www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

A collaborative environment to produce, share and store DNA biomolecular simulations

Ramon Goñi, PhD

3rd EUDAT User Forum

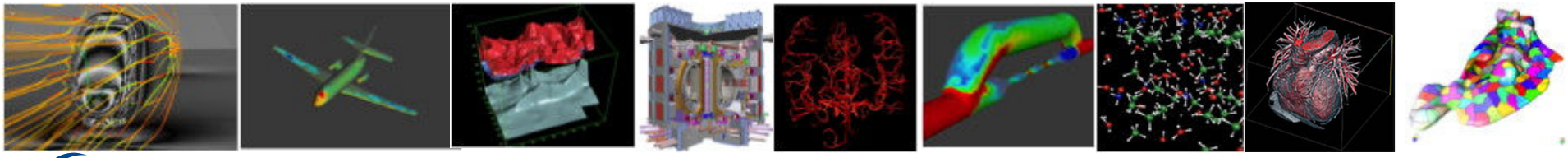
University in Prague 23-24 April 2014

Computer Simulation

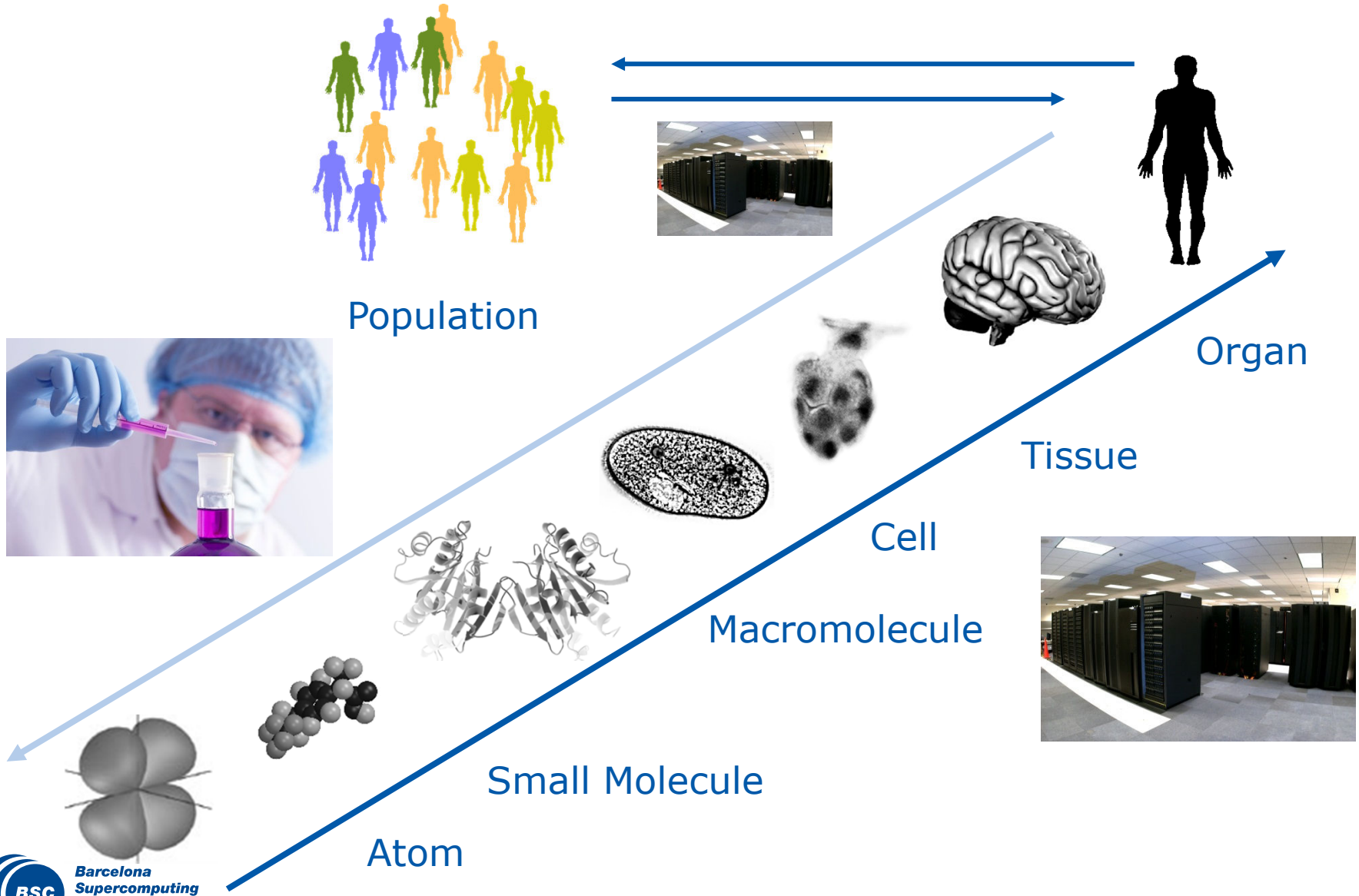
- Why and when we use it
  - To validate a known model
  - As a cost-effective alternative
  - As the only realistic approach to solve a problem

- 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



Life Sciences and Health



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Molecular Simulation

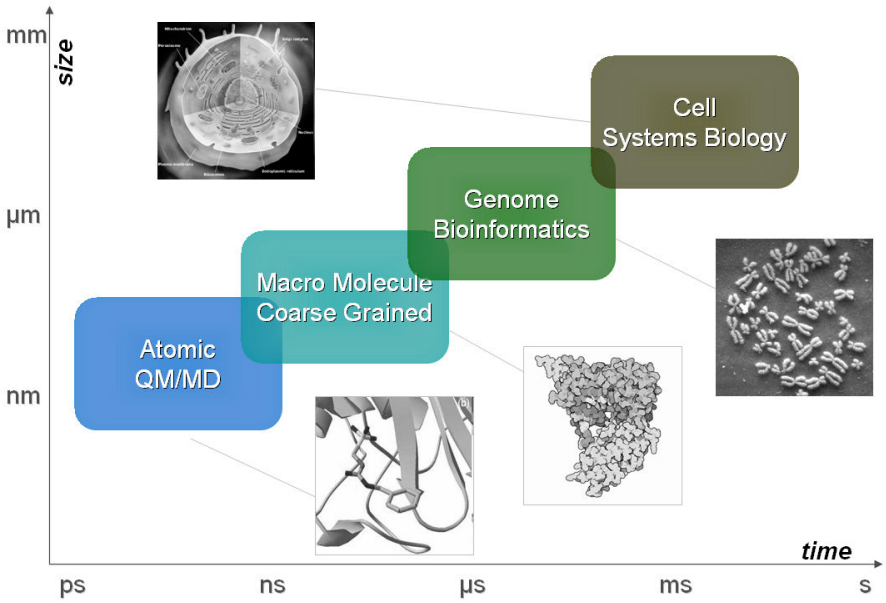
PRECISION

Quantum Mechanics

Molecular Dynamics
(Newton motions)

