Knowledge-based potentials in protein fold recognition

Mehdi Mirzaie[1,*], Mehdi Sadeghi[2, 3]

[1] Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
[2]National Institute of Genetic Engineering and Biotechnology, Tehran, Iran
[3]School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran.

*Corresponding author: e-mail address: mirzaie@ipm.ir (M.Mirzaie)

ABSTRACT

An accurate potential function is essential for protein folding problem and structure prediction. Two different types of potential energy functions are currently in use. The first type is based on the law of physics and second type is referred to as statistical potentials or knowledge based potentials. In the latter type, the energy function is extracted from statistical analysis of experimental data of known protein structures. By increasing the amount of three dimensional protein structures, this approach is growing rapidly. There are various forms of knowledge based potentials depending on how statistics are calculated and how proteins are modeled. In this review, we explain how the knowledge based potentials are extracted by using known protein structures and briefly compare many of the potentials in theory.

Keywords: Knowledge Based Potentials; Reference State; Accessible Surface; Protein Structure; Decoy Structure.

INTRODUCTION

Proteins are macromolecules that are formed by amino acids and linked together with peptide bonds. These biological macromolecules perform a wide variety of functions in organisms such as catalysis reactions and transporting. Also, almost all diseases can be related to the function or malfunction of proteins. The function of proteins is a consequence of their unique three dimensional structures and through their binding to other molecules such as DNA, RNA or proteins.

In 1973, Anfinsen [1] showed that the structure of protein is dictated by its amino acid sequence. In fact, he showed that an unfolded protein could refold to its biologically active conformation. Therefore, the main problem in protein research is modeling and predicting the relationship between sequence and structure. Anfinsen's results led to the thermodynamic hypothesis of folding, which demonstrates that the native structure of proteins fold in the lowest potential energy function. Therefore, based on this hypothesis all studies of proteins including structure prediction, folding simulation and protein design depend on an accurate energy function.

Two different types of potential energy functions are currently in use. The first type is based on the law of physics. In physical energy function, a molecular mechanics force field is used. Molecular mechanics force fields such as AMBER [2-6], CHARMM [7-8], GROMOS [9], ECEPP [10-12] and OPLS [13-14] are parameterized from ab-initio calculation and small molecular structural data. They are essentially summation of pair wise electrostatic and Vander Waals interaction energies, bonds, angles and dihedral angles terms. In addition, terms such as entropy and solvent effect are implicitly included [15-16]. These potential functions are very time consuming, so these functions have been out of favor in protein structure prediction [17].

To reduce computational complexity, second type of potential energy function is used. These types are referred to as statistical potentials, knowledge based potentials, scoring functions or empirical potentials [17]. In this type, the energy function is extracted from statistical analysis of experimental data of known protein structures [18-25]. In the last decades, this approach was rapidly growing as a consequence of the increasing amount of the experimentally determined three dimensional protein structures. There are various forms of statistical energy functions depending on how statistics are calculated and how proteins are

modeled [26], e.g. distance independent contact energies[27-35], solvent accessible surface potential[22,25,31], packing density potentials, distance dependent potentials[20,27,36-50] and angular dependence[31, 51-54]. Recently, combination of these statistics such as distance and orientation are widely used [55-64]. Initially, statistical potentials were based on statistical mechanics and Boltzmann law [27,55,65], but recently employ many other ideas, such as conditional probabilities [39], linear and quadratic programming on various decoy sets [66-68] and information theory [69-72]. The dependence of statistical potentials on structural data base is also studied [73-74].

Most often, statistical potentials use the Boltzmann law to convert the observed frequencies into potentials. These potentials are obtained as the ratio of observed and expected frequencies, where the observed frequencies are the number of occurrence in known protein structures and expected frequencies are the number of occurrence in the absence of any interaction which serve as reference states. Therefore, depending on this consideration, there are various forms of statistical potential functions.

In this review, we explain how the statistical potentials are extracted by using known protein structures and briefly compare many of the potentials in theory.

Derivation of knowledge based potentials

For derivation of knowledge based potentials, at first, the structural representative of protein reduced to the coordinates of C_α, C_β or side chain centers or all atoms. Once the amino acid sequence and reduced structure are given, the protein descriptors are extracted. A descriptor can be, e.g. the distance between pair of atoms, solvent accessible surface area, backbone or side chain dihedral angle, packing density or any other features of protein. Therefore, a protein structure can be represented by a vector $c = (c_1, c_2, \dots, c_d)$ where each c_i is a descriptor.

