P-found GRID – A Distributed Repository for Protein Folding and Unfolding Simulations

Cândida G. Silva

Chemistry Department and

The Center for Neuroscience and Cell Biology

The Folding Problem

- Conversion of a linear sequence of amino acids into a functional tridimensional structure
- BSE, Alzheimer’s or Parkinson’s identified as Protein Folding Disorders
- What are the determinants of protein structure?
- How does a polypeptide fold to its native state?

The answer to these question may provide clues to understand diseases which appear to involve misfolding.

Computational Approaches

- Different simulation methods
  - Molecular dynamics (MD)
  - Monte Carlo based techniques
  - structure-based force fields
  - ...
- using simplified or all-atom protein representations
  - implicit or explicit solvent descriptions
  - in aqueous or organic solution, with or without co-solutes
- for different proteins
  - wild type *vs.* mutant
  - different structural classes or different topologies
  - ...
- mimicking different experimental conditions
  - temperature
  - pressure
  - pH
  - ionic strength
  - ...

An example

The Folding Problem
- Computational Approaches
- An example
- The current situation

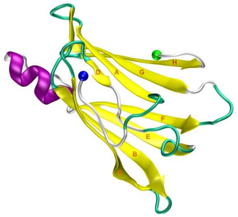
The P-found GRID project

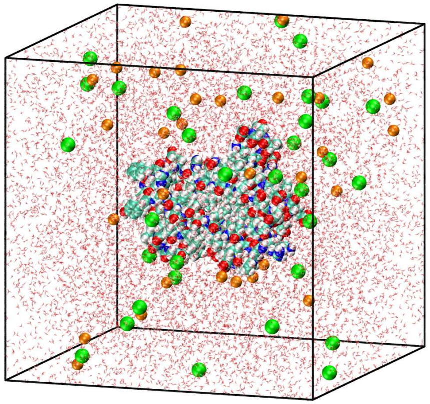
The P-found GRID Architecture

Challenges

Acknowledgements

Task Force





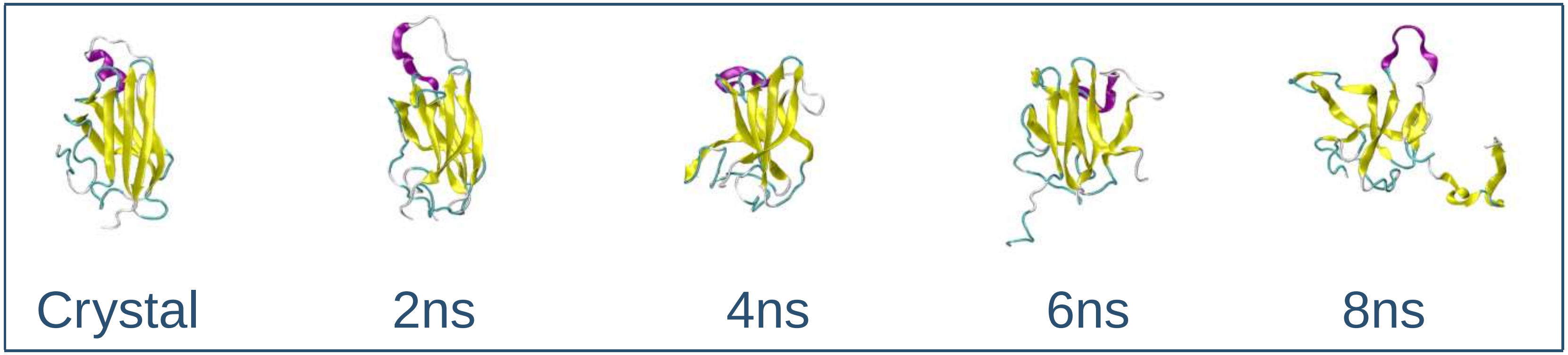
All-atom representation of the solvated L55P-TTR system used in the MD simulation

System description:

- Protein: 1912 atoms
- Water: 3*14137 atoms
- Na^+ Cl^-: 71 ions
- Total: 44394 atoms

- NAMD with CHARMM27 force field
- Simulated time: 8 nsec
- CPU time: ∼12 days/nsec/CPU (@ pentium4 Linux cluster)

