KFC/STBI
Strukturní bioinformatika

06_predikce struktury

Karel Berka

Predikce

- minule jsme se snažili najít způsob, jak spolu budou interagovat malé molekuly a proteiny.
- Ale co dělat, když strukturu proteinu nemame?

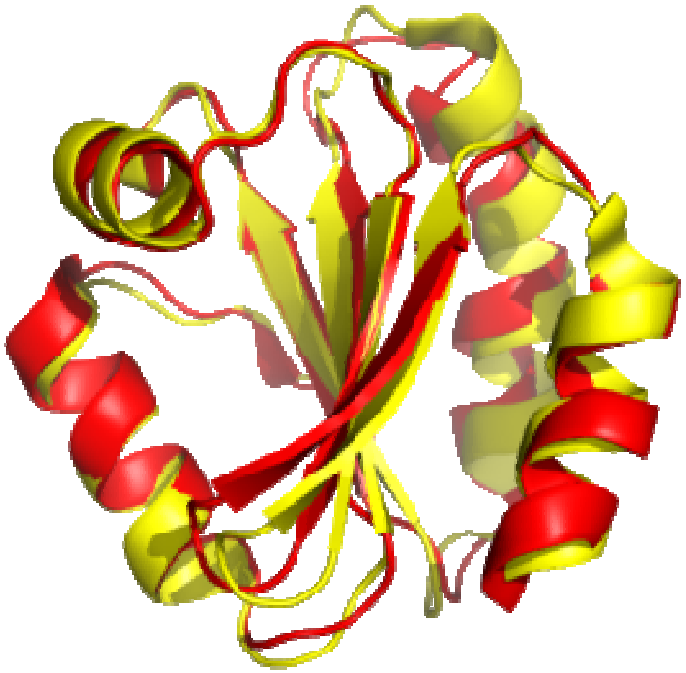
Syllabus

- Strukturní alignment
- Predikce struktury
  - Homologní modelování
    - SwissMODEL
  - threading
    - Modeller, I-TASSER
  - de novo modelování
    - Robbeta, Quark
  - molekulární mechanika
    - skládání proteinů (protein folding)
    - Folding@Home, FoldIt

Strukturní alignment

Jak lze rozeznat strukturní podobnost?

- Podle oka
- Algoritmicky:
  - Strukturní alignment



Structural alignment of thioredoxins
from humans (red) and Drosophila (yellow)
PDBID: 3TRX and 1XWC.

Strukturní alignment

- Pro daný pár molekulárních struktur nalézt korespondenci mezi souřadnicemi atomů, jež povede k nejlepšímu srovnání (alignmentu)
- *Nejlepší* znamená ve smyslu “nejmenší RMSD”
- Nevýhody: alignment pro několik atomů vynikající pro zbytek nevyhovující

Strukturní alignment

- Nutno zohlednit:
  - Počet odpovídajících si aminokyselin
  - RMSD těchto aminokyselin
  - Procento identity v „aligned residues“
  - Počet vnesených „gaps“
  - Velikost těchto dvou proteinů
  - Místa s konzervovanou sekvencí

- Nejsou žádná universální kriteria. Vše závisí na cíli.

Strukturní alignment

- Upozornění
- Jedná se o jiný druh problému než při určení RMSD dvou proteinů při dané korespondenci atomů
- V tomto případě nevíme které atomy porovnávat s kterými. Z toho důvodu je nutno podniknout analýzu všech možných korespondencí
- RMSD se užívá jako nástroj k určení korespondence

Proč dělat strukturní alignment

Struktura se většinou zachovává více než sekvence

1. Homologní proteiny (podobný předek)
   – „zlatý standard“ pro sekvenční alignment

2. Nehomologní proteiny
   – určení obecné či příbuzné substruktury (domén)

1. Klasifikace proteinů do klastrů
   – založených na strukturní podobnosti (CASP, SCOP)

Typy strukturního alignmentu

bodové metody

vlastnosti bodů či vzdáleností k určení korespondencí

- **CE**

  rigidní struktura, hledání největší skupiny sekvenčně řazených ekvivalentních atomů

- **DALI** (Holm, Sander)

  matice vzdáleností pro nalezení podobných vzorků naznačujícím korespondenci bez ohledu na sekvenci

metody založené na sekundární struktuře

vektorová reprezentaci ss k určení korespondence

- **VAST**

  alignment sekundárních struktur

- **FATCAT**

  protein není rigidní, hinges (klouby) se mohou hýbat

CE (Combinatorial Extension)

- z párů proteinových segmentů
- typicky po 8 AA
- porovnání na základě lokální geometrie
- posléze se páry alignmentu rozšiřují, dokud to jde
- výsledky:
  - RMSD
  - z-statistika

(standard) z-score is $z = \frac{x - \mu}{\sigma}$

where:

x is a raw score to be standardized;
μ is the mean of the population;
σ is the standard deviation of the populat

Predikce struktury

Predikce terciárních struktur proteinů

- Predikce struktury
  - vycházející ze známých stabilních struktur
    - Homologní modelování
      - SwissMODEL, I-TASSER
    - threading
      - Modeller,
  - vycházející z fyzikálních modelů
    - de novo modelování (ab initio)
      - Quark, Robbeta, molekulární mechanika
      - skládání proteinů (protein folding)
      - Folding@Home, FoldIt

