ScalaLife

Scalable Software Services for Life Science

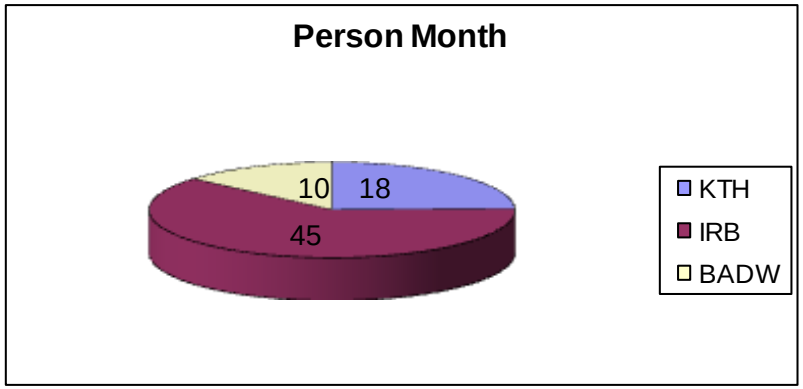
Josep Ll. Gelpí – WP7

IRB - BSC

Participants

- Kungliga Teknisja Hoegskolan (KTH)
- Leibniz-Recehnzentrum (LRZ)
- Institute of Research in Biomedicine (IRB)

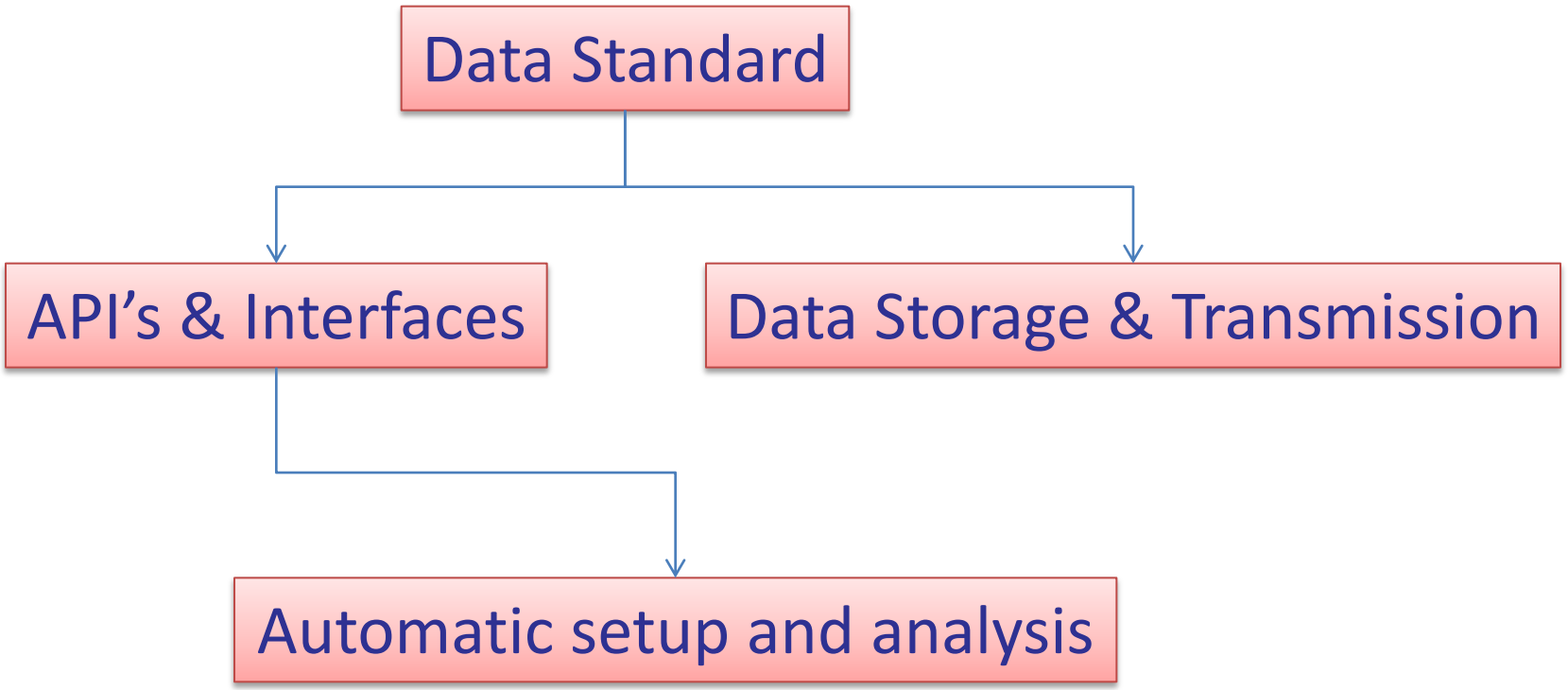
1-38 month activity
Total 73 person month
3 partners involved



Workpackage tasks

- 7.1. Analysis of requirements on data storage and exchange formats (LRZ)
- 7.2. File format standards for data and job description (KTH)
- 7.3. APIs for molecular modeling (IRB)
- 7.4. HT-set-up and analysis tool (IRB)
- 7.5. Simulation databases (IRB)
- 7.6. Management. (IRB)

Interfaces and Data exchange



Data standarization and integration

Scalable Software Services for Life Science

BioXSD

MathML

RCSB PDB PROTEIN DATA BANK

EDAM

CML

BioCatalogue beta "The Life Science Web Service Registry"

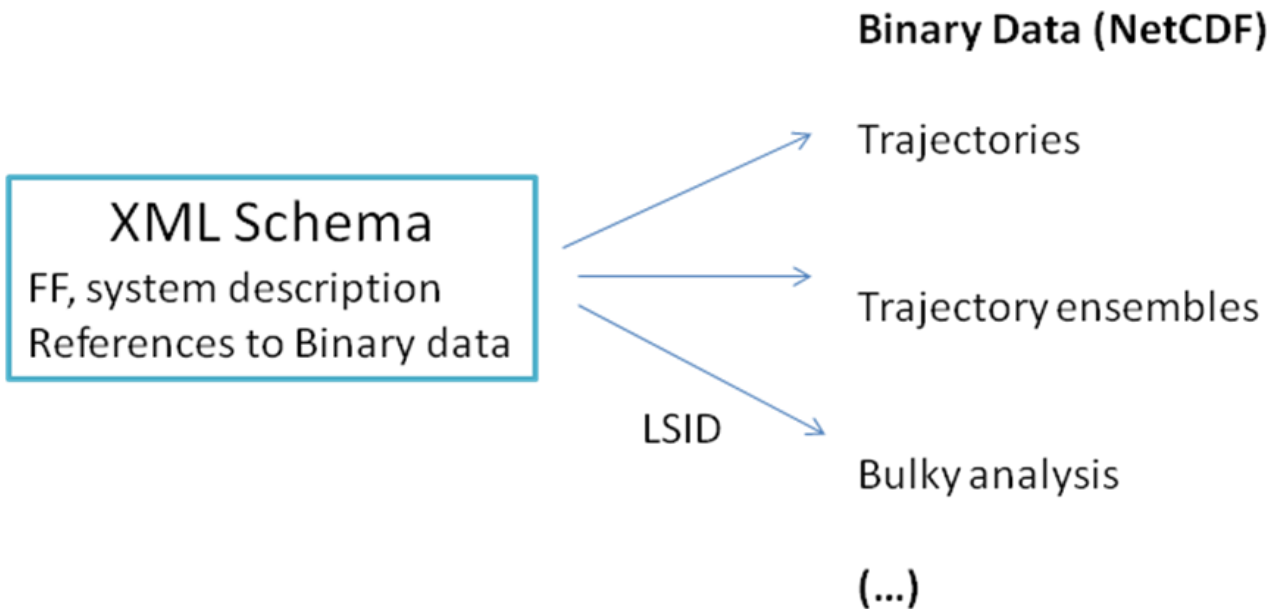
ELIXIR

OBO

moby

Data standards

- XML Based (FF, system and sim. Descriptions)
  - Data ontology required
- Use available standards for binaries (NetCDF, …)
- Allow for distributed storage (LSID)



Data standards

- **FF description:** Equations used, Unified symbols for FF parameters
  - MathML

- **FF atom types and parameters:** Parameter values, incremental datasets

```
<math title="k (x - x_o) ^2 ">
   <mstyle>
     <mi>k</mi>
     <msup>
       <mrow>
         <mo>(</mo><mi>x</mi>
         <mo>-</mo>
         <msub>
           <mi>x</mi><mi>o</mi>
         </msub>
         <mo>)</mo>
       </mrow>
       <mn>2</mn>
    </msup>
    </mstyle>
</math>
```

Data standards

- **Compound/Residue libraries:** Libraries as collection of compounds.