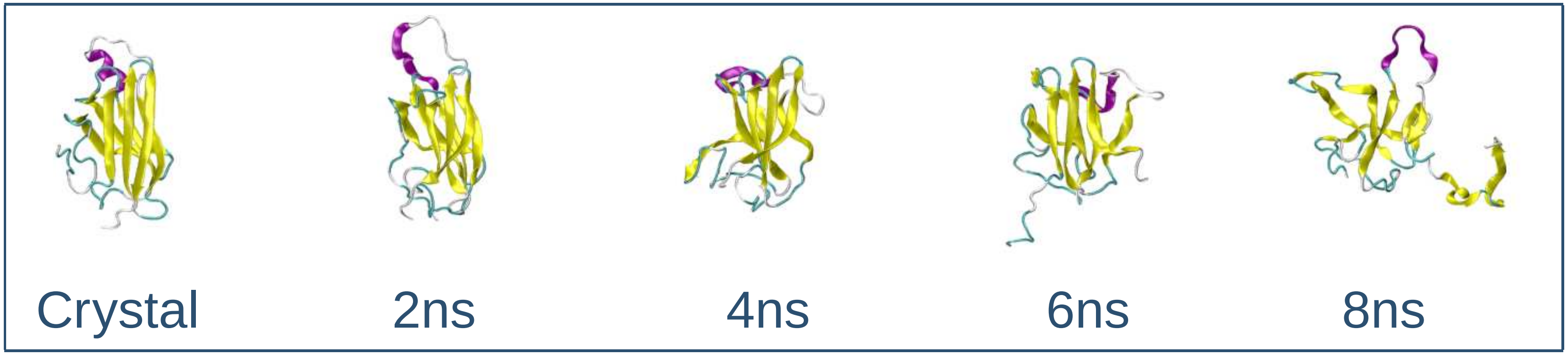
An example...



- Computation run time: 4-6 weeks using 8-12 Pentium-4 CPUs
- Binary file capturing all atoms: $\approx$ 4 GB
- Binary file capturing protein’s atoms: $\approx$ 180 MB

An example...

The Folding Problem
● Computational Approaches
● An example
● The current situation
The P-found GRID project
The P-found GRID Architecture
Challenges
Acknowledgements
Task Force



- Computation run time: 4-6 weeks using 8-12 Pentium-4 CPUs
- Binary file capturing all atoms: ≈ 4 GB
- Binary file capturing protein's atoms: ≈ 180 MB

If multiple simulations in the same or different experimental conditions are required, the data volume increases proportionally.

The current situation

The Folding Problem
- Computational Approaches
- An example
- The current situation

