# Fragment-Based Protein Structure Prediction and Design

David Simoncini

Mathématiques et Informatique appliquées de Toulouse, INRA

November 19, 2014

# Background

### PhD. Thesis, Computer Science

- University of Nice-Sophia Antipolis, France
- Selective pressure in cellular evolutionary algorithms

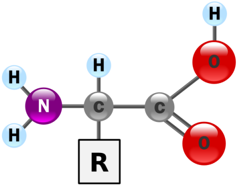
### Postdoctoral fellowship

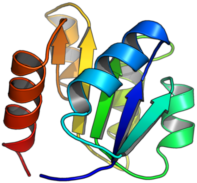
- Zhang IRU, RIKEN, Japan
- Fragment-based protein structure prediction
- Computational protein design

# What are proteins ?

WORLD CLASSIC BODYBUILDING
SOFIA · BULGARIA · NOVEMBER 24-25TH
IFBB 2012
СВЕТОВНО ПЪРВЕНСТВО ПО КЛАСИЧЕСКИ КУЛТУРИЗЪМ
IFBB

# What are proteins ?





## Amino acids (and primary structure)

Building blocks of proteins. A sequence of amino acids is called **primary structure** of a protein. Each amino acid in the sequence is called a **residue**.

## Secondary structure

There are two types of recurrent interactions between residues which lead to two regular patterns in the structure of a protein:
$\alpha$ **helices** and $\beta$ **sheets**

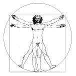
## Tertiary structure

Oh well, the 3D structure is called **tertiary structure**.

# Role of proteins

## In our body

- Enzymes
- Cell signaling and ligand binding
- Structural proteins



## In our industry

- Biofuel production
- Wastewater treatment

## In our diseases

- Alzheimer
- Parkinson
- Type II diabetes
- Cancer

# Experimental determination of protein structures

## Nucleic Magnetic Resonance spectroscopy

- Establish restraints between atoms (distance, angle, orientation)
- Build models that satisfy the restraints

## X-ray crystallography

- Crystallize a protein
- Shoot it with X-rays and get a diffraction map
- Recover the phase of the structure factors, and solve the structure

All the solved protein structures are centralized in the Protein DataBank (PDB).

# Protein structure prediction

All the information that is needed to determine the tertiary structure of a protein is contained in its sequence.

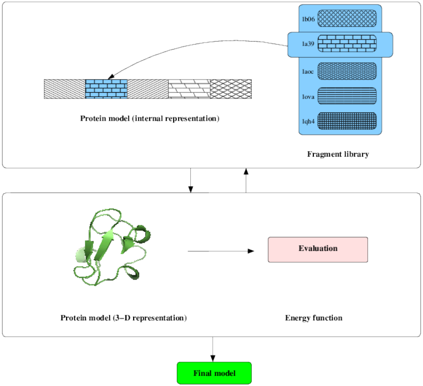
## Comparative modeling

- Identify an homologous protein by sequence alignment
- Generate models with the homologous protein as template

## Ab initio modeling (or free modeling)

Start "from scratch". Fold the models starting from the protein in its extended state, using only its sequence as information

# Fragment-based approaches

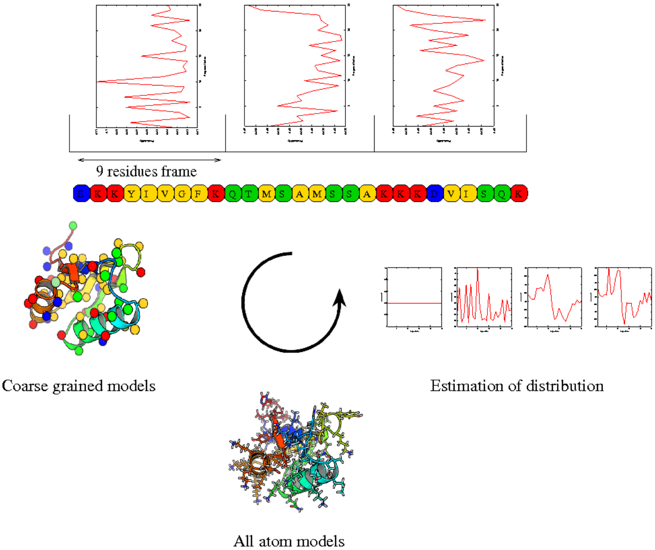


1b06
1a39
1aoc
1ova
1qh4
Protein model (internal representation)
Fragment library
Evaluation
Protein model (3–D representation)
Energy function
Final model