Homologní modelování

- také komparativní, nebo knowledge-based modelování
- strukturu neznámého proteinu sestaví na základě znalosti struktury homologního proteinu

- Swiss-MODEL
- http://spdbv.vital-it.ch/modeling_tut.html

Obecný protokol

- vybrat protein k modelování
- hledání homologů
  - ne – ab initio, ev. threading
  - ano – Clustal W – alignment
- modelování centrálního regionu
  - (analýza očima – odpovídá to experimentu?)
- if ano – modelování loopů (hledání fragmentů
- modelování vedlejších řetězců (rotamery)
- minimalizace (molekulové modelování, dynamika)
- stereochemická kontrola modelu (PROCHECK, Ramachandran plot)

Princip funkce Swiss-MODEL

http://www.expasy.org/spdbv/

- identifikace strukturního templátu
- sekvenční alignment cílové sekvence s templátem
- model building – přeložení AA v templátu aminokyselinami z cílové sekvence
- evaluace kvality modelu – skórovací funkce

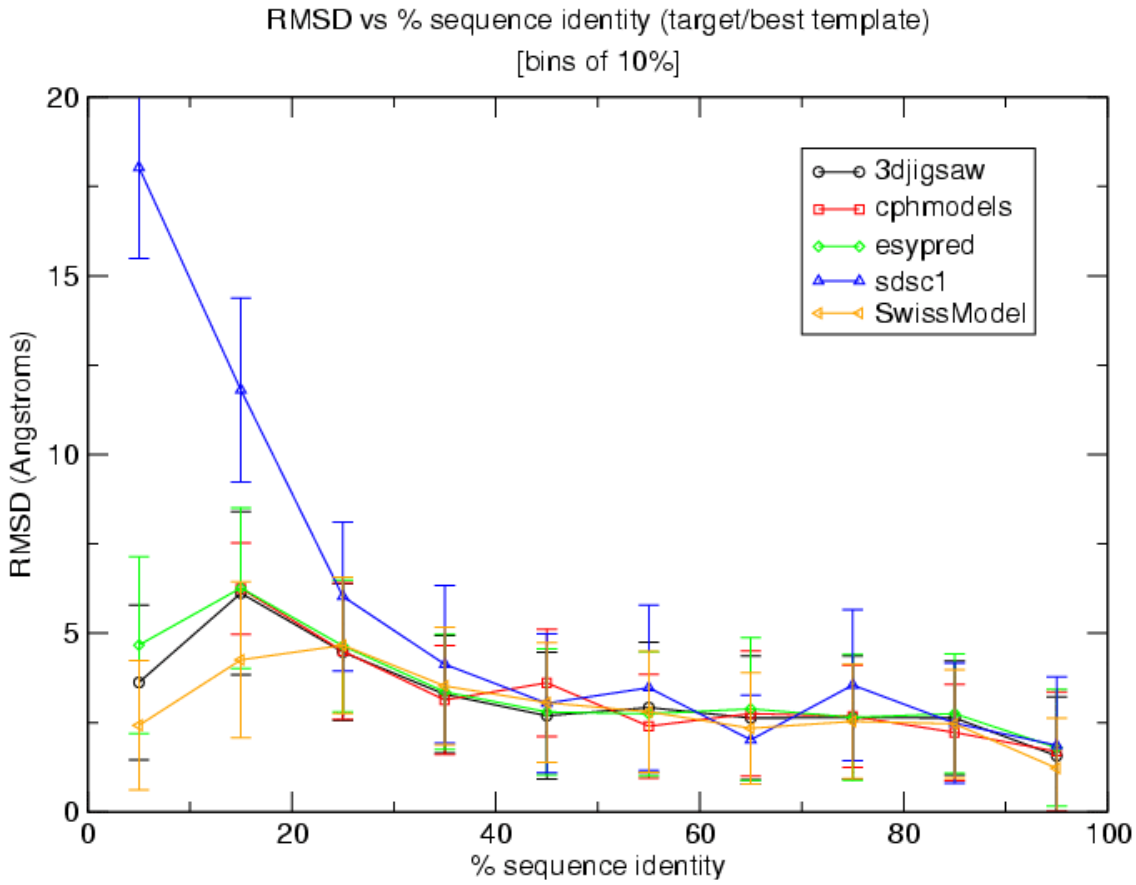
- A to lze opakovat, dokud nejsme spokojeni.