Knowledge based potentials are a simple consequence of the Boltzmann distribution. According to the Boltzmann law, the distribution of protein molecules among the microscopic states at the equilibrium state is related to potential function that means for a microstate *C* (descriptor) the probability of occupancy *P(c)* connects to potential function *E(c)* as follows:

$$P(c) = \frac{\exp(-\frac{E(c)}{RT})}{Z} \quad (1)$$

Where R is Boltzmann constant and T is the absolute temperature measured in Kelvin and Z is the partition function:

$$Z = \sum_c \exp(\frac{-E(c)}{RT})$$

From equation (1):

$$E(c) = -RT\ln(c) - RT\ln Z$$

The efficient knowledge based potential must consider the sequence-structure relation of protein, so the energetic interactions that are independent of the protein sequence and protein structure must be removed. This energetic contribution is referred to as *reference state*. That means

$$\Delta E(c) = E(c) - E'(c)$$

So, the efficient potential energy is

$$\Delta E(c) = -RT\ln\left(\frac{P(c)}{P'(c)}\right) - RT\ln\left(\frac{Z}{Z'}\right)$$

where, $P'(c)$ is the probability of the descriptor *c* in the reference state. *Z* and *Z'* are both constant and usually assumed that *Z=Z'*. So

$$\Delta E(c) = -RT\ln(\frac{P(c)}{P'(c)})$$

By assuming that the probability distribution to each descriptor is independent, we have:

$$\frac{P(c)}{P'(c)} = \prod_i \frac{P(c_i)}{P'(c_i)}$$

Therefore:

$$\Delta E(c) = -RT\sum_i \ln(\frac{P(c_i)}{P'(c_i)})$$

where, $P(c_i)$ and $P'(c_i)$ are the probability of the i^{th} descriptor in native proteins and the reference state, respectively. $P(c_i)$ can be estimated by counting frequency of i^{th} descriptor in data base of native protein structures and $P'(c_i)$ is the probability of the i^{th} descriptor in reference state. Therefore, the choice of reference state is critical and

effective for knowledge based potentials. The portion of i^{th} descriptor energy ΔE_i is

$$\Delta E_i = -\mathrm{RT}\ln\frac{P(c_i)}{P'(c_i)} \quad (2)$$

Distance dependent potentials

Descriptor for distance dependent potentials is distance of interactions such as distance between pair of atoms or residues. The distance of interactions is usually divided into a number of small intervals. We use *(i,j,d)* to represent the k^{th} descriptor c_k . The effective energy function is equal to

$$-RT\ \ln\frac{N_{obs}(i,j,d)}{N_{exp}(i,j,d)}$$

where, $N_{obs}(i,j,d)$ represent the observed number of *(i,j,d)* interacting pairs which can be estimated from database of known protein structures and $N_{exp}(i,j,d)$ represents the expected number of *(i,j,d)* interacting pairs in the reference state which typically results from calculations or simulations.

The major difference of these types of potentials is in the atom type definition and derivation of the reference state. Sippl [20] first proposed a model of reference state which is known as the "uniform density" [75]. He assumed that each pair of contacting atom types in reference state is uniformly distributed along the distance between them. Based on this assumption Samudrala and Moult [39] calculate the $N_{exp}(i,j,d)$ as

$$\frac{N_{obs}(i,j)}{N_{total}}N_{obs}(d)$$

where, $N_{obs}(d)$ is the number of occurrences of interacting of any atom type at distance d and N_{total} is the total number of interacting pairs observations.

Lu and Skolnick [40] employed a quasi-chemical approximation and estimated $N_{exp}(i,j,d)$ as

$$N_{exp}(i,j,d) = \chi_i\chi_j N_{obs}(d)$$

where χ_k is the mole fraction of atom type *k*. The higher population of hydrophobic residues than that of hydrophilic residues at the core of proteins led to unphysical long range repulsion between hydrophobic residues in statistical potential based on Sippl's Assumption [76]. Zhou and Zhou [46] proposed a reference state

$$N_{exp}(i,j,d) = N_{obs}(i,j,d_{cut})\frac{V_d}{d_{cut}}$$

by assuming that atom types can be modeled as ideal gas molecules (called DFIRE for *distance scaled finite ideal gas reference state*). In fact in this model, it is assumed that the distribution of interaction pairs follows the uniform distribution in whole volume of the protein [75]. The DFIRE reference state is as where $d_{cut} = 14.5$ and V_d is the volume of a spherical shell of width Δd at distance d from the center.

In 2006, Shen and Sali used no interacting atoms in a homogeneous sphere (called DOPE for *Discrete Optimized Protein Energy*) as reference state [48]. DOPE and DFIRE were derived from a non-interacting ideal gas reference state with the difference that in DOPE the size effect of proteins takes in to account.

