# Checking the validity of Protein Folding in Lattice

Noël Malod-Dognin and Ivan Todorin

South-West University, Blagoevgrad, Bulgaria.

The $29^{th}$ of April

# Amino-acids properties

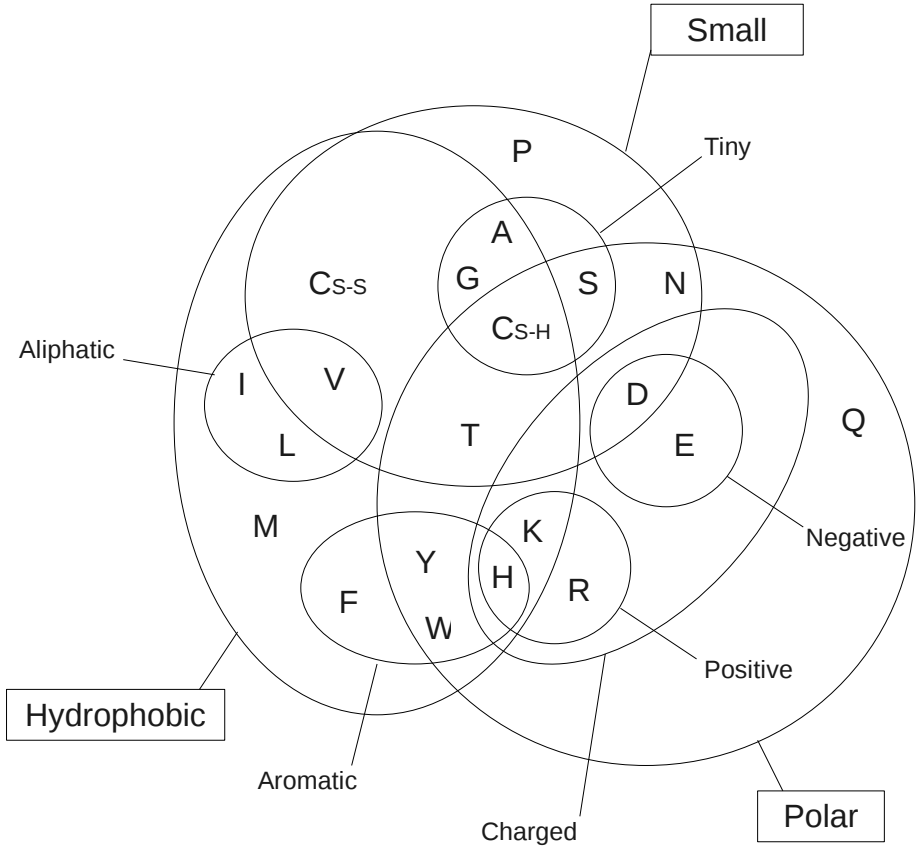



FIGURE: Venn diagram grouping α-amino-acids according to their properties. The most important ones are small, hydrophobic and polar (and their counterparts).

# Hydrophobic and Polar amino-acids

| Name | Symbol | Classification | Name | Symbol | Classification |
|---|---|---|---|---|---|
| Alanine | A | Hydrophobic | Leucine | L | Hydrophobic |
| Arginine | R | Polar | Lysine | K | Polar |
| Asparagine | N | Polar | Methionine | M | Hydrophobic |
| Aspartic Acid | D | Polar | Phenylalanine | F | Hydrophobic |
| Cysteine | C | Polar | Proline | P | Hydrophobic |
| Glutamic Acid | E | Polar | Serine | S | Polar |
| Glutamine | Q | Polar | Threonine | T | Polar |
| Glycine | G | Polar | Tryptophan | W | Hydrophobic |
| Histidine | H | Polar | Tyrosine | Y | Polar |
| Isoleucine | I | Hydrophobic | Valine | V | Hydrophobic |

TABLE: Hydrophobic / Polar classification of the 20 $\alpha$-amino-acids.