Kdy použít Homologní modelování

- pokud je podobnost mezi sekvencí templátu a cílové sekvence dostatečně vysoká

- alespoň

  30% IDENTITY



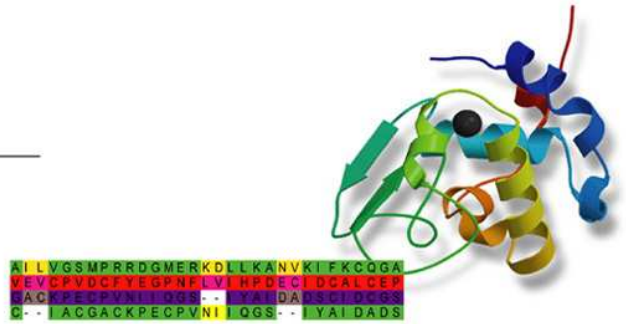
Threading

- nebo také fold recognition
- není zapotřebí homologní struktura
- místo toho se snaží modelovat na známé foldy a snaží se zjistit, který z nich je nejlepší

- Modeller

Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints



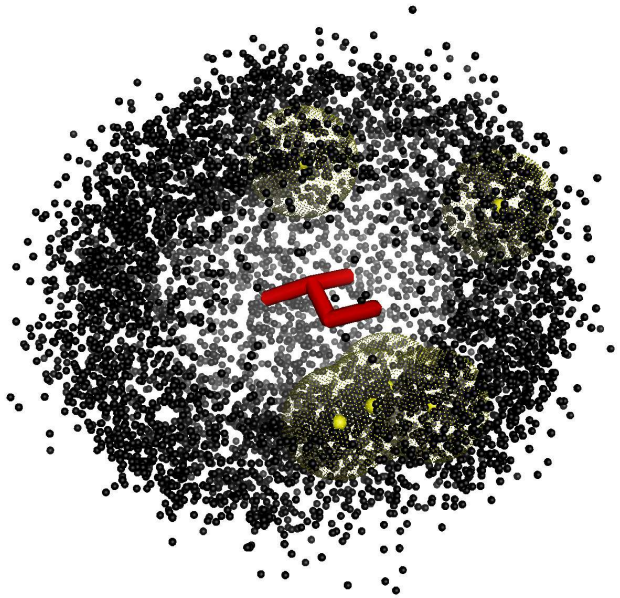
http://salilab.org/modeller/

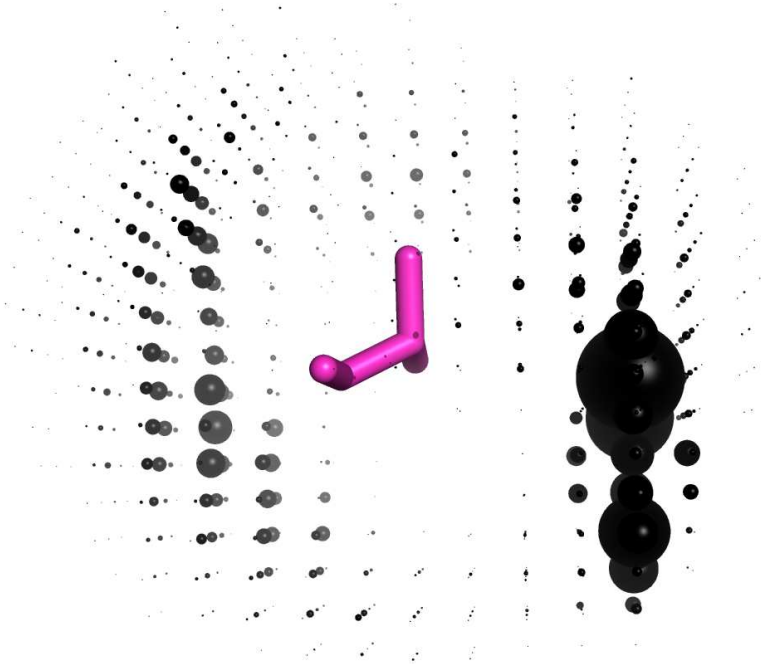
Protein threading

- **Sestavení databáze strukturních templátů**
  - This generally involves selecting protein structures from databases such as PDB, FSSP, SCOP, or CATH, after **removing** protein structures with high sequence similarities.
- **Sestavení skórovací funkce**
  - measure the fitness between target sequences and templates based on the knowledge of the known relationships between the structures and the sequences.
  - A good scoring function should contain mutation potential, environment fitness potential, pairwise potential, secondary structure compatibilities, and gap penalties.
  - The quality of the energy function is closely related to the prediction accuracy, especially the alignment accuracy.
- **Threading alignment**
  - Align the target sequence with each of the structure templates by optimizing the designed scoring function.
  - solving the optimal alignment problem derived from a scoring function considering pairwise contacts.
- **Threading predikce**
  - statistically most probable alignment => threading prediction
  - construct a structure model for the target by placing the **backbone atoms** of the target sequence at their aligned backbone positions of the selected structural template.

threading function

- energetická závislost vložení aminokyseliny na specifické místo
- funkce je vlastně preferencí dotyčné AA pro specifické místo