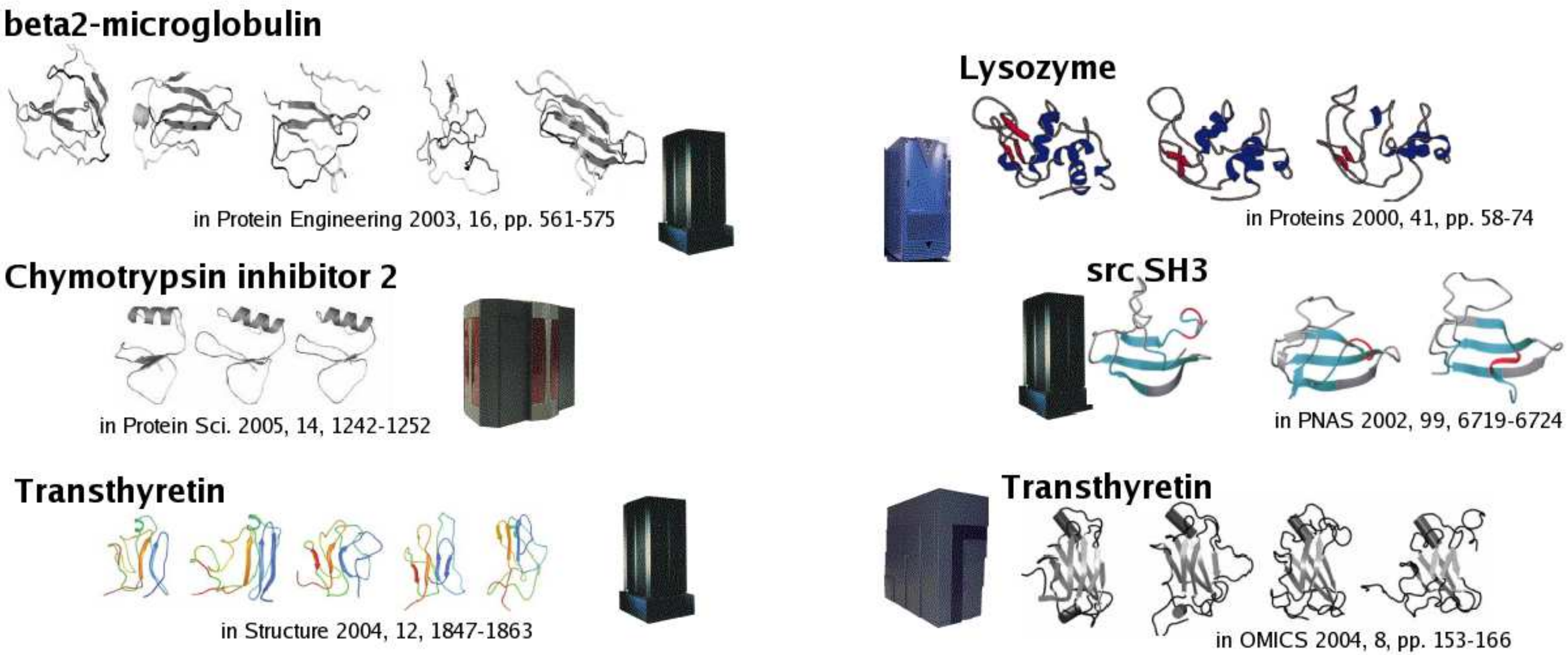
The P-found GRID project

The P-found GRID Architecture

Challenges

Acknowledgements

Task Force



beta2-microglobulin

in Protein Engineering 2003, 16, pp. 561-575

Lysozyme

in Proteins 2000, 41, pp. 58-74

Chymotrypsin inhibitor 2

in Protein Sci. 2005, 14, 1242-1252

src SH3

in PNAS 2002, 99, 6719-6724

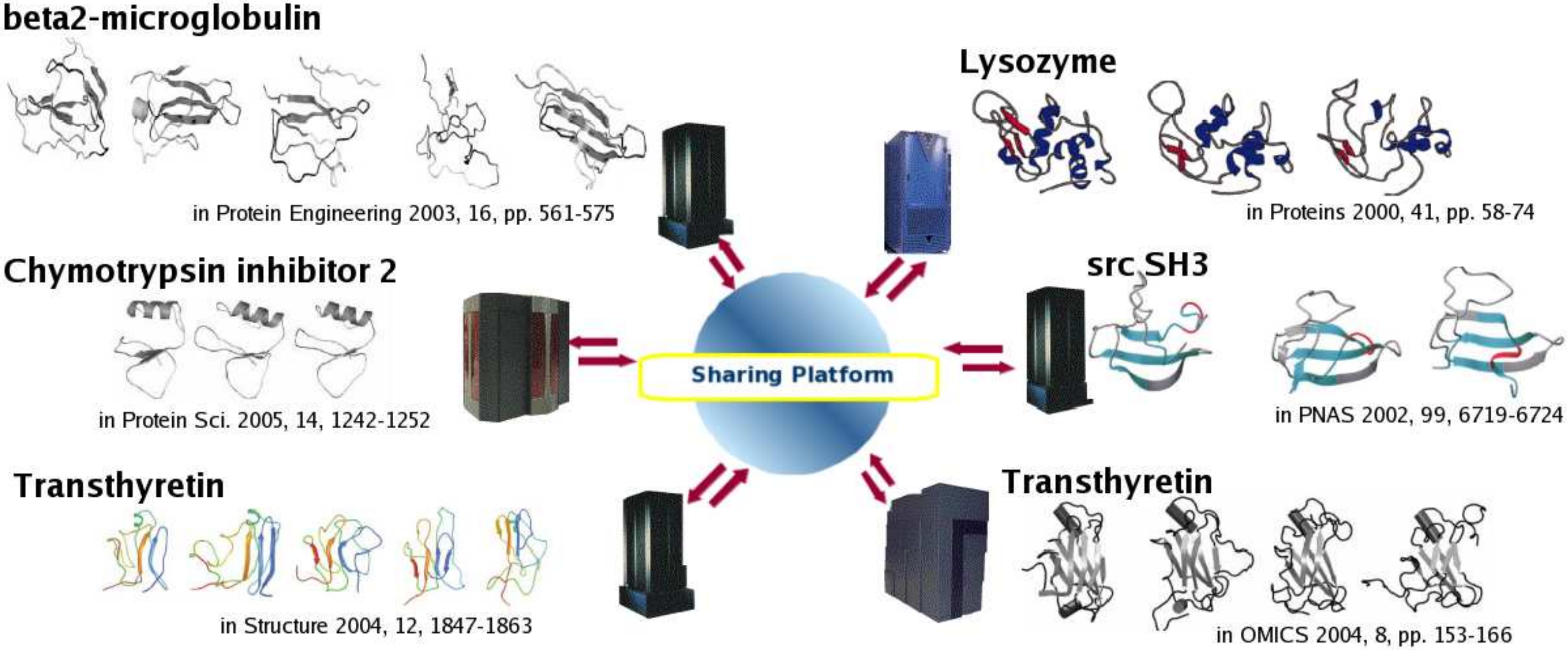
Transthyretin

in Structure 2004, 12, 1847-1863

Transthyretin

in OMICS 2004, 8, pp. 153-166

The current situation



beta2-microglobulin
in Protein Engineering 2003, 16, pp. 561-575
Lysozyme
in Proteins 2000, 41, pp. 58-74
Chymotrypsin inhibitor 2
in Protein Sci. 2005, 14, 1242-1252
src SH3
in PNAS 2002, 99, 6719-6724
Sharing Platform
Transthyretin
in Structure 2004, 12, 1847-1863
Transthyretin
in OMICS 2004, 8, pp. 153-166

