Computational Short Cuts to Protein Structure and Function:

Fold Recognition Methods.

Jarek Meller

Biomedical Informatics

Children’s Hospital Research Foundation

Outline:

- Introduction
- Fold recognition: sequence similarity vs. threading
- Common models and algorithms
- Fold recognition servers and annotation strategies
- Discussion

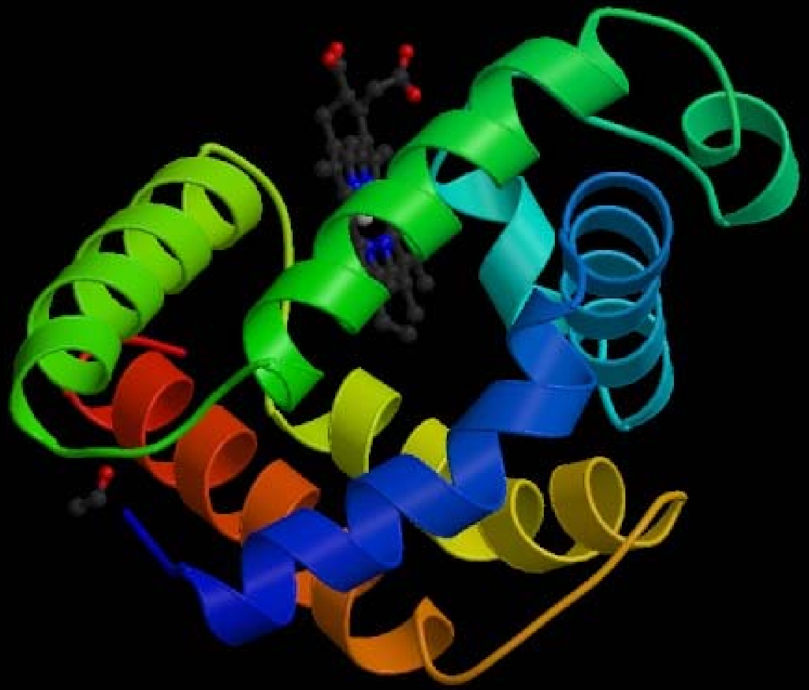
Introduction:

- Protein machinery of life: from sequence to structure to function (from DNA to mRNA to protein sequence to protein structure to protein-protein/DNA/RNA/small molecules interactions to phenotype)

- Deciphering protein structure: experiment vs. simulation (**C**omputer-**A**ided **Sh**ort Cuts = CASH)

- Fold recognition: nature as best computational device

Three lovely proteins: hemoglobin

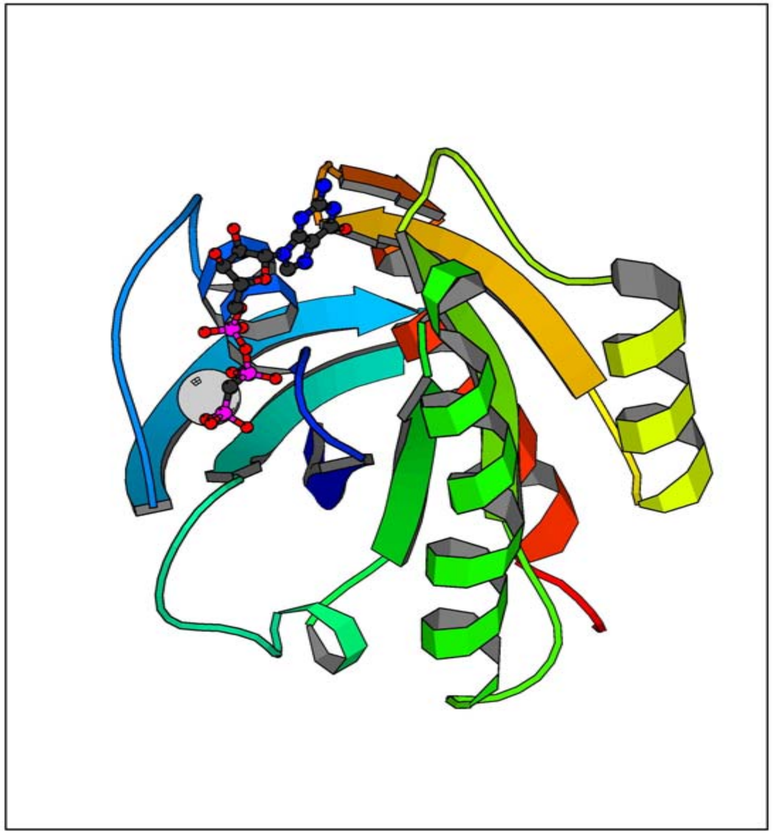


- Four units carrying oxygen
- Sickle-cell anemia: inherited disease
- Glu6 – Val6 mutation causes aggregation