> PDBML: the representation of archival macromolecular structure data in XML. *John Wesbrook, Nobutoshi Ito, Haruki Nakamura, Kim Henrick and Helen M. Berman, Bioinformatics, 21(7), 988-992, 2005.* http://pdbml.pdb.org/

- Covers small molecules, macromolecule residues
- Includes molecular geometry, standard and non-standard residues
- Compatible with PDB's XML

Data standards

- Simulated system
  - Cross-references to biological DBs
  - BioXSD for biological data
  - Physical description using compound library
  - Differences from reference system
- Simulation details
  - Conditions, logs,…
  - PBC boxes
- Simulation results
  - Coordinates, trajectories, velocities, ensembles, …
  - Restart, checkpoints
  - Binary / Compressed /Distributed

Data standards

- Analysis data organized according to data structure
  - Single values
    - Global or averaged
  - 1D,
    - Residue or time based
  - 2D,
    - Contacts, Structure clustering, …
  - Grid data
    - Densities, interaction potentials,…

Data standards

- Distributed approach
  - Large, binary data could be kept at source
  - Perhaps, including analysis software modules
  - LSID: Life Sciences Identifiers
    - Unique
    - Include a standard procedure to locate/retrieve data
    - LSID:<Authority>:<Namespace>:<ObjectID>[:<Version>]

      mmb.pcb.ub.es:ScalaLife:1I6F_GROMACS3.2_G43-SPC_0_TRAJ

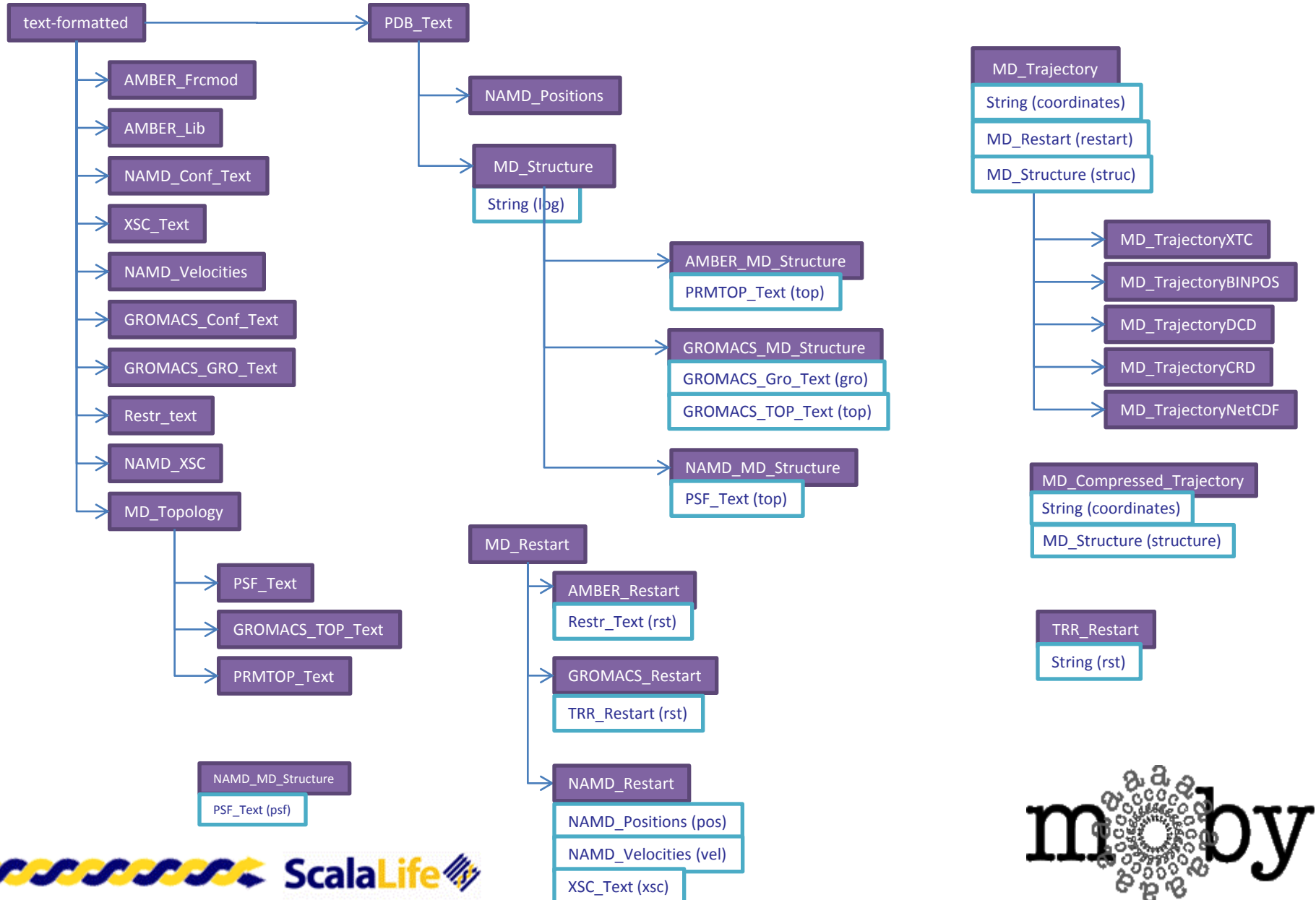
    - Central catalogue of data.
    - Data providers should implement a simple retrieval system

Example of XML (MDmoby)


```
<MD_TrajectoryXTC  id='2ki5' namespace='PDB'>

    <String articleName='coordinates'>
    ...XTC Data…
    </String>

    <GROMACS_Restart  articleName='restart'>
      <TRR_Restart>…TRR data…</TRR_Restart>
    </GROMACS_Restart>

    <GROMACS_MD_Structure articleName="struc">

      <PDB_Text articlename="structure"> ---PDB Data… </PDB_Text>

      <GROMACS_Gro_Text articleName="gro">
        …GRO data…
      </GROMACS_Gro_Text>

      <GROMACS_Top_Text articleName="top">
        …Top data---
      </GROMACS_Top_Text>

      <String articleName='log'>…log data…</String>

    </GROMACS_MD_Structure>

    </MD_TrajectoryXTC>
```

MDMoby ontology

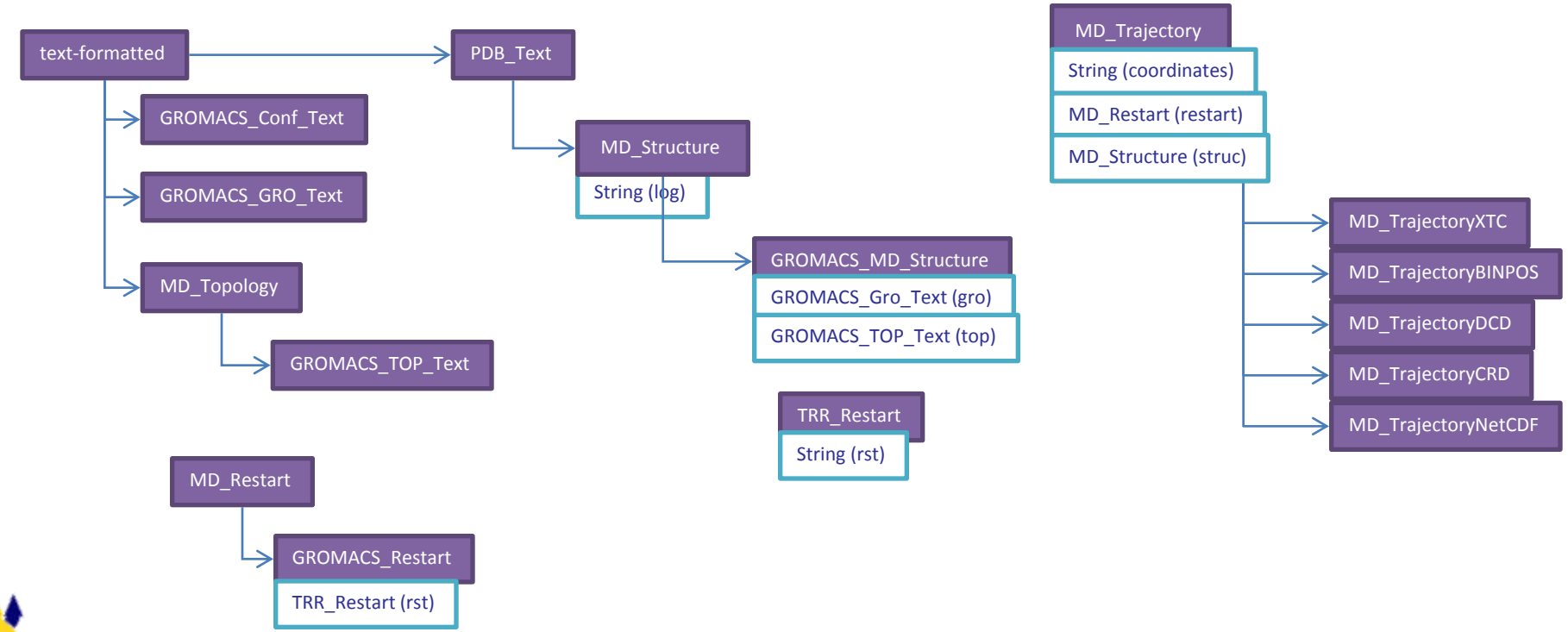
Scalable Software Services for Life Science



