Protein Scoring

Mikhail Karasikov

MIPT, Skoltech, INRIA

Fall 2016

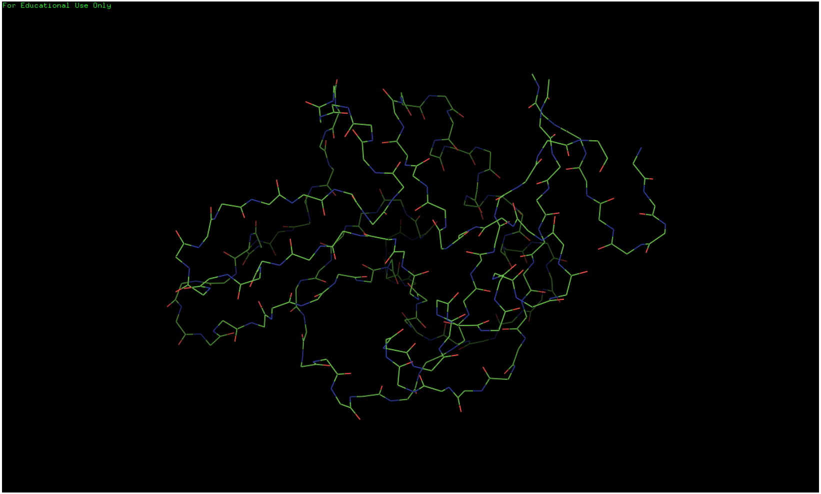
Proteins

- Protein — a sequence of amino acids $\{\text{Ala}, \text{Arg}, \dots\} =: \mathcal{A}$
- Each amino acid consists of atoms
  E.g. (Cysteine):

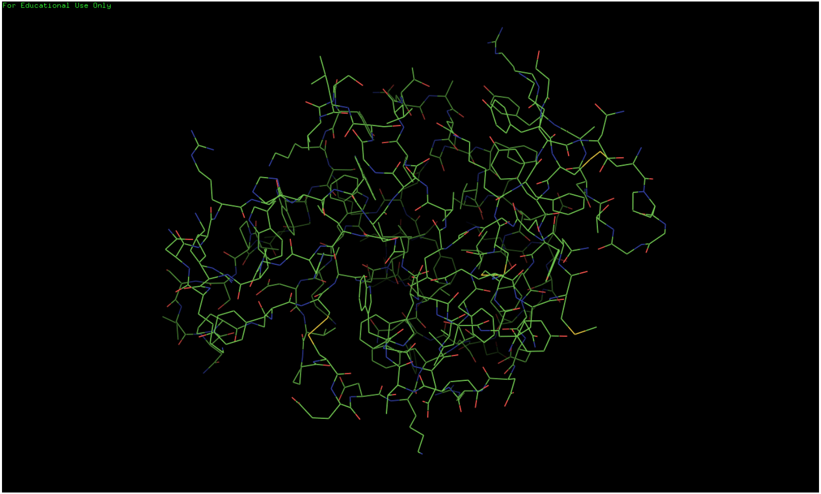
$$[\underbrace{N, C_\alpha, C, H, O}_{\text{backbone part}}, \underbrace{H_\alpha, C_\beta, H_{\beta_1}, H_{\beta_2}, S_\gamma, H_\gamma}_{\text{side-chain}}]$$

- Primary structure — linear sequence of amino acids
- Tertiary structure — 3D structure of protein molecules

Protein backbone



Backbone with side-chains



Rotamer prediction problem statement

Given

Protein backbone

Predict

Rotamers — discretized conformations of side-chains

In other words: predict folding of side-chains.

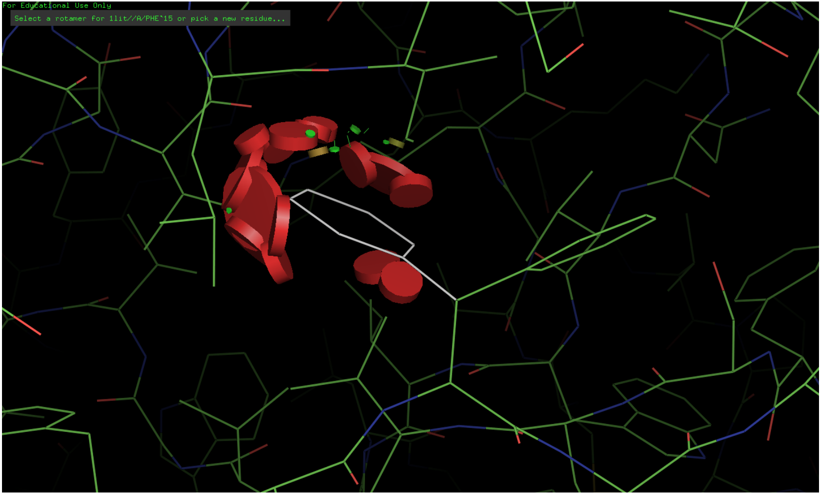
Quality criteria

RMSD-like metrics based on the side-chain geometry

The key

Protein folds according to physical laws, minimizing free energy F

Rotamers



For Educational Use Only

Select a rotamer for 1lit//A/PHE`15 or pick a new residue...

