Published in final edited form as:
*Biochem Biophys Res Commun*. 2010 June 4; 396(3): 736–741. doi:10.1016/j.bbrc.2010.05.005.

# Distance-dependent statistical potentials for discriminating thermophilic and mesophilic proteins

**Yunqi Li** and **Jianwen Fang**[*]
Applied Bioinformatics Laboratory, the University of Kansas, Lawrence, KS 66047, USA

## Abstract

Identification of the characteristic structural patterns responsible for protein thermostability is theoretically important and practically useful but largely remains an open problem. These patterns may be revealed through comparative study on thermophilic and mesophilic proteins that have distinct thermostability. In this study we constructed several distance-dependant potentials from thermophilic and mesophilic proteins. These potentials were then used to evaluate the structural difference between thermophilic and mesophilic proteins. We found that using the subtraction or division of the potentials derived from thermophilic and mesophilic proteins can dramatically increase the discriminatory ability. This approach revealed that the ability to distinct the subtle structural features responsible for protein thermostability may be effectively enhanced through rationally designed comparative study.

**Keywords**

Thermostability; mesophilic proteins; thermophilic proteins; statistical potential

## INTRODUCTION

Thermophiles are organisms live under elevated temperatures as high as 113 °C [1]. Naturally the proteins (TPs) produced by thermophiles are intrinsically more thermostable than their mesophilic counterparts (MPs). Therefore studying the difference between thermophilic and mesophilic proteins may provide key knowledge for designing thermostable proteins, which is not only practically useful but also theoretically important [1,2,3,4].

Numerous studies have been performed to compare thermophilic and mesophilic proteins and resulted in a variety of overall patterns associated with protein thermostability [5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22]. Among them, the simplest factor may be the length of protein chain [13]. The protein chain rigidity/flexibility was also reported to be strongly correlated with protein thermostability [12,16,19]. The number and magnitude of hydrogen bonds, salt bridges and hydrophobic interactions significant affects the protein thermostability [11,15,17,18]. Glyakina et al. reported that the packing of external residues is strongly correlated to the thermostability of thermophilic and mesophilic proteins [22]. It was also discovered that the amino acid compositions of thermophilic and mesophilic

[*]Correspondence to: Jianwen Fang, Applied Bioinformatics Laboratory, University of Kansas, 2034 Becker Dr., Lawrence, KS 66047, USA. jwfang@ku.edu; Phone: +1-785-864-3349.

proteins are significantly different [14,20,21]. In summary, these comparative studies have greatly enriched our understanding of protein thermostability.

In this study, we took a different approach from previous studies to understanding the difference of thermophilic and mesophilic proteins by developing statistical potentials separately for each group of proteins. Numerous statistical potentials have been constructed but, to the best of our knowledge, have not considered thermophilic and mesophilic proteins separately. These potentials are based on pair-wised interactions in protein structures and can be at the residue level [23,24] or the atomic level [10,25,26], distance dependent [27] or distance independent [28]. These potentials have been successfully applied to a broad range of tasks in protein science, such as the selection of native-like decoys from nonnative conformations [8,27] and the prediction of mutation induced stability changes [8,9,10]. The rationale of the present study is that constructing potentials separately for thermophilic and mesophilic proteins may allow us to identify subtle difference between thermophilic and mesophilic proteins. Such potentials may be used in a number of applications such as discriminating thermophilic and mesophilic proteins, identifiying hotspots in proteins [29]. Therefore, we carried out this study by formulating several distance dependent, pair-wised statistical potentials.

In this paper, we will firstly describe the construction of training and testing datasets, and several distance-dependent, pair-wised statistical potentials constructed using either statistical or theoretical approaches. The distribution of residue pairs will then be analyzed. We will provide the results of comparing these potentials with several classical potentials for discriminating thermophilic and mesophilic proteins.