Side chain packing is very important determinant for protein structure [77]. Described knowledge based potentials are limited in their ability to describe side chain packing. Recently, the orientation dependent and all atom potential (called OPUS-PSP) have been proposed. In this potential function a set of 19 rigid body blocks were extracted from the chemical structures of 20 amino acids to capture the feature of side chain packing. [63]

Protein is a continuous sequential chain of amino acids and reference state should correctly reflect and counteract the chain connectivity effect. The ideal gas reference state cannot be able to capture this feature [78]. Recently, Cheng etal, proposed a more physical reference state that is based on free rotating and self avoiding chain model [79].

In 2007, Rykunov and Fiser proposed a reference state which atoms assumed to be random [80]. Consequently, a good model to approximate such model would be a system with randomized particles.

Zhang and Zhang, proposed a random walk ideal chain as the reference state [78]. This reference state was derived from a linear freely jointed chain model, which can be considered as the segment of an ideal polymer chain performing a random walk in three dimensional spaces.

Some other definitions, such as the use of decoys were also suggested [66-68].

Recently, Mirzaie et al introduced a knowledge based distance dependent force

extracted from knowledge based potentials [50].

Interaction center in statistical potentials

Various representations for interaction centers have been introduced. Two major representations are residue or atom level. In residue level, CA, CB or side chain centroids are used and then extended alphabet based on occurring amino acids extracted [27, 42-43, 45, 81-82]. Side chain to backbone and side chain to side chain residue potential is also introduced [54].

In successful energy function two interaction centers per residue namely CA atoms and the side chain center of mass (CA atom in Glycine) were considered [83]. CB atoms were also used for representation [84].

Another representation is all heavy atoms. In a strict physicochemical point of view, all the atoms with different environments, connectivity and chemical nature would be different. Among all the 20 amino acids the total number of heavy atoms is 167 and the number of nonequivalent heavy atoms is 98[50]. Some models such as DFIRE use 167 residue specific atomic types [47], but to reduce this number and raise observed frequencies, various atom type definitions have been considered [44, 61, 50, 85]. For example, four atom types containing carbon, nitrogen, oxygen and sulfur were considered by Mirzaie, et al [50], a total of 19 different atom types were used by Ferrada and Melo [86], a total of 40 atom types were considered for all heavy atoms in the 20 amino acids by Melo and Feytmans [44]. All pairs of non-hydrogen atoms in each of the 20 amino acids ignoring the N-terminal and C-terminal nitrogen and oxygen atoms containing 158 residue dependent atom types were considered by Shen and Sali [48] and Mingyang, et al [63].

Also, a clustering algorithm was used to group atom types by similarity. In fact, from an initial 167 atom types, 12 different atom types were extracted [32].

Accessible surface potentials

Solvent effects and hydrophobic interactions are known as important characteristic for protein folding [87]. So, calculation of the solvent accessible surface area is very important to estimate solvation energies [88, 89]. Zarei, et al presented a method for prediction of protein surface accessible with information theory [90], and Naderi Manesh, et al presented a method based on residue type and conformational states [91].

The accessible surface of an interaction center is defined as the number of interaction centers within a sphere around it and the radius of the sphere is the distance range of the potential. This type of potentials has been described by Sippl method [22, 25].

Contact potentials

Contact potentials are the simplest description of pair-wise potentials. They are similar to the distance dependent potentials. These types of potentials have two values; below the contact distance threshold, energy is assigned to interaction and above it energy is zero.

In the contact potential, the distance between the centers of interacting pairs are calculated and the observed frequency of contacts between residues converts to free energy using Boltzmann equation. In this way, two problems may be encountered. First, when an atom or center of mass is selected for each residue, calculated potential is independent of orientation of the side chains and when the distance between two atoms of two residues is equal to the distance of two atoms of other residues in other positions, the same potentials are assigned to them although the orientation of two residue side chains may be quite different. Second, the atoms of two residues may not have direct contact with each other and some atoms may be located in an interval close to them [92].

Delaunay tessellation is a geometrical tool that is used in protein structure analysis. In fact, voronoi tessellation partitions the space into convex poly-topes called voronoi polyhedra. For a given protein, the voronoi polyhedra is the region of space around an atom, such that all points of this region are closer to this atom than to any other atoms of the protein. A group of four atoms, whose voronoi polyhedra meet at one vertex, forms another basic topological object called, the Delaunay tessellation simplex. So we can consider two atoms are in contact, if they are two vertices of an edge in a simplex [50].

In 2010, Arab et al [92] developed a new approach to calculate a knowledge-based potential energy using pair-wise residue contact area. They calculate the parts of each pair-wise residue area which are in contact in 2Å by rolling a probe ball of different sizes

around the atoms of a residue to determine the contact area of each pair. This pair-wise contact area is used to determine statistical contact area preference between each residue pairs, when a contact area preference estimates a sum of energetic interactions and structural constraints.