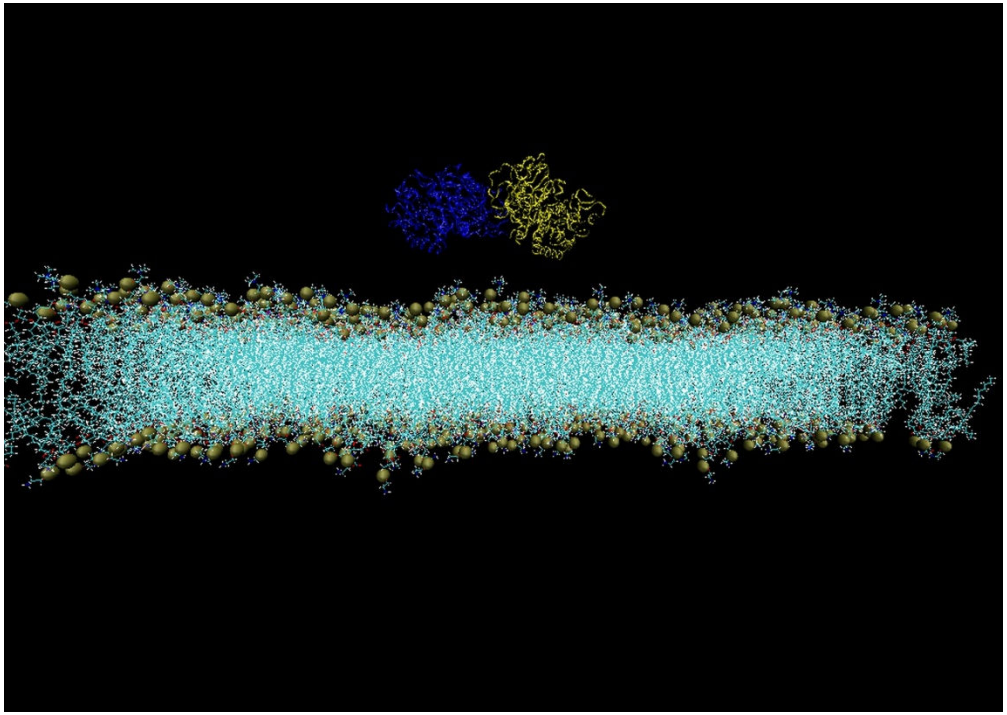
Corse Grained – MD
(Pseudo-atoms)

SPEED

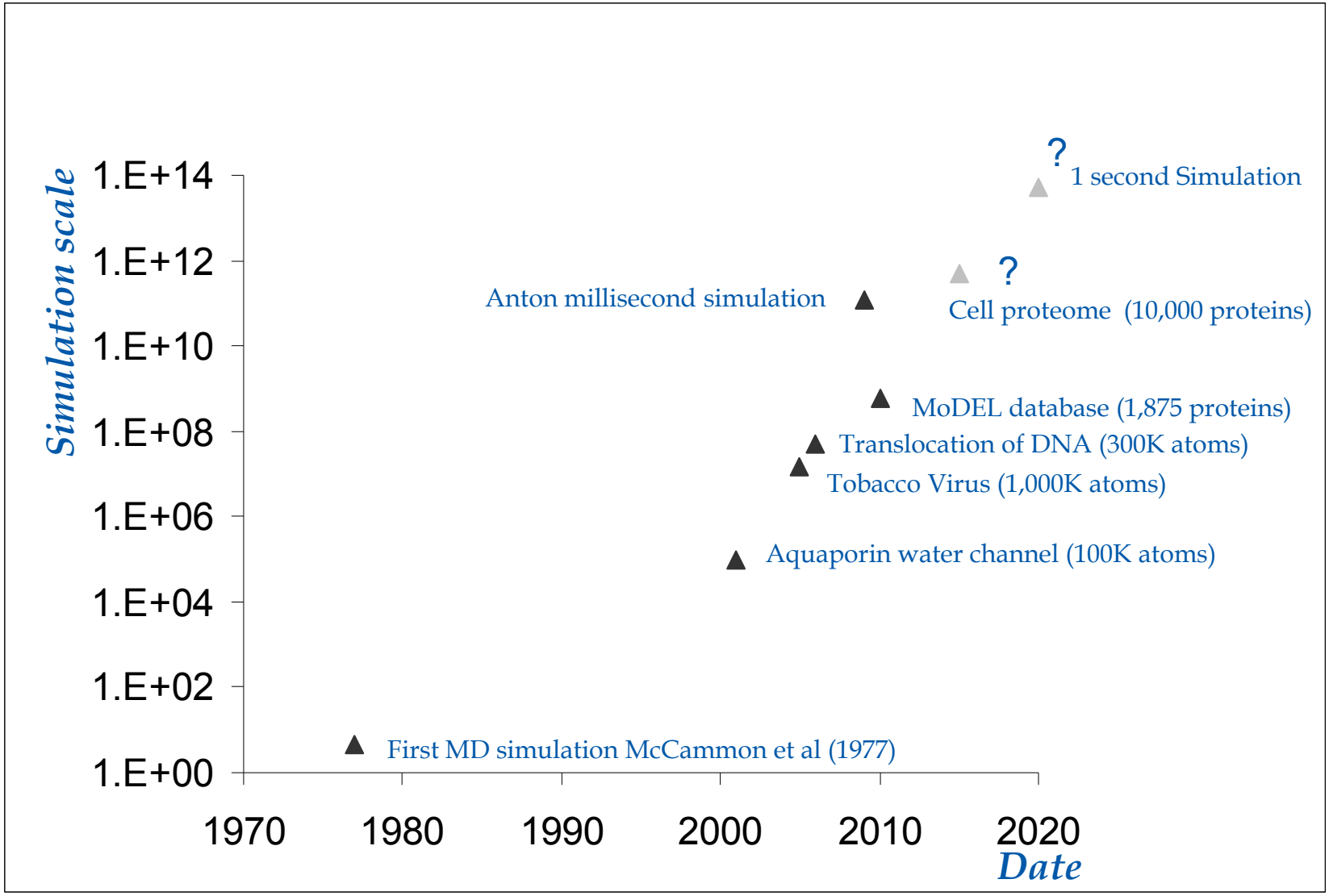


Current limitations in MD

- Size of the system
  - Typically: 10^4 - 10^5 particles
  - Flagship: 10^7
- Simulation length (10^4 particles)
  - Typically: 10^1 – 10^2 ns
  - Using HPC: μs
  - Using Anthon: ms



Trend in MD



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Simulation scale for long and large simulations is calculated multiplying simulation length (nanoseconds) with size (number of atoms).

Global view of Simulation Information

Metadata

Application. Version, Forcefield, Simulation time, solvent, etc.

Pre-process

Output of solvation, equilibration, etc.

Trajectory

Wet/Dry

Post-process

RMSD, B-Factors, PCA, Cavities, etc.

External links

Links to of Protein, Structure, Small Molecule, Domain ,etc.

Big Data Challenge

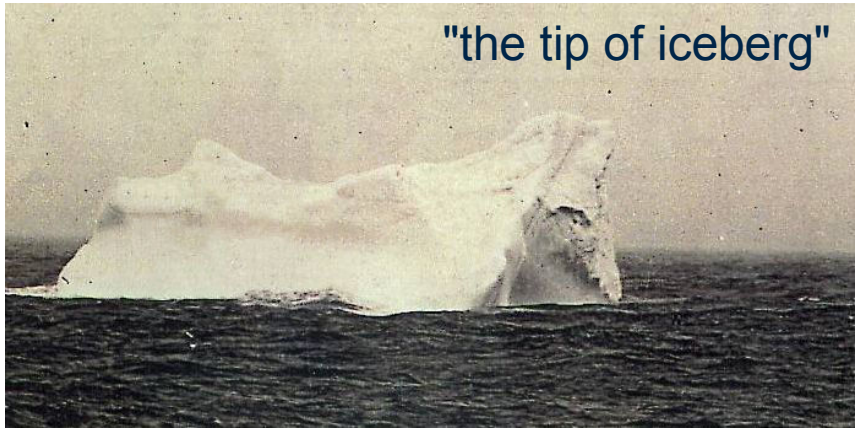


Photo of the iceberg which was probably rammed by the RMS Titanic.