The P-found GRID project

The Folding Problem

The P-found GRID project

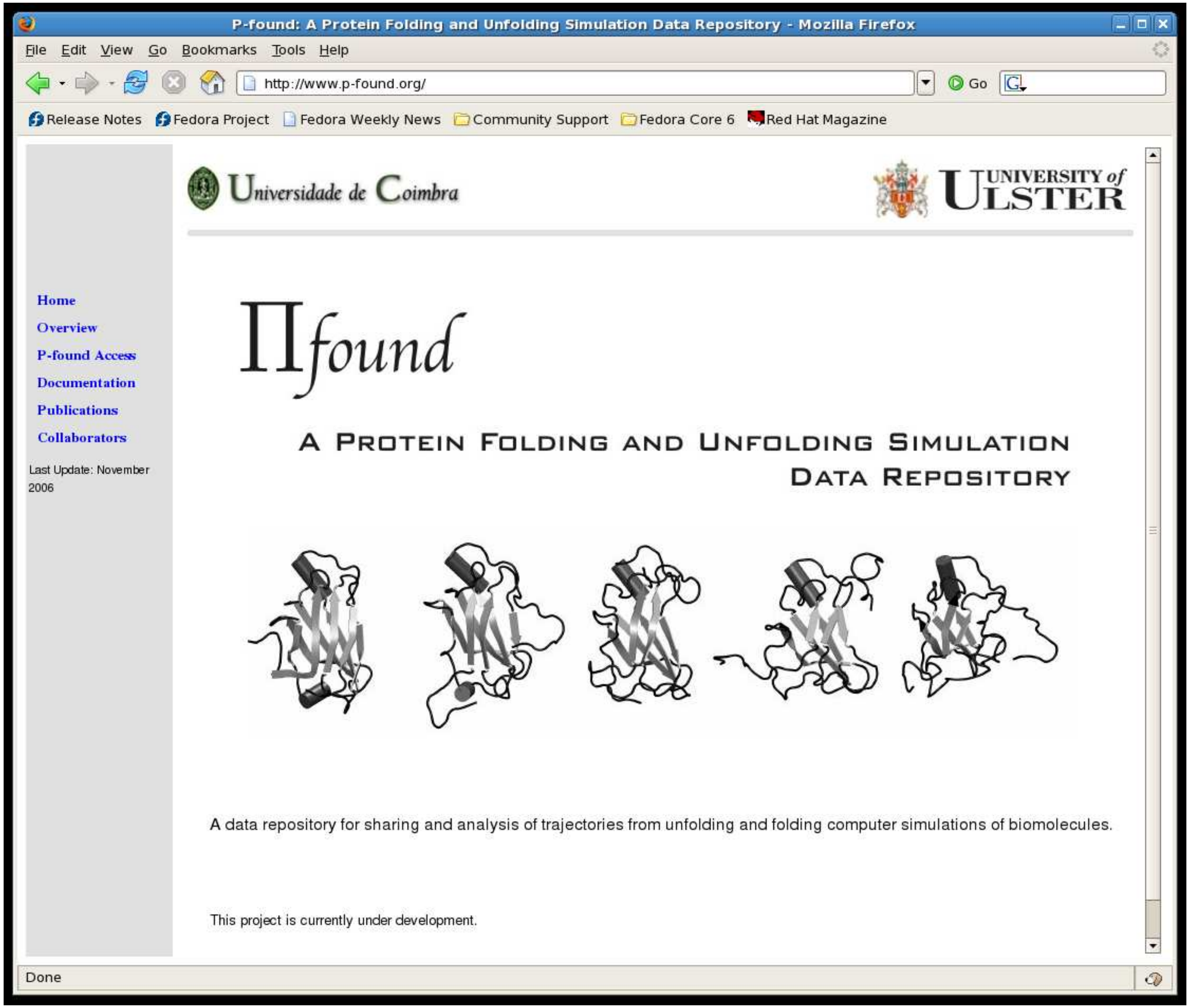
- Objectives
- The data
- User profiles

The P-found GRID Architecture

Challenges

Acknowledgements

Task Force



Last Update: November 2006

Universidade de Coimbra

UNIVERSITY of ULSTER

Πfound

A PROTEIN FOLDING AND UNFOLDING SIMULATION DATA REPOSITORY

A data repository for sharing and analysis of trajectories from unfolding and folding computer simulations of biomolecules.

This project is currently under development.

Objectives

1. Sharing of simulation data
   - Raw simulation data
   - Calculated molecular property data
   - Provenance data
   - Metadata
2. Analysis and data mining of molecular property data
3. Dynamic deployment and application of proprietary programs for calculating molecular properties and for analyzing molecular property data

The data

The Folding Problem

The P-found GRID project

- Objectives
- The data
- User profiles

The P-found GRID Architecture

Challenges

Acknowledgements

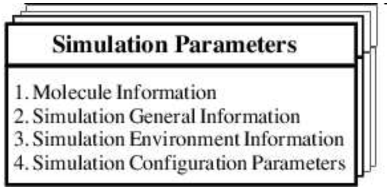
Task Force

Simulation Raw Data

Record the atomic positions of all atoms in the protein along the trajectory

Derived molecular property data

Represent different molecular properties of the protein simulation



Simulation Parameters

1. Molecule Information
2. Simulation General Information
3. Simulation Environment Information
4. Simulation Configuration Parameters

Provenance data

Record the parameters of the processes, tools and other aspect which led to the creation of the simulation raw data

Metadata

Convey the content and structure of the repository to users so that they can efficiently navigate and use P-found.

User profiles

The Folding Problem

The P-found GRID project

- Objectives
- The data
- User profiles

The P-found GRID Architecture

Challenges

Acknowledgements

Task Force

Information users

Browse the data stored

Perform searches

Visualize graphical representations of the molecular properties data

User profiles

Information users

- Browse the data stored
- Perform searches
- Visualize graphical representations of the molecular properties data