Composite potentials

Some potential combine various terms, including hydrogen bonding, torsion angle, solvation, pair-wise potential in residue level, or all atom level and some terms of physical energy function. For example, ROSETTA scoring function [93,94] combines sequence dependent and sequence independent features of protein. A composite residue level potential was introduced that combined contact and local sequence structure descriptions [95]. In 2007, a hydrophobic potential of mean force, generalized Born function and Amber 99 force field were combined [96]. The distribution of pseudo-bonds, pseudo-angle, pseudo-dihedrals and distances between centers of interactions was converted into potentials of mean force in PC2CA model [83]. The QMEAN scoring function is a successful composite potential which consist of 6 different terms; distance dependent pair-wise potentials, solvation potential, torsion angle potential, secondary structure specific and solvent accessibility[97-98].Another composite potential, in residue level [62] and all atom version [63] combines distance dependent pair-wise potential with orientation preference, hydrogen bonding, packing, three body interactions and salvation terms.

Validation of potentials functions

For assessing potential function, it must be able to distinguish protein native structure from protein-like decoys. In fact the energy of native structure must be lowest among energy of decoy structures. Another challenging test of energy function is discrimination of near native structures that means, in absence of the native structure, the energy of structure with minimum RMSD to native structure must be minimum.

In addition, the correlation coefficient between the energy of a model and RMSD should be close to 1, i.e., proteins with high RMSD have high energy.

The *Z-score* of the native structure in the decoys set is equal to

$$Z - score = \frac{\langle E_{decoys} \rangle - E_{native}}{\sigma_{decoys}}$$

In which E_{native} is the energy calculated for native structure and $\langle E_{decoys} \rangle$ and σ_{decoys} are the average and the standard deviation of energy distributions of decoys proteins, respectively. For a good energy function, *Z-score* should be high.

One of the most popular decoy data sets is available in the Decoys'R'us database under the category 'multiple'(http://dd.compbio.washington.edu). This data set contains the ig_structal_hires, 4state_reduced, fisa_casp3, fisa, vhp_mcmd, semfold, hg_structal, lmds, ig_structal and lattice_ssfit. These decoys are obtained with different methods and are very appropriate for assessment of model.

The 4state-reduced decoy set contains 7 different proteins. For each protein, 632 to 689 native like conformations are present in the dataset. This decoy set was generated using a four-state off lattice model with a conformational relaxation method [99].

The fisa and fisa_casp3 decoy sets with four and six targets (500–1400 models per target), respectively, were obtained using a combination of a Bayesian scoring function and a simulated annealing protocol [100,101].

The ig_structal, hg_structal and ig_structal_hires decoy sets contain immunoglobulins (ig) or globins (hg) created by homology modeling.

The largest lattice_ssfit decoy set, containing 2000 decoys for each of the eight targets, was generated using a tetrahedral lattice model with the all-atom ENCAD energy function [99]. The ranges of RMSD from native for all 8 proteins in the set are larger than 4Å.

The lmds set includes decoys with RMSD's less than 10 Å. lmds decoy set with 215–500 models for each one of 10 primarily short targets, was obtained by a local optimization method and a reduced ENCAD energy function [103].

The semfold set includes a very large number of decoys for each of the 6 proteins. In some cases RMSD from native are in range 3 Å to 5 Å. This decoy set was generated by fragment insertion method. This decoy set is the most challenging, since it has more than 10000 decoys for each of the six targets. The vhp_mcmd decoy set has been generated by molecular dynamics simulations.

The ROSETTA decoy set presented by Baker and coworkers [104,105] contains 20

random models and 100 lowest scoring models from 10,000 decoys, which were generated for 58 small proteins using ROSETTA de novo structure predictions followed by all-atom refinement. This data set is downloadable at http://depts.washington.edu/bakerpg.

The I-TASSER decoy set includes the atomic structure decoys generated for 56 non-homologous small proteins. The backbone structures were first generated by the I-TASSER ab initio modeling by Wu et al. [106], where for each protein target 12,500–32,000 conformations were taken from the trajectories of 3 lowest-temperature replicas of the simulations. A full set of I-TASSER decoys is downloadable at http://zhanglab.ccmb.med.umich.edu/decoys.

DISCUSSION

The results of energy functions depend on the presence or absence of the native structure [80]. Often, all atom potential on the presence of native structure have very good performance, but in the absence of native structure, residue based, all atom and composite potentials perform competitively [80]. For example, the composite potential QMEAN6 [97,98] and residue level potential [84] are the best performing potentials, where the native structure were absent while some potentials have better performance in the presence of native structure.