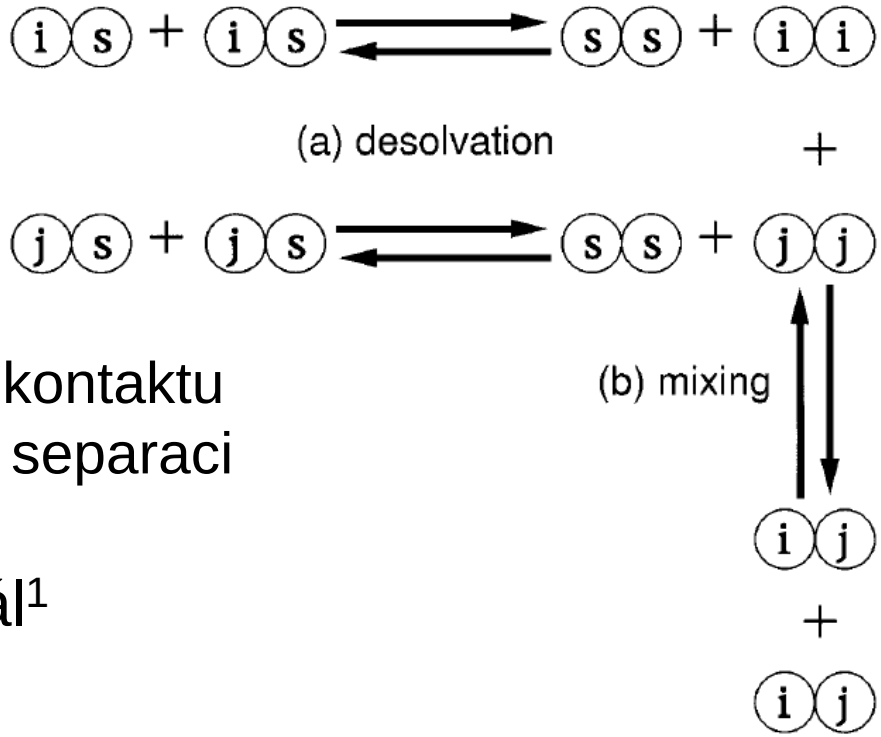




threading function

- Boltzmannův vztah

$$w_{ij}(r) = -kT \ln\left(\frac{\rho_{ij}(r)}{\rho^*}\right)$$



  - w_{ij} – energie
  - $\rho_{ij}(r)$ – sledovaná hustota kontaktu
  - ρ^* - referenční hustota při separaci

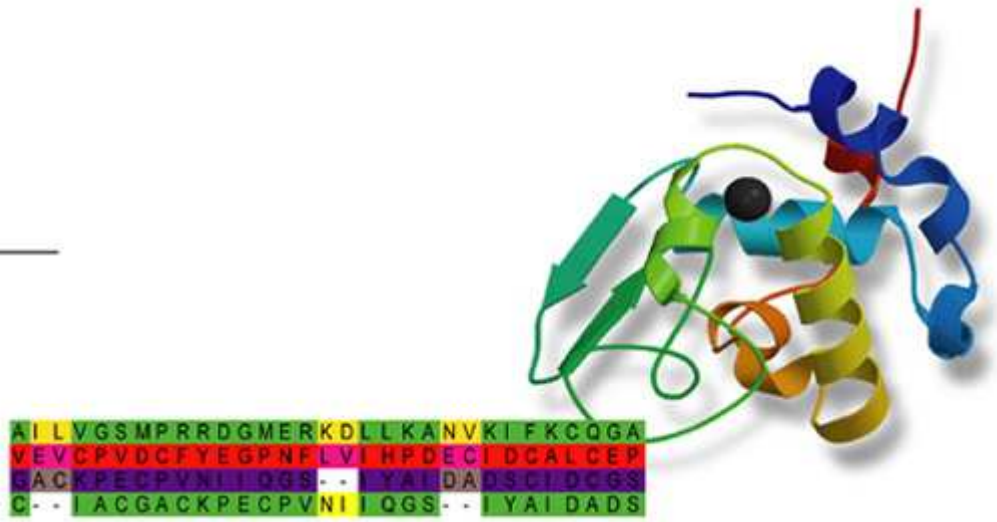
- Fenomenologický potenciál[1]

$$e_{ij} = A \cdot \ln \frac{\overline{n_{ij}} \cdot \overline{n_{oo}}}{\overline{n_{io}} \cdot \overline{n_{jo}}}$$

[1] Miyazawa, S. & Jernigan, R.L. *PROTEINS,***1999**, (36)357–369

Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints



- homologní modelování s constraints (např. NMR, EM, apod)
  - používá tzv. i-Sites (krátké kousky, pro které zná strukturu)

http://salilab.org/modeller/

Comparative protein modelling by satisfaction of spatial restraints. Šali A, Blundell TL. *J Mol Biol.* 1993 Dec 5;234(3):779-815.

Kdy použít threading?

- Pokud nemám dost sekvenční identity k jednomu templátu
- nejlépe po jednotlivých doménách
- nejlépe zkusit několik programů a porovnat, který fold je nejčastější – konsensus
- použít další znalosti o proteinu (funkce) – opět to může napomoct vybrat správný fold.