## How does it work ?

- Cut target sequence in small overlapping pieces
- Find fragments of structures in the PDB that match the small pieces
- Build a library of fragments (usually 25 per fragment window)
- Assemble the fragments and minimize an energy function

# EdaFold



9 residues frame

Coarse grained models

Estimation of distribution

All atom models

# Experiments

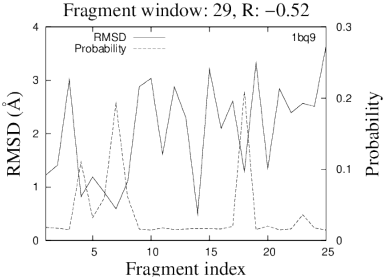
## Benchmark

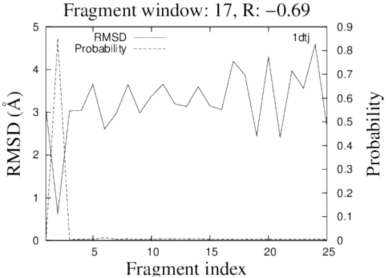
- 20 proteins
- 200,000 models generated by protein
- 4 iterations (50,000 models generated at each iteration)
- 200,000 models generated with Rosetta Modeling Suite for comparison

## How do we evaluate

- We analyze the impact of the EDA
- We look at the energy of the models
- We look at how close we are from the native structure (the real structure)

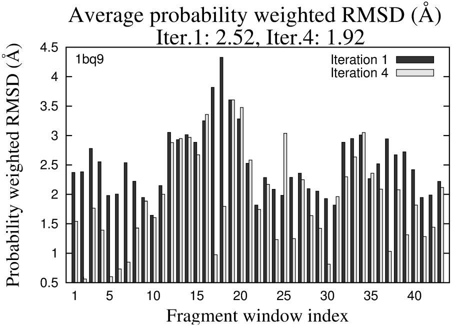
# Probabilities of selecting fragments





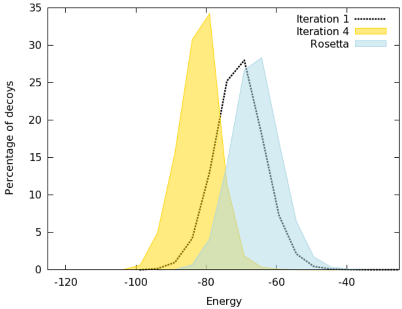
One window contains 25 fragments. The Root Mean Square Deviation (RMSD) of each fragment to the native one is plotted as well its probability of being selected.

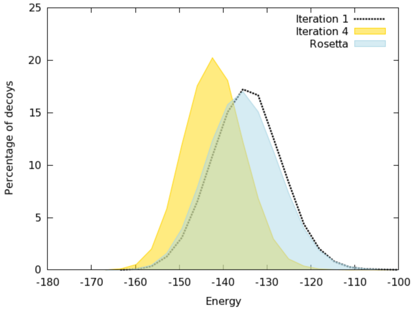
# Probabilities of selecting fragments



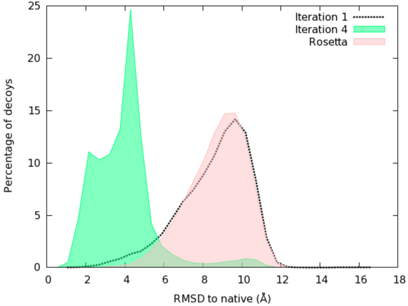
The RMSD of each fragment to the native one is weighted by its probability of being selected. We plot the average over 25 fragments for each window.