Three lovely proteins: gramicidin

- Transmembrane ion channel
- Bacteria killer - antibiotic

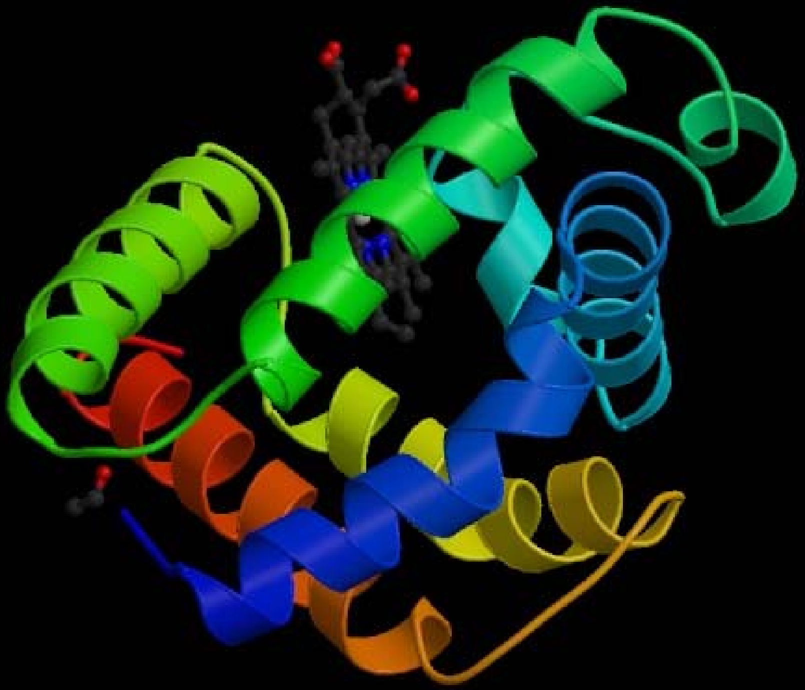
Three lovely proteins: ras p21



- Molecular switch based on GTP hydrolysis
- Cellular growth control and cancer
- Ras oncogene: single point mutations at positions Gly12 or Gln61

Significance of Protein Folding Problem

V
L
S
E
G
E
W
Q
L
V
L
V
.
.
.



⇄ O_2

Sequence

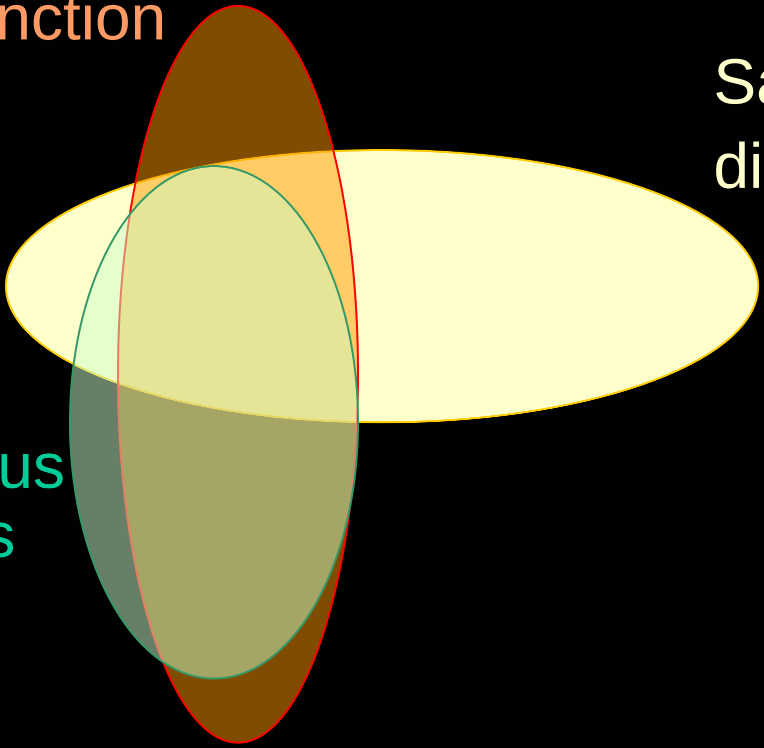
folds into a 3D

structure

to perform a

function

Sequence → Structure → Function



Same fold, different function

Same function, different fold

Homologous sequences

Sequence → Structure → Function

- Continuous nature of folds, multiple functions
- SCOP: up to 7 folds per function and up to 15 functions per fold
- Divergent (common ancestor) vs. convergent (no ancestor) evolution
- PDB: virtually all proteins with 30% seq. identity have similar structures, however most of the similar structures share only up to 10% of seq. identity !

