Physics of threading.

Leonid Mirny
Eugene Shakhnovich
**Model**

**Computation of Z score:**
For a contact potential $E = \sum_{ij} U(a_i, a_j) \Delta ij$
Z score can be computed when average frequency $\langle \Delta ij \rangle$ of every contact $ij$ and covariance $\text{cov}(\Delta ij, \Delta kl)$ are known:

$$Z = \frac{(E_{nat} - \langle E \rangle)}{\sigma(E)} = \sum_{ij} U(a_i, a_j) \langle \Delta ij - \langle \Delta ij \rangle \rangle$$

**Protein structure**
- backbone: chain of connected $C_\alpha$ atoms.
- side chains: spheres of different radii centered at $C_\beta$ atom.

**Energy function:**
- contact cutoff is different for different residues
- we consider:
  - side chain – side chain ($E_{\beta\beta}$)
  - side chain – backbone ($E_{\beta\alpha}$)
  - backbone – backbone interactions ($E_{\alpha\alpha}$)

$$E = E_{\beta\beta} + E_{\beta\alpha} + E_{\alpha\alpha}$$

**Pairwise contact potential for each type of interaction**

$$E_{ij} = \sum_{ij} U(a_i, a_j) \Delta ij (R(a_j) + R(a_i))$$

where
- $a_i$ - amino acid sequence, i.e. identity of residue in position $i$
- $U(u, v)$ - energy of interaction between residues $u$ and $v$
- $\Delta ij(R) = 1$ if residues $i$ and $j$ are at distance less than $R$ = 0 elsewhere
- $R(u)$ - is radius of interaction for residue $u$

**TOTAL NUMBER OF PARAMETERS:**
231 interaction energies and 21 residue radii
This is achieved by maximizing

$$|\langle Z \rangle_{harm}| = \left| \frac{\sum_{m=1}^{M} \frac{1}{Z_m}}{M} \right|$$

Poor Potential

$$\langle Z \rangle_{harm}(Poor) = \frac{3}{\frac{1}{-1} + \frac{1}{-5} + \frac{1}{-4.5}} = -2.1$$

Energy spectra
Good Potential

\[
\langle Z \rangle_{harm}(Good) = \frac{3}{-3} + \frac{1}{-3.5} + \frac{1}{-4.5} = -3.57
\]

\[
|\langle Z \rangle_{harm}(Good)| > |\langle Z \rangle_{harm}(Poor)|
\]
Convergence

Correlation between potentials obtained for different number of proteins shows that procedure converges to a single potential.

Correlation between potential obtained for database of 100→ proteins and databases containing of 1..30 proteins.
Real Proteins

Using our procedure we derive potential for real proteins from PDB.
We study convergency of the method by deriving potential from databases containing different number of proteins.

Native sequence are NOT RANDOM!
Native sequences are NOT VERY WELL DESIGNED!
Values of $Z$ scores obtained for our potential is greater by modulus than $Z$ scores computed using other published potentials.
**Sampling technique:**

**Monte Carlo Threading**

Monte Carlo (MC) threading is used to find an alignment of a sequence with a structure that minimizes the energy.

MC makes random moves aligning sequence to a structure. Each move is accepted with probability 1 if it decreases the energy or with probability $\exp(-\delta E/T)$ if it increases the energy.

$T$ – MC temperature is an optimization parameter

- No penalties for gaps!

Instead we apply the following constraints:

- No gaps in the middle of secondary structure elements are allowed
- Each region of matches should be longer than 4 residues
- Gaps can be inserted between any pair of residues which are closer than 6Å apart in space.

---

**Moves of Monte Carlo threading**

- **Merge two fragments**
  - IDVLL
  - FPHN1

- **Split a fragment**
  - GVDAKISMDIlnK

- **Move a fragment**
  - GVDAKISMDIlnK

---
Monte Carlo Threading

- We allow gaps in both sequence and structure.
- No energy penalties for gaps!
- Fragments of continues matches \( \geq 6 \) residues.
- Sampling is controlled by temperature \( T \).
Alignment

\[ I \]
\[ J \]

\[
\begin{array}{cccccccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 \\
1 & 2 & 3 & 4 & 5 & 6 & 7 & 10 & 11 & 12 & 13 & 14 & 15 \\
1 & 2 & 3 & 4 & 5 & - & 6 & 7 & 10 & 11 & - & 12 & 13 & 14 \\
\end{array}
\]
Model

Protein of length \( I \) has a sequence \( a_1, a_2, \ldots a_I \) and distances \( r_{ii'} \) between \( C_\beta \) atoms of residues \( i \) and \( i' \).

