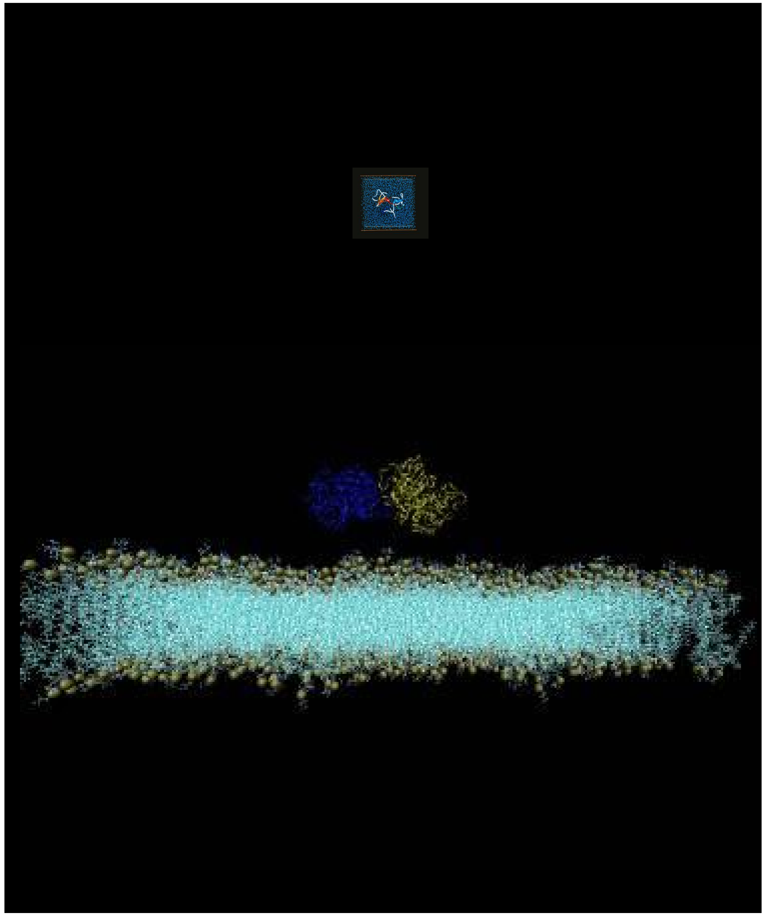


CAL, FJL, MO: J. Phys.Chem 99, 11591-11599 (1995).


A
RMSD
4
3
2
1
0
0.2
0.4
0.6
0.8
1
1.2
B
0.6 μs
0 μs
1.2 μs
C
28
24
20
16
12
8
groove width
0
0.2
0.4
0.6
0.8
1
1.2
time
D
100
80
60
population
0
0.2
0.4
0.6
0.8
1
1.2
time

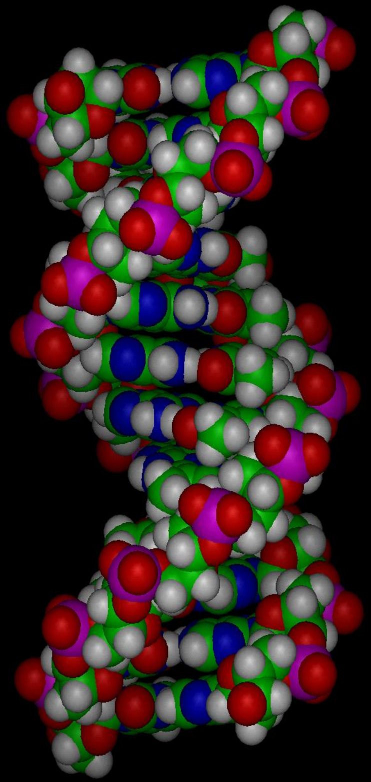
Current limitations in MD

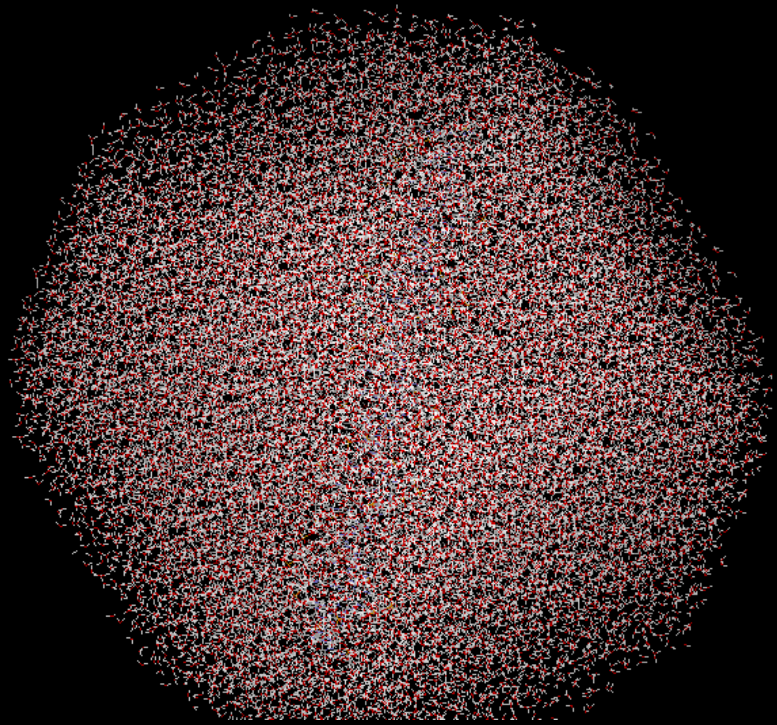
- Size of the system
  - Typically: 10^4 - 10^5 particles
  - Flagship: 10^6
- Simulation length
  - Typically: 10^1 – 10^2 ns
  - Using HPC: μs
  - For 10^4 particles



Longer and bigger,...

Data scales as 3N*T
N= number of atoms
T= time length
Ex. Scale up $3x10^3$



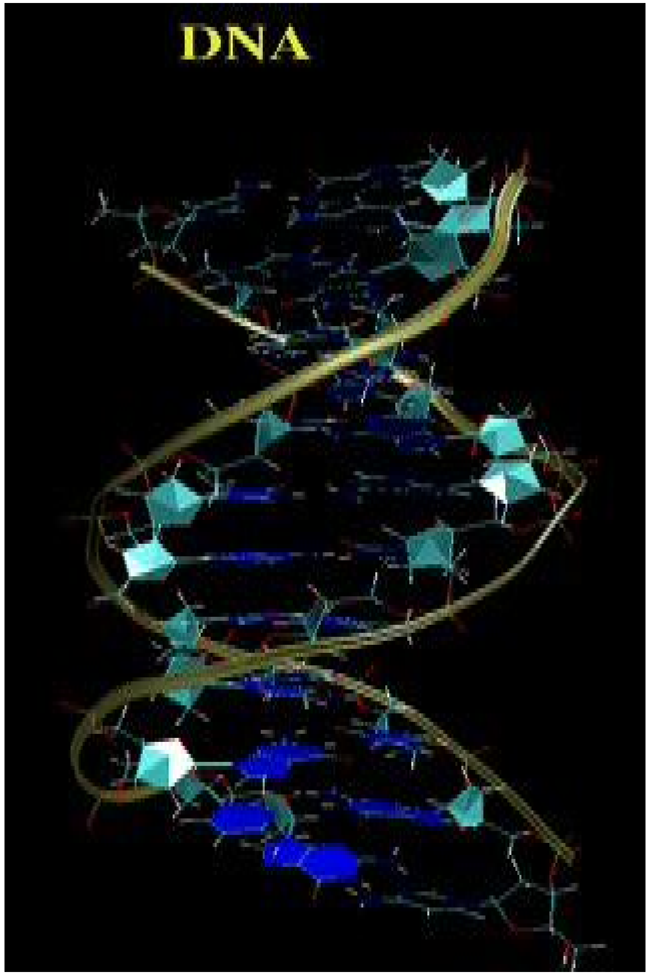


10 ns

Individual trajectories 0.1 Tb

1000 ns

Compression tools (PCAZIP)



