I86 – Alfold: a Statistical Potential for Protein Folding

Andres Colubri, Jessica Lynn Raddatz – University of Chicago, Chicago, USA
David Angulo, DePaul University, Chicago, USA

1. Introduction

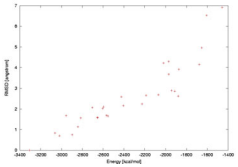
a) We describe an all-atom statistical potential, **Alfold**, that incorporates environmental effects through backbone hydrogen bond, solvation and pair-wise terms.
b) We present a number of software packages to mine the PDB, rank decoy sets and run Monte-Carlo folding simulations.

2. Objectives

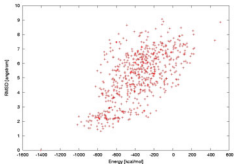
a) To construct an all-atom statistical potential, **Alfold**, that models detailed backbone and side-chain packing.
b) To improve the performance of previous statistical potentials by accounting explicitly for hydrogen bond and solvation.
c) To derive a mean-force version of **Alfold** to use in "coarse grained" folding simulations.

4. Results

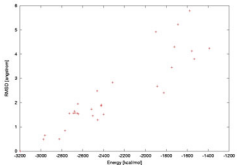
a) Fold recognition: The decoy sets from "Decoys "R" us" website were ranked with **Alfold**. Four Energy-RMSD plots are shown here. In all cases, the energy minimum corresponds to the native structure.



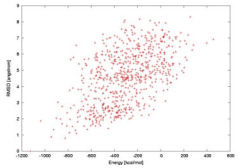
1BAB-B from "hg-structal" set



1CTF from "4state-reduced" set

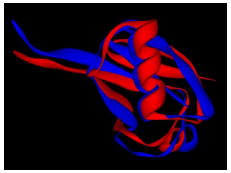


1HDA-A from "hg-structal" set

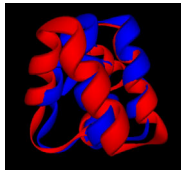


2CRO from "4state-reduced" set

b) Ab-initio folding: The mean force version of **MiniFire** was used for energy minimization (**MiniFire** is another all-atom statistical potential **[4]**). Below, two predicted structures (blue) are superposed to the native conformations (red).



3.0Å prediction for 1UBI (76 aa)



2.8Å prediction for 1O82 (70 aa)

3. Methods

a) Statistical model: We use a non-redundant subset of the PDB to derive the probability function $P(structure|at_1,...at_M)$. We expand a previous statistical approach **[1]** to the all-atom case:

$$P(structure \mid at_1,...,at_M) = \frac{P(structure)}{P(at_1,...,at_M)} P(at_1,...,at_M \mid structure) \cong \left(P_{HB}(\{\delta_{ij},\theta_{ij},\phi_{ij},e_{ij}\}_{ij})\right) P(at_1,...,at_M \mid structure) =$$

Atom indices

Realistic protein structures have buried backbone hydrogen bonds **[2]** which define the secondary structure elements (helices/strands). Based on this, we assume that $P(structure)$ is solely determined by the geometry (δ_{ij}, θ_{ij}, ϕ_{ij}) and solvent exposure (e_{ij}) of the hydrogen bonds:

$$P(structure) = \prod_{i,j} P(\delta_{ij},\theta_{ij},\phi_{ij},e_{ij}) = P_{HB}(\{\delta_{ij},\theta_{ij},\phi_{ij},e_{ij}\}_{ij})$$

Expansion of joint probabilities:

$$P(x_1,x_2,...x_n) = \prod_i P(x_i) \prod_{i,j} \frac{P(x_i,x_j)}{P(x_i)P(x_j)} \prod_{i,j,k} ...$$

$$= P_{HB} \prod_i P(at_i \mid structure) \prod_{i,j} \frac{P(at_i,at_j \mid structure)}{P(at_i \mid structure)P(at_j \mid structure)} \prod_{i,j,k} ... \approx$$

$$\approx P_{HB} \prod_i P(at_i \mid e_i) \prod_{i,j} \frac{P(at_i,at_j \mid d_{ij},e_{ij})}{P(at_i \mid d_{ij},e_{ij})P(at_j \mid d_{ij},e_{ij})} = P_{HB}P_{solv}P_{pair}$$

Single atom probabilities are dependent only on the solvent exposure (e_i). Pair-wise probabilities depend on the atoms distance d_{ij} as well.