Recently, some different approaches for discrimination of native structure from decoys have been presented. For example, graph theoretic properties including average degree and clustering coefficient of protein graph have been used to perceive the difference between correctly folded proteins and decoys [107,108]. Decoy discrimination method using wavelet analysis [109] and a simple approach based on network pattern of conserved hydrophobic residues have been also presented [110]. An approach by using machine learning and neural network and evolutionary information [111], and a physical method based on force is also presented [50]. The majority of knowledge based potentials are pair-wise, but multi body potentials were also reported [112-117].

REFERENCES

1.Anfisen CB. Principles that govern folding of protein chains. Science 1973; 181:223-230.

2. Cornell WD, Cieplak P, Bayly CI, Gould I R, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA, A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules, J. Am. Chem. Soc 1995; 117:5179–5197.

3. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple amber force fields and development of improved protein backbone parameters, Proteins 2006; 65: 712–725.

4. Wang JM, Cieplak P, Kollman PA. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules?, J. Comput. Chem 2000; 21:1049–1074.

5. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, et al. A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc 1984; 106:765-784.

6. Weiner SJ, Kollman PA, Nguyen DT, Case DA. An all atom force-field for simulations of proteins and nucleic-acids, J. Comput. Chem 1986; **7**:230–252.

7. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A Program for Macromolecular Energy, Minimization and Dynamics Calculations. J Comp Chem 1983; 4:187-217.

8. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem 1998; 102:3586–3616.

9. Scott WRP, Hunenberger PH, Tironi IG, Mark AE, Billeter SR, Fennen J , et al. The GROMOS biomolecular simulation program package.J. Phys. Chem.1999; 103: 3596–3607.

10. Momany FA, Mcguire RF, Burgess AW, Scheraga HA. Energy parameters in polypeptides. 7. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen-bond interactions, and intrinsic torsional potentials for naturally occurring amino-acids. J. Phys. Chem. 1975; 79: 2361–2381.

11. Nemethy G, Pottle MS, Scheraga HA. Energy parameters in polypeptides. 9. Updating of geometrical parameters, nonbonded interactions, and hydrogen-bond interactions for the naturally-occurring amino-acids, J. Phys. Chem 1983; 87:1883–1887.

12. Sippl MJ, Nemethy G, Scheraga HA. Intermolecular potentials from crystal data. 6. Determination of empirical potentials for O-

H=O=C hydrogen-bonds from packing configurations. J. Phys. Chem 1984; 88: 6231–6233.
13. Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 1996; 118: 11225- 11236.
14. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J. Phys. Chem. B 2001; 105: 6474–6487.
15. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999; 35:133-152.
16. Ponder JW, Case DA. Force fields for protein simulations. Advanced in Protein Chemistry 2003; 66:27- 85.
17. Melo F, Feytmans E. Scoring functions for protein structure prediction. In: Schwede T, Peitsch M, editors. Computational structural biology. World Scientific Publishing Co. Pte. Ltd.: Singapore; 2008. pp. 61–88.
18. Tanaka S, Schraga HA. Statistical mechanical treatment of protein conformation .I. Conformational properties of amino acids in proteins. Macromolecules 1976; 9:142-159.
19. Tanaka S, scheraga H A. Medium and long rang interaction parameters between amino acids for predicting three dimensional structures of proteins. Macromolecules 1976; 9: 945-950.
20. Sippl MJ. Calculation of conformational ensembles from potentials of mean force: An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990; 213:859–883.
21. Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol 1995; 5:229-235.
22. Sippl MJ. Boltzmann's principle, knowledge based mean force and protein folding. An approach to the computational determination of protein structures. Journal of Computer Aided Molecular Design 1993; 7:473-501.
23. Sippl MJ, Weitckus S. Detection of native like models for amino acid sequences of unknown three dimensional structures in a data base of known protein conformations. Proteins 1992; 13:258-271.
24. Melo F, Sanchez R, Sali A. Statistical potentials for fold assessment. Protein Sci 2002; 11: 430–448.
25. Melo F, Feytmans E. Assessing protein structures with nonlocal atomic interaction energy. J Mol Biol 1998; 277: 1141–1152.
26. Zhou Y, Zhou H, Zhang C, Liu S. What is a desirable statistical Energy Function for Proteins andHow Can It Be Obtained? Cell Biochemistry and Biophysics 2006; 46:165-174.
27. Miyazawa S, Jernigan RL. Estimation of Effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromole 1985; 18:534-552.
28. DeBolt SE, Skolnick J. Evaluation of atomic level mean force potentials via inverse folding and inverse refinement of protein structures: atomic burial position and pairwise non-bonded interactions. Protein Eng 1996; 9:637-655.
29. Zhang C, Vasmatzis G, cornette J, Delisi C. determination of atomic desolvation energies from the structures of crystallized proteins. Journal of molecular biology. 1997; 267:707-726.
30. Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair potentials based on week sequence fragment similarity. Proteins 2000; 38:3-16.
31. Zhou H, Zhou Y. Single body knowledge based energy score combined with sequence profile and secondary structure information for fold recognition. Proteins 2004; 55:1005-1013.
32. McConkey BJ, Sobolev V, Edelman M. Discrimination of native protein structures using atom-atom contact scoring. Proc Natl Acad Sci USA. 2003; 100(6):3215.
33. Pokarowski P, Kloczkowski A, Jernigan RL, Kothari NS, Pokarowska M, Kolinski A: Inferring ideal amino acid interaction forms from statistical protein contact potentials. Proteins 2005 ; 59(1)**:**49.
34. Zhang C, Kim SH. Environment-dependent residue contact energies for proteins. Proceedings of the National Academy of Sciences 2000; 97(6)**:**2550.
35. Solis AD, Rackovsky S. Information and discrimination in pairwise contact potentials. Proteins. 2008; 71:1071–1087.
36. Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, et al. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials from potentials of mean force. J mol Biol 1990; 216:167-180.