ScalaLife

moby

A zoom in GROMACS,…



Data Standards

- Validation protocol
  - 1st draft of XML schema agreed by WP7/Scalalife partners
  - Expert consultants community selected
    - Scalalife partners
    - MD / QM developers
    - Visualization / analysis developers
    - Key MD users (alpha users community?)
  - Feedback from consultants discussed and incorporated
  - Approval by Scalalife and release

Data Standards

- Open questions
  - Existing standards to use
  - Transferability of parameters/FF
  - GROMACS oriented?
  - Integration in BioXSD / MLPDB / OBO / …
  - …

Operations

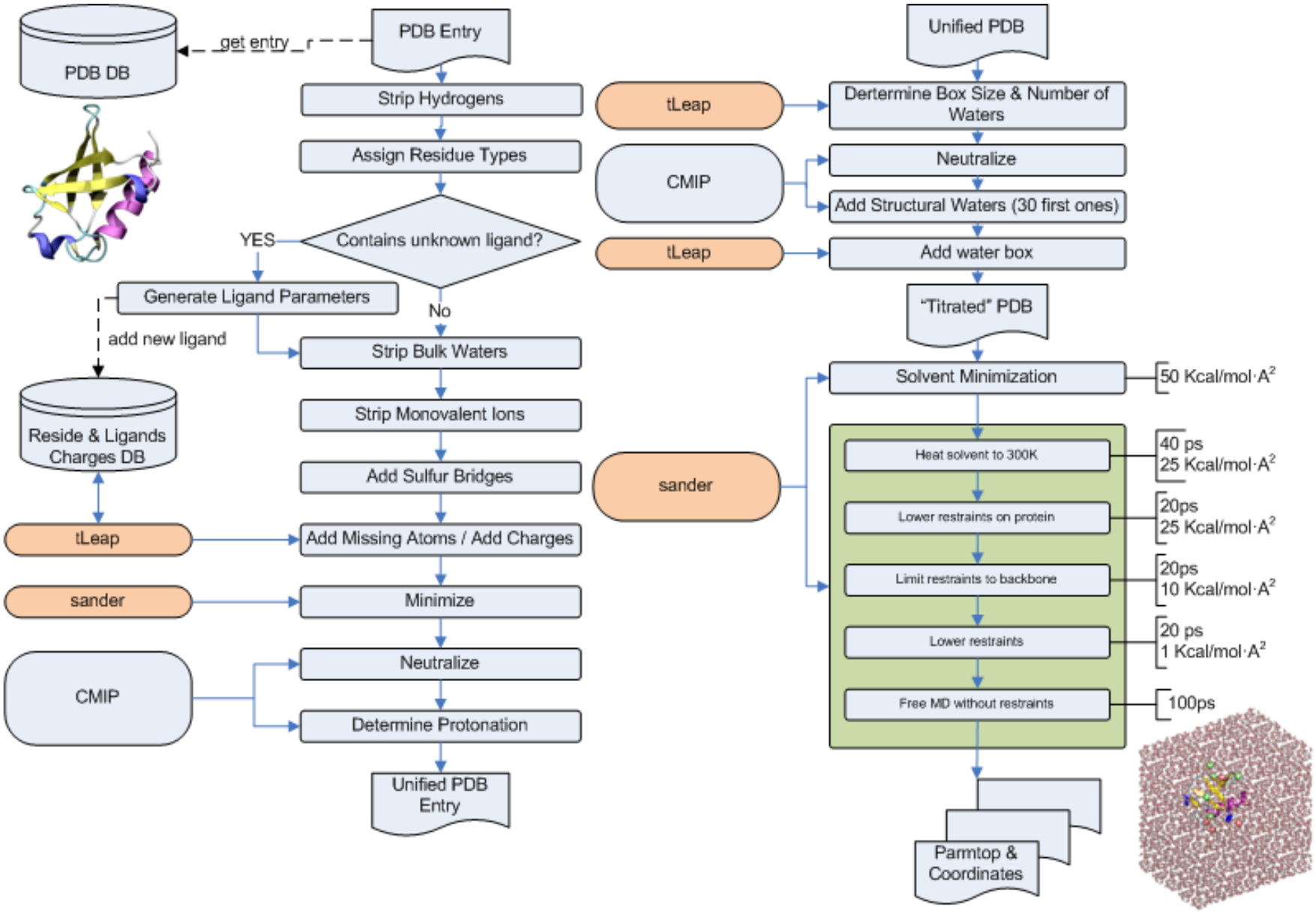
- Aim: Define API's adapted to standard datatypes, to be developed and optimized at the lower level.
- MD or analysis codes culd be built on top the these universal MD API's
- Requires Operations Ontology
  - Highly modular and hierarchical set of operations with well defined input and output
- Building an ontology requires agreement !!

WP7 – Operations

Scalable Software Services for Life Science

-System setup
- *is a* Structure checking
  - *is input of* *PDB*
  - *is output of* *PDB* *#Corrected structure*
  - *is a* chirality check
    - *is a* Thr chirality check
    - *is a* CA chirality check
- *is a* Residue Protonation
  - *is input of* *PDB*
  - *is output of* *PDB* *#Protonated structure*
  - *is a* CMIP based
  - *is a* ProtpKa based

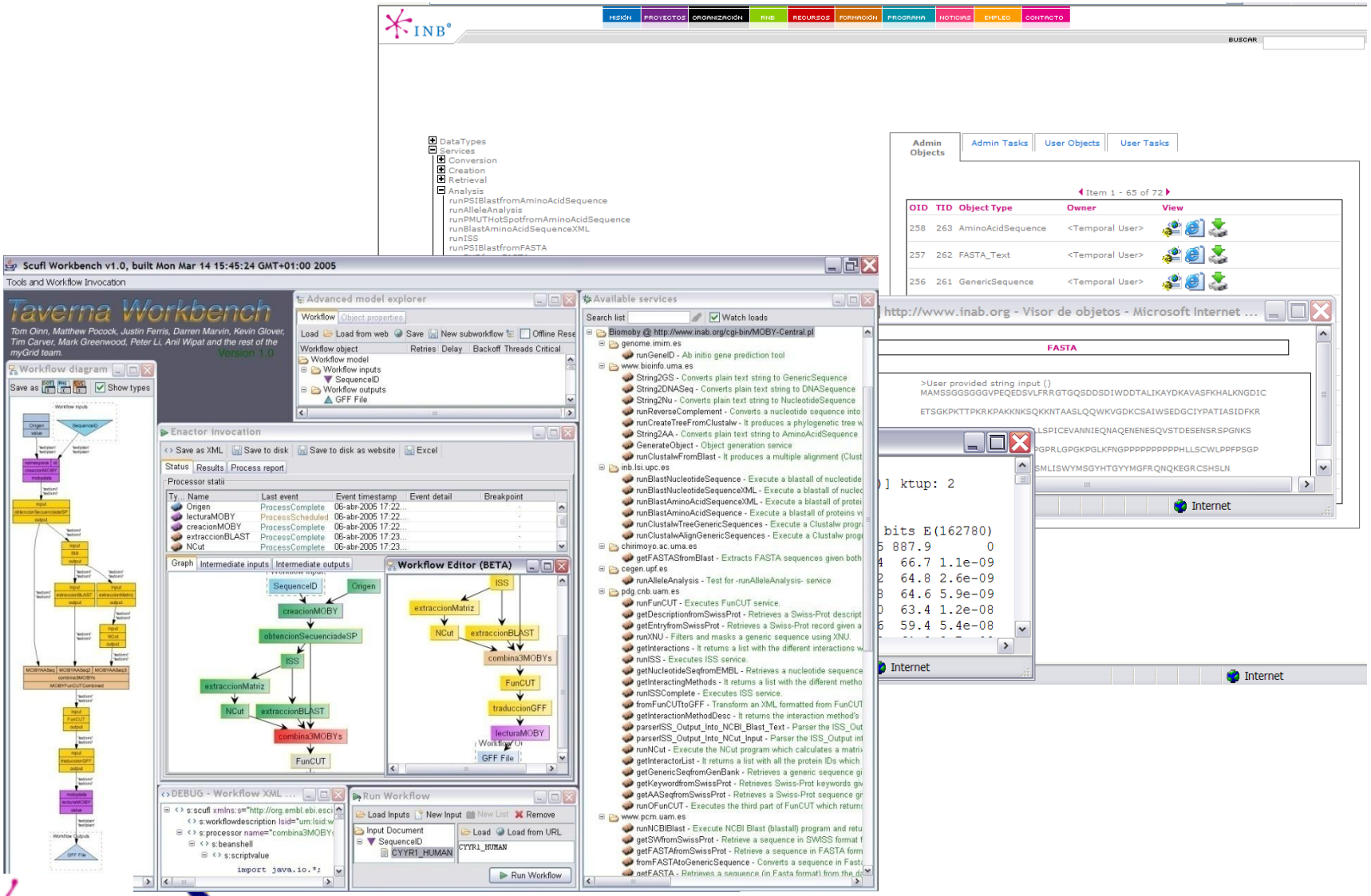
-Simulation
- *is a* NPT Simulation
  - *is input of* PBC-Solvated system
  - *is input of* SimulationSettings
  - *is output of* MD trajectory
  - *is output of* String # Simulation log
  - *is output of* PBC-Solvated system # Restart data