Diagonalization of covariance matrix

Eigenvectors

Select essential space

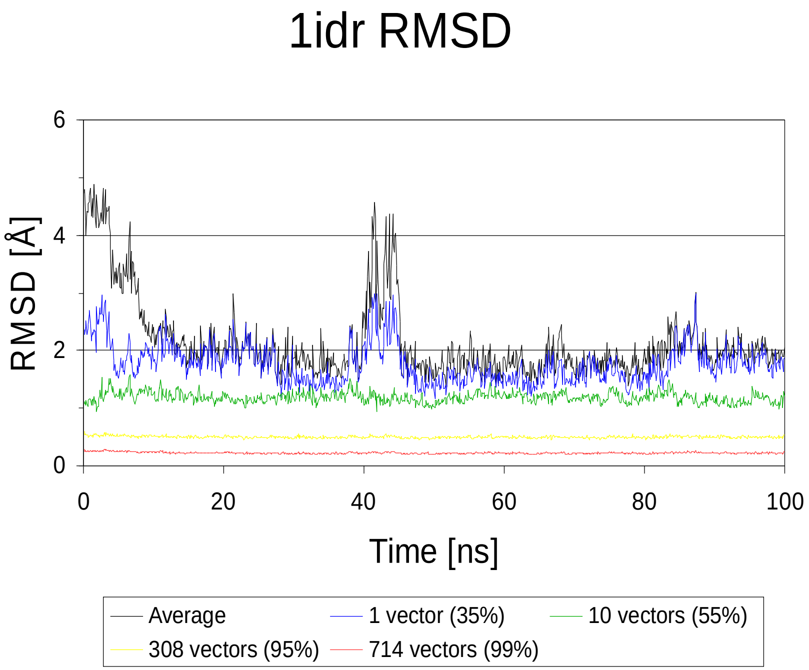
Quality threshold

Project cart coord. Essential space

Store projections

Col. C.Laughton

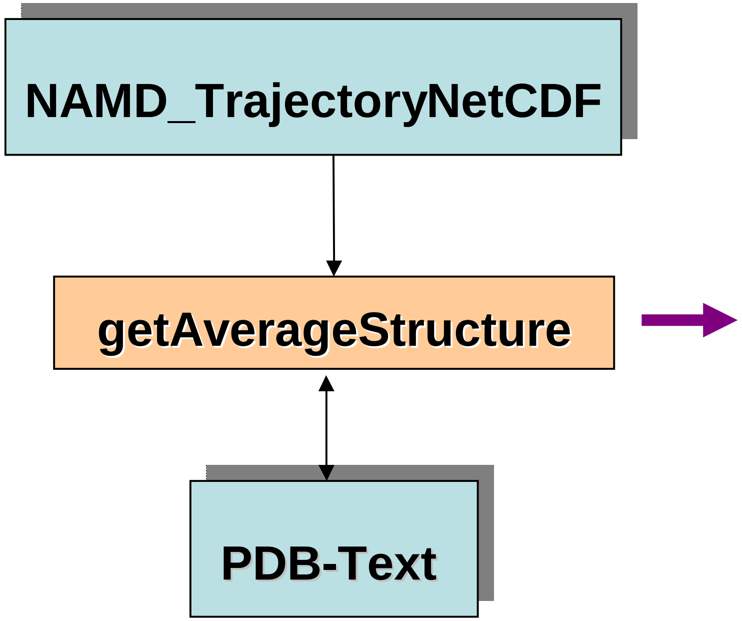
PCAZIP data reduction proteins



	95% cutoff	
Protein	RMSd	File size
1ark	0.36	8.5
1cei	0.36	7.8
1sr0	0.45	6.0
2gb1	0.29	10.0
3ci2	0.36	8.6
2icb	0.33	8.8
1idr	0.50	5.1

Web-based MD Trajectory Analysis Toolkit

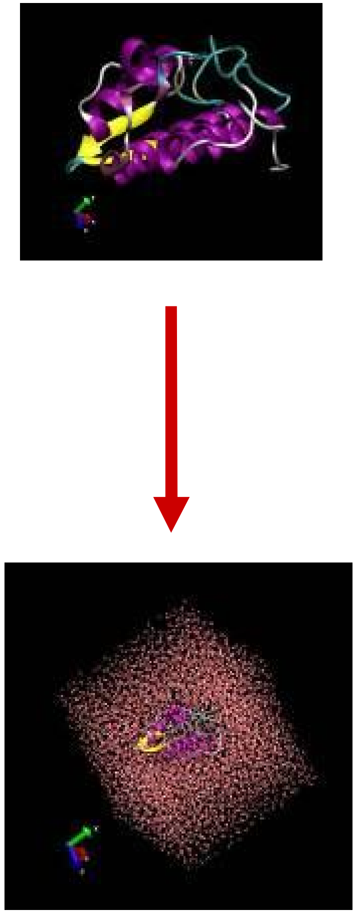
- From Reference Format NetCDF.

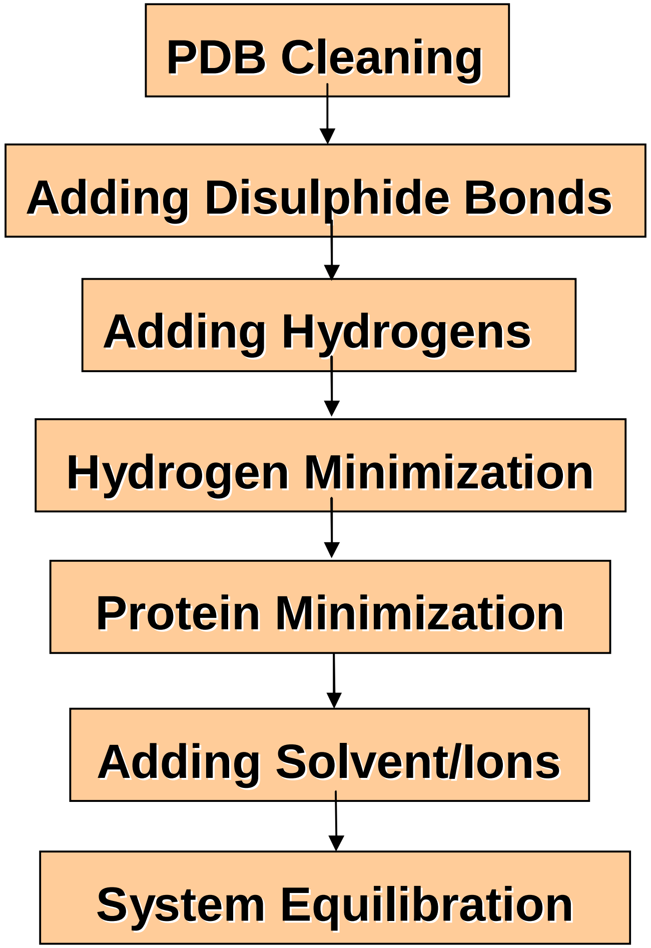


- ***Bfactor***
- ***Bfactor Per Residue***
- ***Rms***
- ***Rms Per Residue***
- ***Superposition***
- ***Average Structure***
- ***Radius of Gyration***
- ***SASA***
- ***GaussianRmsd***
- ***Esential dynamics***
- ***etc.***