European Exascale Software Initiative

Efficient data management, quick and fast interaction with the computer and flexibility in access to computer resources are in many fields of life sciences at least as important as total theoretical peak power.

Only a few part of simulation data is visible

- Simulations unpublished
- Failed simulations necessary to generate the good one
- Simulations “lost” in system files
- Simulation repeated in fragmented in laboratories systems
- Errors identified too late (lack of interactivity)
- Backups, copies to share, temporal copies to transfer files
- … and the cost of maintaining information in disk systems (power)

First attempts

- Two initial attempts to provide software infrastructure to build biosimulation databases
  - BioSimGrid (2004)
  - P-Found (2006)
- They both shared the philosophy of having a central repository of trajectories that would allow obtaining a comprehensive view of biomolecular structure.
- At the time when these projects were started, computer power was still limited,

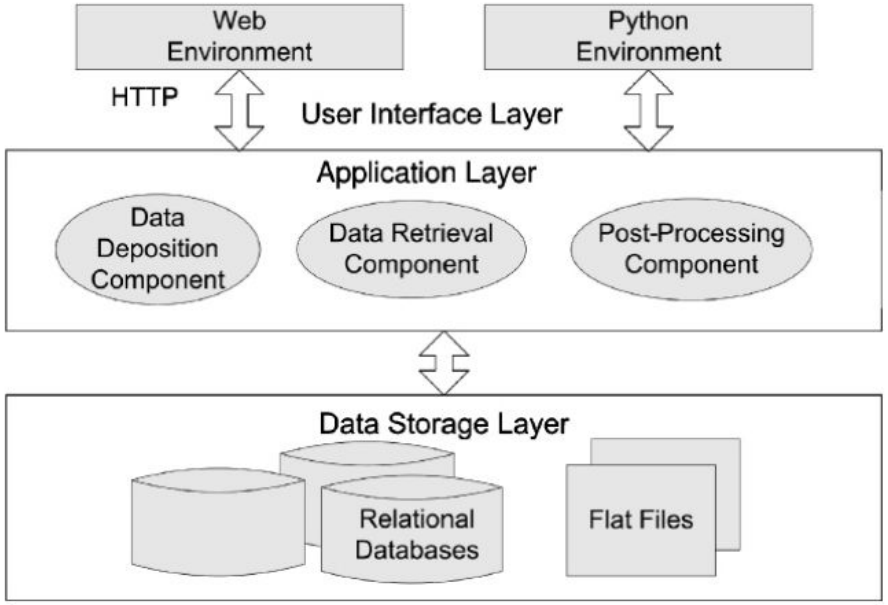
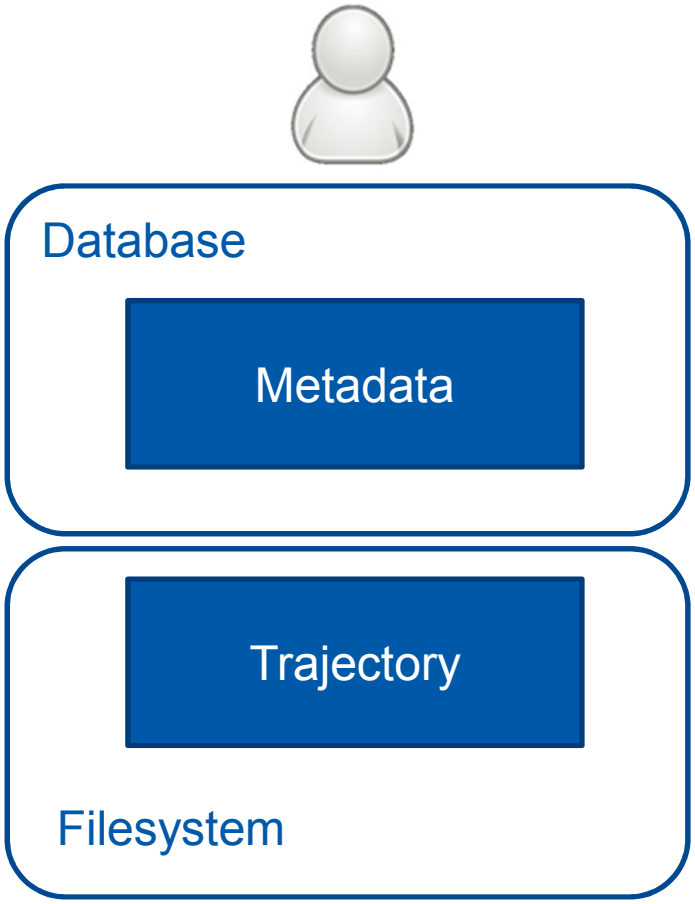
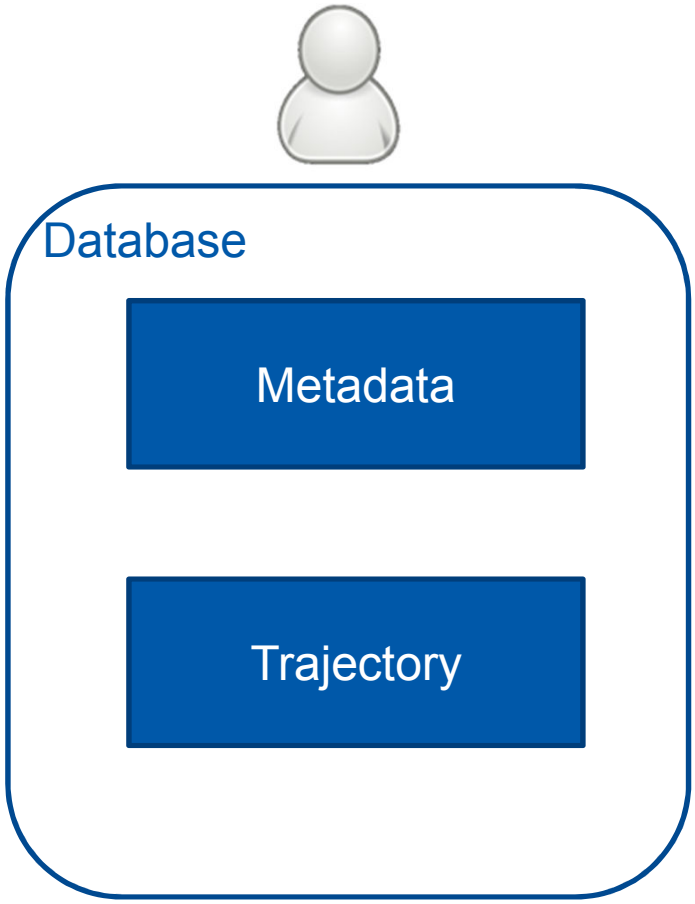


Fig. 1 The architecture of BioSimGrid.