Mathematical formulation

m — sequence length,
$n_k < \infty$ — number of rotamers for k-th amino acid,
$r_k \in \{1, \dots, n_k\} =: \mathcal{R}_k$ — indices of rotamers, $\mathcal{R} = \times_{k=1}^{m} \mathcal{R}_k$,
$U_{kl}(r_k, r_l)$ — symmetrical potentials of pairwise interactions,

Potential energy minimization:

$$\sum_{k=1}^{m} \sum_{l=1}^{m} U_{kl}(r_k, r_l) \to \min_{(r_1,\dots,r_m)\in\mathcal{R}} \tag{1}$$

Drawbacks:

- There are potentials of higher orders
- Actually, it is not free, but potential energy minimization

Problem statement for protein design

Given

Protein backbone

Find

Primary structure that folds to the target protein structure

Quality criteria

Depends on particular problem statement

- computational time
- similarity of primary structure and the native structure
- consistency with predicted secondary structure:

$$L(3D \xrightarrow{f}_{\varepsilon} 1D \to_{\delta} 2D,\ 3D \to_{0} 2D) \to \min_{f}.$$

Notation

m — number of residues,
$a_k = 1, \ldots, 20$ — amino-acids,
$n = \sum_{k=1}^{m} n_k$ — dimension of the search space,
$E_{kl}(a_k, a_l)$ — energy.

Protein design optimization problem:

$$\sum_{k=1}^{m} \sum_{l=1}^{m} E_{kl}(a_k, a_l) \to \min_{(a_1, \ldots, a_m) \in \mathcal{A}^m} \tag{2}$$

Reduction to boolean Quadratic Programming

Problem 2 can be reduced to BQP

$$\begin{aligned} \underset{\vec{x}\in\{0,1\}^n}{\text{minimize}} \quad & \vec{x}^{\mathsf{T}}\mathbf{Q}\vec{x} \\ \text{subject to} \quad & \mathbf{A}\vec{x} = \vec{1}_m, \end{aligned} \tag{3}$$

where

$$[\mathbf{Q}]_{ij} = E_{ij}(a_i, a_j),$$

$$\mathbf{A} = \left[\begin{array}{ccccccccccc} 1 & \cdots & 1 & 0 & \cdots & 0 & \cdots\cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & \cdots\cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \cdots\cdots & \vdots & \ddots & \vdots \\ \underbrace{0}_{} & \cdots & 0 & 0 & \cdots & 0 & \cdots\cdots & 1 & \cdots & 1 \end{array}\right] .$$

$$\underbrace{\quad\quad\quad}_{20} \quad \underbrace{\quad\quad\quad}_{20} \quad\quad\quad \underbrace{\quad\quad\quad}_{20}$$

Final optimization problems

Rotamer prediction 1

$$\sum_{k=1}^{m}\sum_{l=1}^{m} U_{kl}(r_k, r_l) \to \min_{(r_1,\dots,r_m)\in\mathcal{R}}$$

Protein design 2

$$\sum_{k=1}^{m}\sum_{l=1}^{m} E_{kl}(a_k, a_l) \to \min_{(a_1,\dots,a_m)\in\mathcal{A}^m}$$

But we do not know actual potentials U_{kl} and E_{kl}!

Another look

1. $(r_1, \dots, r_m)$ and $(a_1, \dots, a_m)$ can be treated as proteins

$$P \in \mathcal{P}$$

2. energy potentials can be treated as protein scoring functions

$$\sum_{k=1}^{m} \sum_{l=1}^{m} U_{kl}(r_k, r_l) =: S_1(r_1, \dots, r_m)$$

$$\sum_{k=1}^{m} \sum_{l=1}^{m} E_{kl}(a_k, a_l) =: S_2(a_1, \dots, a_m)$$

Introduced notation

Rotamer prediction 1

$$S_1(r_1, \ldots, r_m) \to \min_{(r_1,\ldots,r_m)\in\mathcal{R}}$$

Protein design 2

$$S_2(a_1, \ldots, a_m) \to \min_{(a_1,\ldots,a_m)\in\mathcal{A}^m}$$

So, the problem is to score proteins $P \in \mathcal{P}$.
Here we can apply machine learning!

Protein scoring

For each native structure P_0 a set of decoy structures $\mathcal{D}$ is given:

$$\mathcal{D} = \{P_1, \dots, P_m\} \subset \mathcal{P}$$

Find

Scoring

$$(i_1, \dots, i_m) : \; P_{i_m} \preceq \dots \preceq P_{i_1} \prec P_0.$$

The problem is to train protein scoring function

$$S : \; \mathcal{P} \to \mathbb{R}.$$

Then

$$S(P_0) < S(P_{i_1}) \leqslant \ldots \leqslant S(P_{i_m}).$$

Performance estimation

First, we have to define the actual score function $S^*(P)$.