# The Research Collaboratory for Structural Bioinformatics Protein Data Bank (PDB)

- The main freely available protein structure database
- http ://www.pdb.org

---

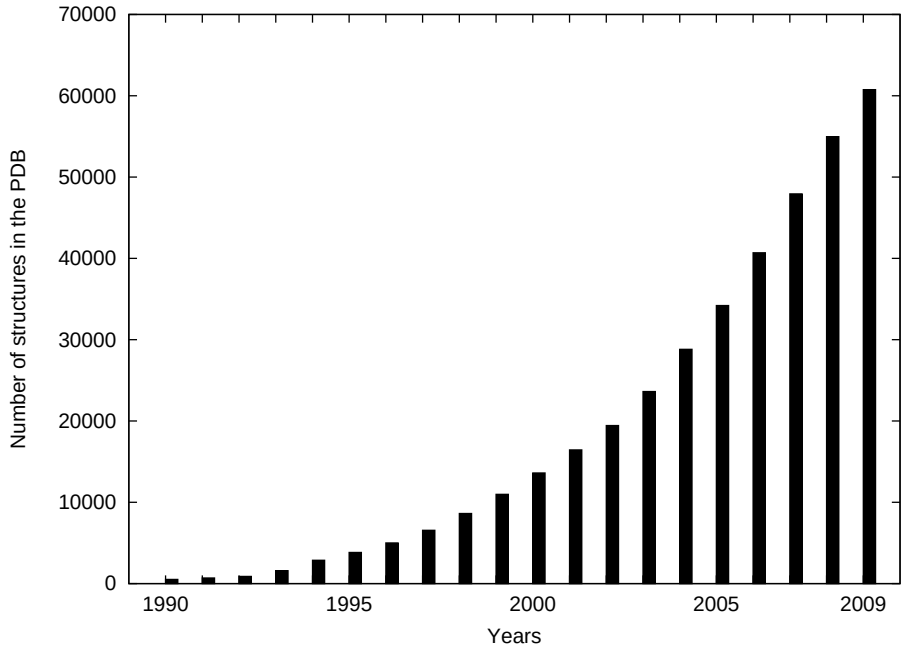



FIGURE: Data growth in the PDB during the last 20 years. This April 2010, there are 64781 structures in the PDB.

# Available data in the PDB

- 3D coordinates of the amino-acid atoms (.ent or .pdb files)
- Amino-acid sequences (fasta files)

# The Sokol set

| Protein | Length | Species |
|---|---|---|
| 1bpi | 58 | Cow (Bos taurus) |
| 5pti | 58 | Cow (Bos taurus) |
| 1knt | 58 | Human (Homo sapiens) |
| 2knt | 58 | Human (Homo sapiens) |
| 1era | 62 | Sea snake (Laticauda semifasciata) |
| 3ebx | 62 | Sea snake (Laticauda semifasciata) |
| 6ebx(A) | 62 | Sea snake (Laticauda semifasciata) |

TABLE: 7 small protein chains. The four first are from the “Small Kunitz-type inhibitors & BPTI-like toxins” SCOP family, the last three are from the “Snake venom toxins” family.

# The Skolnick set

| | SCOP Family | Length | Proteins |
|---|---|---|---|
| 1 | CheY-related | 120-130 | 1b00A, 1dbwA, 1natA, 1ntrA, 3chyA |
| | | | 1qmp(A,B,C,D), 4tmy(A,B) |
| 2 | Plastocyanin | 97-105 | 1bawA, 1byo(A,B), 1kdiA, 1ninA |
| | /azurin-like | | 1plaA, 2b3iA, 2pcyA, 2pltA |
| 3 | Triosephosphate | 243-256 | 1amkA, 1aw2A, 1b9bA, 1btmA, 1htiA |
| | isomerase (TIM) | | 1tmhA, 1treA, 1triA, 1ydvA, 3ypiA, 8timA |
| 4 | Ferritin | 158-191 | 1b71A, 1bcfA, 1dpsA, 1fhaA, 1ierA, 1rcdA |
| 5 | Fungal ribonucleases | 104 | 1rn1(A,B,C) |

TABLE: It contains 40 protein chains from 33 proteins, classified by SCOP in five differant families.

# Other possible benchmarks

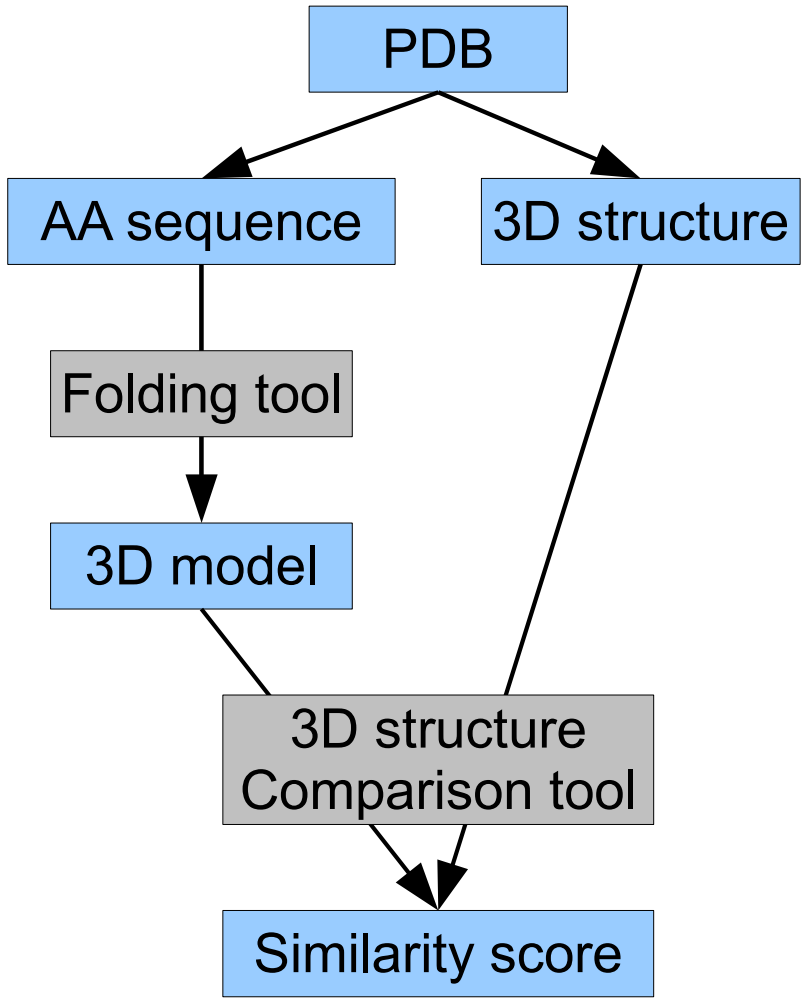
## The CASP competition (Critical Assessment of protein Structure Prediction)