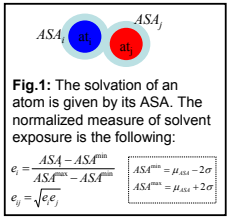


Fig.1: The solvation of an atom is given by its ASA. The normalized measure of solvent exposure is the following:

$$e_i = \frac{ASA - ASA^{min}}{ASA^{max} - ASA^{min}}$$

$$e_{ij} = \sqrt{e_i e_j}$$

$$ASA^{min} = \mu_{ASA} - 2\sigma$$

$$ASA^{max} = \mu_{ASA} + 2\sigma$$

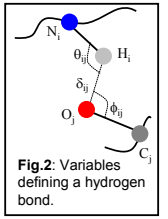


Fig.2: Variables defining a hydrogen bond.

b) Flowchart for the construction of Alfold:

Data mining stage – Protein Cruncher **[A]**
1. Create non-redundant set of protein structures from PDB.
2. Calculate atom distances, bond angles and ASA parameters.

↓

Statistical analysis stage – Alfold scripts **[B]**
1. Bin distances, angles and ASA's.
2. Calculate P_{HB}, P_{solv} and P_{pair}.
3. $E(x) = -RT \ln(P_{HB}P_{solv}P_{pair})$

↓

Testing stage
1. Fold recognition: rank structures from Decoy 'R' us website **[C]**
2. Ab-initio folding by Monte Carlo sampling – OOPS **[D]**

c) Potential of mean force: For *ab-initio* folding simulations, we consider only the backbone and C^β atoms **[3]**. Therefore, we need to average the all-atom version of **Alfold:**

$$E_{AA}(x) = -RT \ln\left(P_{HB}P_{solv}P_{pair}\right)$$

$$E_{CB}(\tilde{x}) = \left\langle E_{AA}(x) \right\rangle_{side-chains}$$

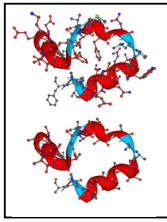


Fig.3: All-atom and C^β-only representations of villin (1VII). YAPView **[E]** was used to create these figures.

5. Conclusions

a) Alfold successfully recognizes the best structures from large decoy sets. The Energy-RMSD correlation is high and in most of the cases the native structure has the lowest energy.
b) Additional tests are being done to quantify the relative importance of the hydrogen-bond and solvation terms.
c) The side-chain average procedure used in *ab-initio* coarse grained simulations works remarkably well for small (up to 100 residues) globular proteins.

6. References

[1] Simons, KT, Kooperberg, C, Huang, E and Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using Simulated Annealing and Bayesian scoring functions. *J Mol Biol.* (1997), 268:209-225.
[2] Fernandez A, Sosnick TR, Colubri A. Dynamics of hydrogen bond desolvation in protein folding. *J Mol Biol.* (2002) 321:659-75.
[3] Colubri, A. Prediction of Protein Structure by Simulating Coarse-grained Folding Pathways, a Preliminary Report. *J Biomol Struct Dyn.* (2004) 21:625-38.
[4] Min-yi Shen. Personal communication.

7. Links

[A] Protein Cruncher: a package for PDB mining.
http://sosnick.uchicago.edu/aifoldlab/ProtCruncher/ProtCruncher.html
[B] Alfold scripts. http://sosnick.uchicago.edu/aifoldlab/OOPS/Alfold/Alfold-0.5.tar.gz
[C] Decoy 'R' us website: repository of protein decoys. http://dd.stanford.edu/
[D] OOPS is an Open Protein Simulator:
http://sosnick.uchicago.edu/aifoldlab/OOPS/OOPS.html
[E] YAPView: Yet Another Protein Viewer:
http://sosnick.uchicago.edu/aifoldlab/YAPView/YAPView.html