Ab initio modeling

- ab initio = bez předchozích znalostí (templátu)
- masivní hledání správné konformace a k tomu fyzikální (pseudo-fyzikální) energetická funkce na popis volné energie

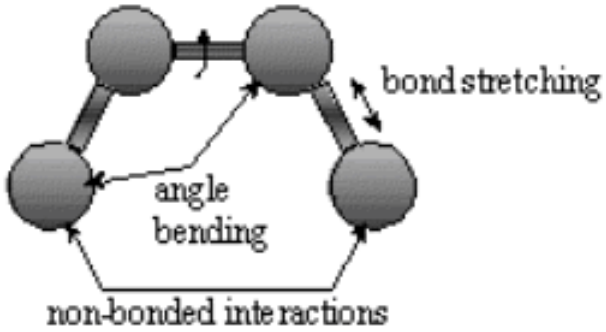
- http://robetta.bakerlab.org/
- *ab initio* and comparative models of protein domains

- Nejméně přesné ale jediné použitelné, pokud neznám templát

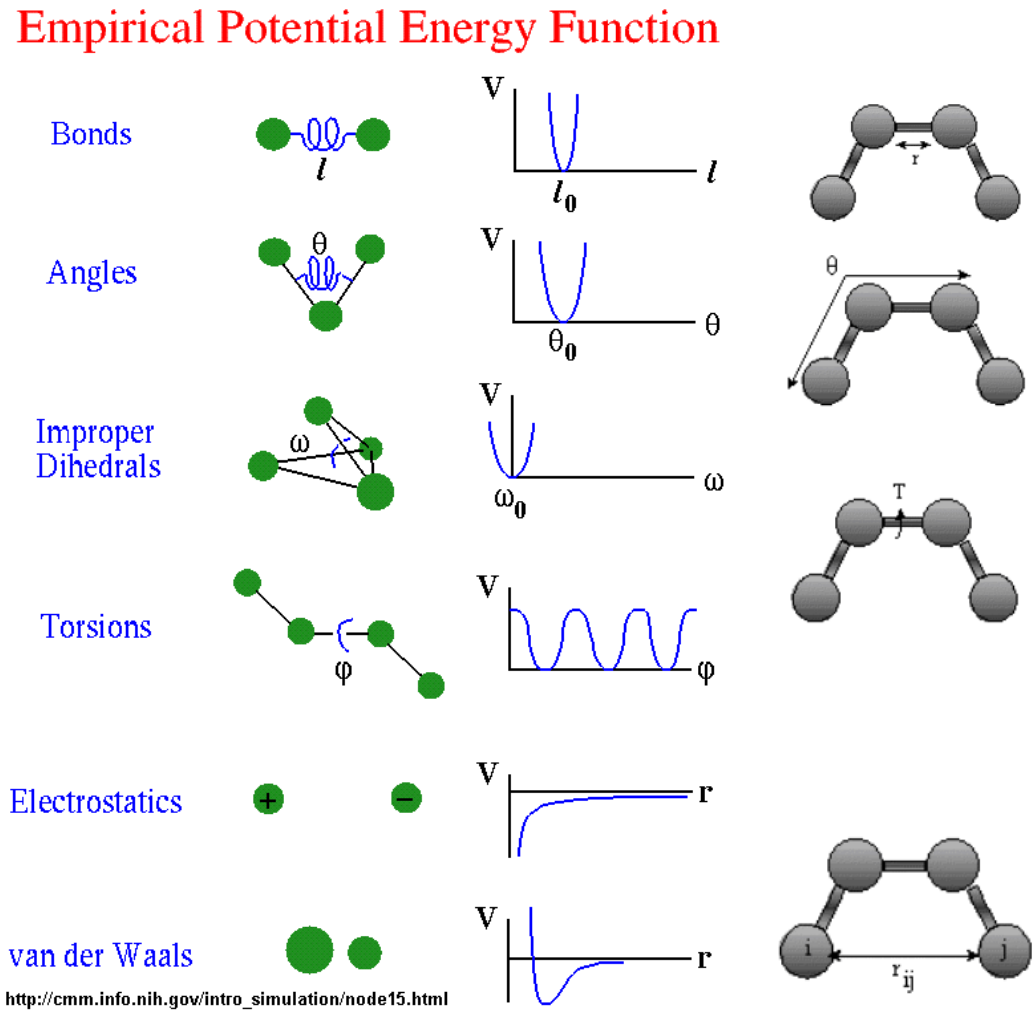
Molekulová mechanika

celková energie je funkcí vzájemné pozice atomů

$$E = f(\mathbf{R}) = E_b + E_a + E_t + E_c + E_{vdw}$$



Silová pole (Force-field)



$$E_b = \frac{k_r}{2}\left(r - r_0\right)^2$$

$$E_a = \frac{k_\theta}{2}\left(\theta - \theta_0\right)^2$$

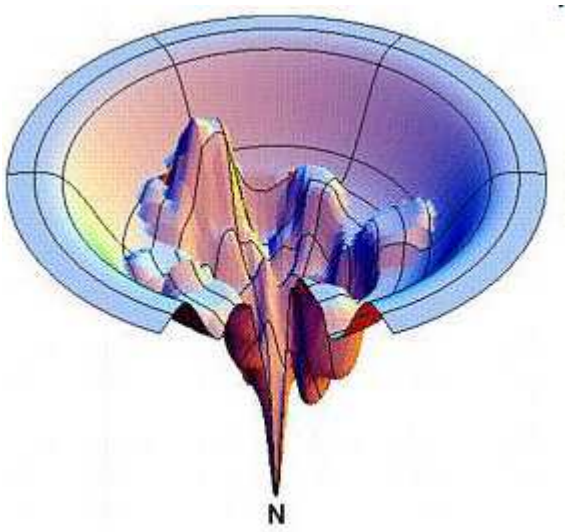
$$E_t = \frac{k_\theta}{2}\left(1 + \cos\left(n\phi - \phi_0\right)\right)$$