www.columbia.edu/~rost/Papers/1997_evolution/paper.html (B. Rost)
www.bioinfo.mbb.yale.edu/genome/foldfunc/ (H. Hegyi, M. Gerstein)

Classifications of protein shapes and families:

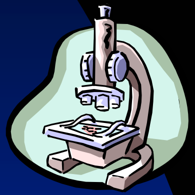
- SCOP (Structural Classification of Proteins, scop.berkeley.edu, Murzin et. al.):

  548 folds (major structural similarity in terms of secondary structures e.g. globin-like, Rossman fold); 1296 families (clear evolutionary relationship or homology e.g. globins, Ras)

- CATH (Class, Architecture, Topology, Homologous Superfamily, www.biochem.ucl.ac.uk/bsm/cath/, Orengo et. al):

  35 architectures (gross arrangment of secondary structures e.g. non-bundle, sandwich); 580 topologies (connectivity of secondary structures e.g. globin-like, Rossman fold); 1846 families (clear homology, same function)

Deciphering protein structure and function:



- Experiment (X-ray, NMR): months

Experiments can be lengthy and costly. Therefore computational methods are often used to focus and facilitate experimental research.

- Atomistic (physical principles based) simulations: weeks
- Homology based modeling: hours
- Sequence similarity based annotations: seconds

Computational complexity price of accurate models:

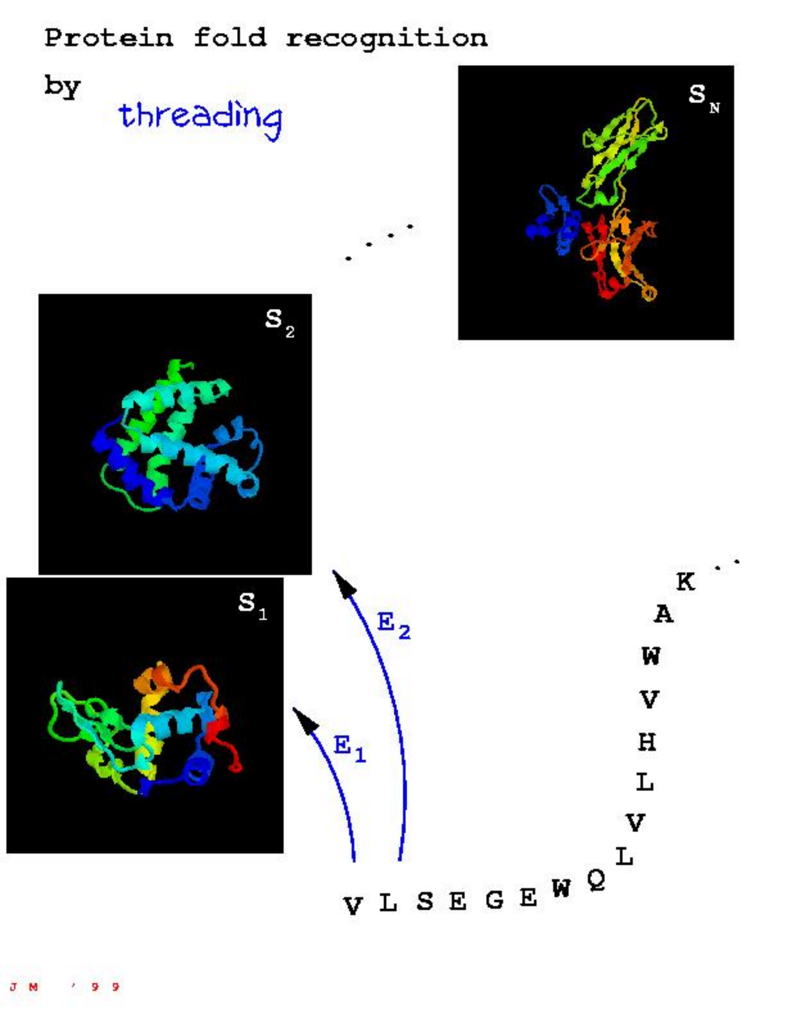
- Huge search problem - scaling with size in protein folding:

  No. of conformations $\sim 10^{n}$
- Rugged energy landscape and local minima problem

Nature performs these “computations” efficiently and one can use solutions provided by nature as templates:

from protein folding to protein recognition.