## MATERIALS AND METHODS

### Datasets

We used the following steps to create a collection of non-redundant thermophilic and mesophilic proteins structures:

1. A list of organisms with known optimal growth temperature (OGT) was collected from PGTdb [30], the UCSC archaea gnome database (http://archaea.ucsc.edu/), and several recent published literatures [31,32,33,34,35,36,37]. We considered organisms with OGT of 50 °C or higher as thermophiles. Therefore, this definition includes both hyperthermophiles and thermophiles as commonly defined in literature [34]. All remaining organisms, except those known as other types of extremophiles such as psychrophile, halophile, acidophile and alkaliphile, were considered as mesophiles.
2. All protein structures in Protein Data Bank (PDB, http://www.pdb.org) were downloaded on Jan 5th, 2009 and sorted into thermophilic and mesophilic proteins according to their source organisms defined in the previous step. The PDB entries without known source organisms (i.e. synthetic proteins or protein models) were not used in this study. The PDB entries with chains from both thermophile and mesophile were also excluded. The protein structures were further filtered by R-factor (≤ 0.25) and resolution (≤ 2.0 Ǻ) using PISCES [38].
3. Membrane proteins, according to the SCOP classification [39], were removed. We also excluded chains with less than 50 residues or more than 800 residues.
4. To reduce the redundancy in the dataset, we clustered all remaining protein sequences using Blastclust [40] and only kept the longest chain in each cluster. The sequence identity threshold was set to 30% and minimum length coverage was 0.9.

The final dataset has 1020 thermophilic protein chains from 90 organisms extracted from 996 PDB entries and 4977 mesophilic protein chains from 790 organisms extracted from 4742 entries. This dataset named as the training dataset was then used for the generation of statistical potentials.

For evaluating the correlation between the statistical potentials and optimal growth temperatures (OGT) of a set of proteins, we grouped all proteins in the training dataset according to the OGT of their original organisms, in 5 °C bins from 0 °C to 105 °C. The proteins in each bin were then clustered using Blastclust (sequence identity threshold = 30%) and the longest protein in each cluster was selected. Finally, we obtained 1538 proteins from 121 organisms distributed on 64 distinct OGTs.

In order to build a representative blind testing dataset, we retrieved the newly deposited protein structures between Jan. 2009 and Jan. 2010. We filtered these proteins using the same 1), 2) and 3) steps used in collecting of the training dataset and obtained 331 TPs and 3770 MPs. We then clustered these newly collected proteins using Blastclust with sequence identity threshold of 90% and excluded those proteins sharing higher than 30% sequence identity to any protein in the training dataset. The final blind testing dataset contains 66 TPs and 672 MPs.

## Reference states

All distance-dependant statistical potentials based on atomic pairs can be generalized as:

$$u(i,\ j,\ r) = -k_B T \ln \frac{N_{\text{obs}}(i,\ j,\ r)}{N_{\exp}(i,\ j,\ r)} \qquad (1)$$

where $(i, j)$ are all possible atomic pairs in proteins and $r$ is the distance between atoms $i$ and $j$. $N_{\text{obs}}(i,j,r)$ is the number of observed atomic pairs of $i$ and $j$ in a distance shell of $r \pm \Delta r$. $N_{\exp}(i,j,r)$ is the expected number of atomic pairs of $i$ and $j$ in the same distance shell in a reference state. The choice of the reference state determines the performance of a statistical potential [8].

We proposed three reference states, including two statistical and one theoretical approaches, to construct the potentials. The two reference states based on statistical approaches are

$$N_{\exp}(i,\ j,\ r) = \sum_{i,j} N_{\text{obs}}(i,\ j,\ r) \qquad (2)$$

$$N_{\exp}(i,\ j,\ r) = \sum_{i,j,r} N_{\text{obs}}(i,\ j,\ r) \qquad (3)$$

Physically, the statistical potential based on the reference state defined in Eq. 2 represents the propensity of an atomic pair $i$ and $j$ located in a spatial shell with mean separation of $r$. The potential based on Eq. 3 is proportional to the probability to find a pair within a sphere with radius of $r$.