Strategies

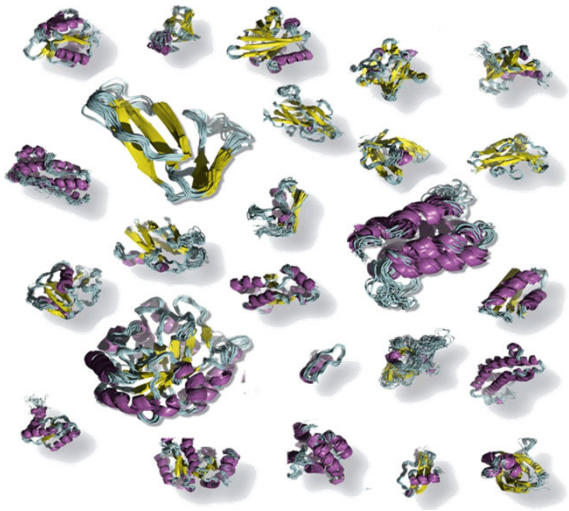


Heavy data file transfer



Poor Scalability
Analysis tools not ready

First repositories (2010)



MySQL

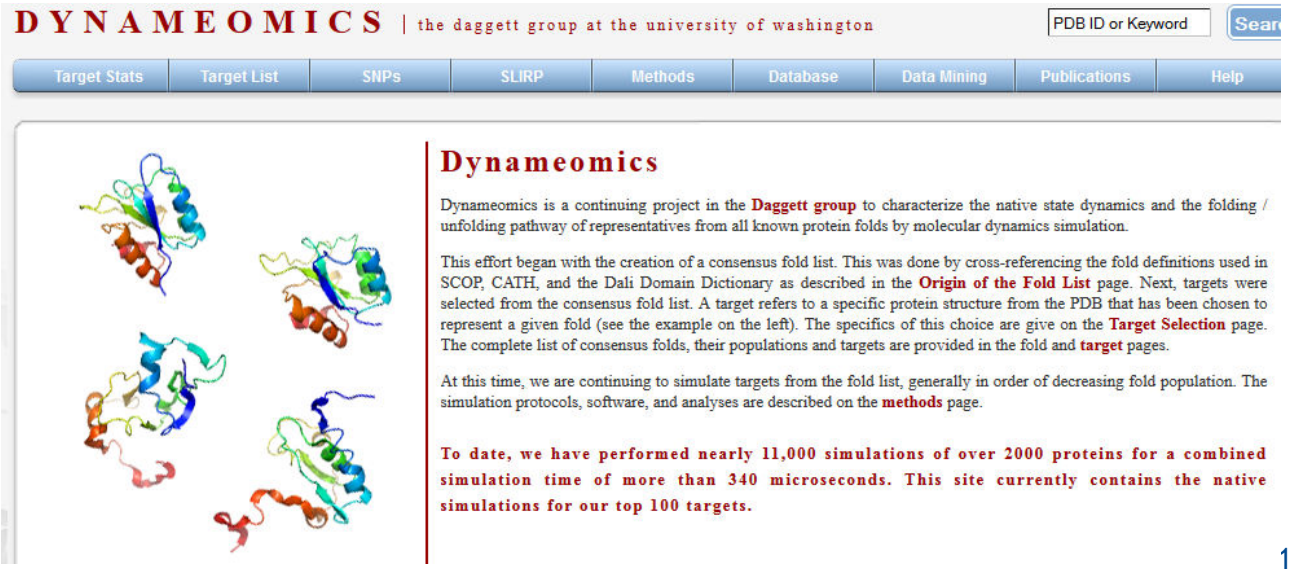
MODEL (Molecular Dynamics Extended Library)

mmb.pcb.ub.es/MODEL

- more than 1,800 entries (~20 Tb)
- covers around 40% of PDB structures, 8% of UniProtKB sequences, 29% of Human UniProtKB sequences and 33% of DrugBank proteins

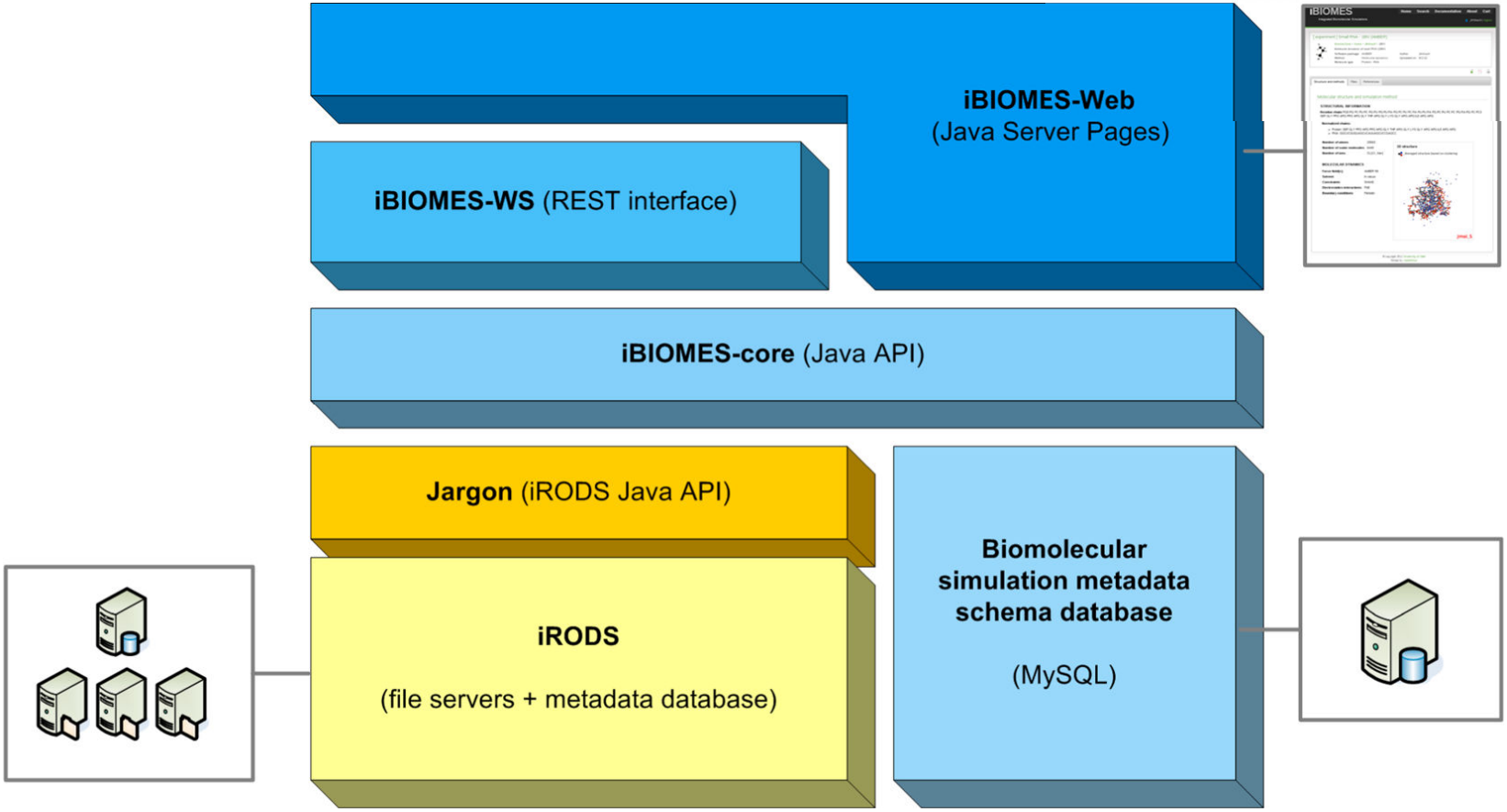
OLAP (SQL)