Data consumer users

- Download molecular properties data

User profiles

The Folding Problem

The P-found GRID project

- Objectives
- The data
- User profiles

The P-found GRID Architecture

Challenges

Acknowledgements

Task Force

Information users

Browse the data stored

Perform searches

Visualize graphical representations of the molecular properties data

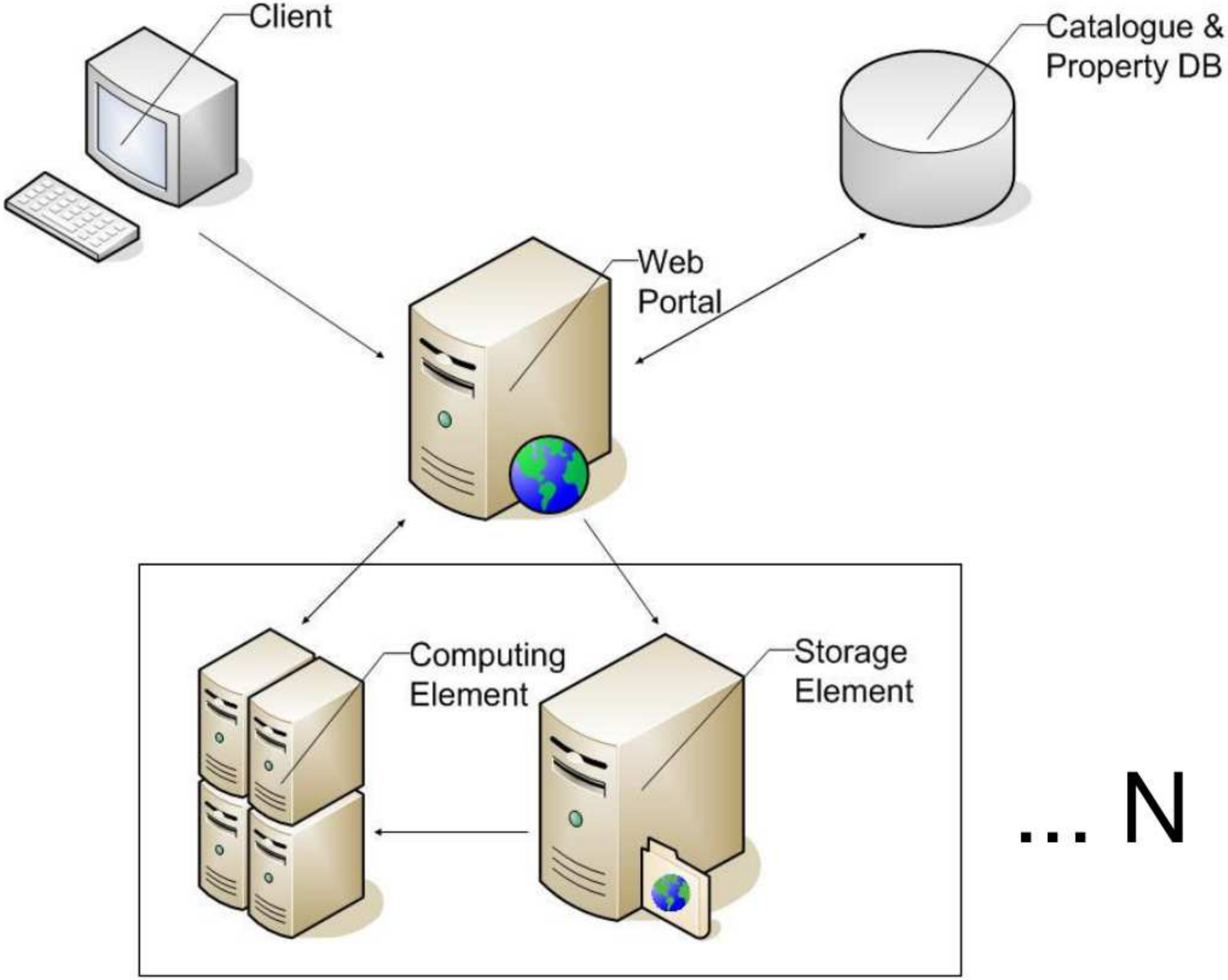
Data consumer users

Download molecular properties data

Data provider users

Upload simulation data

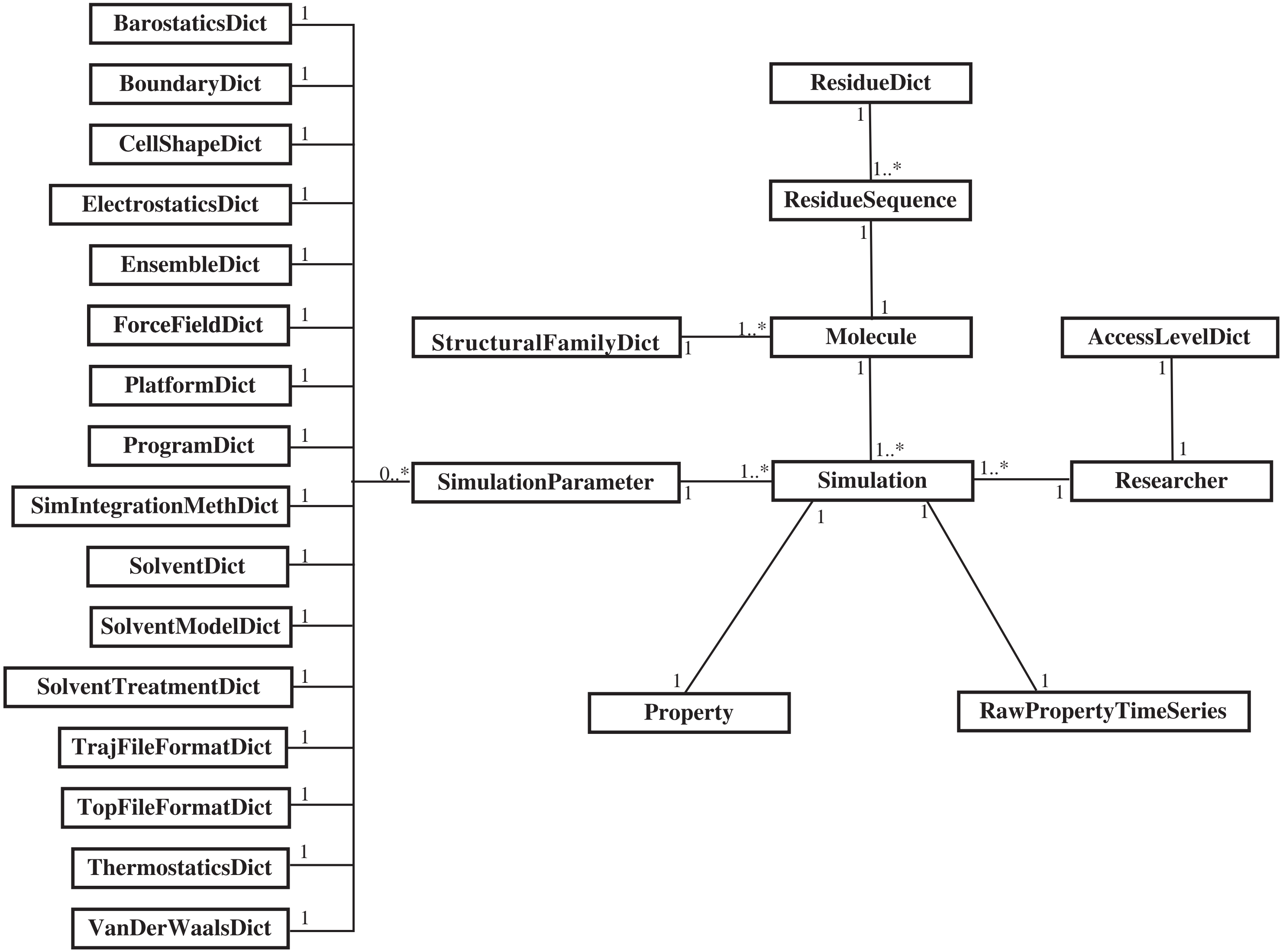
The P-found GRID Architecture



The Catalogue & Property DB

- Stores four different types of data
  - Simulation files catalogue
  - Molecular property data
  - Simulation description data
  - P-found GRID management information
- Implemented in PostgreSQL
  - Powerful, open source relational database system
  - Strong reputation for reliability, data integrity, and correctness
  - Supported within the **Globus Toolkit Framework**

The Catalogue & Property DB Model



Storage and Computing Elements

The Folding Problem

The P-found GRID project

The P-found GRID Architecture

- The Catalogue & Property DB
- The Catalogue & Property DB Model
- Storage and Computing Elements
- The P-found GRID Web Portal
- The P-found GRID Application

