

WP4: Case Study

Virtual Laboratory for Protein Structure Prediction

Sara Montagna
`sara.montagna@unibo.it`

Ingegneria Due
ALMA MATER STUDIORUM—Università di Bologna a Cesena

Modena 30th June, 2008



- 1 Overview
- 2 The biological problem
 - The biology of protein folding
 - Protein structure prediction
- 3 The VLab
 - An AOSE perspective
 - The VLab Requirements
- 4 A scenario
 - An examples of the system's structure
- 5 Conclusions



MEnSA project - Work Package 4

Objective

- Engineering a case study for
 - investigating and validating the methodologies and infrastructures
 - verifying the effectiveness and usability of the abstractions and tools identified
 - stressing the synergy between the methodological framework and the developed infrastructure



Virtual Laboratory (VLab) for Protein Structure Prediction

- Bioinformatics domains of interest
 - proteomic data analysis
- Towards a Virtual Lab
 - distributed execution of in-silico experiments
 - automatic analysis of data
 - management of large distributed data set
 - distributed cooperation and coordination among heterogeneous actors (biologists, modellers...)
- The Virtual Lab is a software systems with complex features
- Agents and MAS
 - for modelling and engineering this complex Virtual Lab



Outline

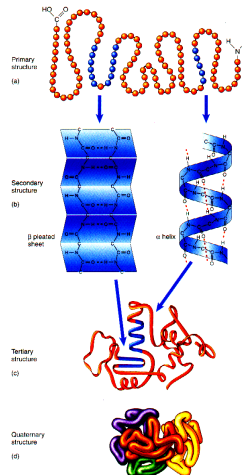
- 1 Overview
- 2 The biological problem
 - The biology of protein folding
 - Protein structure prediction
- 3 The VLab
 - An AOSE perspective
 - The VLab Requirements
- 4 A scenario
 - An examples of the system's structure
- 5 Conclusions



The Biological Process: Protein Folding

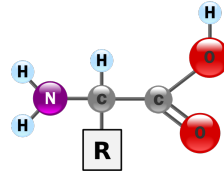
Protein folding is the physical process by which a polypeptide folds into its characteristic three-dimensional structure [Wikipedia]

- Protein structures
 - primary structure
 - secondary structure
 - tertiary structure
 - quaternary structure

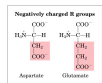
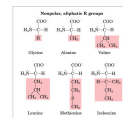
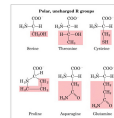
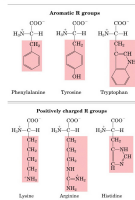


Amino Acids Are the Subunits of Proteins

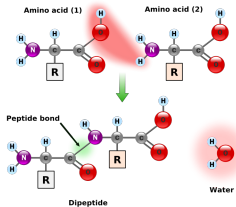
- Basic building blocks of a protein
- All composed by:
 - carboxyl group (COOH)
 - amino group (H₂N)
 - side chain (R)
- 20 different types with different side chains along with their chemical properties



- weak acids, weak bases, hydrophiles, hydrophobes



The Primary Structure



The *primary structure* is a long chain of amino acids linked together by peptide bonds.

This polypeptide constitutes the protein specific primary sequence

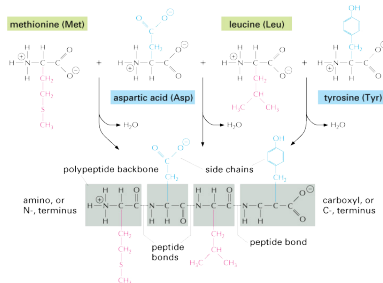


Figure 4-2 Essential Cell Biology, 2/e. (© 2004 Garland Science)

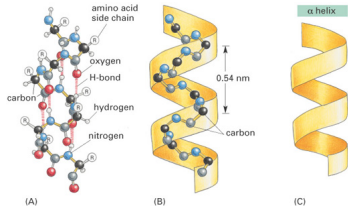


The Secondary Structure

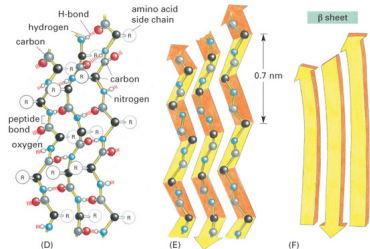
The ordered sequence of amino acids in a protein confer regular conformational forms upon that protein.