37. Tobi D, Elber R. Distance dependent, pair potential for protein folding: Results from linear optimization. Proteins 2000; 41:40-56.
38. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992; 358:86-89.
39. Samudrala R, Moult J. An all atom distance dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 1998; 275:895-916.
40. Lu H, Skolnick J. A distance dependent atomic knowledge based potential for improved protein structure selection. Proteins 2001; 44:223-232.
41. Rojnuckarin A, Subramaniam S. Knowledge based interaction potentials for proteins. Proteins 1999; 36:54-67.
42. Bahar I, Jernigan RL. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. JMolBiol. 1997; 266(1):195.
43. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol. 1996; 256(3):623-644.
44. Melo F, Feytmans E. Novel knowledge-based mean force potential at atomic level. JMolBiol. 1997; 267(1):207.
45. Miyazawa S, Jernigan RL. An empirical energy potential with a reference state for protein fold and sequence recognition. Proteins. 1999; 36(3):357.
46. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure derived potentials of mean force for structure selection and stability prediction. Protein Science 2002; 11: 2714-2726.
47. Chi Zhang, Song Liu, Hongyi Zhou, Yaoqi Zhou. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. Protein Science 2004; 13(2):400-411.
48. Shen M, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci 2006; 15:2407-2524.
49. Ferrada E, Vergara IA, Melo F. A knowledge-based potential with an accurate description of local interactions improves discrimination between native and near-native protein conformations. Cell Biochem Biophys. 2007; 49:111–124.
50. Mirzaie M, Eslahchi Ch, Pezeshk H, Sadeghi M. a distance dependent atomic knowledge based potential and force for discrimination of native structures from decoys. Proteins 2009; 77(2):454-463.
51. Rooman MJ, Kocher JPA, Wodak SJ. Prediction of protein backbone conformation based on seven structure assignments. Influence of local interactions. Journal of molecular biology 1991; 211:961- 979.
52. Rooman MJ, Kocher JPA, Wodak SJ. Extracting information on folding from the amino acid sequence: accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. Biochemistry 1992; 31:10226-10238.
53. Kocher JPA, Rooman MJ, Wodak SJ. Factors influencing the ability of knowledge based potentials to identify native sequence structure matches. Journal of Molecular Biology 1994; 235:1598-1613.
54. Fang Q, Shortle D. A consistent set of statistical potentials for quantifying local side-chain and backbone interactions. Proteins. 2005; 60(1):90.
55. Pohl FM. Empirical protein energy maps. Nat New Biol. 1971; 234(52):277.
56. Bagci Z, Kloczkowski A, Jernigan RL, Bahar I. The origin and extent of coarse-grained regularities in protein internal packing. Proteins: Structure, Function, and Bioinformatics. 2003; 53(1):56–67.
57. Buchete NV, Straub JE, Thirumalai D. Orientational potentials extracted from protein structures improves native fold recognition. Protein Science. 2004; 13(4):862.
58. Buchete NV, Straub JE, Thirumalai D. Continuous anisotropic representation of coarse-grained potentials for proteins by spherical harmonics synthesis. J Mol Graph Model. 2004; 22(5):441.
59. Mukherjee A, Bhimalapuram P, Bagchi B. Orientation-dependent potential of mean force for protein folding. Journal of Chemical Physics. 2005; 123
60. Misura KM, Morozov AV, Baker D. Analysis of anisotropic side-chain packing in proteins and application to high-resolution structure prediction. J Mol Biol. 2004; 342:651–64.
61. Miyazawa S, Jernigan RL. How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins? J Chem Phys. 2005; 122:024901
62. Wu Y, Lu M, Chen M, Li J, Ma J. OPUS-Ca: a knowledge based potential function requiring only Calpha positions. Protein Sci. 2007; 16(7):1449–1463.