$$E_c = \frac{1}{4\pi\varepsilon_0\varepsilon_r}\frac{q_i q_j}{r_{ij}}$$

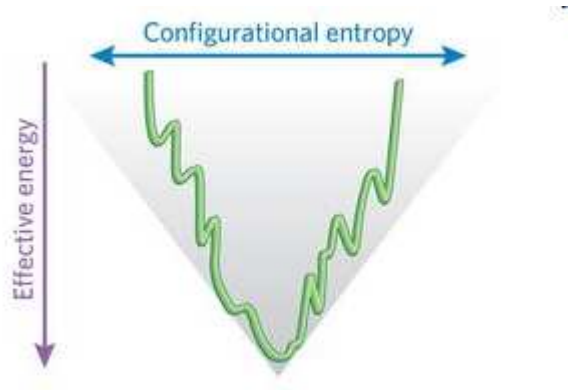
$$E_{vdw} = -2\varepsilon_{ij}\left(\frac{r_{ij}^*}{r_{ij}}\right)^6 + \varepsilon_{ij}\left(\frac{r_{ij}^*}{r_{ij}}\right)^{12}$$

Konformace

- Povrch potenciální energie (PES)
  - bariéry, minima
  - globální vz. lokální



- Pohyb po PES v proteinech
  - folding funnel
  - metody:
    - optimalizace energie
    - molekulární dynamika
    - simulované žíhání
  - Monte Carlo



Použití MM

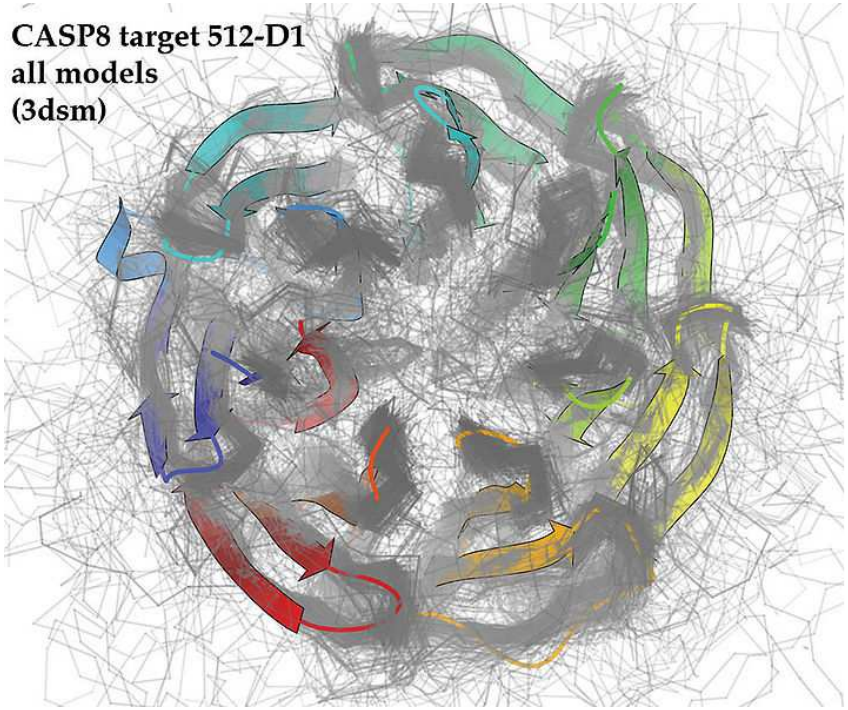
- v principu by měla jít použít na hledání globálního minima (všechny konformace a oskórování) – **protein folding**
- příliš náročné na výpočetní prostředky,
- používá se pouze u malých proteinů ke studiu např. **kinetiky skládání**, apod.
- případně k **rafinaci modelu** získaného homologním modelováním

Kontrola kvality

CASP

CASP

- **Critical Assessment of Techniques for Protein Structure Prediction**
- Comparison with prepublished x-ray data
- no prior information for predictors (double-blind)



- CASP10 will be public at 12/1/2013

http://predictioncenter.org/casp9/index.cgi
Proteins: Structure, Function, and Bioinformatics
Volume 77, Issue S9, Pages 1-228 (2009)

CASP

- tertiary structure prediction (all CASPs)
- secondary structure prediction (dropped after CASP5)
- prediction of structure complexes (CASP2 only; a separate experiment - CAPRI - carries on this subject)
- residue-residue contact prediction (starting CASP4)
- disordered regions prediction (starting CASP5)
- domain boundary prediction (CASP6-CASP8)
- function prediction (starting CASP6)
- model quality assessment (starting CASP7)
- model refinement (starting CASP7)
- high-accuracy template-based prediction (starting CASP7)