11,000 simulations of over 2,000 proteins



iRODS based

iBIOMES
Integrated biomolecular simulations
julien.thibault@utah.edu



i·R·O·D·S
Integrated Rule-Oriented Data System

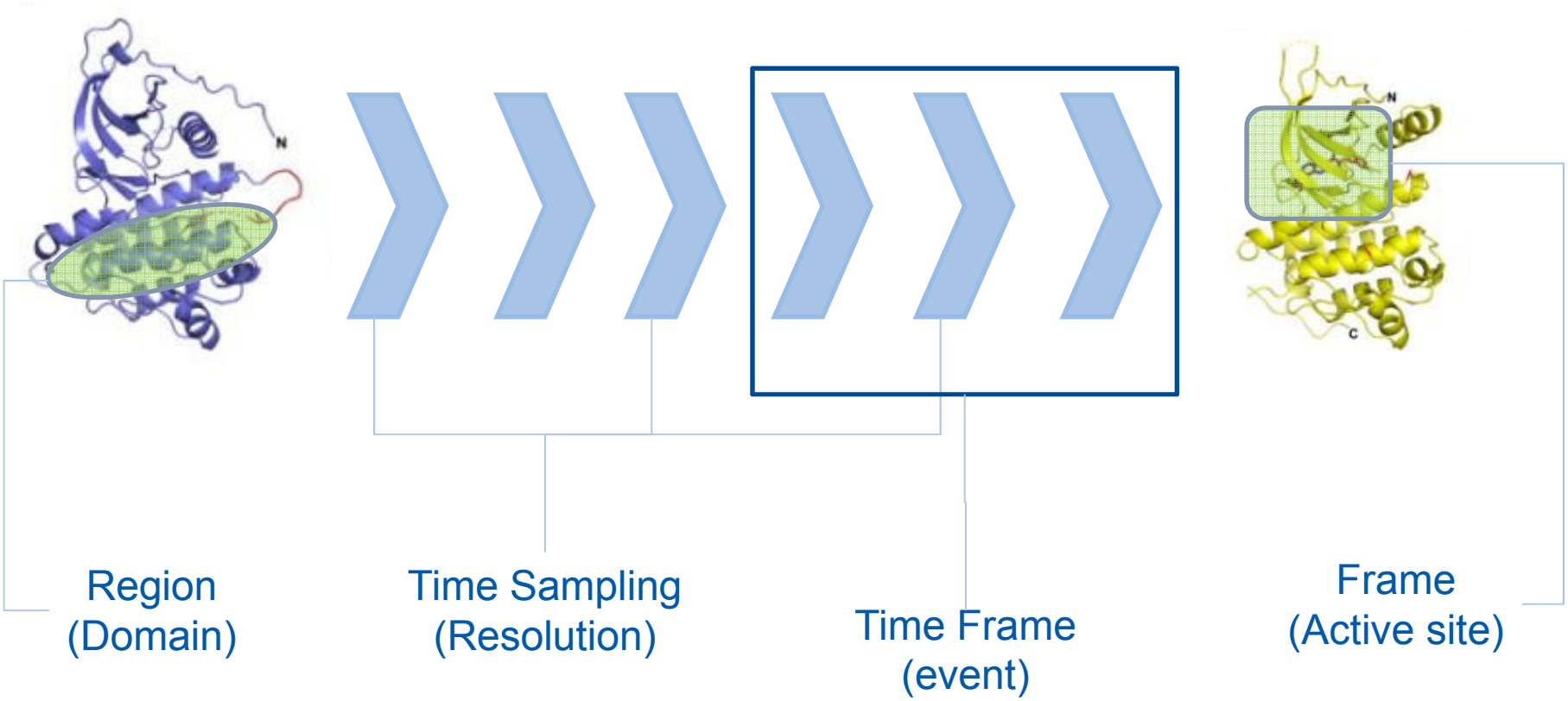
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Thibault, J.C. et al, J. Chem. Inf. Model. 2013, 53 (3), 726-736

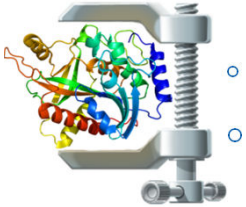
Data Analysis

First snapshot

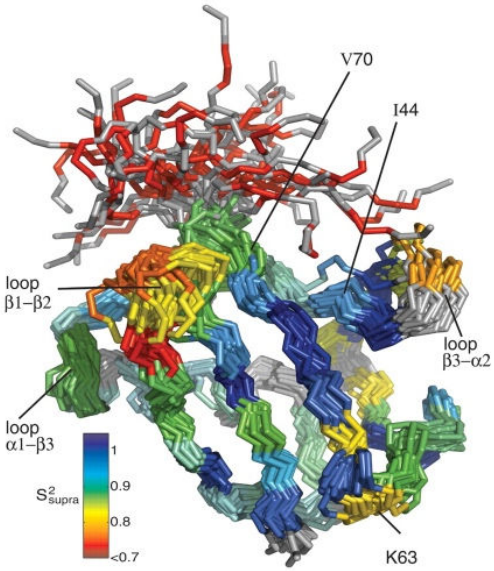
Last snapshot

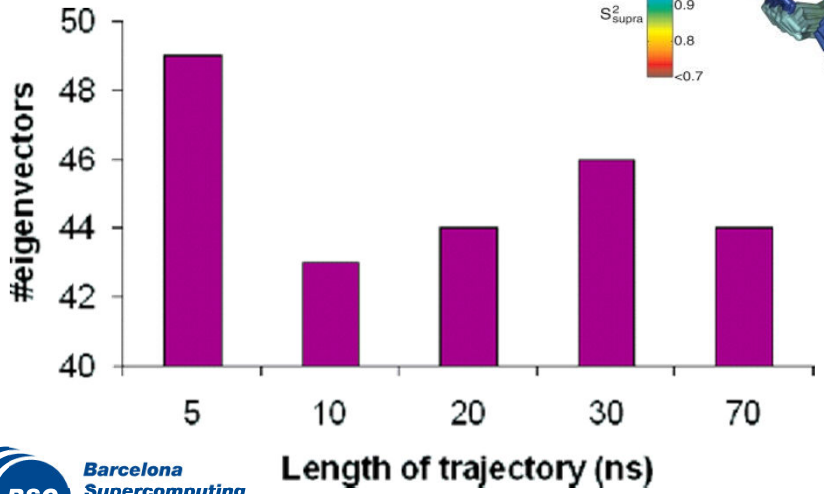


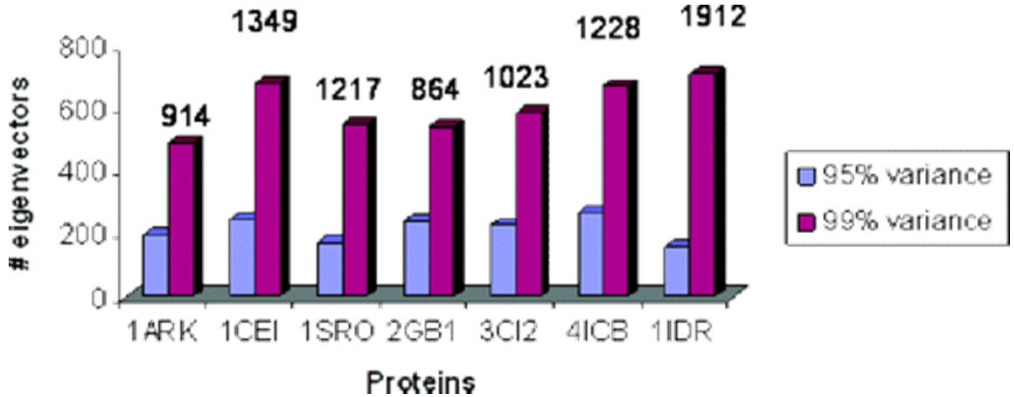
PCA Compression



- Compression through principal components selection
- T. Meyer et al. J. Chem. Theor. Comp. 2006, 2 251-258







Cooperative Biomolecular Simulation project

- The Ascona B-DNA Consortium (ABC) was set up following discussions during a conference in Ascona (Switzerland) in 2001.
- Its aim is to study the effect of base sequence on the structure, dynamics and interactions of DNA using molecular dynamics simulations.
- The initial goal of ABC was to generate a database containing structural and dynamic information on all unique tetra nucleotide sequences (2005)
- 1Tb of data! (compressed)