*The **secondary structure** is the general three-dimensional form of local segments of the protein polypeptidic sequence*

• α -helix



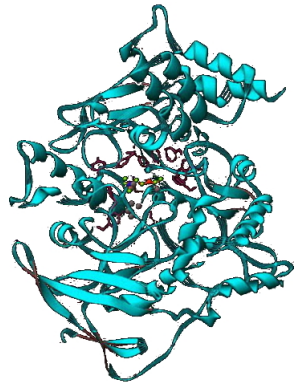
• β -sheet



The Tertiary Structure

*The **tertiary structure** refers to the complete three-dimensional structure of the polypeptide units of a given protein.*

- Spatial relationship of different secondary structures to one another within a polypeptide chain
- How these secondary structures themselves fold into the three-dimensional form of the protein.



Summerising: Protein Structure

- The shape of a protein is specified by its amino acid sequence
- Protein folds into a conformation of lowest energy
- The α -helix and the β -sheet are common folding patterns
- The three dimensional structure represents a conformation with particular chemical properties



Outline

- 1 Overview
- 2 The biological problem
 - The biology of protein folding
 - Protein structure prediction
- 3 The VLab
 - An AOSE perspective
 - The VLab Requirements
- 4 A scenario
 - An examples of the system's structure
- 5 Conclusions



The Importance of the Three-Dimensional Structure

The chemical properties are essential for protein function but...

- ...The folding mechanism is still not clear
 - the number of possible protein structures is extremely large
 - the physical basis of protein structural stability is not fully understood
 - the experimental determination of the three dimensional structure of a protein is often very difficult and expensive
- The primary sequence of the protein is often known

Protein structure prediction is of high importance.



Bioinformatics: Application of Computational Techniques

Protein structure prediction methods (manly from AI)

- Search strategies along the space of possible structures (search strategy and evaluation function)
- Neural networks for pattern classification

Problems

- The problem size is very high: thousands, millions, billions of possibles structures
 - best the use of metaheuristic (Tabu Search, Simulated Annealing, Evolutionary Computation, Ant Colony Optimization) with the support of physical, chemical or biological informations
- The evaluation function is extremely difficult to be calculated accurately
 - use of highly simplified models of biological phenomena

The VLab

- System that can support biologists and bioinformatics in the problem of protein structure prediction.
 - sharing the data
 - integration of multiple users and algorithms efforts
 - interactive and cooperative work



Outline

- 1 Overview
- 2 The biological problem
 - The biology of protein folding
 - Protein structure prediction
- 3 The VLab
 - An AOSE perspective
 - The VLab Requirements
- 4 A scenario
 - An examples of the system's structure
- 5 Conclusions



Why this Case Study for MEnSA Project

- The system is massively knowledge-based, as it requires the definition of specific and rigorous ontologies
- The system requires the management of multiple processes and tasks for the elaboration of the knowledge and data at different levels:
 - integrating heterogeneous skills and methods
 - enabling the interaction with human users who are involved in the processes as external observers, controllers and analysts.
- Works on this line [Addis et al., 2007] [Bortolussi et al., 2007]



Outline

- 1 Overview
- 2 The biological problem
 - The biology of protein folding
 - Protein structure prediction
- 3 The VLab
 - An AOSE perspective
 - The VLab Requirements
- 4 A scenario
 - An examples of the system's structure
- 5 Conclusions



First Requirement

Hybrid or parallel algorithms have improved search patterns obtaining reasonably good approximate solutions [Milano and Roli, 2004].

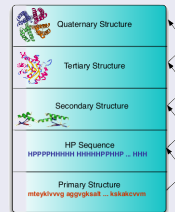
Implementation of methods to

- Define the models and strategies (search strategies, neural net...) parameters and tuned them through a learning process
 - Integrate models and algorithms
 - Define the problem's model
 - a set of variables, $X = [x_1, \dots, x_n]$
 - variable domains, $D_i \forall i = 1, \dots, n$
 - constraints among variables
 - an *evaluation function*
- Elaborate the knowledge
- Produce a possible candidate solution
- Identify new information on the protein structure

Second Requirement

Organization and representation of data and knowledge

- Shared data structure where informations
 - are accessible by all the users
 - are organized in hierarchical levels
- At each level
 - data contain more and more informations
 - data are partial solutions to the problem