63. Mingyang L, Athanasios DD, Jianpeng M. OPUS-PSP. An orientation dependent statistical all atom potential derived from side chain packing. J Mol Biol.2008; 376:288-301.
64. Liang S, Zhou Y, Grishin N, Standley DM. Protein side chain modeling with orientation dependent atomic force fields derived by series expansions. J. Comput. Chem 2011.in press.
65. Levitt M, Warshel A: Computer simulation of protein folding. Nature 1975; 253(5494):694.
66. Rajgaria R, McAllister SR, Floudas CA: A novel high resolution Calpha--Calpha distance dependent force field based on a high quality decoy set. Proteins 2006; 65(3)**:**726-741.
67. Rajgaria R, McAllister SR, Floudas CA: Distance dependent centroid to centroid force fields using high resolution decoys**.** Proteins 2008; 70(3)**:**950-970.
68. Qiu J, Elber R: Atomically detailed potentials to recognize native and approximate protein structures. Proteins: Structure, Function, and Bioinformatics 2005; 61(1)**:**44-55.
69. Jaynes ET. Information theory and statistical mechanics. Phys Rev 1957; 106:620-630.
70. Cline MS, Karplus K, Lathrop RH, Smith TF, Rogers RG, Jr, Haussler D. Information-theoretic dissection of pairwise contact potentials. Proteins. 2002; 49:7–14.
71. Solis AD, Rackovsky S. Improvement of statistical potentials and threading score functions using information maximization. Proteins 2006; 62(4)**:**892.
72. Solis AD, Rackovsky S. Information and discrimination in pairwise contact potentials. Proteins. 2008; 71:1071–1087.
73. Furuichi E, Koehl P. Influence of protein structure databases on the predictive power of statistical pair potentials. Proteins. 1998; 31:139–149.
74. Zhang C, Liu S, Zhou H, Zhou Y. The dependence of all-atom statistical potentials on structural training database. Biophys J. 2004; 86:3349–3358.
75. Ying Xu, Dong Xu, Liang J. Computational Methods for Protein Structure Prediction and Modeling: Volume 1: Structure Prediction 2010. Springer; 1st Edition.
76. Thomas P. D, Dill K. A. Statistical potentials extracted from protein structures .how accurate are they? J. Mol. Bio 1996; 257:457-469.
77. Habibi M, eslahchi Ch, sadeghi M, Pezashk H. The interpretation of protein structures based on graph theory and contact map, Open Access Bioinformatics 2010; 2: 127–137.
78. Zhang J, Zhang Y. A novel side chain orientation dependent potential derived from random walk reference state for protein fold selection and structure prediction. PloS ONE 2010; 5(10):.
79. Cheng J, Pei JF, Lai LH. A free rotating and self avoiding chain model for deriving statistical potentials based on protein structures. Biophysical Journal 2007; 92:3868-3877.
80. Rykunov D, Fiser A. Effects of amino acid composition, finite size of proteins, and sparse statistics on distance-dependent statistical pair potentials. Proteins: Structure, Function, and Bioinformatics. 2007; 67(3):559–568.
81. Reva BA, Finkelstein AV, Sanner MF, Olson AJ. Residue-residue mean-force potentials for protein structure recognition. Protein Eng 1997; 10(8)**:**865.
82. Fitzgerald JE, Jha AK, Colubri A, Sosnick TR, Freed KF. Reduced Cbeta statistical potentials can outperform all-atom potentials in decoy identification. Protein Sci 2007; 16(10)**:**2123-2139.
83. Fogolari F, Pieri L, Dovier A, Bortolussi L, Giugliarelli G, Corazza A, et al. Scoring predictive models using a reduced representation of proteins: model and energy definition. BMC Struct Biol. 2007; 7:15.
84. Rykunov D, Fiser A. New statistical potential for quality assessment of protein models and a survey of energy functions. BMC bioinformatics 2010; 11:128.
85. Summa CM, Levitt M, Degrado WF. An atomic environment potential for use in protein structure prediction. JMolBiol. 2005; 352(4):986.
86. Ferrada E, Melo F. Effective knowledge-based potentials. Protein Science 2009; 18: 1469-1485.
87. Baldwin L. Making a network of hydrophobic clusters. Science 2002; 295(5560):1657–8.
88. Eisenberg D, McLachlan A D. Solvation energy in protein folding and binding. Nature 1986; 319(6050):199–203.
89. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Bio 1971; 55(3);379-400.
90. Zarei R, Arab Sh, Sadeghi M. A Method for Protein Accessibility Prediction based on Residue Types and Conformational States,