Members

US

D. Beveridge, Wesleyan U., USA
T. Bishop, Tulane U., USA
D. Case, Rutgers U., USA
T. Cheatham, U. Utah, USA
A. Pérez, USA
R. Osman, Mount Sinai, NY, USA

APAC

B. Jayaram, IIT New Delhi, India
T. Singh, IIT New Delhi, India

EUROPE

J. Curuksu, EPFL, Switzerland
F. Lankas, Prague, Czech Rep.
C. Laughton, Nottingham, UK
R. Lavery, IBCP, France
J. Maddocks, EPFL, Switzerland
A. Michon, IBCP, France
M. Orozco, Barcelona, Spain
D. Petkeviciute, EPFL, Switzerland
N. Spackova, Brno, Czech Rep.
J. Sponer, Brno, Czech Rep.
K. Zakrzewska, IBCP, France

Results

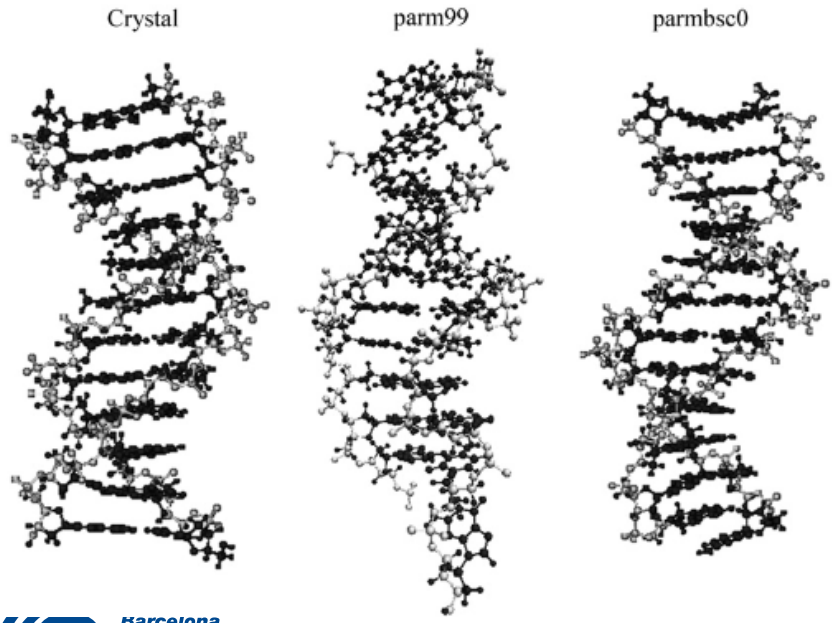
Biophysical Journal Volume 92 June 2007 3817–3829 3817

Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers

Alberto Pérez,*† Iván Marchán,*† Daniel Svozil,‡¶ Jiri Sponer,§¶ Thomas E. Cheatham III,‖ Charles A. Laughton,** and Modesto Orozco*†,††

*Molecular Modeling and Bioinformatics Unit, Institut de Recerca Biomèdica & Instituto Nacional de Bioinformática, Parc Científic de Barcelona, Barcelona 08028, Spain; †Computational Biology Program, Barcelona Supercomputer Centre, Edifici Torre Girona, Barcelona 08028, Spain; ‡Institute of Organic Chemistry and Biochemistry, Center for Biomolecules and Complex Molecular Systems, Academy of Sciences of the Czech Republic, 166 10 Prague 6, Czech Republic; §Institute of Biophysics, Academy of Sciences of the Czech Republic, 612 65 Brno, Czech Republic; ¶Faculty of Science, Masaryk University, 611 37 Brno, Czech Republic; ‖Departments of Medicinal Chemistry, Pharmaceutical Chemistry and Pharmaceutics and Bioengineering, University of Utah, Salt Lake City, Utah 84112; **School of Pharmacy and Centre for Biomolecular Sciences, University of Nottingham, Nottingham NG7 2RD, United Kingdom; and ††Departament de Bioquímica i Biología Molecular, Facultat de Biología, Universitat de Barcelona, Barcelona 08028, Spain

ABSTRACT We present here the parmbsc0 force field, a refinement of the AMBER parm99 force field, where emphasis has been made on the correct representation of the α/γ concerted rotation in nucleic acids (NAs). The modified force field corrects overpopulations of the α/γ = (g+,t) backbone that were seen in long (more than 10 ns) simulations with previous AMBER parameter sets (parm94-99). The force field has been derived by fitting to high-level quantum mechanical data and verified by comparison with very high-level quantum mechanical calculations and by a very extensive comparison between simulations



A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA

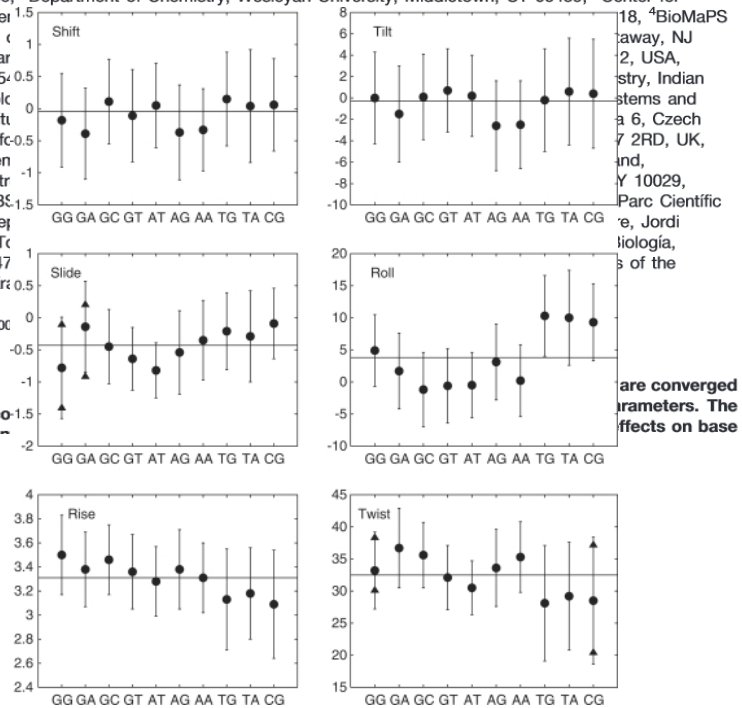
Richard Lavery[1,*], Krystyna Zakrzewska[1], David Beveridge[2], Thomas C. Bishop[3], David A. Case[4], Thomas Cheatham III[5], Surjit Dixit[6], B. Jayaram[7], Filip Lankas[8], Charles Laughton[9], John H. Maddocks[10], Alexis Michon[1], Roman Osman[11], Modesto Orozco[12], Alberto Perez[12], Tanya Singh[7], Nada Spackova[13] and Jiri Sponer[13]

[1]Institut de Biologie et Chimie des Protéines, CNRS UMR 5086/Université de Lyon, 7 passage du Vercors, 69367 Lyon, France, [2]Department of Chemistry, Wesleyan University, Middletown, CT 06459, [3]Center for Computational Scie 18, [4]BioMaPS Institute and Dept. (away, NJ 08854-8087, [5]Depar 2, USA, [6]Zymeworks Inc., 5 stry, Indian Institute of Technol stems and Biomolecules, Instit a 6, Czech Republic, [9]Centre fo 7 2RD, UK, [10]Institut de Mathén nd, [11]Department of Str Y 10029, USA, [12]Joint IRB-BS Parc Científic de Barcelona, Jose e, Jordi Girona 31, Edifici T iología, Avgda Diagonal 647 s of the Czech Republic, Kr