Assigning fold and function utilizing similarity to experimentally characterized proteins:



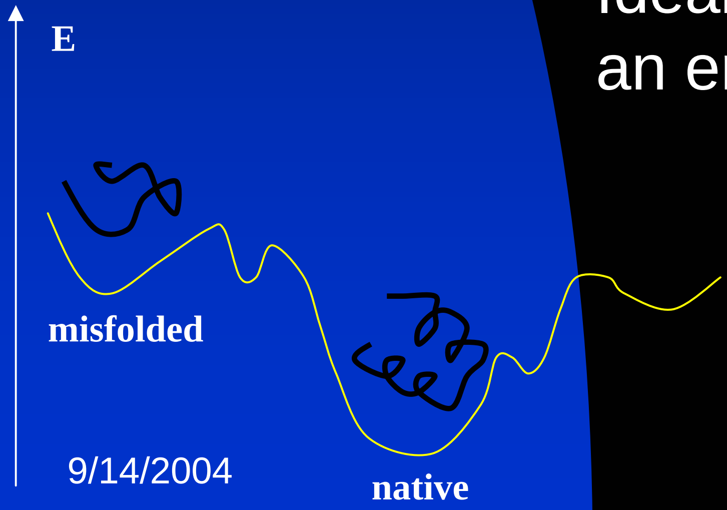
- Sequence similarity: BLAST and others
- Beyond sequence similarity: matching sequences and shapes (threading)

Sequence to structure matching (threading) may detect distantly related proteins due to conservation of structure.

In practice fold recognition methods are often mixtures of sequence matching and threading.

D.Fischer and D. Eisenberg, Curr. Opinion in Struct. Biol. 1999, 9: 208

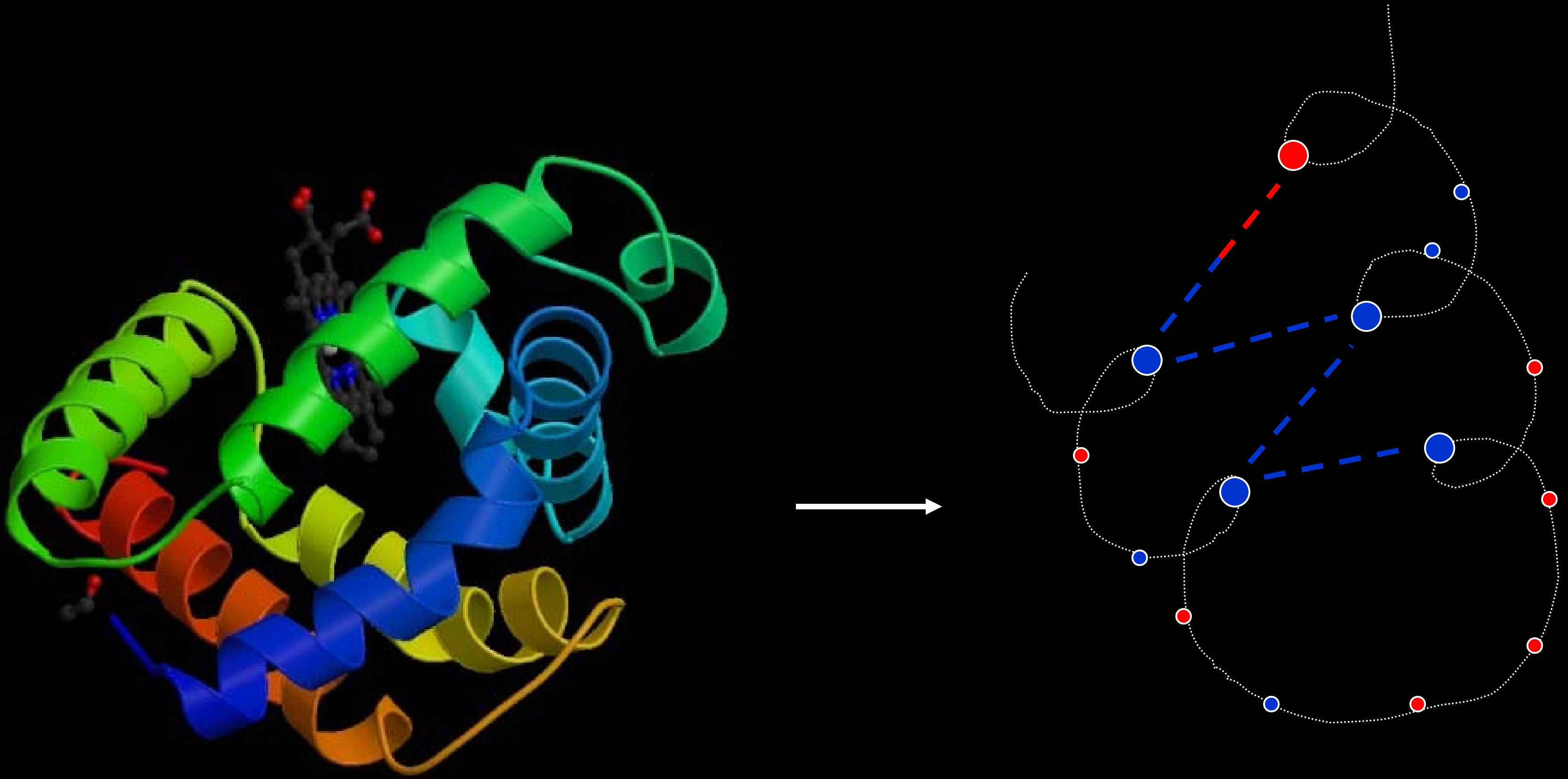
We need a scoring (energy) function to distinguish native structure from misfolded structures.