The third reference state was originally proposed in DFIRE potentials developed by Zhou and Zhou [8], which was defined as:

$$N_{\exp}(i,\ j,\ r)=(r/r_{cut})^{1.61}(\Delta r/\Delta r_{cut})N_{obs}(i,\ j,\ r_{cut}) \tag{4}$$

where the $r_{\mathrm{cut}}$ equals to 14.5 Å, and the width of distance shell Δ*r* is dependent on distance, which will be described later.

Based on the definition of Eq. 1, the potential for a given protein can be calculated through

$$E=\frac{1}{2L}\sum_{i,j,r}u(i,\ j,\ r) \tag{5}$$

where *L* is the number of residues in the protein and the summation over all atomic pairs appeared in the structure except those pairs from the same residue. To simplify the description, we label the potential utilizing different reference states defined in Eq. 2, Eq. 3 and Eq. 4 as E1, E2 and DFIRE* respectively. DFIRE* was labeled as such because it uses a training dataset different from the original DFIRE.

## The distance shell and other settings

We used two different types of distance shells. The first (Bin1) was adopted from DFIRE, where the width of distance shell, Δ*r*, was set to 2 Å for $r < 2$ Å, 0.5 Å for 2 Å $< r <$ 8 Å, and 1 Å for 8 Å $< r <$ 15 Å. For the second setting (Bin2) of distance shell, Δ*r* was set to 1 Å for $r <$ 11 Å and 2 Å for 11 Å $< r <$ 21 Å.

Another important issue of developing statistical potentials is how to model amino acid residues. In this study, we used two representations: the first counts backbone heavy atoms and $C_\beta$ atoms while the other only counts Cα atoms. The atoms pairs from same residue were excluded in the calculation of potentials for the first representation. Both representations were used when the Bin1 distance shell setting was used to calculate the potentials. Only Cα atoms were considered when the distance shell setting of Bin2 was used.

Each protein would have two potentials, one based on the TPs and the other based on MPs. We then studied whether the potential difference or the ratio of potentials at the protein level could provide better performance.

## Other potentials

We used a number of well-known potentials for comparison. The DFIRE potential [8], a statistical potential derived from representative native structures, has been used in native-like protein structures modeling and selection. It also showed good performance in discriminating proteins from the organisms surviving under different thermal environments [41,42,43,44,45]. Local structure entropy (LSE) [46], derived from representative protein domains, has shown strong correlation with protein thermostability. It was reported that the number of residues in each protein chain has strong correlation with the thermodynamic parameters representative for protein thermostability [13]. We also evaluated the discriminative of two classical energy terms. One is CHARMM22 [47], a molecular mechanical force field based on quantum mechanics for protein structure modeling, and the other is DSSP hydrogen bonding (DSSP_HB) potential [48] which is tightly correlated with secondary structures.

### Discriminatory performance evaluation

To evaluate the performance of these potentials, we created the receiver operating characteristic (ROC) curve for the classification of TP and MP based on the potential values. ROC is a plot of the true-positive rate (sensitivity) against the false-positive rate (1 - specificity). The area under an ROC curve represents the trade-off between sensitivity and specificity. In general, an area of 1 represents a perfect prediction model, and an area over 0.9 is considered excellent. An area between 0.8 and 0.9 is considered good, whereas the range of 0.7 to 0.8 is fair. We also report the overall accuracy of the classification. Accuracy is defined as

$$\mathrm{ACC}=\frac{\mathrm{tp}+\mathrm{tn}}{\mathrm{tp}+\mathrm{tn}+\mathrm{fp}+\mathrm{fn}} \tag{7}$$

where tp stands for true positive, tn for true negative, fp for false positive, and fn for false negative. A true case represents the class of a protein has been correctly classified. A positive case represents the class of thermophilic proteins. We calculated the accuracy at 80% specificity.