Molecular Dynamics Workflow

MDweb





PDB Cleaning

Adding Disulphide Bonds

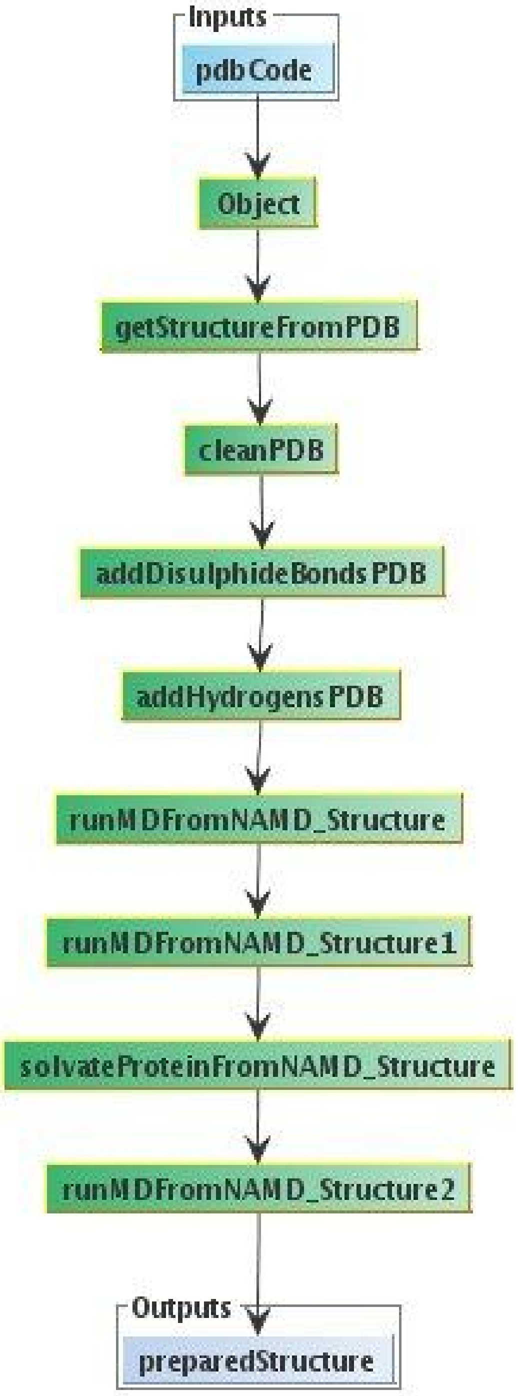
Adding Hydrogens

Hydrogen Minimization

Protein Minimization

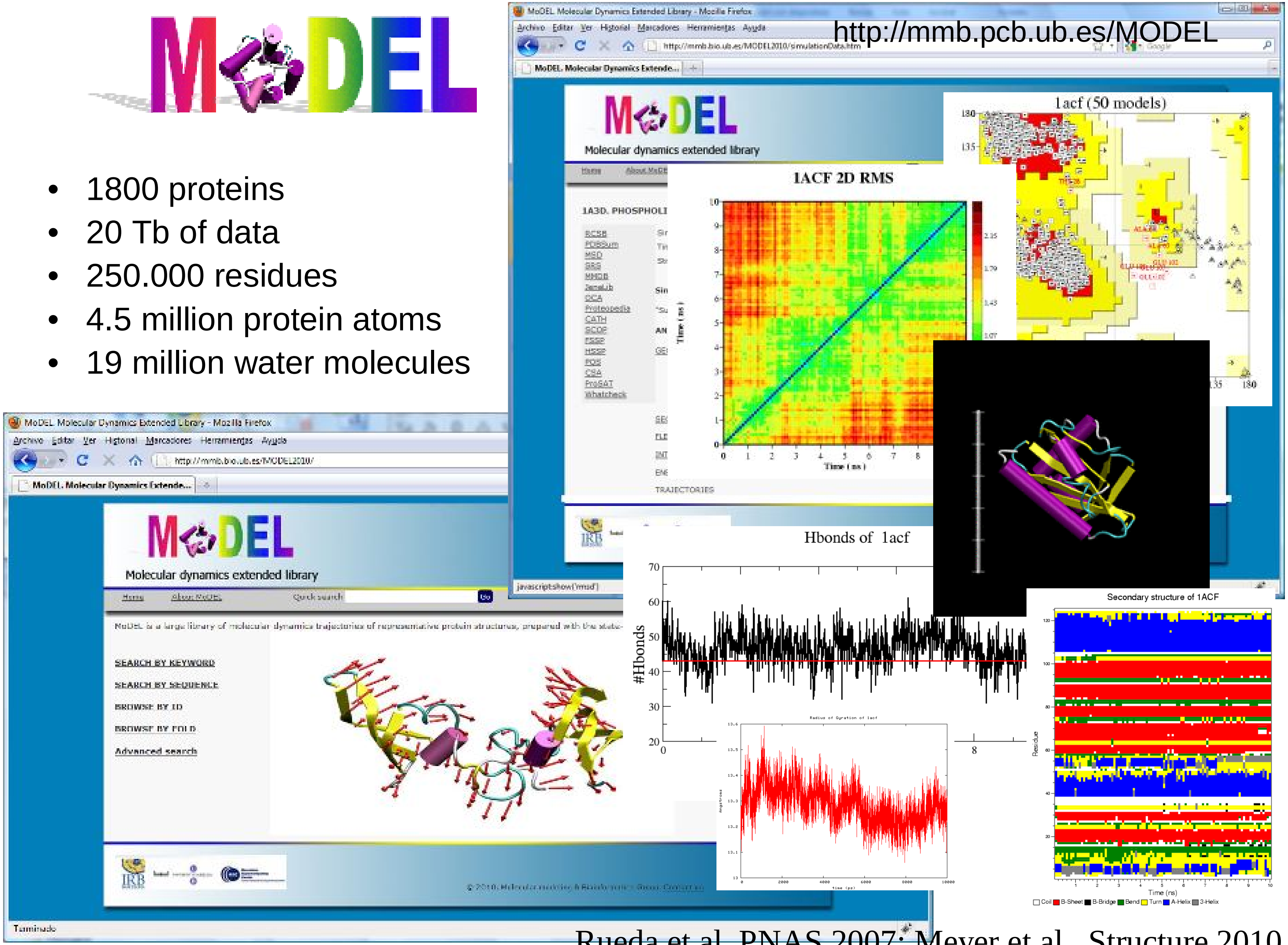
Adding Solvent/Ions

System Equilibration



MoDEL

- 1800 proteins
- 20 Tb of data
- 250.000 residues
- 4.5 million protein atoms
- 19 million water molecules