Challenges

Acknowledgements

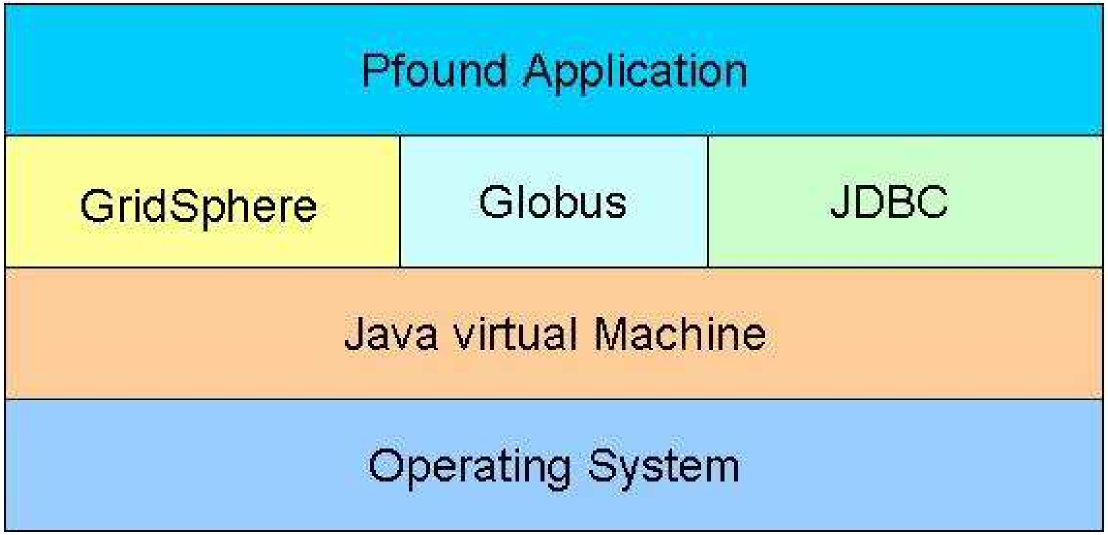
Task Force

- Modular components of the P-found GRID system
- Storage Element
  - Stores simulations raw data
  - Globus Toolkit 4.0 (GridFTP)
- Computing Element
  - Computation of molecular properties
  - Geographically close to simulation data
  - Globus Toolkit 4.0 (GRAM)
  - VMD

The P-found GRID Web Portal

- Provides a friendly interface between the end-user and the P-found GRID system
  - Coordinates the submission process of a new simulation
    - Input of simulation parameters
    - Upload of files
    - Standard moelcular properties calculation, job submission and gathering
  - Browse simulation and properties
  - Coordinate other properties generation
  - Allow data mining on properties and files
- Developed within the Gridsphere web portal framework

The P-found GRID Application



Challenges for the future

- Global accessibility to the data repository
- Development of new data mining tools for study and comparison of multiple simulations
- Prepare the system to accommodate simulation for methods other than molecular dynamics

Challenges for the future

- Global accessibility to the data repository
- Development of new data mining tools for study and comparison of multiple simulations
- Prepare the system to accommodate simulation for methods other than molecular dynamics

Identification of high-level rules for discrimination among folding and unfolding processes in amyloidogenic and different structural classes of proteins

Acknowledgements

VMD Visual Molecular Dynamics

John Stone

BioCoRE Biological Collaborative Environment

Kirby Vandivort

FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

SFRH/BD/16888/2004 (PhD scholarship)
GRID/GRI/81809/2006 (Project I & D - *Iniciativa Nacional GRID*)

Task Force

University of Coimbra

Department of Chemistry and CNC

Rui M. M. Brito

Cândida G. Silva

Nuno Loureiro-Ferreira

Carlos J. V. Simões

Department of Physics and LCA

Pedro Vieira Alberto

Miguel Afonso Oliveira

Department of Informatics Engineering

Pedro Furtado

Ricardo Antunes

João Pedro Costa

University of Ulster

School of Biomedical Sciences

Werner Dubitzky

Vitaliy Ostropytskyy

Martin Swain

Olivier Riché

Daniel Berrar

Critical Software, S.A.

Nuno Cunha

Sérgio Cruz

João Brito

Sérgio Carvalho

University of Minho

Department of Informatics

Paulo J. Azevedo

João Luís Sobral

University of Porto

Faculty of Engineering

Rui Camacho