Best results

- Template–based

  I-Tasser

- Ab Initio

  QUARK

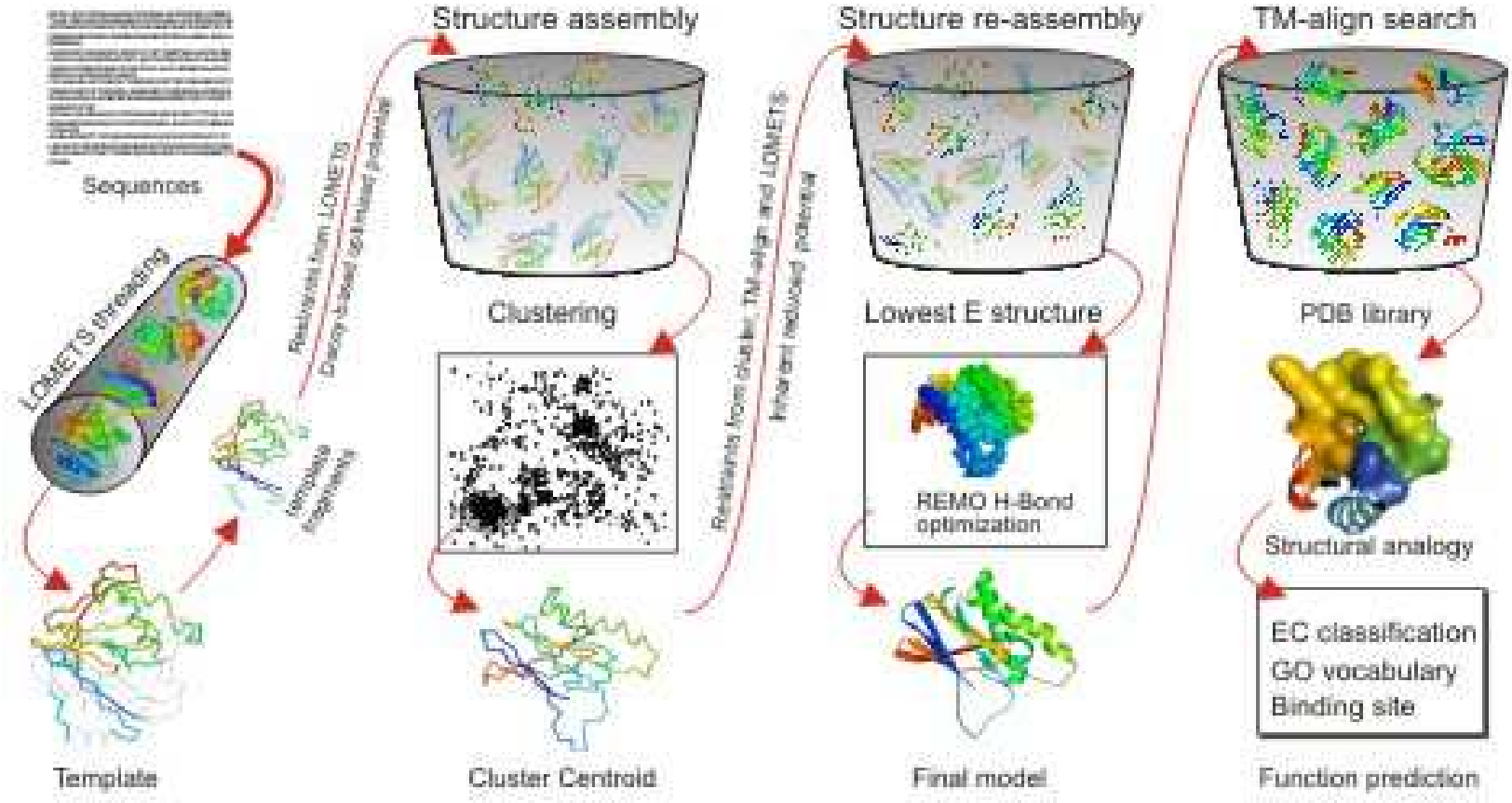
oboje Zhang Lab – University of Michigan

http://zhanglab.ccmb.med.umich.edu

I-TASSER

I-TASSER ONLINE
Protein Structure & Function Predictions

- Best automated server for prediction of 3D structure



- http://zhanglab.ccmb.med.umich.edu/I-TASSER/

I-TASSER

I-TASSER ONLINE
Protein Structure & Function Predictions

1. LOMETS
   – **metaserver** pro 8 metod predikce terciarni struktury fragmentu
2. složení fragmentů z identifikovaných templátů
   - replica-exchange Monte Carlo simulace
   - threading nepřeložených regionů (loops)
     pomocí ab initio modeling
3. SPICKER – clustrovani nejlepsich vysledku
4. znovu LOMETS dle clusteru
5. TM-align