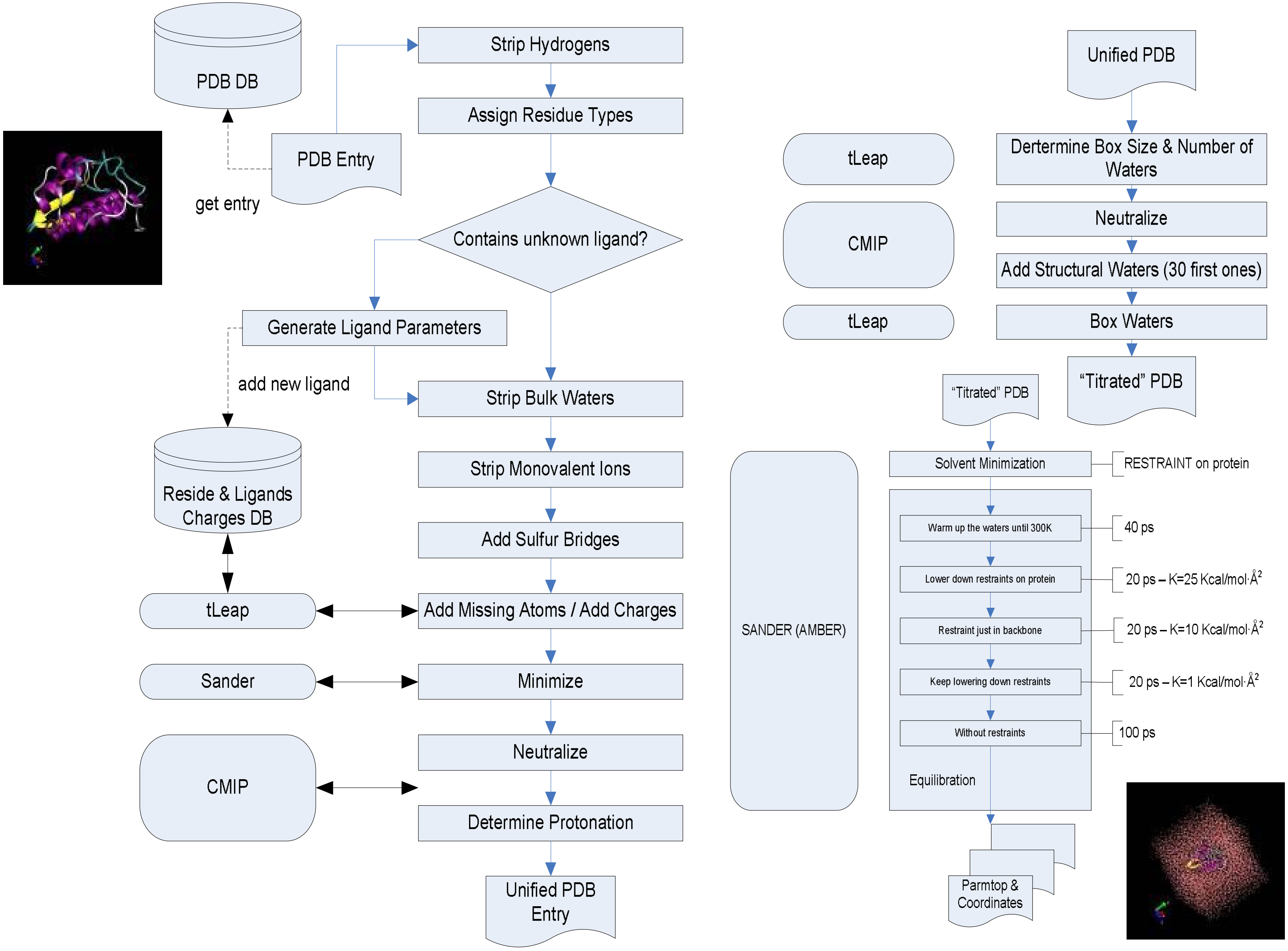


MoDEL Molecular Dynamics Extended Library - Mozilla Firefox
Archivo Editar Ver Historial Marcadores Herramientas Ayuda
http://mmb.bio.ub.es/MODEL2010/simulationData.htm
MoDEL Molecular Dynamics Extende...
Molecular dynamics extended library
Home
About MoDEL
1A3D. PHOSPHOLI
RCSB
PDBSum
MSD
SRS
MMDB
JenaLib
OCA
Proteopedia
CATH
SCOP
FSSP
HSSP
PQS
CSA
ProSAT
Whatcheck
TRAJECTORIES
javascript:show('rmsd')
1ACF 2D RMS
Time (ns)
1acf (50 models)
Hbonds of 1acf
#Hbonds
Radius of Gyration of 1acf
Secondary structure of 1ACF
Residue
Time (ns)
□ Coil 🟥 B-Sheet ■ B-Bridge 🟩 Bend 🟨 Turn 🟦 A-Helix ▨ 3-Helix
http://mmb.bio.ub.es/MODEL2010/
Quick search
SEARCH BY KEYWORD
SEARCH BY SEQUENCE
BROWSE BY ID
BROWSE BY FOLD
Advanced search
Terminado

MoDEL is a large library of molecular dynamics trajectories of representative protein structures, prepared with the state-

http://mmb.pcb.ub.es/MODEL

Rueda et al. PNAS 2007; Meyer et al., Structure 2010



PDB DB

get entry

PDB Entry

Strip Hydrogens

Assign Residue Types

Contains unknown ligand?

Generate Ligand Parameters

add new ligand

Reside & Ligands Charges DB

Strip Bulk Waters

Strip Monovalent Ions

Add Sulfur Bridges

tLeap

Add Missing Atoms / Add Charges

Sander

Minimize

Neutralize

CMIP

Determine Protonation

Unified PDB Entry

Unified PDB

tLeap

Dertermine Box Size & Number of Waters

CMIP

Neutralize

Add Structural Waters (30 first ones)

tLeap

Box Waters

"Titrated" PDB

"Titrated" PDB

SANDER (AMBER)