1. RMSD

$$S^*(P_i) = \text{RMSD}(P_i, P_0)$$

2. TM-score (Template modelling score)

$$\max\left[\frac{1}{L_{\text{target}}}\sum_i^{L_{\text{aligned}}}\frac{1}{1+\left(\frac{d_i}{d_0(L_{\text{target}})}\right)^2}\right]$$

3. GDT-TS (Global distance test, total score)
4. GDT-HA (Global distance test, high accuracy)

Then we estimate:

- Loss, Z-score
- Pearson/Spearman correlation

Two approaches

1. **Single-model QA**
   - Computationally efficient
   - Have far from perfect quality
2. Consensus-model QA

$$S(P_i) = \frac{1}{|\mathcal{P}|} \sum_{P \in \mathcal{P}} \rho(P, P_i)$$

   - More precise
   - Hard to compute

Methods

1. **Machine learning**
   - Features extraction
   - Allows using 2D information
   - Robust to errors in side-chain positions
2. Statistical potentials

   $\mathcal{A}$ — atoms

   $\mathsf{AT} = \{\mathsf{at}_1, \dots, \mathsf{at}_m\}$ — atom types

   $\mathsf{at} : \ \mathcal{A} \to \mathsf{AT}$

$$S(\mathsf{at}(a_i), \mathsf{at}(a_j), r_{ij}) \propto -kT \log \hat{p}(\mathsf{at}(a_i), \mathsf{at}(a_j), r_{ij})$$

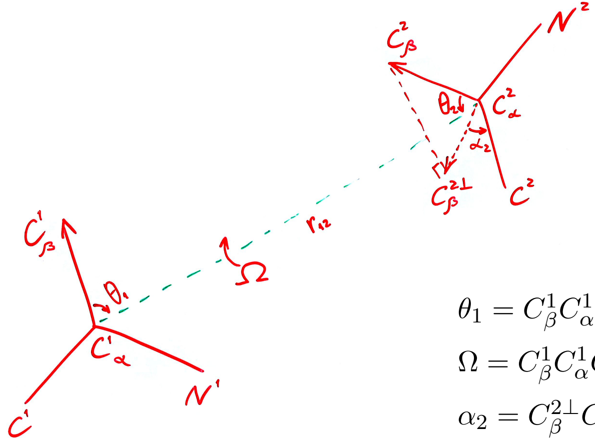
$$S(P) = \sum_{a_i \neq a_j} S(\mathsf{at}(a_i), \mathsf{at}(a_j), r_{ij})$$

Single-model QA

1. **Coarse-grained** model
   Uses only backbone conformation
   - Applied first to predict backbone conformation
   - Computationally efficient
   - Robust to errors in side-chain positions
2. All-atoms model
   Uses all protein's atoms
   - Applied on the stage of refinement
   - Usually more precise

1. Reduced representation terms
    - Predicted secondary structure penalty
    - Solvent accessibility
    - Predicted contact map
    - Sheet formation
    - Backbone repulsion
    - Centroid repulsion
    - Residue environment potential
    - Context independent pair-wise potential
    - Context dependent pair-wise potential
    - Compactness
2. All-atom terms
    - Side-chain hydrogen bonding
    - Van der Walls forces
    - Solvation effects
    - Electrostatic interactions

Geometrical Features



$$\theta_1 = C^1_\beta C^1_\alpha C^2_\alpha$$

$$\Omega = C^1_\beta C^1_\alpha C^2_\alpha C^2_\beta$$

$$\alpha_2 = C^{2\perp}_\beta C^2_\alpha C^2$$

Geometrical Features

Featurization:

$$\{P_0, P_1, \dots, P_m\} \mapsto \{\vec{x}_0, \vec{x}_1, \dots, \vec{x}_m\}$$

Learning:

1. Classification

$$y_0 := -1;\ y_i := 1,\ 1 \leqslant i \leqslant m.$$

2. Regression

$$y_i := S^*(P_i),\ 0 \leqslant i \leqslant m$$

3. Learning to Rank

$$P_{i_m} \preceq \cdots \preceq P_{i_1} \prec P_0$$

Results

Таблица: Top 1, Top 5, Spearman correlation

	Logistic Regression	Ridge Regression
Tasser	0.75 / 0.82 / 0.61	0.16 / 0.41 / 0.72
Tasser Original	0.84 / 0.91 / 0.10	0.73 / 0.79 / 0.22
Rosetta	0.93 / 0.97 / 0.62	0.14 / 0.48 / 0.73
Rosetta Original	0.00 / 0.05 / 0.03	0.14 / 0.31 / 0.17
Modeller	0.80 / 0.85 / 0.69	0.25 / 0.40 / 0.78
Modeller Original	0.90 / 0.90 / 0.49	0.55 / 0.65 / 0.74