I-TASSER ONLINE
Protein Structure & Function Predictions

energy terms

contact_cut.comm: Residue contact cutoff parameters
contact_profile.comm: Side-chain contacts environment profile
contact3.comm: Orientation-dependent side-chain contact potential
CA13.comm: Short-range C-alpha correlation of (i,i+2)
CA14.comm: Short-range C-alpha correlation of (i,i+3)
CA15.comm: Short-range C-alpha correlation of (i,i+4)
CA14s.comm: Short-range C-alpha correlation of (i,i+3) for strands
CA14h.comm: Short-range C-alpha correlation of (i,i+3) for helices
CA15s.comm: Short-range C-alpha correlation of (i,i+4) for strands
CA15h.comm: Short-range C-alpha correlation of (i,i+4) for helices
CB.comm: C-beta positions
sidechain.comm: Sidechain center positions

Quark

- ab initio protein folding and protein structure prediction
- construct correct protein 3D model from amino acid sequence only.
- QUARK models are built from a small fragments (1-20 residues long) by replica-exchange Monte Carlo simulation under the guide of an atomic-level knowledge-based force field.
- Since no global template information is used in QUARK simulation, the server is suitable for proteins which are considered without homologous templates.

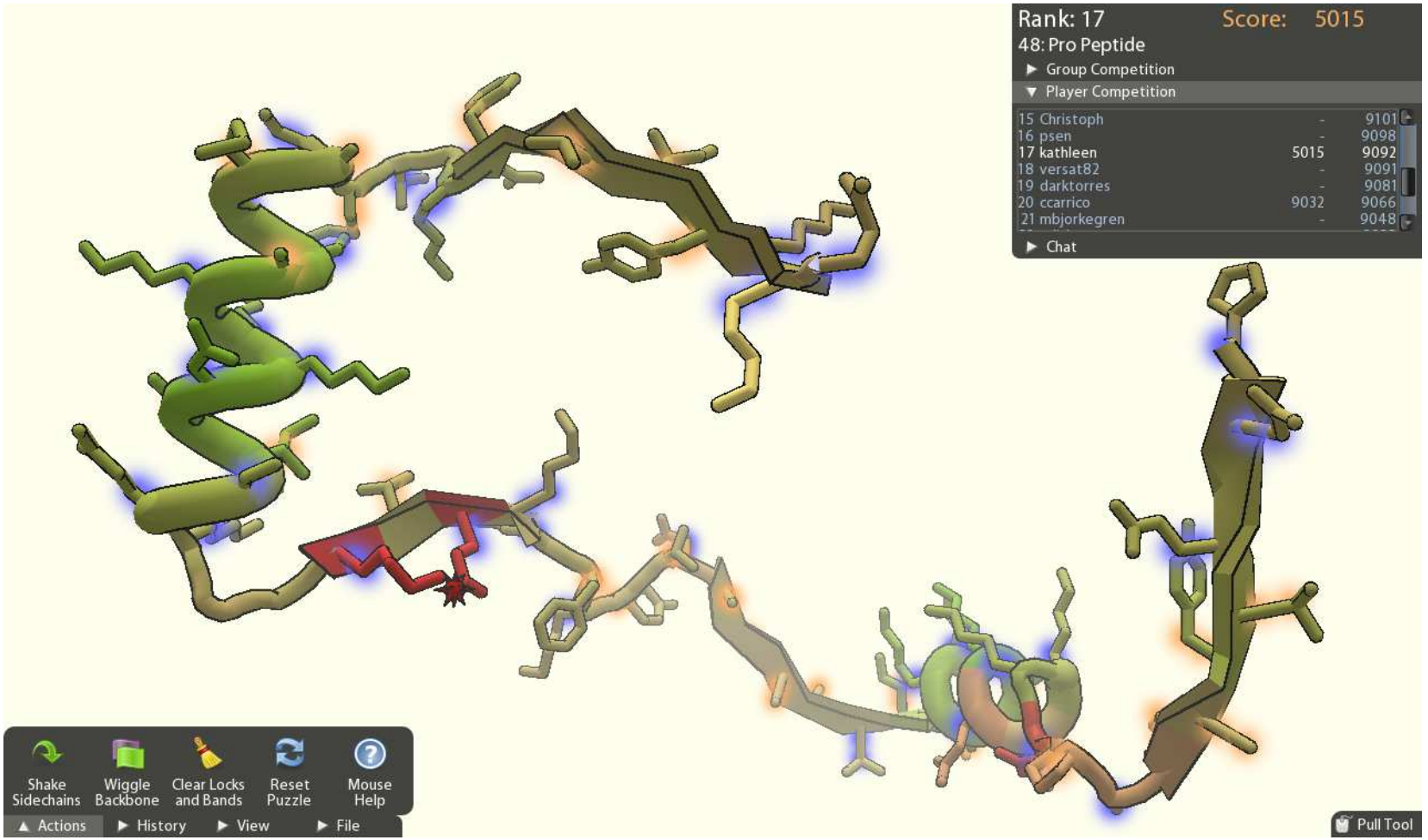
http://zhanglab.ccmb.med.umich.edu/QUARK/

D. Xu, Y. Zhang, Ab Initio Protein Structure Prediction using QUARK: II. Algorithm Flow and Benchmark Result

And now something completely different...

FoldIt

- skládání proteinů jako hra



- http://fold.it/portal/