Solvent Minimization

RESTRAINT on protein

Warm up the waters until 300K

40 ps

Lower down restraints on protein

20 ps – K=25 Kcal/mol·Å^2

Restraint just in backbone

20 ps – K=10 Kcal/mol·Å^2

Keep lowering down restraints

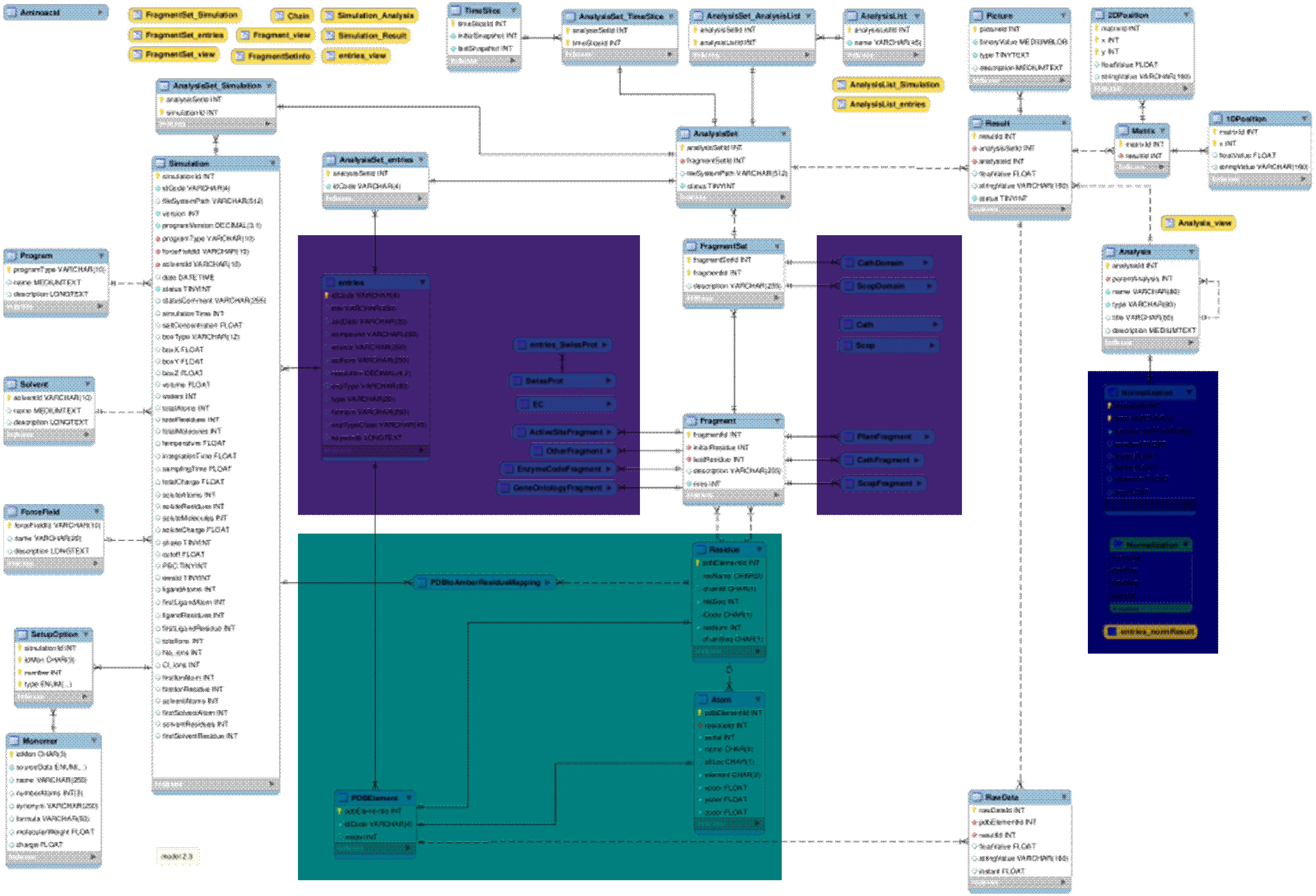
20 ps – K=1 Kcal/mol·Å^2

Without restraints

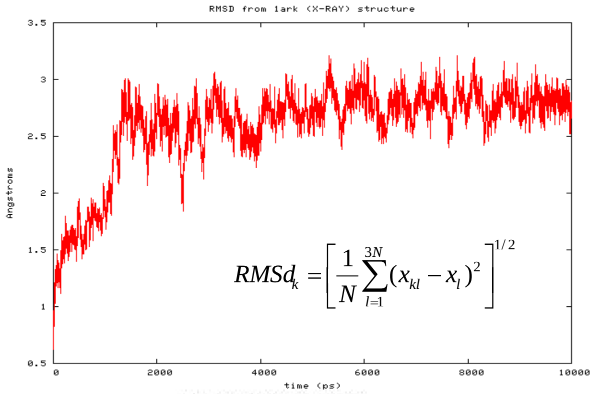
100 ps

Equilibration

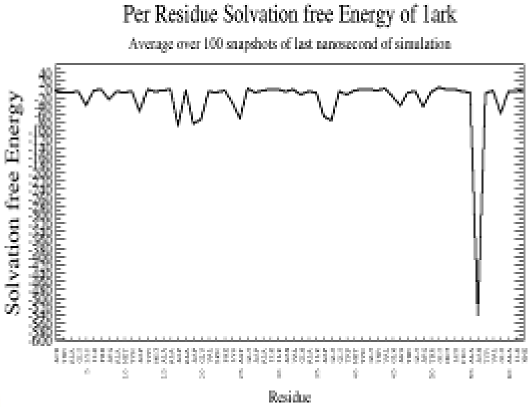
Parmtop & Coordinates

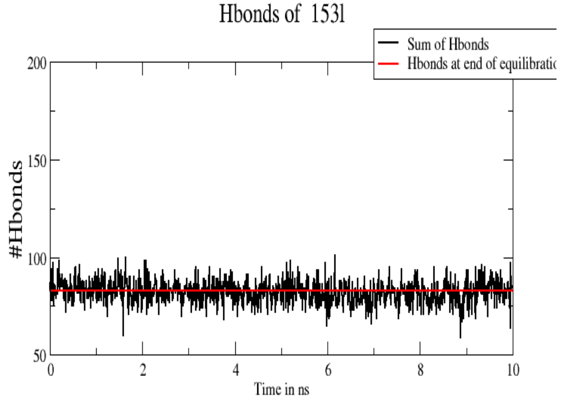


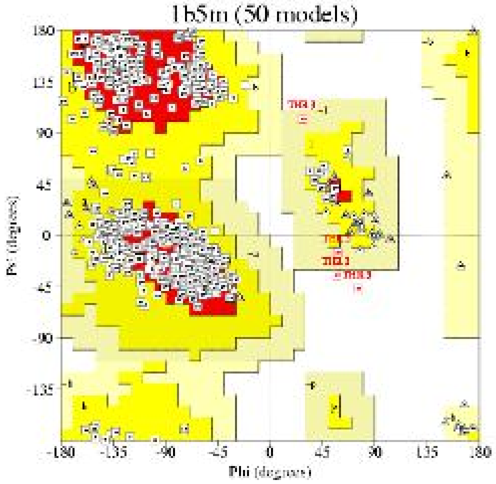
AminoAcid
FragmentSet_Simulation
Chain
Simulation_Analysis
FragmentSet_entries
Fragment_view
Simulation_Result
FragmentSet_view
FragmentSetInfo
entries_view
TimeSlice
timeSliceId INT
initialSnapshot INT
lastSnapshot INT
AnalysisSet_TimeSlice
analysisSetId INT
timeSliceId INT
AnalysisSet_AnalysisList
analysisSetId INT
analysisListId INT
AnalysisList
analysisListId INT
name VARCHAR(45)
AnalysisList_Simulation
AnalysisList_entries
Picture
pictureId INT
binaryValue MEDIUMBLOB
type TINYTEXT
description MEDIUMTEXT
2DPosition
matrixId INT
x INT
y INT
floatValue FLOAT
stringValue VARCHAR(150)
AnalysisSet_Simulation
analysisSetId INT
simulationId INT
AnalysisSet
analysisSetId INT
fragmentSetId INT
fileSystemPath VARCHAR(512)
status TINYINT
Result
resultId INT
analysisSetId INT
analysisId INT
floatValue FLOAT
stringValue VARCHAR(150)
status TINYINT
Matrix
matrixId INT
resultId INT
1DPosition
matrixId INT
x INT
floatValue FLOAT
stringValue VARCHAR(150)
Analysis_view
Simulation
simulationId INT
idCode VARCHAR(4)
fileSystemPath VARCHAR(512)
version INT
programVersion DECIMAL(3,1)
programType VARCHAR(10)
forceFieldId VARCHAR(10)
solventId VARCHAR(10)
date DATETIME
status TINYINT
statusComment VARCHAR(256)
simulationTime INT
saltConcentration FLOAT
boxType VARCHAR(12)
boxX FLOAT
boxY FLOAT
boxZ FLOAT
volume FLOAT
waters INT
totalAtoms INT
totalResidues INT
totalMolecules INT
temperature FLOAT
integrationTime FLOAT
samplingTime FLOAT
totalCharge FLOAT
soluteAtoms INT
soluteResidues INT
soluteMolecules INT
soluteCharge FLOAT
pbsep TINYINT
cutoff FLOAT
PBC TINYINT