Ideally, each misfolded structure should have an energy higher than the native energy, i.e. :

$$E_{misfolded} - E_{native} > 0$$

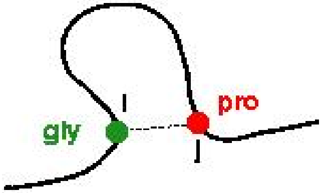
Reduced Representations of Protein Structure:



Each amino acid represented by a point in the 3D space; simple contact model – two amino acids in contact if their distance smaller than a cutoff.

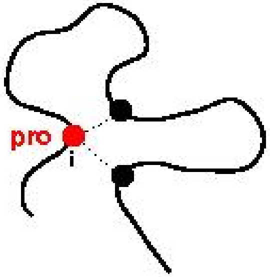
Possible functional form:

1. PAIRWISE



$$\sum_{i<j} \varepsilon_{ij}$$

2. SIMPLE PROFILE (counting neighbors to a site)



$$\sum_{i} \varepsilon_{i}(n)$$

How to choose an energy function?

- Functional form:

  # contact potential?

  # profile model?

  Accuracy vs. efficiency (R.H. Lathrop: protein threading problem with contact potentials is NP-complete, Protein Eng. 7, 1994).

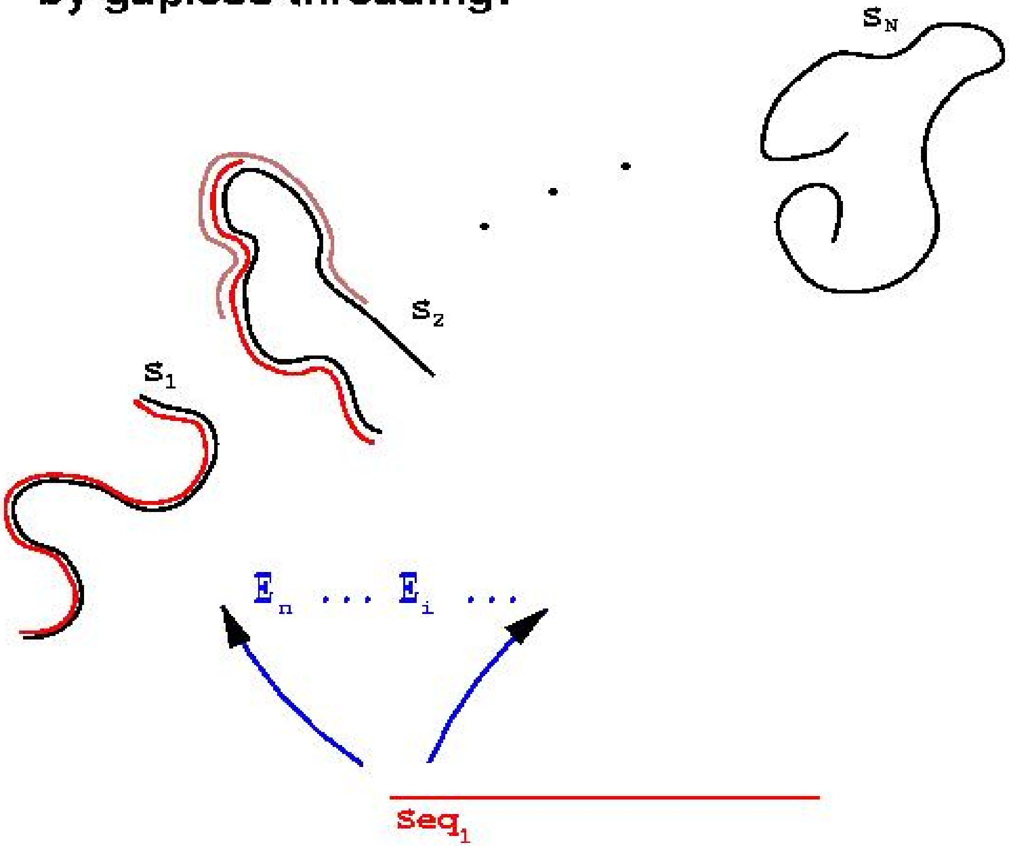
- Optimization of parameters:

  # Linear Programming!

  # $E_{decoy} - E_{native} > 0$

  V.N. Mairov & G.M. Crippen, JMB 227, 1992.

Creating decoy structures (inequalities) by gapless threading:

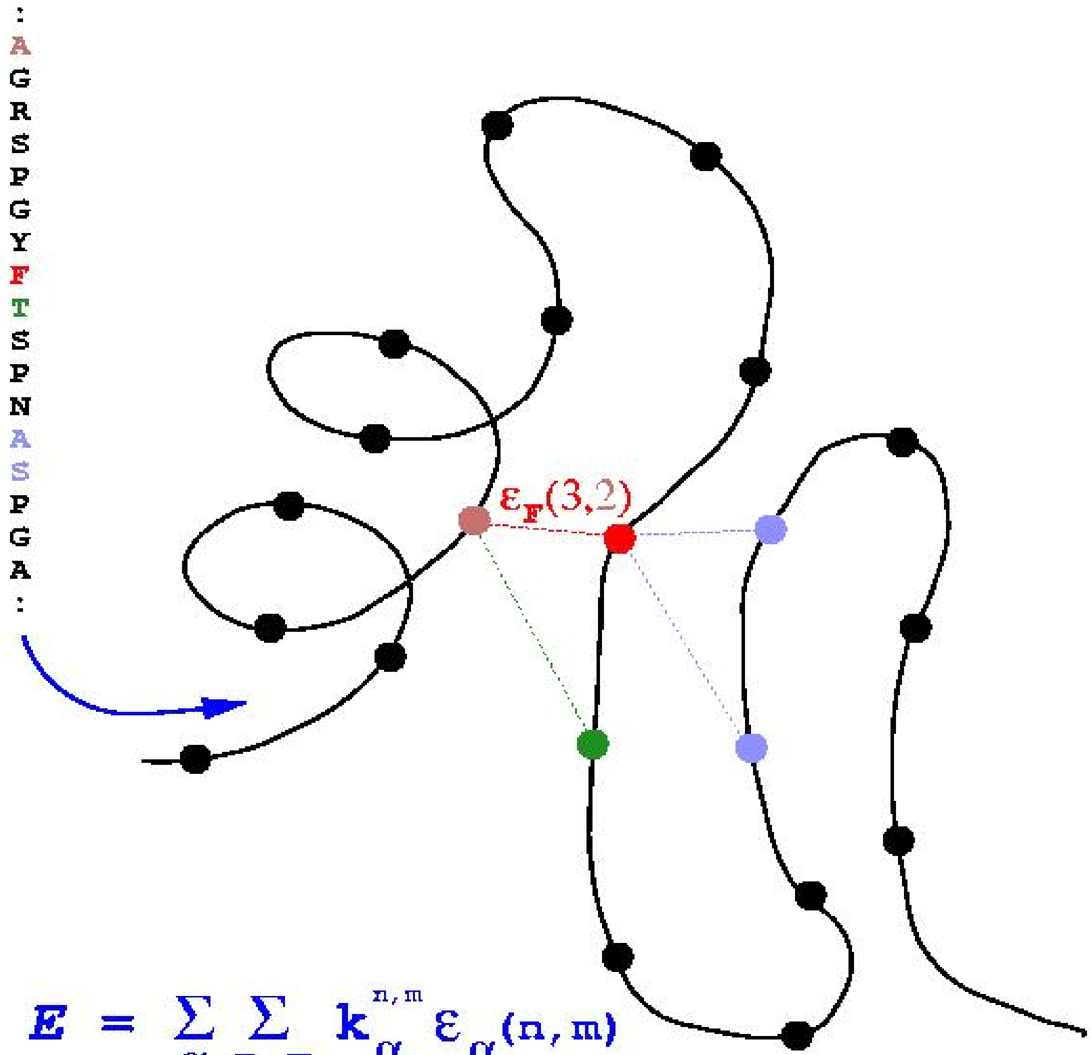


$$E_i - E_n = \sum_{\alpha} (k_{\alpha}^{i} - k_{\alpha}^{n})\,\varepsilon_{\alpha} > 0$$

n - native structure; i - decoy structures

Threading Onion Model

with the first and second contact shells (THOM_2)



$$E = \sum_{\alpha} \sum_{n,m} k_{\alpha}^{n,m} \varepsilon_{\alpha}(n,m)$$

Contact between a site of n neighbours
and occupied by an amino acid of type α
with a site of m neighbours contributes $\varepsilon_{\alpha}(n,m)$

Methodological kit:

- Dynamic programming: optimal string matching
- Neural networks: secondary structure predictions (PsiPRED, Jones DT, JMB 292: 195)
- Hidden Markov Models: family profiles, secondary and tertiary structure prediction (TMHMM by A. Krogh and co-workers, http://www.cbs.dtu.dk/krogh/refs.html)
- Monte Carlo: suboptimal solutions (Mirny LA, Shakhnovich EI, Protein Structure Prediction By Threading. Why It Works Why It Does Not, JMB 283: 507)

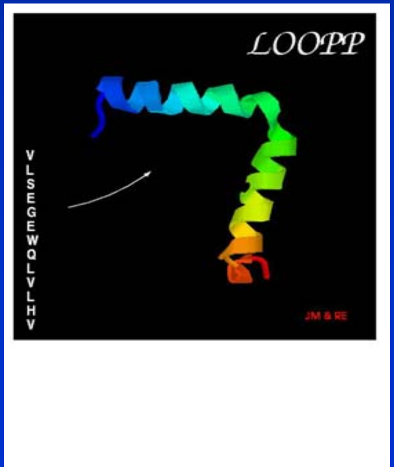
Fold recognition servers:

- PsiBLAST (Altschul SF et. al., Nucl. Acids Res. 25: 3389)
- Live Bench evaluation (http://BioInfo.PL/LiveBench/1/) :

1. FFAS (L. Rychlewski, L. Jaroszewski, W. Li, A. Godzik (2000), Protein Science 9: 232) : seq. profile against profile
2. 3D-PSSM (Kelley LA, MacCallum RM, Sternberg JE, JMB 299: 499) : 1D-3D profile combined with secondary structures and solvation potential
3. GenTHREADER (Jones DT, JMB 287: 797) : seq. profile combined with pairwise interactions and solvation potential

- LOOPP: annotations of “orphan” sequences

http://www.tc.cornell.edu/CBIO/loopp



Annotations Strategies

- Use first sequence methods (with polypeptide chains if possible) and remember: profile methods (e.g. PsiBLAST, SAM) are much more sensitive than pairwise alignments! (Park et. al., “Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods.” JMB 284: 1201)
- Still nothing? Submit your sequence to transmembrane prediction (more than 90% reliability) and secondary structure prediction servers (70 to 80% reliability). (e.g. TMHMM by A. Krogh et. al., PsiPRED, D.T. Jones, JMB 292: 195)
- Having a reasonably good feeling about different domains on your beloved protein submit alternative queries to fold recognition servers. Use all trustworthy servers and pay attention to their estimates of statistical significance.
- Re-evaluate: check consistency with expected sequence motifs, active sites, disulphide bridges etc., validate predictions using all the knowledge about your protein! Use consensus, but without rejecting biologically interesting conclusions.