Third Requirement

Human users interaction

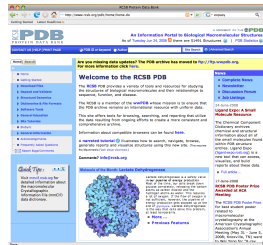
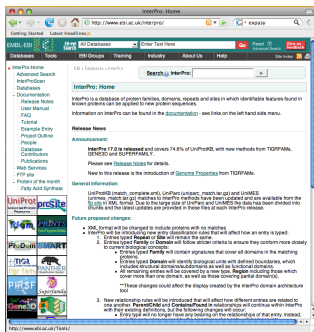
- Interface the VLab with biologists, bioinformatics for
 - adding knowledge on the protein's properties
 - giving feedbacks on the partial solutions collected
 - selecting the algorithms
 - designing models and search strategies



Fourth Requirement

Internet resources exploitation

- Acquire knowledge and data spread among different Internet sources
- Filter and organise them on the shared data structure



How the Overall System Looks Like

- A set of independent modules, named knowledge sources, which contain specific knowledge about a problem domain.
- A shared data structure through which the knowledge sources communicate to each other.
- A set of algorithms and models for the elaboration of the knowledge.
- A set of functionality for allowing the activity of human users (the experts).
- A control mechanism which determines the order in which the knowledge sources will operate on the data.



Outline

- 1 Overview
- 2 The biological problem
 - The biology of protein folding
 - Protein structure prediction
- 3 The VLab
 - An AOSE perspective
 - The VLab Requirements
- 4 A scenario
 - An examples of the system's structure
- 5 Conclusions

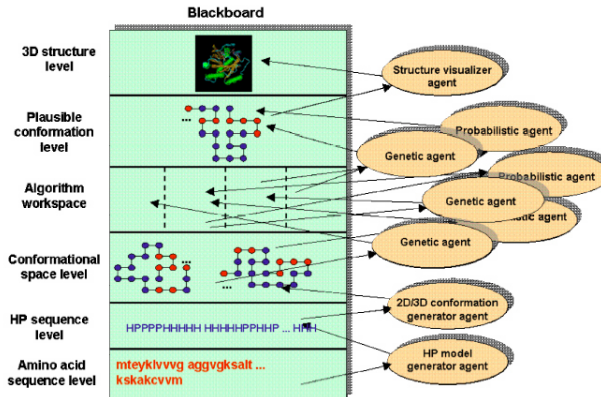


Shared Data Structure Organisation

- At each level incorporation of biological knowledge, solutions provided by algorithms, feedbacks from experts on
 - first knowledge on the primary sequence of the protein
 - second hydrofobic-hydrophilic (HP) model
 - third n random protein conformations on 2D or 3D lattice
 - fourth small set of plausible protein conformations
 - fifth the optimal protein conformation

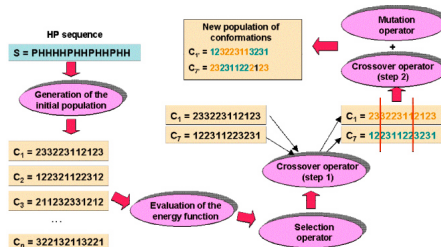


How the VLab Looks Like



Metaheuristic for Protein Folding Prediction

Genetic algorithms



- Exploration of the molecular energy surface
- Optimization of the energy function E
- Identification of the best set of parameters
 - population size
 - mutation
 - recombination
 - local search operators



Conclusions

Already done

- The biological problems and the necessity of a VLab
- The Vlab's requirements

To do

- Refining the system's requirements according to your feedbacks
- Representing the case study with the different methodologies
- ... Doing the same job with the **new** meta-model and methodology



Bibliography



Addis, A., Armano, G., Mascia, F., and Vargiu, E. (2007).

Protein secondary structure prediction through a cooperative multiagent learning approach.

System and Information Sciences Notes, 2(1):122–125.



Bortolussi, L., Dovier, A., and Fogolari, F. (2007).

Agent-based protein structure prediction.

Journal of Multi Agents and Grid Systems, 3(2).



Milano, M. and Roli, A. (2004).

Magma: a multiagent architecture for metaheuristics.

IEEE Transactions on Systems, Man, and Cybernetics, Part B, 34(2):925–941.



WP4: Case Study

Virtual Laboratory for Protein Structure Prediction

Sara Montagna
`sara.montagna@unibo.it`

Ingegneria Due
ALMA MATER STUDIORUM—Università di Bologna a Cesena

Modena 30th June, 2008