Held every two years since 1994, the 2010 edition is ongoing.

- About 120 new protein structures (not published).
- Corresponding sequences are given to each competitor (expert and/or automatic server) which returns 3D models.
- At the end of the competition, competitors are ranked according to the quality of the returned models

http ://predictioncenter.org

# Framework


PDB
AA sequence
3D structure
Folding tool
3D model
3D structure
Comparison tool
Similarity score

# Generating 3D models in lattices

## Constraint-based Protein Structure Prediction (CPSP) Tools

- HPStruct
  - Computes optimal structures of 3D lattice proteins in the HP-model
  - http ://cpsp.informatik.uni-freiburg.de
- Need to convert amino-acids sequences into HP sequences
- Returns movements sequences that can be converted into 3D models (.pdb)

# Comparing 3D models and real structures

## Contact map overlap maximisation

A_purva (Andonov, Yanev and Malo-Dognin, 2008)

- Need to convert 3D models into contact maps
- Returns similarity scores

# Results on the Sokol set

| Protein chain | $\|E\|$ real 3D structure | $\|E\|$ fcc model | *NCC* | Similarity |
|---|---|---|---|---|
| 1bpi | 180 | 246 | 74 | 0.35 |
| 5pti | 186 | 246 | 76 | 0.35 |
| 1knt | 177 | 249 | 70 to 71 | 0.33 |
| 2knt | 184 | 249 | 71 to 72 | 0.33 |

| Protein chain | $\|E\|$ real 3D structure | $\|E\|$ cubic model | *NCC* | Similarity |
|---|---|---|---|---|
| 3ebx | 197 | 433 | 107 to 111 | 0.34 |
| 6ebx | 196 | 433 | 108 to 111 | 0.34 |
| 1knt | 177 | 400 | 94 | 0.33 |
| 2knt | 184 | 400 | 94 | 0.32 |

TABLE: Similarity beteween real structure and predicted models.

# Results on the Skolnick set

| Protein chain | $\lvert E \rvert$ real 3D structure | $\lvert E \rvert$ fcc model | *NCC* | Similarity |
|---|---|---|---|---|
| 1nin | 366 | 508 | 121 to 139 | 0.27 |
| 1pla | 332 | 462 | 135 to 136 | 0.34 |
| 1rn1a | 320 | 512 | 108 to 120 | 0.26 |
| 1rn1b | 320 | 512 | 110 to 128 | 0.26 |
| 1rn1c | 320 | 512 | 110 to 128 | 0.26 |
| 2b3i | 337 | 463 | 109 to 133 | 0.27 |
| 2pcy | 344 | 468 | 102 to 128 | 0.25 |

| Protein chain | $\lvert E \rvert$ real 3D structure | $\lvert E \rvert$ cubic model | *NCC* | Similarity |
|---|---|---|---|---|
| 1rn1a | 320 | 860 | 176 to 192 | 0.29 |
| 1rn1b | 320 | 860 | 176 to 192 | 0.29 |
| 1rn1c | 320 | 860 | 176 to 191 | 0.29 |

TABLE: Similarity beteween real structure and predicted models.

# About HP sequences

Many proteins, with different AA sequences, possess identical HP sequences

- 1rn1 (A,B,C)
- 1knt and 2knt
- 5pti and 1bpi
- 6ebx and 3ebx
- ...

# Conclusion

HP models :

- Too much compact structures
- not similar to real 3D structures
- HP folding in lattice with sidechain models ?

HP sequences :

- Show potential for sequence alignment

  (or maybe filtering purpose)