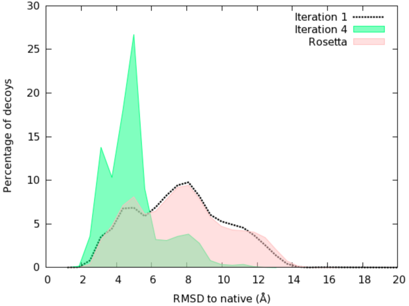
# Energy minimization



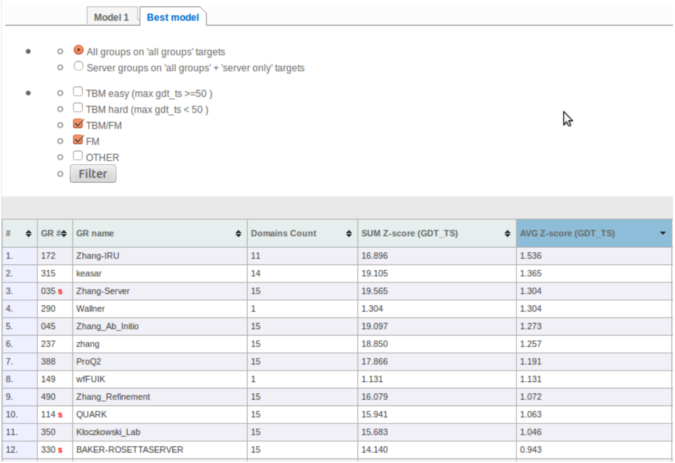


# RMSD to native structure





# CASP 10

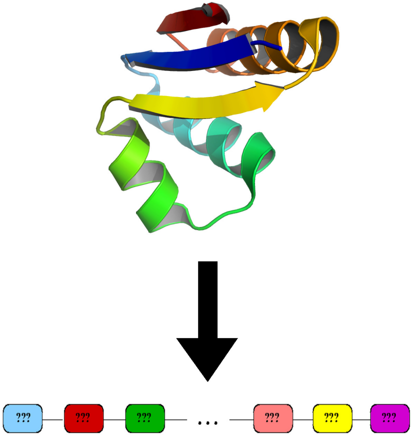


- ● All groups on 'all groups' targets
  - ○ Server groups on 'all groups' + 'server only' targets
- ☐ TBM easy (max gdt_ts >=50 )
  - ☐ TBM hard (max gdt_ts < 50 )
  - ☑ TBM/FM
  - ☑ FM
  - ☐ OTHER

| # | GR # | GR name | Domains Count | SUM Z-score (GDT_TS) | AVG Z-score (GDT_TS) |
|---|---|---|---|---|---|
| 1. | 172 | Zhang-IRU | 11 | 16.896 | 1.536 |
| 2. | 315 | keasar | 14 | 19.105 | 1.365 |
| 3. | 035 s | Zhang-Server | 15 | 19.565 | 1.304 |
| 4. | 290 | Wallner | 1 | 1.304 | 1.304 |
| 5. | 045 | Zhang_Ab_Initio | 15 | 19.097 | 1.273 |
| 6. | 237 | zhang | 15 | 18.850 | 1.257 |
| 7. | 388 | ProQ2 | 15 | 17.866 | 1.191 |
| 8. | 149 | wfFUIK | 1 | 1.131 | 1.131 |
| 9. | 490 | Zhang_Refinement | 15 | 16.079 | 1.072 |
| 10. | 114 s | QUARK | 15 | 15.941 | 1.063 |
| 11. | 350 | Kloczkowski_Lab | 15 | 15.683 | 1.046 |
| 12. | 330 s | BAKER-ROSETTASERVER | 15 | 14.140 | 0.943 |

# Computational Protein Design

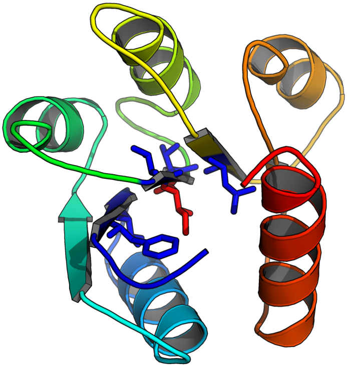
Protein design is the problem of finding an amino acid sequence compatible with a given protein backbone.



## Default method

Minimize a score function that captures residue interactions in the structure. 20 values: the 20 natural residues. Size of the search space: $20^n$ where $n$ is the length of the sequence to predict.