ewald TINYINT
ligandAtoms INT
firstLigandAtom INT
ligandResidues INT
firstLigandResidue INT
solvAtoms INT
Na_ions INT
Cl_ions INT
firstIonAtom INT
firstIonResidue INT
solventAtoms INT
firstSolventAtom INT
solventResidues INT
firstSolventResidue INT
AnalysisSet_entries
analysisSetId INT
idCode VARCHAR(4)
FragmentSet
fragmentSetId INT
fragmentId INT
description VARCHAR(255)
CathDomain
ScopDomain
Cath
Scop
Program
programType VARCHAR(10)
name MEDIUMTEXT
description LONGTEXT
entries
idCode VARCHAR(4)
title VARCHAR(252)
accDate VARCHAR(20)
compound VARCHAR(252)
source VARCHAR(250)
authors VARCHAR(250)
resolution DECIMAL(4,2)
expType VARCHAR(60)
type VARCHAR(20)
formula VARCHAR(250)
expTypeClass VARCHAR(45)
keywords LONGTEXT
entries_SwissProt
SwissProt
EC
Analysis
analysisId INT
parentAnalysis INT
name VARCHAR(45)
type VARCHAR(45)
title VARCHAR(45)
description MEDIUMTEXT
Solvent
solventId VARCHAR(10)
name MEDIUMTEXT
description LONGTEXT
ActiveSiteFragment
OtherFragment
EnzymeCodeFragment
GeneOntologyFragment
Fragment
fragmentId INT
initialResidue INT
lastResidue INT
description VARCHAR(255)
rmsd INT
PlainFragment
CathFragment
ScopFragment
Normalization
Normalization
entries_normResult
ForceField
forceFieldId VARCHAR(10)
name VARCHAR(20)
description LONGTEXT
Residue
pdbElementId INT
resName CHAR(3)
chainId CHAR(1)
resSeq INT
iCode CHAR(1)
residueId INT
chainSeq CHAR(1)
PDBtoAmberResidueMapping
SetupOption
simulationId INT
idMon CHAR(3)
number INT
type ENUM(...)
Atom
pdbElementId INT
residueId INT
serial INT
name CHAR(4)
altLoc CHAR(1)
element CHAR(2)
xcoor FLOAT
ycoor FLOAT
zcoor FLOAT
Monomer
idMon CHAR(3)
sourceData ENUM(...)
name VARCHAR(250)
numberAtoms INT(3)
synonym VARCHAR(250)
formula VARCHAR(50)
molecularWeight FLOAT
charge FLOAT
PDBElement
pdbElementId INT
idCode VARCHAR(4)
modelId INT
RawData
rawDataId INT
pdbElementId INT
resultId INT
floatValue FLOAT
stringValue VARCHAR(150)
instant FLOAT
model2.3

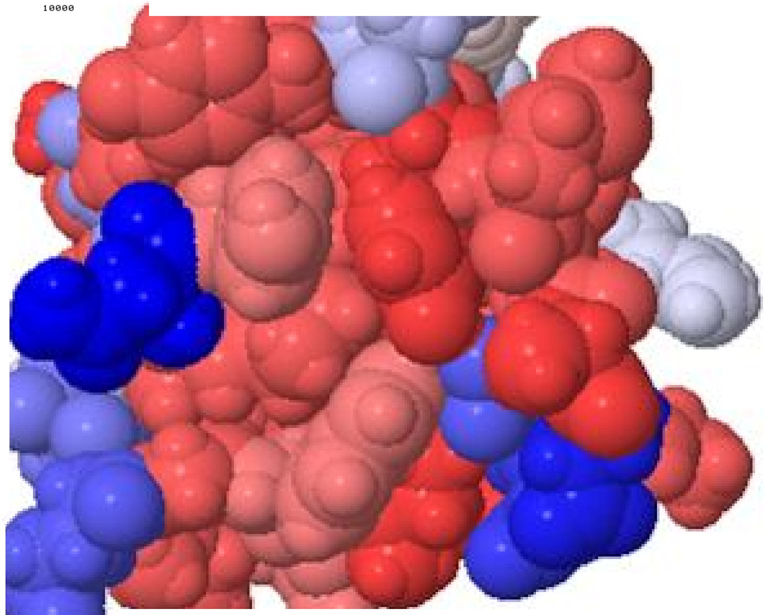


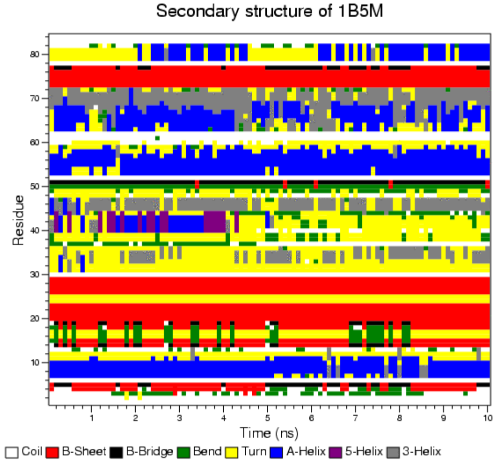
$$RMSd_k = \left[\frac{1}{N}\sum_{l=1}^{3N}(x_{kl}-x_l)^2\right]^{1/2}$$