PDB DB
get entry_r
PDB Entry
Strip Hydrogens
Assign Residue Types
YES
Contains unknown ligand?
No
Generate Ligand Parameters
add new ligand
Strip Bulk Waters
Reside & Ligands Charges DB
Strip Monovalent Ions
Add Sulfur Bridges
tLeap
Add Missing Atoms / Add Charges
sander
Minimize
CMIP
Neutralize
Determine Protonation
Unified PDB Entry
Unified PDB
tLeap
Dertermine Box Size & Number of Waters
CMIP
Neutralize
Add Structural Waters (30 first ones)
tLeap
Add water box
"Titrated" PDB
Solvent Minimization
50 Kcal/mol·A2
sander
Heat solvent to 300K
40 ps
25 Kcal/mol·A2
Lower restraints on protein
20ps
25 Kcal/mol·A2
Limit restraints to backbone
20ps
10 Kcal/mol·A2
Lower restraints
20 ps
1 Kcal/mol·A2
Free MD without restraints
100ps
Parmtop & Coordinates

Automatic MD

- Automatic setup and analysis
  - Build on top of API's operations
  - Adapted to a wide selection of end users
    - Black-box workflows for non-expert
      - Ex: Full MD Setup using XXX FF including solvation
      - Ex. Stability after residue mutation
    - Detailed worflow control for experts
  - Runnable from
    - Web interfaces (MDWeb)
    - Web Services (MDMoby)
    - Command-line scripting (MDMoby – MobyLite API)

WS clients


INB
MISIÓN
PROYECTOS
ORGANIZACIÓN
RNB
RECURSOS
FORMACIÓN
PROGRAMA
NOTICIAS
EMPLEO
CONTACTO
BUSCAR
DataTypes
Services
Conversion
Creation
Retrieval
Analysis
runPSIBlastfromAminoAcidSequence
runAlleleAnalysis
runPMUTHotSpotfromAminoAcidSequence
runBlastAminoAcidSequenceXML
runISS
runPSIBlastfromFASTA
Admin Objects
Admin Tasks
User Objects
User Tasks
Item 1 - 65 of 72
OID TID Object Type Owner View
258 263 AminoAcidSequence <Temporal User>
257 262 FASTA_Text <Temporal User>
256 261 GenericSequence <Temporal User>
Scufl Workbench v1.0, built Mon Mar 14 15:45:24 GMT+01:00 2005
Tools and Workflow Invocation
Taverna Workbench
Tom Oinn, Matthew Pocock, Justin Ferris, Darren Marvin, Kevin Glover, Tim Carver, Mark Greenwood, Peter Li, Anil Wipat and the rest of the myGrid team.
Version 1.0
Workflow diagram
Advanced model explorer
Available services
Enactor invocation
Workflow Editor (BETA)
DEBUG - Workflow XML ...
Run Workflow
http://www.inab.org - Visor de objetos - Microsoft Internet ...
FASTA
Internet

```
my $pdbStruct = new PDB__Text($pdbCode,"PDB");
$pdbStruct->content($pdbContent,1);

my $cleaned = cleanStructureFromPDBText ('structure' => $pdbStruct) ->
{'structure'};

my $pdb_h = getGROMACS_MD_StructureFromPDBText (
  'structure' => $cleaned, 'forcefield' => $forcefield ) -> {'structure'};

my $min_h = optimizeStructureFromGROMACS_MD_Structure (
  'structure' => $pdb_h, 'minimize' => 500,'md_type' => 1) -> {'structure'};

my $min_h2 = optimizeStructureFromGROMACS_MD_Structure (
  'structure' => $min_h,'minimize' => 500,'md_type' => 2) -> {'structure'};

my $solv = solvateStructureFromGROMACS_MD_Structure (
  'structure' => $min_h2,'ions' => 'true', 'boxsize' => 0.8, 'boxtype' =>
'octahedron', 'ionic_concentration' => 0.05) -> {'structure'};

my $eq = runMDFromGROMACS_MD_Structure_async (
  'structure' => $solv, 'time' => 1,'timestep' => 0.001, 'md_type' =>
3,'persistent' =>  1) -> {'structure'};

my $md = runMDFromMD_TrajectoryXTC_async (
  'structure' => $eq,'time' => 5, 'md_type' => 4,'timestep' => 0.004,
'coordFreq' => 2,'persistent' => 1) -> {'structure'};

my $md_traj_base64 = $md->coordinates;

my $md_traj_dcd = decode_base64($md_traj_base64);
```

MDWeb. Molecular Dynamics on Web - Mozilla Firefox

Archivo Editar Ver Historial Marcadores Herramientas Ayuda

http://mmb.pcb.ub.es/MDWeb-inb238/openProject.php

LSID

Más visitados Comenzar a usar Firefox Google Reader (22) MoDEL (Molecular Dy... Import to Mendeley Recuérdalo todo | Ever... Últimas noticias

Desactivar Cookies CSS Formularios Imágenes Información Varios Resaltar Tamaño Herramientas Código fuente Op

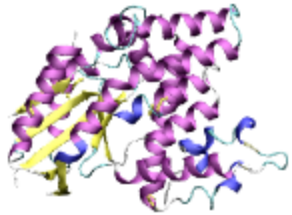
MDWeb. Molecular Dynamics on Web

MDWeb

Molecular Dynamics on Web

User: Josep Ll. Gelpi

Home | Start new project | Close workspace | Help



test 2ki5_a (MDWeb4d80f45acca7c)

Last modification on: 16/03/2011 18:33

Disk Usage: 1.9 MB

Stored structures

Click on structure title to deploy the toolbox.

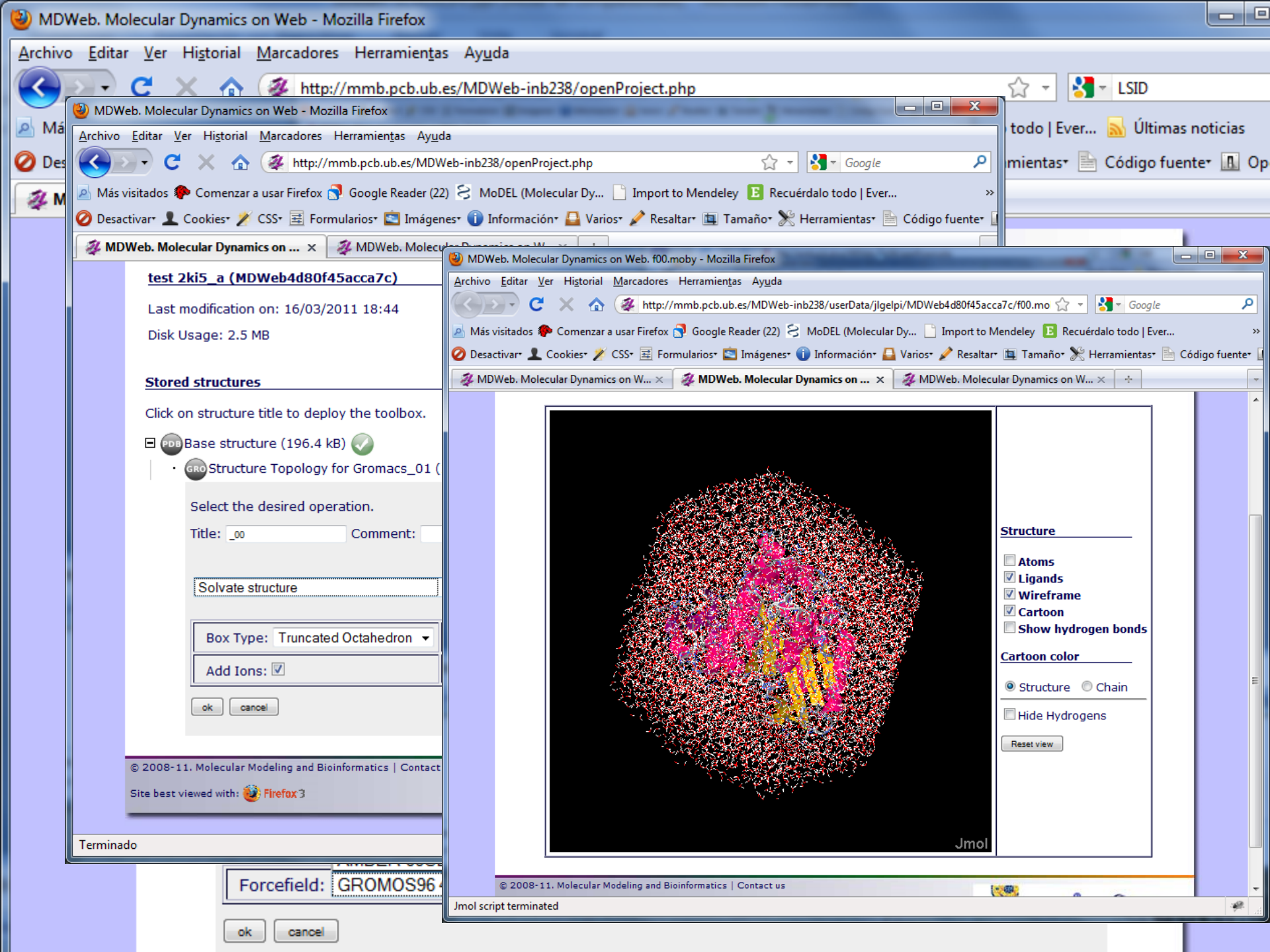
- Base structure (196.4 kB)

Select the desired operation.