Received August 17, 20

ABSTRACT

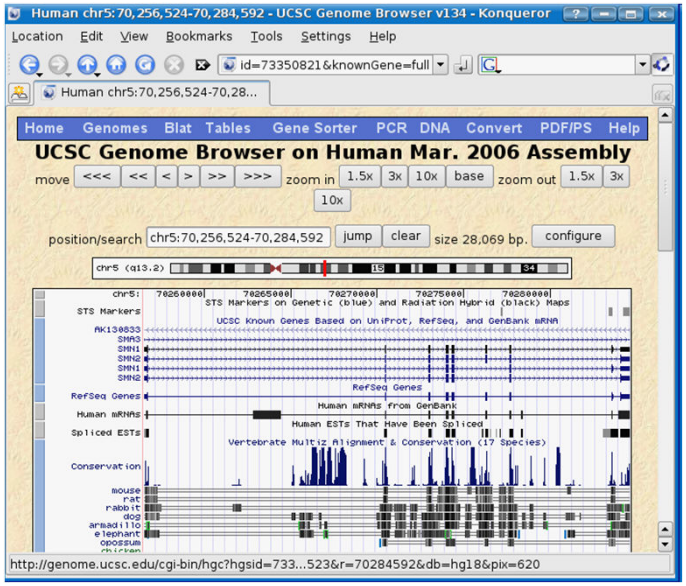
It is well reco are converged rameters. The ffects on base

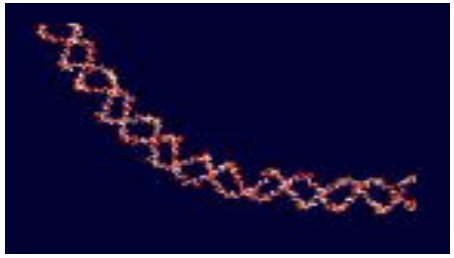


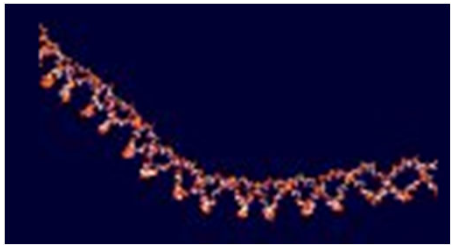
DNA sequence and DNA structure

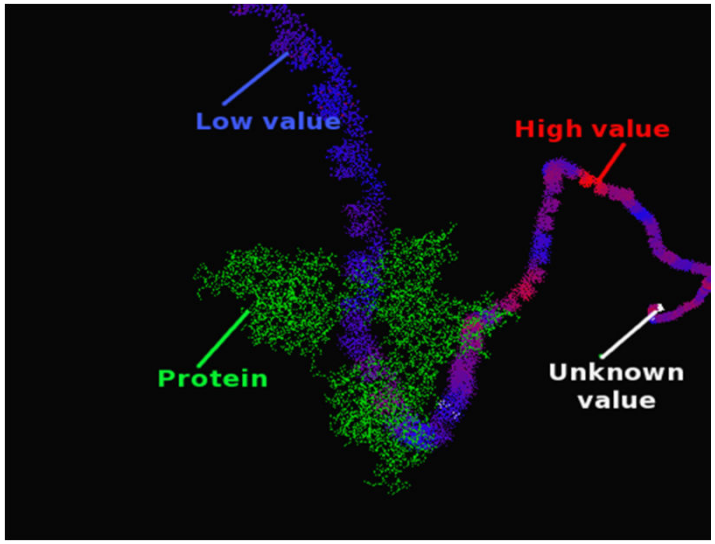
```
<3> 701~724       AGCCTTGTCATCCGTATC-TTTCAA----
<4> 132~154       AGCCTTGTCATCCGTATC-TTTCA-----
<8> 264~288       -GCCTTGTCATCCGTATC-TTTCAACG--
<7> 70~95         --CCTTGTCATCCGTATC-TTTCAACGTG
<7> 344~366       --CCTTGTCATCCGTATC-TTTCAAC---
<0> 304~326       ---CTTGTCATCCGTATCTTTTCAAC---
<2> 482~497       ---CTTGTCATCCGTATC-T---------
<2> 522~537       ----TTGTCATCCGTATC-TT--------
<5> 440~461       ------GTCATCCGTATC-TTTCAACGTG
<6> 16~37         ------GTCATCCGTATC-TTTCAACGTG
<1> 910~925       --------CATCCGTATC-TTTCAA----
<6> 986~1000      --------CATCCGTATC-TTTCA-----
<1> 1051~1069     ---------ATCCGTATC-TTTCAACGTG


<1> 28~68         AACAAGCA-A-ACTTTTATCCATGGTCGTGGTACAGAGGGGTC
<4> 333~373       AACAAGCA-A-ACTTTTATCCATGGTCGTGGTACAGAGGGGTC
<8> 154~193       AACAAGCA-A-ACTTTTATCCATGGTCGTGGTACAGAGGGGT-
<6> 615~647       AACAAGCAGA-ACTTTTATCCATGGTCGTGGTAC---------
<4> 502~533       AACAAGCA-ACCCTTTTATCCATGGTCGTGGTA----------
<1> 844~872       AACAAGCA-A-ACTTTTATCCATGGTCGTGG------------
<8> 194~220       ---------A-ACTTTTATCCATGGTCGTGGTACAGA------
<5> 451~480       ------------CTTTCA-ACGTGGTCGTGGTACAAAGGGGTC
```









Results

7220–7230 Nucleic Acids Research, 2013, Vol. 41, No. 15 Published online 12 June 2013
doi:10.1093/nar/gkt511

Unravelling the hidden DNA structural/physical code provides novel insights on promoter location

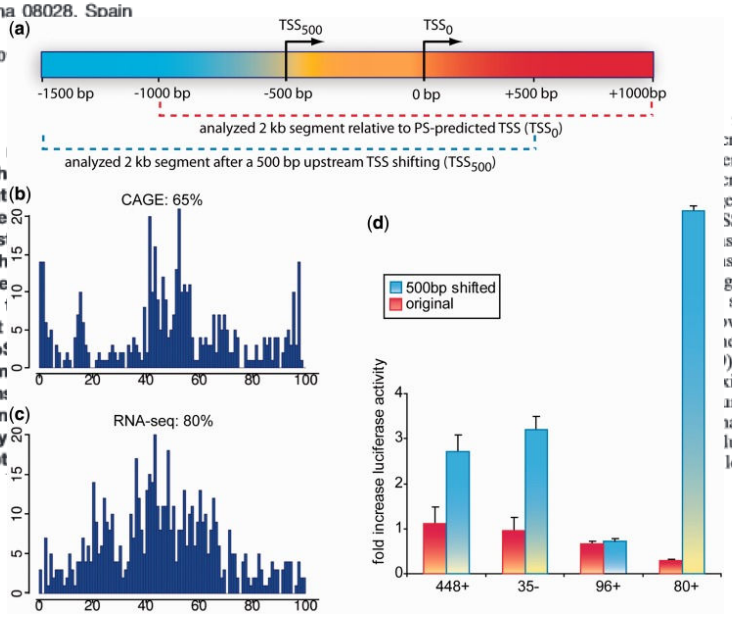
Elisa Durán[1,2], Sarah Djebali[3], Santi González[2,4], Oscar Flores[1,2], Josep Maria Mercader[2,4], Roderic Guigó[3], David Torrents[2,4], Montserrat Soler-López[1,2] and Modesto Orozco[1,2,4,5,*]