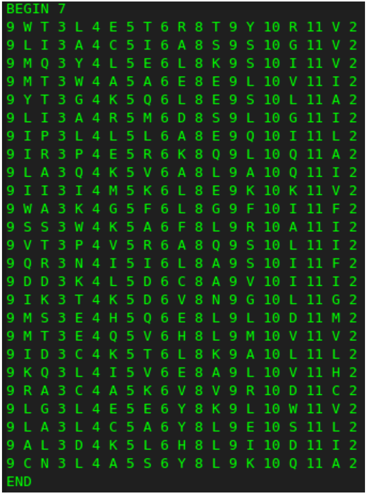
# Extracting knowledge to guide the search



## Our approach

- Establish an interaction profile for our query structure
- Find interaction matches in a database for each sequence position
- Rank matches according to surface overlap

# Interaction library



```
BEGIN 7
9 W T 3 L 4 E 5 T 6 R 8 T 9 Y 10 R 11 V 2
9 L I 3 A 4 C 5 I 6 A 8 S 9 S 10 G 11 V 2
9 M Q 3 Y 4 L 5 E 6 L 8 K 9 S 10 I 11 V 2
9 M T 3 W 4 A 5 A 6 E 8 E 9 L 10 V 11 I 2
9 Y T 3 G 4 K 5 Q 6 L 8 E 9 S 10 L 11 A 2
9 L I 3 A 4 R 5 M 6 D 8 S 9 L 10 G 11 I 2
9 I P 3 L 4 L 5 L 6 A 8 E 9 Q 10 I 11 L 2
9 I R 3 P 4 E 5 R 6 K 8 Q 9 L 10 Q 11 A 2
9 L A 3 Q 4 K 5 V 6 A 8 L 9 A 10 Q 11 I 2
9 I I 3 I 4 M 5 K 6 L 8 E 9 K 10 K 11 V 2
9 W A 3 K 4 G 5 F 6 L 8 G 9 F 10 I 11 F 2
9 S S 3 W 4 K 5 A 6 F 8 L 9 R 10 A 11 I 2
9 V T 3 P 4 V 5 R 6 A 8 Q 9 S 10 L 11 I 2
9 Q R 3 N 4 I 5 I 6 L 8 A 9 S 10 I 11 F 2
9 D D 3 K 4 L 5 D 6 C 8 A 9 V 10 I 11 I 2
9 I K 3 T 4 K 5 D 6 V 8 N 9 G 10 L 11 G 2
9 M S 3 E 4 H 5 Q 6 E 8 L 9 L 10 D 11 M 2
9 M T 3 E 4 Q 5 V 6 H 8 L 9 M 10 V 11 V 2
9 I D 3 C 4 K 5 T 6 L 8 K 9 A 10 L 11 L 2
9 K Q 3 L 4 I 5 V 6 E 8 A 9 L 10 V 11 H 2
9 R A 3 C 4 A 5 K 6 V 8 V 9 R 10 D 11 C 2
9 L G 3 L 4 E 5 E 6 Y 8 K 9 L 10 W 11 V 2
9 L A 3 L 4 C 5 A 6 Y 8 L 9 E 10 S 11 L 2
9 A L 3 D 4 K 5 L 6 H 8 L 9 I 10 D 11 I 2
9 C N 3 L 4 A 5 S 6 Y 8 L 9 K 10 Q 11 A 2
END
```

## Method

- Construct a library of interactions
- Randomly select one fragment per window
- Mutate all residues described in one fragment
- Relax and score the structure
- Update the probabilities to select each fragment

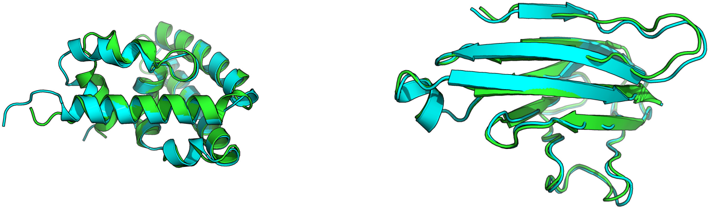
# The search engine: Shades

## Structural Homology Approach for DESign

- Source code in C++
- Uses Rosetta Macromolecular Modeling Suite to model proteins and score them [a]
- Uses an Estimation of Distribution algorithm for sampling: typically 25 iterations, 50,000 models generated
- Parallelized with MPI