Title: _00 Comment:

List of Operations:

List of Operations:
Check for disulphide bonds
Clean PDB
Fix Side Chains
Mutate residue
Amber FULL MD Setup
Amber MD Setup

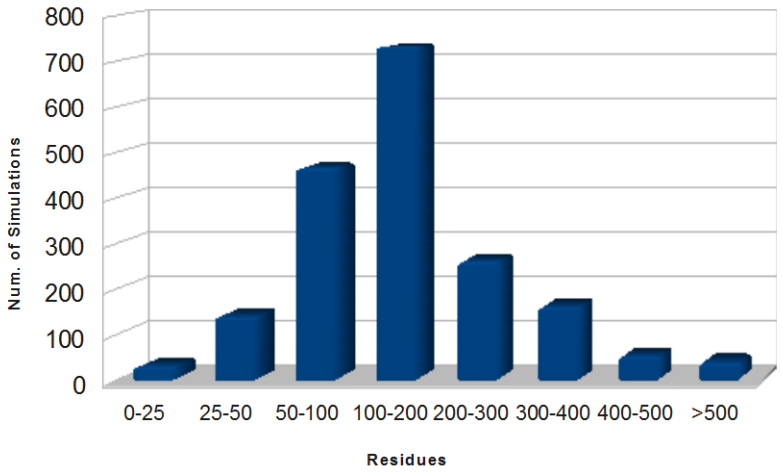

MDWeb. Molecular Dynamics on Web - Mozilla Firefox
Archivo Editar Ver Historial Marcadores Herramientas Ayuda
http://mmb.pcb.ub.es/MDWeb-inb238/openProject.php
MDWeb. Molecular Dynamics on Web - Mozilla Firefox
Archivo Editar Ver Historial Marcadores Herramientas Ayuda
http://mmb.pcb.ub.es/MDWeb-inb238/openProject.php
Más visitados Comenzar a usar Firefox Google Reader (22) MoDEL (Molecular Dy... Import to Mendeley Recuérdalo todo | Ever...
Desactivar Cookies CSS Formularios Imágenes Información Varios Resaltar Tamaño Herramientas Código fuente
MDWeb. Molecular Dynamics on ...
test 2ki5_a (MDWeb4d80f45acca7c)
Last modification on: 16/03/2011 18:44
Disk Usage: 2.5 MB
Stored structures
Click on structure title to deploy the toolbox.
Base structure (196.4 kB)
Structure Topology for Gromacs_01 (
Select the desired operation.
Title: _00
Comment:
Solvate structure
Box Type: Truncated Octahedron
Add Ions:
ok
cancel
© 2008-11. Molecular Modeling and Bioinformatics | Contact
Site best viewed with: Firefox 3
Terminado
MDWeb. Molecular Dynamics on Web. f00.moby - Mozilla Firefox
Archivo Editar Ver Historial Marcadores Herramientas Ayuda
http://mmb.pcb.ub.es/MDWeb-inb238/userData/jlgelpi/MDWeb4d80f45acca7c/f00.mo
Más visitados Comenzar a usar Firefox Google Reader (22) MoDEL (Molecular Dy... Import to Mendeley Recuérdalo todo | Ever...
Desactivar Cookies CSS Formularios Imágenes Información Varios Resaltar Tamaño Herramientas Código fuente
MDWeb. Molecular Dynamics on W...
MDWeb. Molecular Dynamics on ...
MDWeb. Molecular Dynamics on W...
Jmol
Structure
Atoms
Ligands
Wireframe
Cartoon
Show hydrogen bonds
Cartoon color
Structure
Chain
Hide Hydrogens
Reset view
© 2008-11. Molecular Modeling and Bioinformatics | Contact us
Jmol script terminated
Forcefield: GROMOS96
ok
cancel

API's & automatic MD

- Open questions
  - Programming library vs. Web Services library
  - Grid/Cloud
  - Access in HPC
  - Integration with bioinformatics (non structural)
  - Collection of "black-box" workflows

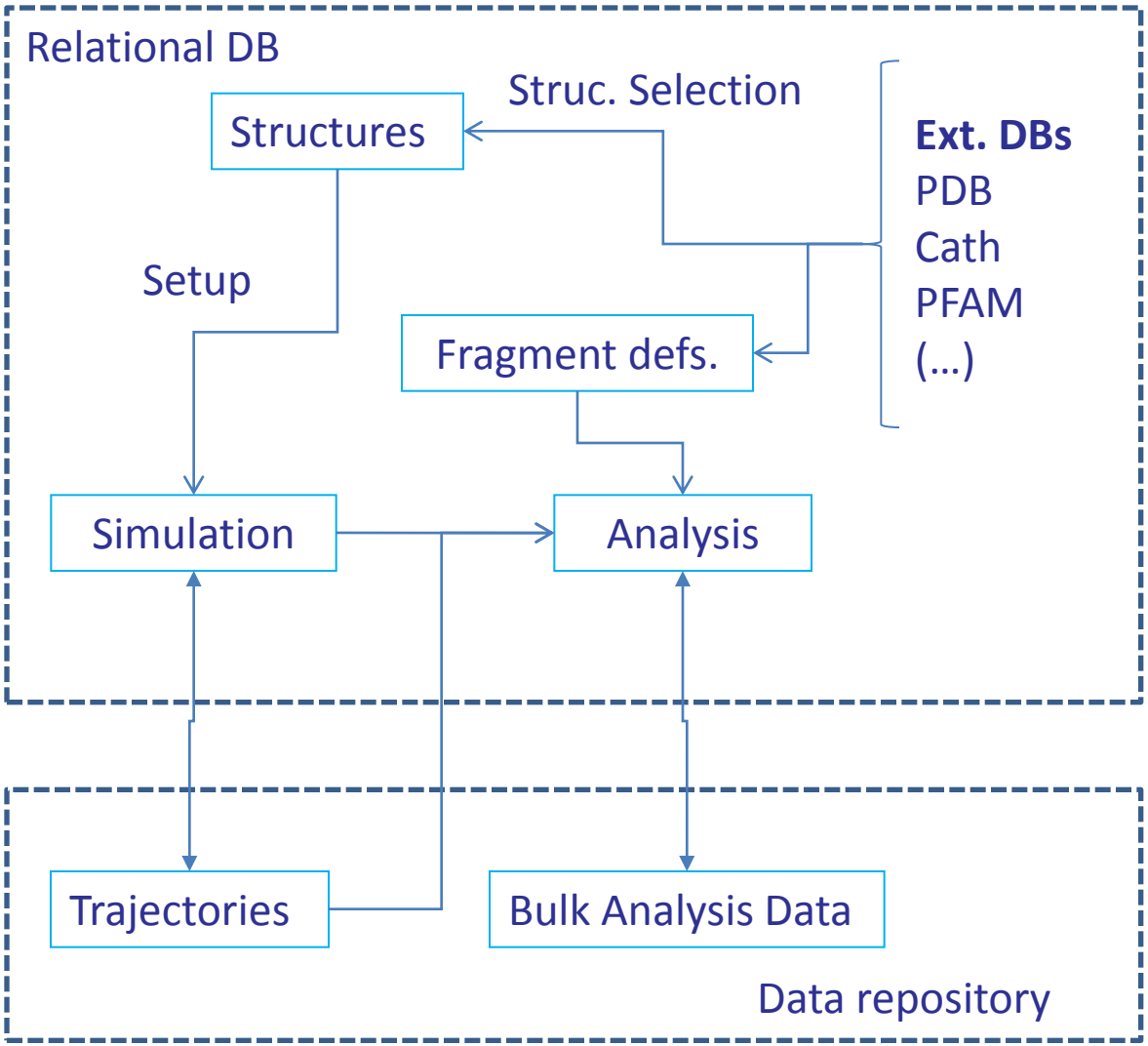
Requeriments on data Storage

- Reference Example MoDEL project
  - 1875 simulations, 1595 systems
  - Disk usage:
    - Total: **18Tb**
    - Raw Solvent trajectories: **14Tb**
    - Dry trajectories: **1.7Tb**
    - Analysis: **1.1Tb**
    - Setup: **1.2Tb**
    - MySQL DB: **23Gb**