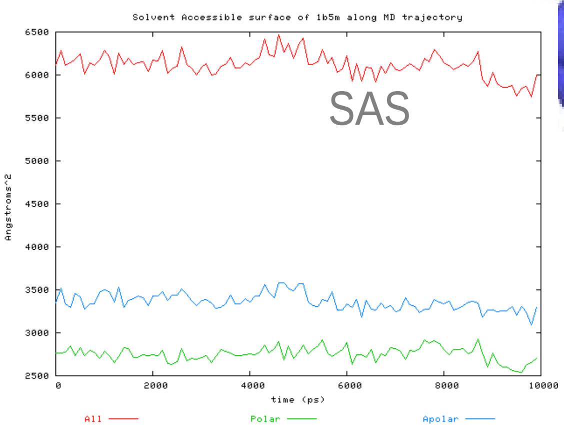


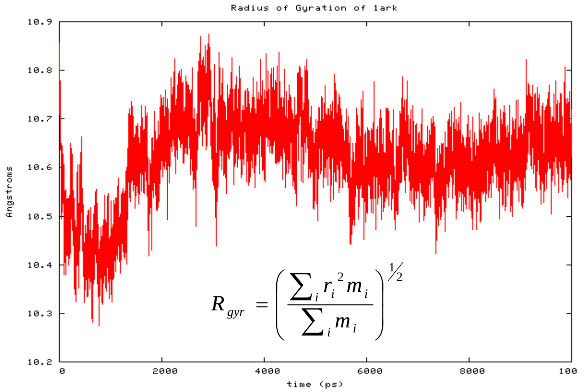




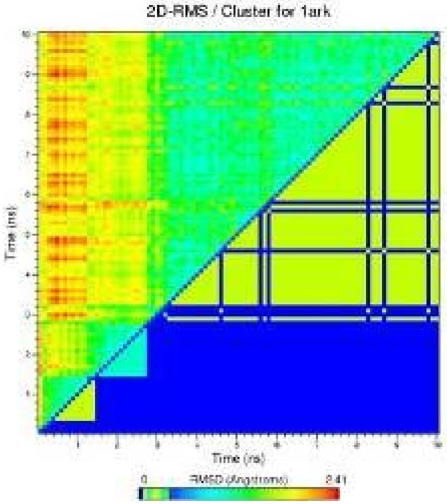


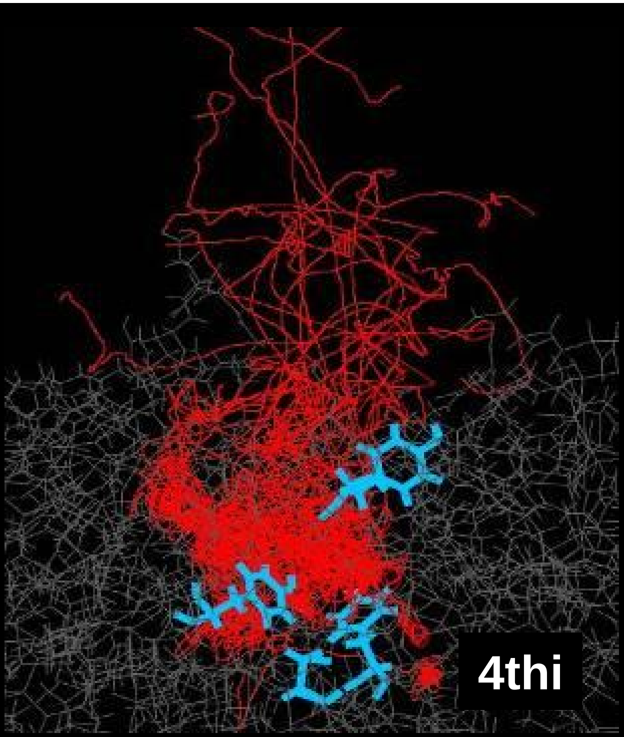




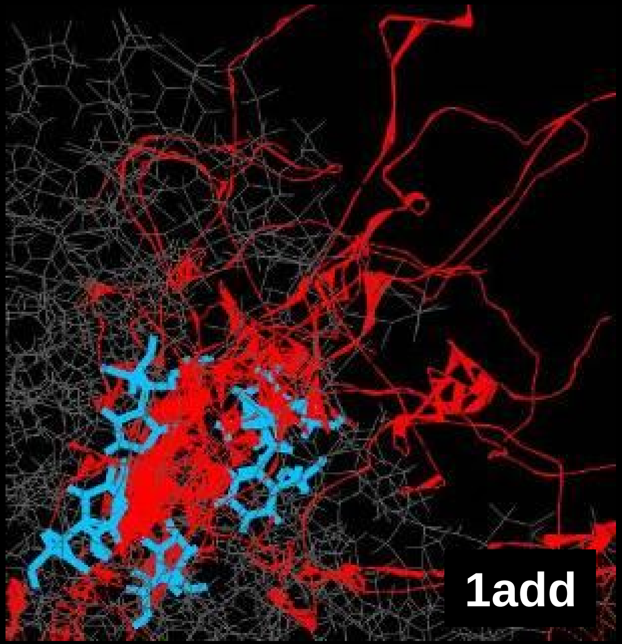


$$R_{gyr} = \left(\frac{\sum_i r_i^2 m_i}{\sum_i m_i}\right)^{1/2}$$

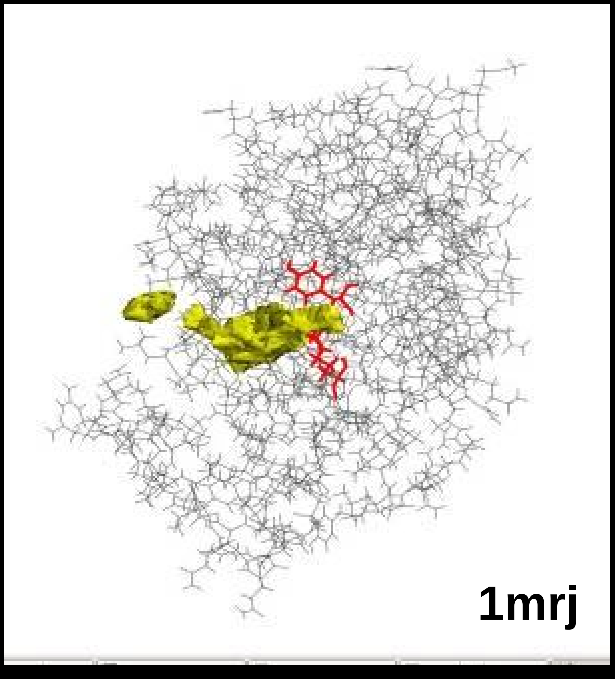




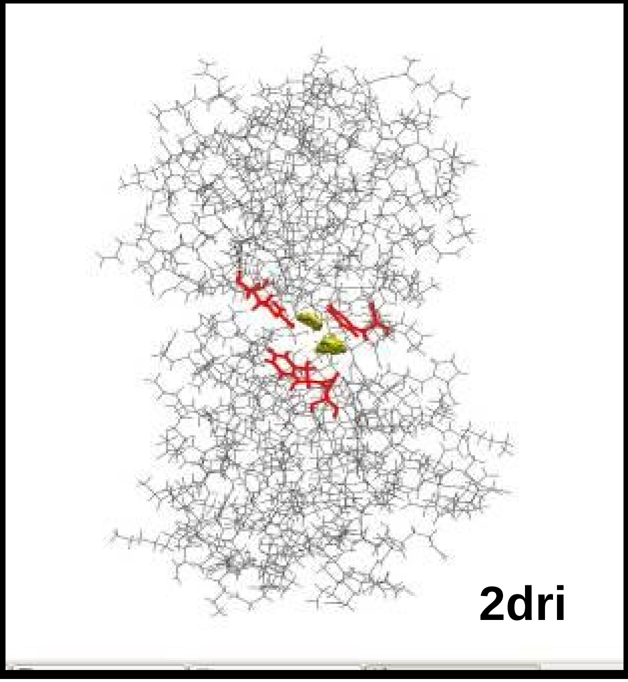