- Pairwise contact potential

\[
E(\mathbf{B}, \Delta) = \sum_{i,i' > i+2}^I B_{ii'} \Delta_{ii'}
\]

\[
\Delta_{ii'} = \begin{cases} 
1 & \text{if } i \text{ and } i' \text{ are in contact: } r_{ii'} < R_{cut} \\
0 & \text{otherwise}
\end{cases}
\]

\( R_{cut} = 8 \text{Å} \).

- Residues-residue interactions

**Go-model** \( B_{ii'} = -\Delta_{ii'} \).

**Potential** \( B_{ii'} = U(a_i, a_{i'}) \), where \( U(u,v) \) is a potential (i.e. \( 20 \times 20 \) matrix of residue-residue interactions)
How to find a sequence-structure alignment with the lowest energy?

**Problem:** Finding an optimal alignment (with gaps) of a sequence to a structure is NP-hard and no exact algorithm is known.

**Idea!** Use Monte Carlo procedure to

1. search for the optimal alignment and

2. sample possible (suboptimal) alignments.

MC also allows systematic study of the energy landscape!
Plan

- Introduction: What is threading?

- Model and Methods: How it works?

- Approach of an "ideal potential".

- Important tests.

- Thermodynamics and kinetics of threading.

- Conclusions.
Monte Carlo Threading

1. Start with a random alignment.

2. Select a fragment at random

3. Try to
   - expand/shrink
   - shift
   - split/merge

   the fragment

4. Compute $\Delta E$

5. Accept or reject the move according to Metropolis (at temperature $T$).

Move set

A.

B.

C.

D.
How to test threading when no exact potential is known?

Idea! use an approach of “ideal potential”.

- **Go model.** Make all native contacts attractive.

- **Sequence Design.** Using a reasonable potential design “native” sequences to provide a large energy gap to the native structure.
**Example:** Ubiquitin and three proteins with similar structures (analogs)

- **GUA** C-Raf1 fragment  \( \text{SeqID} = 14\% \)
  \( dRMS = 2.2\text{Å}(58 \text{ res}) \ Q = 0.82(69 \text{ res}) \)

- **IGD** G-protein  \( \text{SeqID} = 4\% \)
  \( dRMS = 2.8\text{Å}(46 \text{ res}) \ Q = 0.69(52 \text{ res}) \)

- **FRR** Ferredoxin I  \( \text{SeqID} = 11\% \)
  \( dRMS = 3.6\text{Å}(63 \text{ res}) \ Q = 0.53(60 \text{ res}) \)

**control** **PLC** Plastocyanin

\( dRMS = 5.6\text{Å}(50 \text{ res}) \ Q = 0.44 \) (60 res)
Thermodynamics of threading
Transition to the optimal alignment

- Close template (Self, UBI-GUA) → cooperative (1st-order-like) transition.
- Distant template (UBI-IGD, UBI-FRR) → non-cooperative transition.
Equilibrium distribution of alignments

Structure-structure

Threading

Ubi-Ubi

Ubi-Gua

Ubi-Gld
Ensemble of alignments

Structural  Threading

UBI-UBI
self-alignment

UBI-GUA
close template

UBI-IGA
instant template
Folding vs Threading

Exact potential

Real (noisy) potential

Random alignments
Almost correct alignments
Correct alignments

Native
Folding through a similar structure
Threading through a moderately similar structure

Almost correct alignments
Correct alignments

Folding through a similar structure
Threading through a moderately similar structure
Conclusions

- Monte Carlo threading is very efficient in finding the optimal alignment and sampling possible alignments.

- 1st- or 2nd-order phase transitions are observed depending on the degree of similarity between the analog and the native structure (and other factors).

- 1st order transition is an indicator of a very good prediction!

- The optimal alignment is **stable** to errors in potential when structural similarity is high and **unstable** when the similarity is low.

- To find the right alignment one needs
  - very accurate potential and
  - a known protein structure very similar to the native one.