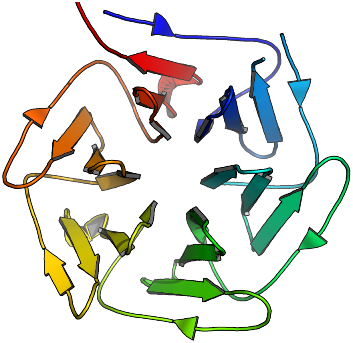
[a] https://www.rosettacommons.org/

# Full redesign of proteins



**Protein structure recovery**
*Designer proteins obtained with a prototype of Shades and assessed with I-TASSER.*

# Application: symmetrical beta propellers



**How did Nature create this ?**

Hypothesis: by gene duplication and fusion

Problem: no structure with the exactly same sequence in each blade

# Application: symmetrical beta propellers

キラリ☆研究開発
漫画・はやのん理系漫画制作室
第155回・世界最小のピザ!? 人工たんぱく質設計の世界（前編）
はやのん理系漫画制作室WEBサイト http://www.hayanon.jp
横浜市にある横浜市立大学大学院生命医科学研究科にやってきました〜!
なんでもここで「世界最小のピザ」をつくることに成功したのだとか……!?
ピザと言っても本物の食べるピザじゃないですよ〜 私たちが取り組んでいるのは「人工たんぱく質設計」なんです
私たちの研究グループは完全回転対称プロペラ型の人工たんぱく質を計算によって設計して それが実際に製造できることを証明しました!※
"PIZZA6"
ピザに似ている!?
回転させると……!?
けっこう ぶ厚い ピザだった!?
※2014年10月7日 Proceedings of the National Academy of Sciences of the United States of Americaに掲載
横浜市立大学大学院 生命医科学研究科 ジェレミー・テイム教授
ジェレミー先生の研究室メンバー 博士課程後期1年 寺田大樹さん
博士課程後期3年 野口大貴さん
直径は約5ナノメートル! ピザに似ているので「ピザ型人工たんぱく質」と名づけました!
理化学研究所 ライフサイエンス技術基盤研究センター アルノウト ヴァット博士

## Method highlights

- Generate a symmetrical backbone from a single blade with symmetric docking (RosettaDock)
- Derive a phylogenetic tree from the blade sequences to predict possible ancestral blade sequences (FastML)
- Map each sequence to the symmetrical backbone and relax the structure (Rosetta Relax)

Voet, A. R. D. *et.al.* (2014)
Computational design of a self-assembling symmetrical -propeller protein.
PNAS, published ahead of print October 6, 2014, doi:10.1073/pnas.1412768111

# Conclusion

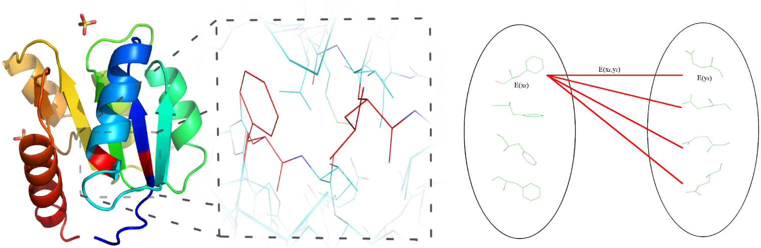
## Protein structure prediction

- Fragment-based approaches still are the best at free modeling
- The problem is still far from being solved
- Protein design can help

## Computational Protein design

- Few positive results
- Very hard to validate any new approach
- Many doors to open

# Upcoming



**CPD as a cost function network**

Interactions can be decomposed as a sum of one-body and two-body energy terms.

$$E_S = \sum_{i=1}^{n} E(x_i^{r_1}) + \sum_{i=1}^{n-1} \sum_{j>i} E(x_i^{r_1}, y_j^{r_2})$$

# Thank you for your attention