4thi



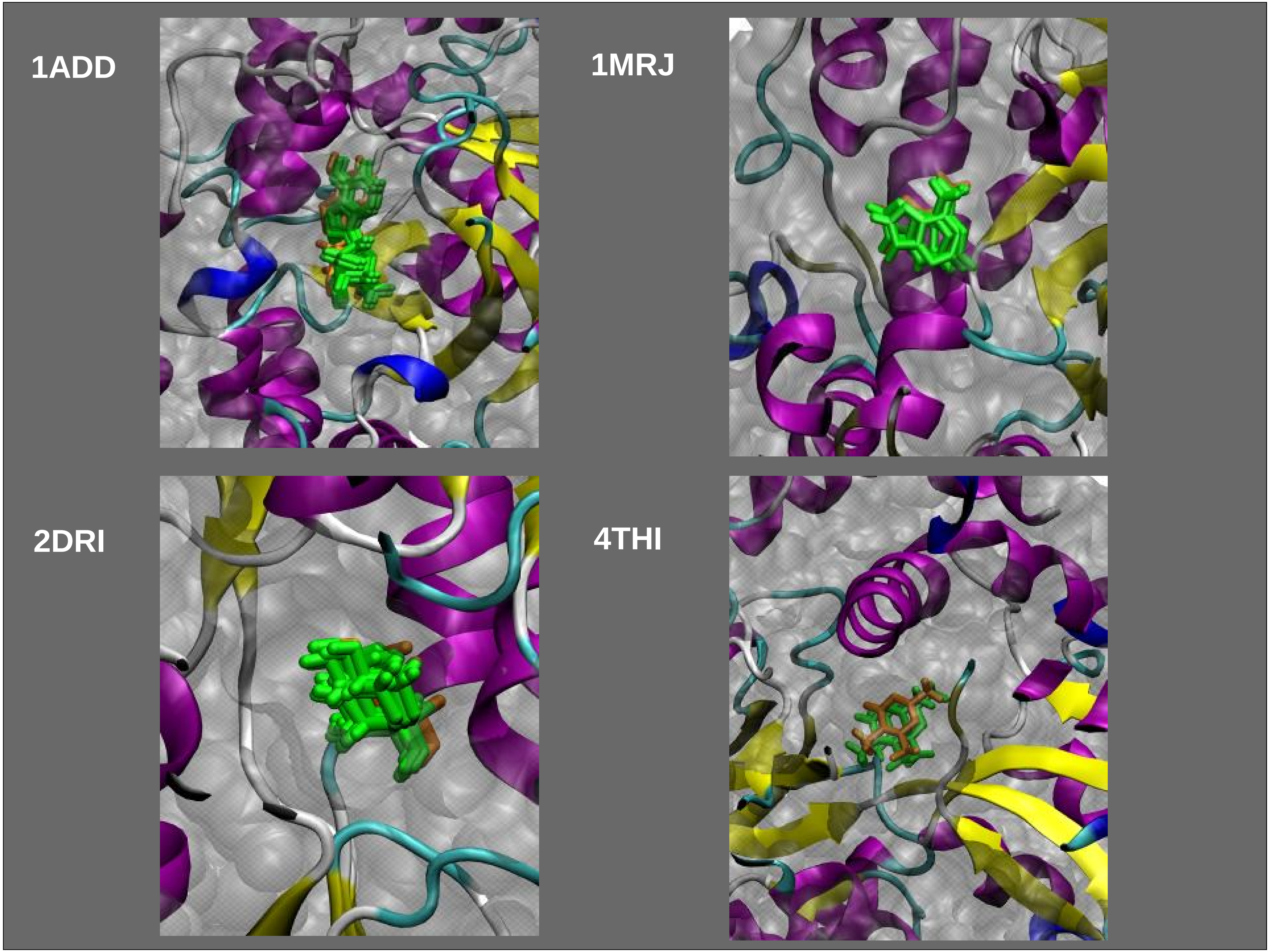
1add



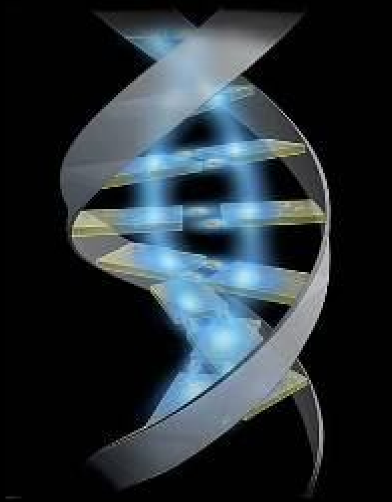
1mrj

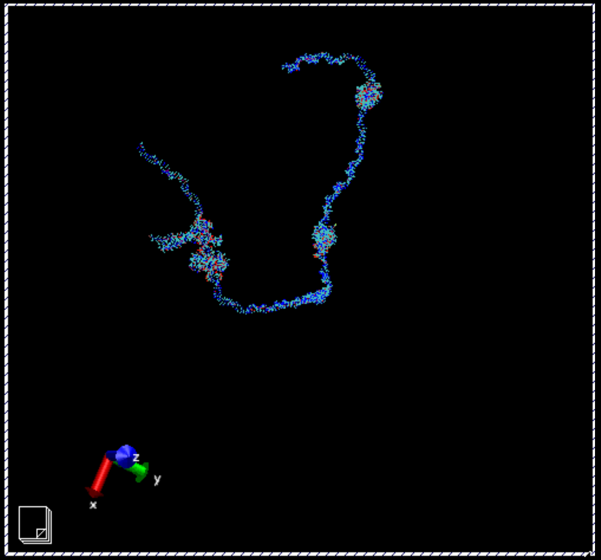


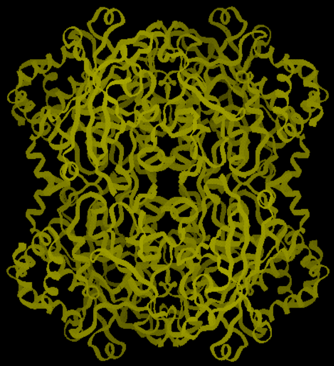
2dri

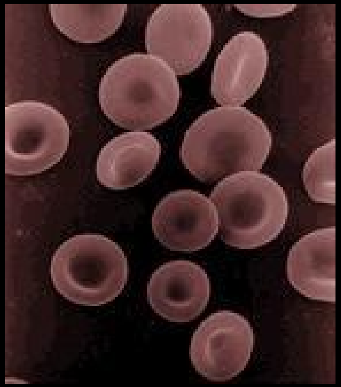

1ADD
1MRJ
2DRI
4THI

Moving atomistic simulations to cell-scale









270 million
Hemoglobin
molecules