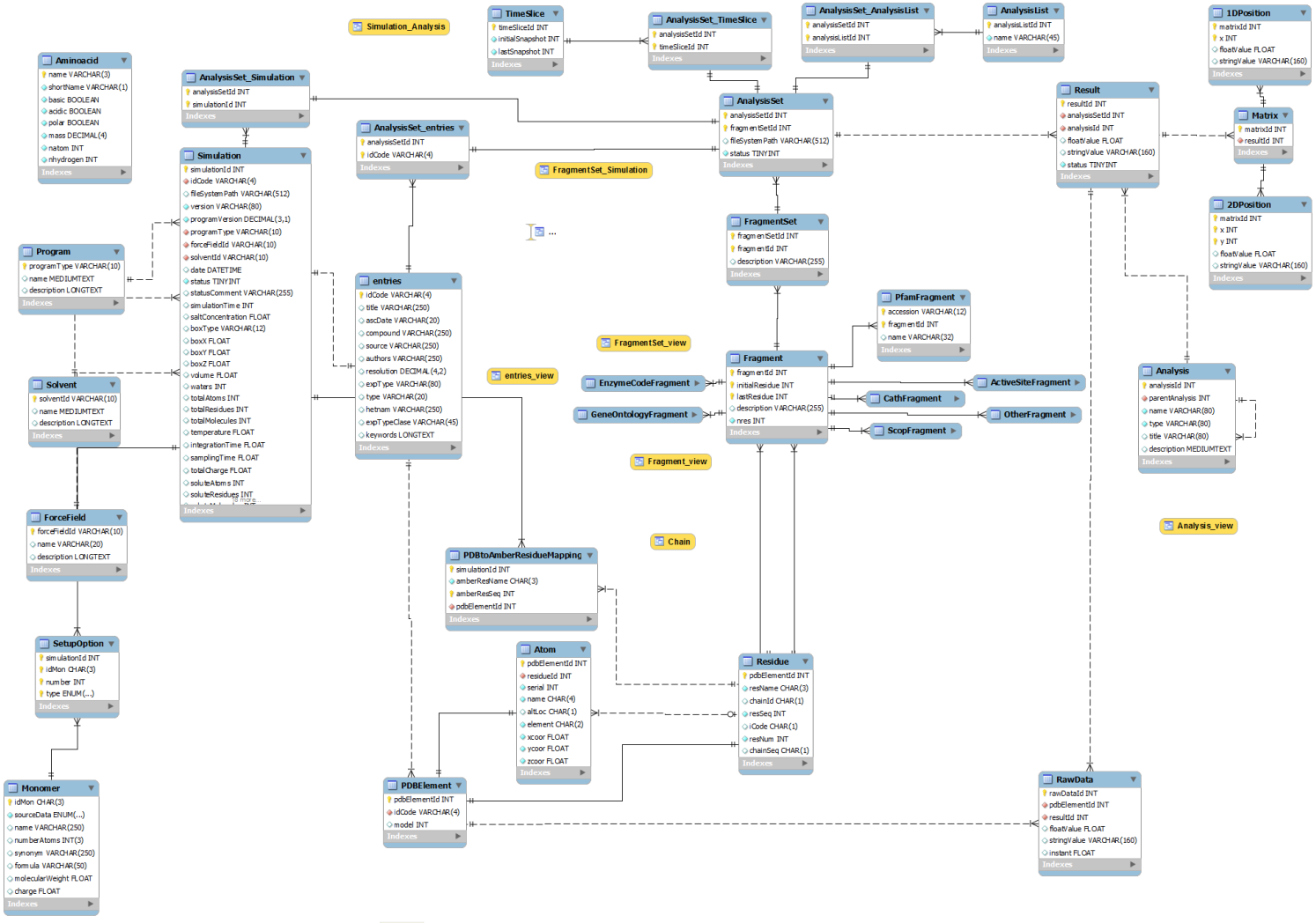


- **Rueda M . et al. PNAS 2007, 104, 798-801**
- **Meyer T. et al. Structure 2010. 18, 1399-1409**
- **http://mmb.pcb.ub.es/MoDEL**