Computational Biology and Chemistry 2007; 31: 384–388
91. Naderi-Manesh H, Sadeghi M, Arab Sh, Moosavi movahedi AA. Prediction of protein surface accessibility with information theory. PROTEINS 2001; 42: 452-459.
92. Arab S, Sadeghi M, Eslahchi C, Pezeshk H. A .A pairwise residue contact area-based mean force potential for discrimination of native protein structure. BMC Bioinformatics 2010; 11:16.
93. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D.Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999; 34(1)**:**82.
94. Rohl CA, Strauss CE, Misura KM, Baker D. Protein structure prediction using Rosetta. Methods Enzymol 2004; 383:66-93.
95. Zhang J, Chen R, Liang J. Empirical potential function for simplified protein models: Combining contact and local sequence-structure descriptors. Proteins: Structure, Function, and Bioinformatics. 2006; 63(4):949–960.
96. Matthew S L, Nicolas L F, Teresa H G. Hydrophobic Potential of Mean Force as a Solvation Function for Protein Structure Prediction. Structure 2007; 15(6): 727-740.
97. Benkert P, Tosatto SC, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins 2008; 71(1):261-277.
98. Benkert P, Kunzli M, Schwede T. QMEAN server for protein model quality estimation. Nucleic Acids Research 2009; 37(2):510-514.
99. Park B, Levitt M. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol 1996; 258:367–392.
100. Simons KT, Kooperberg C, Huang E, 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.
101. Simons KT, Ruczinski I, Kooperberg C, Fox B A, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence- independent features of proteins. Proteins 1999; 34: 82–95.
102. Xia Y, Huang ES, Levitt M, Samudrala R. Ab initio construction of protein tertiary structures using a hierarchical approach. J Mol Biol 2000; 300: 171–185.
103. Keasar C, Levitt M. A novel approach to decoy set generation: Designing a physical energy function having local minima with native structure characteristics. J Mol Biol 2003; 329: 159–174.
104. Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read AJ, Baker D. High-resolution structure prediction and the crystallographic phase problem. Nature 2007; 450: 259–U257.
105. Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, Khare S, Tyka MD, Bhat D, Chivian D, Kim DE, Sheffler WH, Malmström L, Wollacott AM, Wang C, Andre I, Baker D. Structure prediction for CABP7 targets using extensive all-atom refinement with Rosetta@home. Proteins-Structure Function and Bioinformatics 2007; 69: 118–128.
106. Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol. 2007; 5: 17.
107. Taylor T, Vaisman II: Graph theoretic properties of networks formed by the Delaunay tessellation of protein structures.2006; Phys Rev E Stat Nonlin Soft Matter Phys: 73(4):041925.
108. Küçükural A, Sezerman U, Erçil A . Discrimination of native folds using networks properties of protein structures. APBC 2008, Kyoto JAPAN
109. Minxin Chen, Bingchuan Liu,Wenying Yan,Bairong Shen. Wavelet Transform Based Protein Decoy Discrimination. IEEE Proc. ICBBE 2009.
110. Muppirala UK, Li Z. A Simple Approach for Protein Structure Discrimination Based on the Network Pattern of Conserved Hydrophobic Residues. Protein Eng., Design, and Selection 2006; 19(6): 265-275.
111. Tan CW, Jones DT. Using neural networks and evolutionary information in decoy discrimination for protein tertiary structure prediction. BMC Bioinformatics 2008; 9: 94.
112. Singh RK, Tropsha A, Vaisman II. Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. J Comput Biol. 1996; 3(2):213-21.
113. Zheng W, Cho SJ, Vaisman II, Tropsha A. A new approach to protein folds recognition based on Delaunay tessellation of protein structure. Pac Symp Biocomput. 1997:486-97.

114. Carter CW Jr, LeFebvre BC, Cammer SA, Tropsha A, Edgell MH. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. J Mol Biol 2001; 311(4):625-38.
115. Krishnamoorthy B, Tropsha A: Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. Bioinformatics 2003; 19(12)**:**1540-1548.
116. Ngan SC, Inouye MT, Samudrala R. A knowledge-based scoring function based on residue triplets for protein structure prediction. Protein Engineering Design and Selection 2006; 19(5)**:**187.
117. Masso M, Vaisman II: Accurate prediction of enzyme mutant activity based on a multibodystatistical potential. Bioinformatics 2007; 23(23)**:**3155-3161.