[1]Institute for Research in Biomedicine (IRB Barcelona), Barcelona 08028, Spain, [2]Joint IRB-BSC Research Program on Computational Biology, Barcelona 08028, Spain, [3]Bioinformatics and Genomics Group, Center for Genomic Regulation and Universitat Pompeu Fabra, Barcelona 08003, Spain, [4]Barcelona Supercomputing Center, Barcelona 08034, Spain and [5]Department of Biochemistry and Molecular Biology, University of Barcelona, Barcelona 08028, Spain

Received October 23, 20

ABSTRACT

Although protein
promoter regions h
as a critical regulat
location of promote
start sites (TSSs), s
perform a compreh
promoter sequence
dicted TSSs along
defined by distinct
mented in our Pro
representative sam
subjected to extens
analyses. Interestin
be transcriptionally
cific sequence mot



Deniz et al. BMC Genomics 2011, 12:489
http://www.biomedcentral.com/1471-2164/12/489

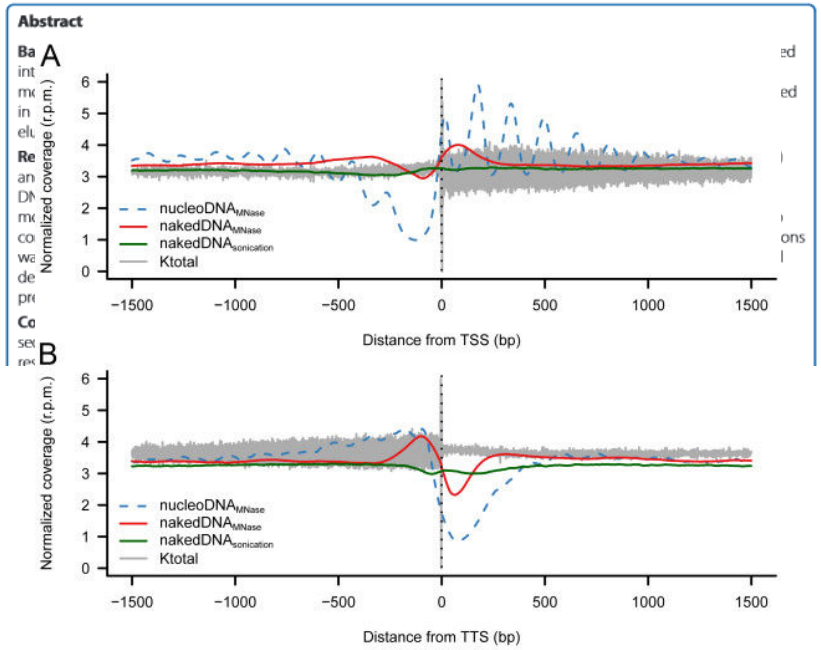
BMC Genomics

RESEARCH ARTICLE Open Access

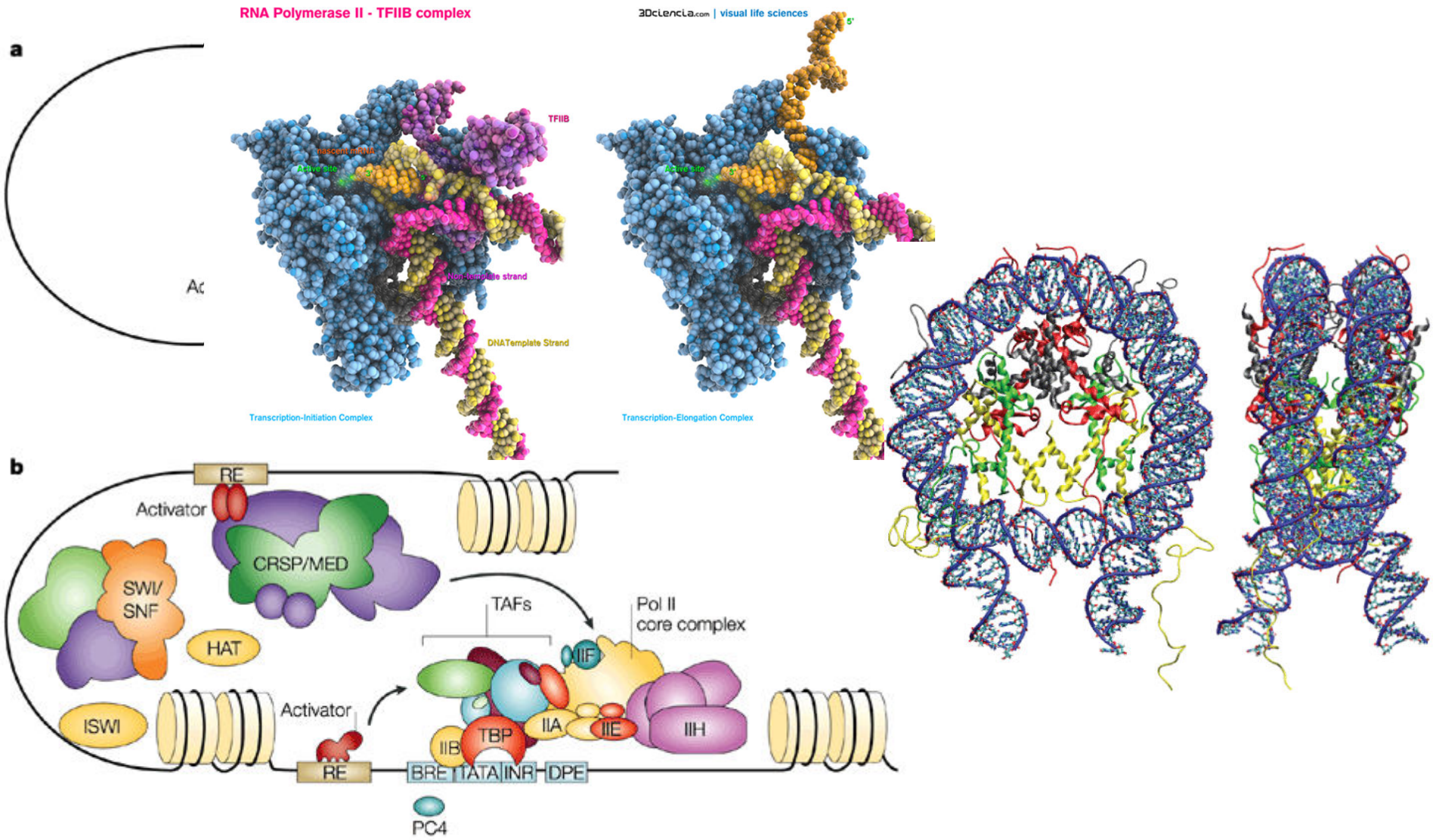
Physical properties of naked DNA influence nucleosome positioning and correlate with transcription start and termination sites in yeast

Özgen Deniz[††], Oscar Flores[††], Federica Battistini[1], Alberto Pérez[2], Montserrat Soler-López[1] and Modesto Orozco[1,3,4*]

Abstract



Next Steps (ABC 2014)



Nature Reviews | Molecular Cell Biology

ABC Meeting 2014: New set up

- Big trajectory files
  - Trajectory Format
  - Compression & I/O efficiency

- Data standardization
  - Data distribution
  - Efficient interoperability

- Data Management
  - Distributed environment
  - Scalability

ABC & EUDAT

- Consortium members are data producers, we need a partner to store results and guarantee open access in the long term. /B2SAFE/
- Data production is computer intensive (HPC), data analysis too. The consortium is not centralized in a single region. We are looking for a solution that optimizes data transfer.
- Simulation data generates heavy trajectory files /B2STAGE/, but also a large set of small unstructured files with metadata, pre-processing information and analysis results /B2SHARE/.

In the long term

- Build an integrated HPC-Big data solution for MD (not only DNA)
- … and integration with Life Sciences databases ('omics, etc.)

EUDAT
European Data Infrastructure