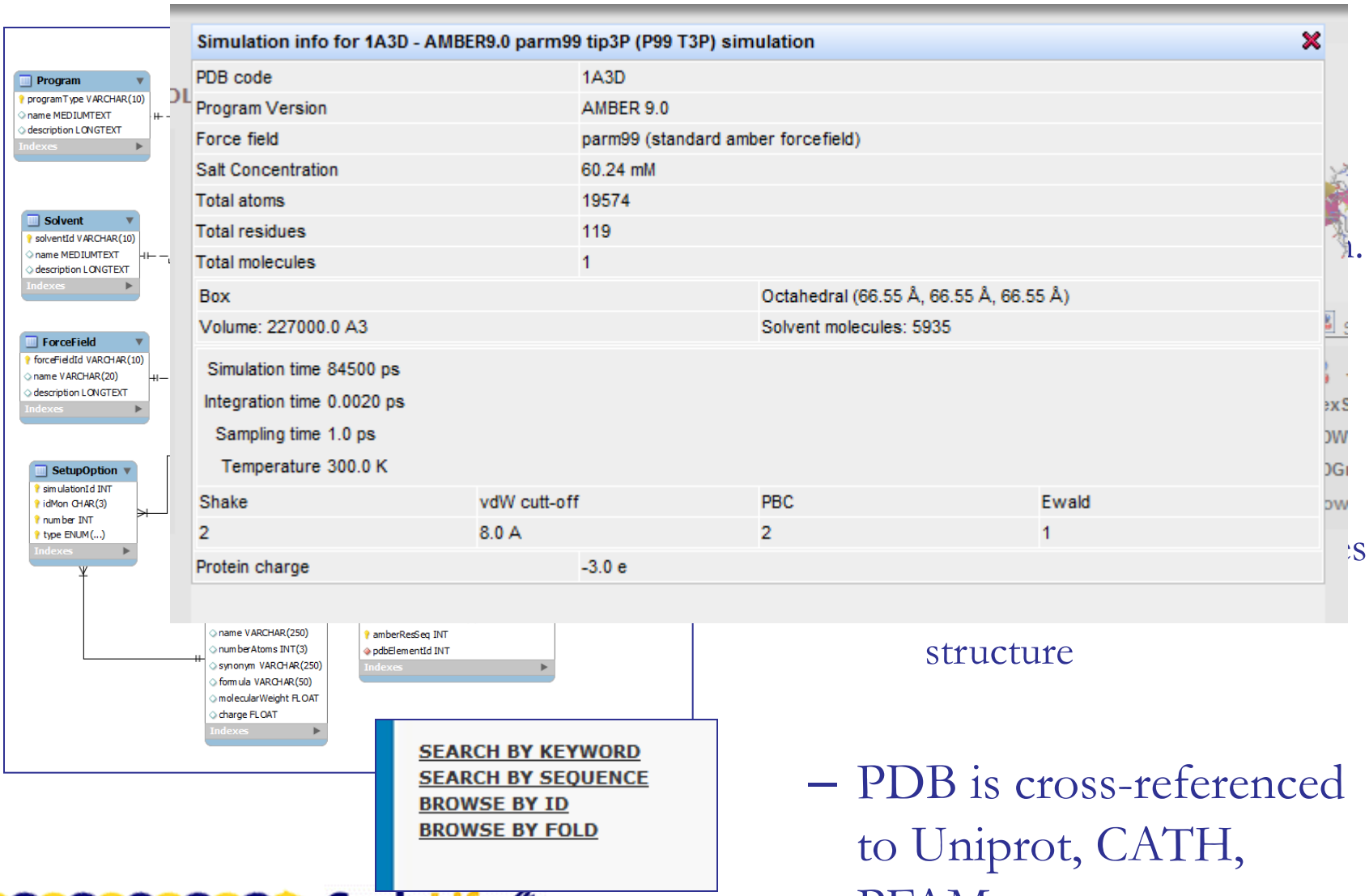
Data structure



Relational DB



Relational DB



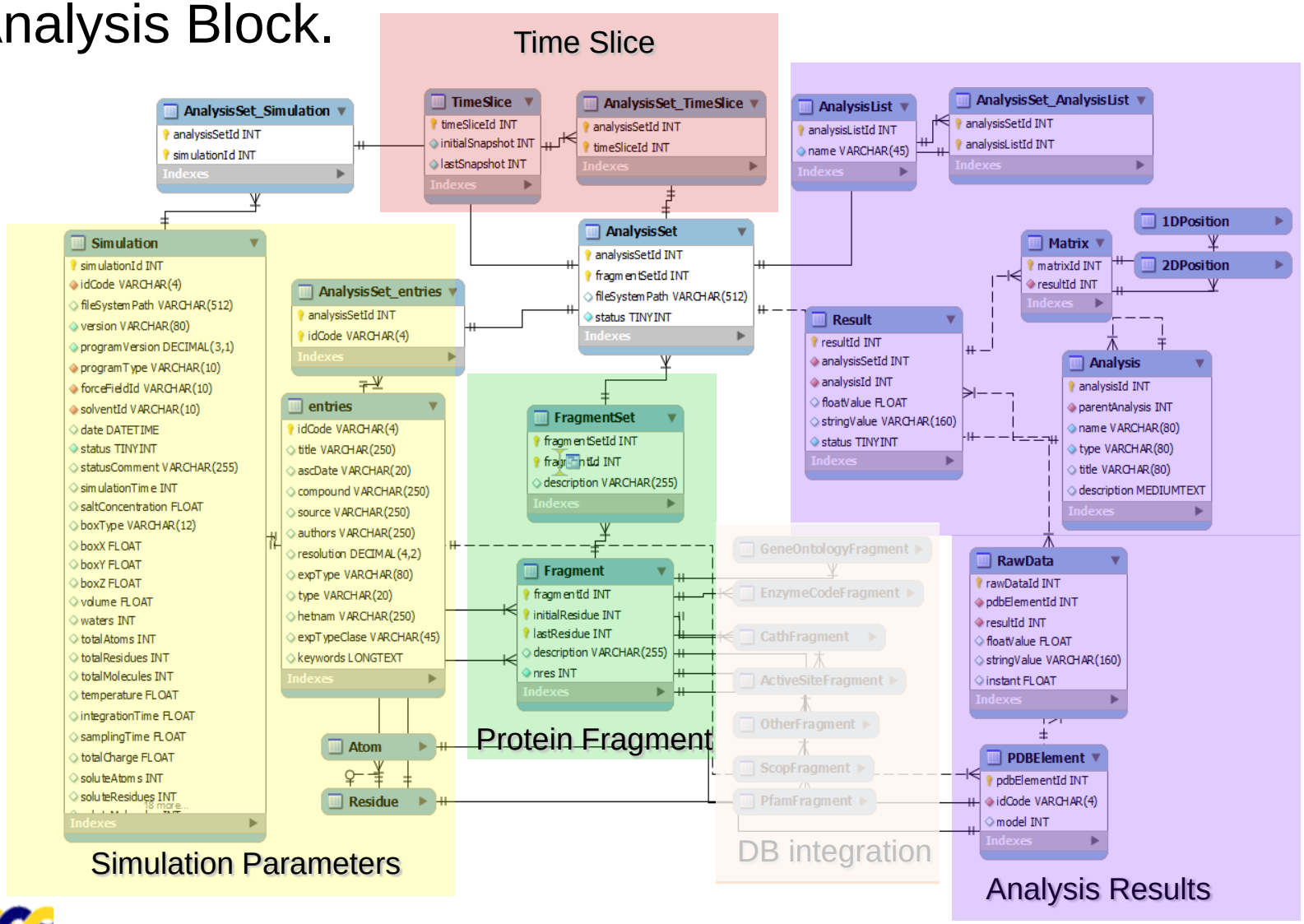
structure

– PDB is cross-referenced to Uniprot, CATH, PFAM, etc.

Analysis - DB

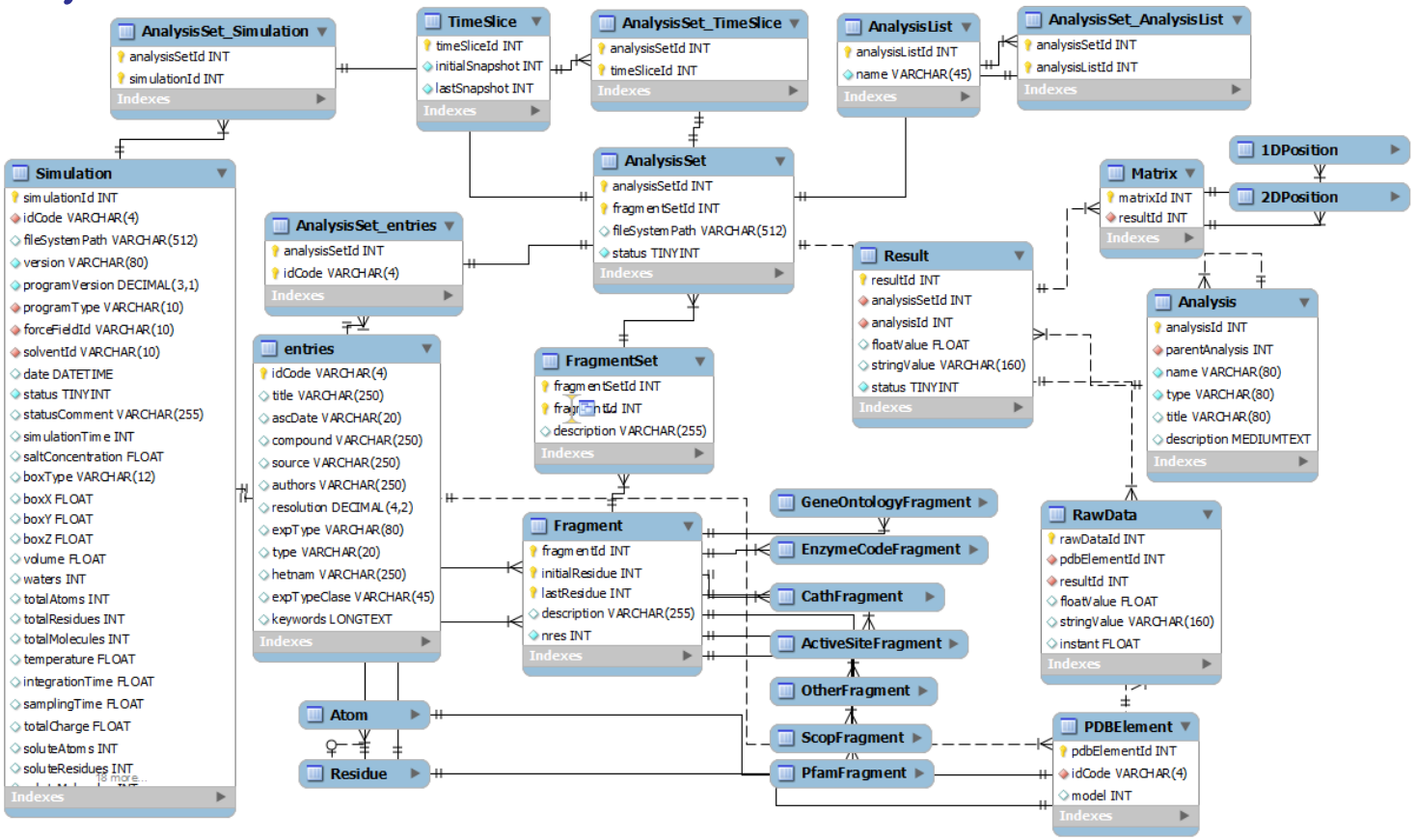
Scalable Software Services for Life Science

Analysis Block.



Relational DB

- Analysis Block

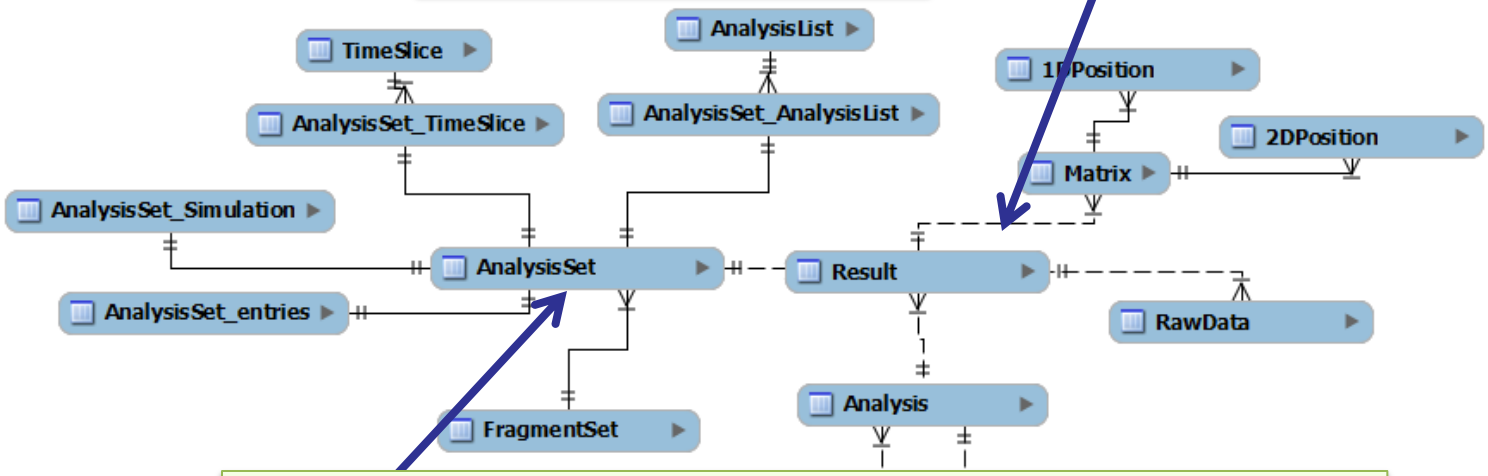


Relational DB

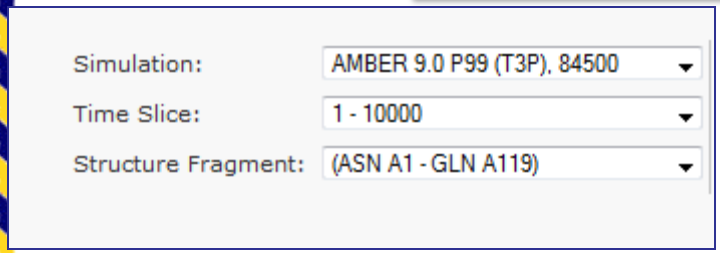
- Analysis Block. Analysis Selection

Predefined lists, possible personalization

Analysis Data. Common data structure for all analysis

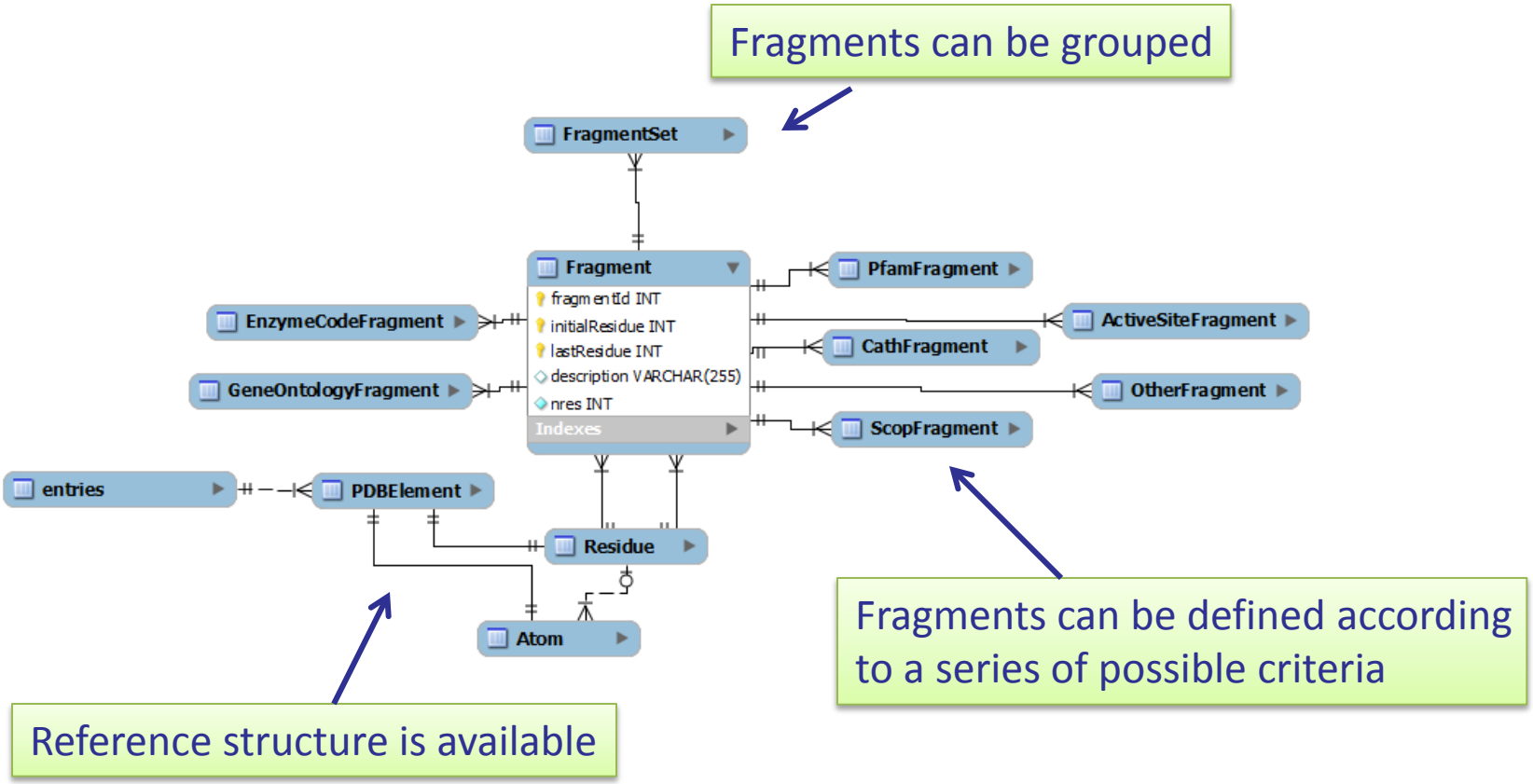


Available Analysis: Simulation + Time Slice + FragmentSet

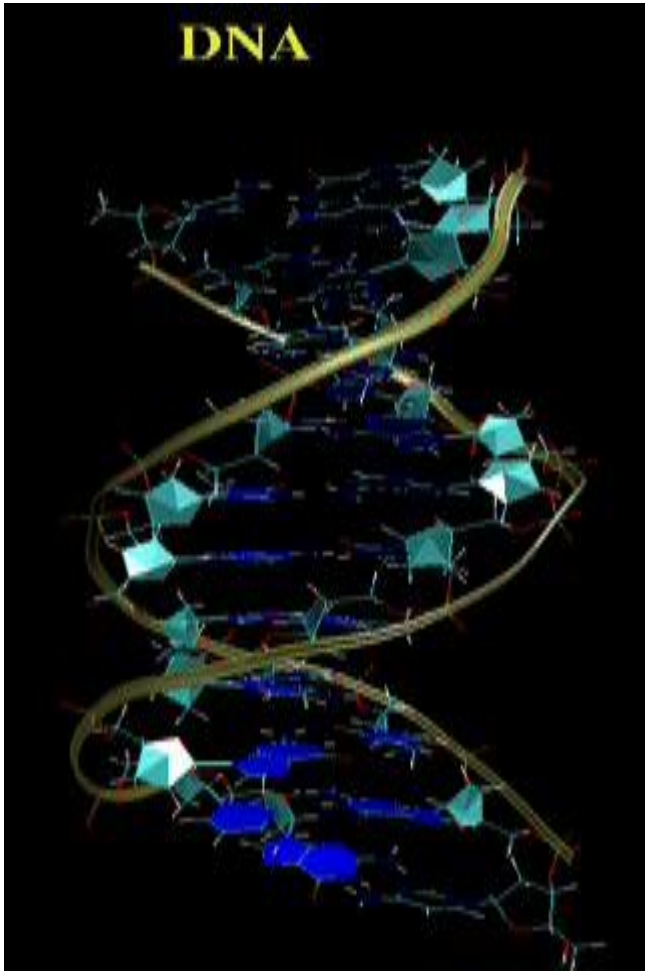


Relational DB

- Analysis Block. Fragment Selection



High-compression tools (PCAZIP)



Diagonalization of covariance matrix

Eigenvectors

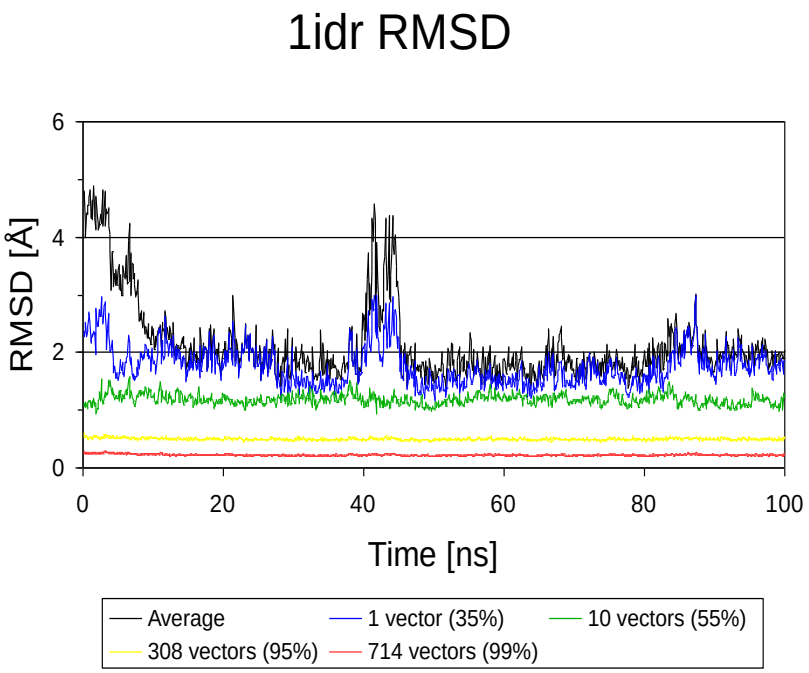
Select essential space

Quality threshold

Project cart coord Essential space

Store projections

PCAZIP data reduction



	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

Data storage & transmission

- Open questions
  - Distributed approach
  - Distributed analysis software
  - Long term storage vs recalculation
  - Compression of raw solvent trajectories
  - Integration with ELIXIR core databases