## RESULTS AND DISCUSSION

### Overall distance dependent distribution of residue pairs

The distributions of residue pairs against different spatial distance bins were plotted in Figure S1 in supplementary material. Each curve represents one of the 210 types of residue pairs. Generally, each curve can be roughly divided into four regions separated by the valleys. Region I (≤ 4.1 Å) may reveal neighboring residue information, i.e. dipeptide patterns. Region II (from 4.1 Å to 7.2Å) is largely contributed to residue pairs constrained by backbone hydrogen bond [49], salt bridge [18,50,51] and disulfide bond [52]. Region III (from 7.2 Å to 11 Å) largely contains residue pairs which neighboring residues hydrogen bonded or disulfide bonded, or their side chains are frozen by hydrogen bonds or salt bridges [18,50,53]. Region IV (> 11 Å) corresponds to the long range correlation of residues pairs.

The residue pairs distributions of TPs and MPs have very similar patterns. This finding suggests that the difference between TPs and MPs, in terms of potentials, may be quite subtle.

### Evaluating discriminating ability

We calculated both TP and MP potentials as well as the difference and the ratio of these potentials of each protein. Based these values, areas under ROC (AUC) and accuracies at 80% specificity were then calculated (Table 1). The results clearly show that using potential difference or the potential ratio greatly improved the discriminating ability of the potentials (Figure 1). We found that the potentials from either single set of proteins are not capable of discriminating TPs and MPs (AUC close to 0.5). The discriminating ability of potentials with backbone and $C_\beta$ atoms is often better than that of Cα based potentials. This is consistent with previous reports [8,25,26]. The discriminating ability of E1 and E2 is very similar, consistent with the observation in Figure S1, where almost all residue pairs have very similar distance dependent distribution patterns. Therefore the spatial distance-dependent preference is highly correlated to the probability of finding a residue pair in a sphere. The different sizes of the distance shells did not significantly affect the discriminative ability. The variances of potential values between thermophilic and mesophilic proteins are less than 3% (see Table S1 in supplementary material), indicating the structural difference between TP and MP is quite small.

To estimate the robustness of the potentials, we performed the same analysis on the testing dataset (Table 2). Clearly the overall performance was similar to that from the training dataset. However, it seems that the E1 and E2 are more robust than DFIRE* in discriminating TPs and MPs unseen in the training.

### Why the difference and ratio of TP and MP potentials have dramatically enhanced discriminating ability?

We illustrated our explanation why the difference and ratio of TP and MP potentials have dramatically enhanced discriminating ability in Figure 2. Since the differences between TPs and MPs are quite subtle, the spaces of TP and MP proper features are significantly overlapped. It is reasonable to believe that the thermostability of a given protein is correlated to the difference or ratio of its occupied spaces unique to TPs and MPs. Therefore protein A in Figure 2 has higher thermostability than protein B. The discriminating ability diminishes if we use the potentials developed using a single dataset.

### Comparing with other potentials/properties

We evaluated several other potentials and protein properties for their discriminating ability using ROC curves as shown Figure 3. The AUC for protein length, DFIRE potential, CHARMM22 potential, DSSP_HB energy, FOLDX potential and LSE potential are 0.521, 0.612, 0.528, 0.614, 0.676 and 0.660 respectively. The chain length has little discriminatory ability, although it was reported that thermophilic protein normally has shorter chain than that of mesophilic proteins [37], which was proposed for predicting protein stability [13]. The DFIRE* based on either MP or TP showed very similar discriminating ability as that of DFIRE, indicating using different training dataset did not cause significant difference. CHARMM22 force field, widely used to refine protein structures closer to native conformations, didn’t show significant ability for discriminating MPs and TPs. FOLDX, developed for predicting mutation induced stability changes, and LSE, a statistical potential to correlate protein stability, only showed moderate discriminating abilities.

### Correlation of the potentials and OGTs

We also studied the correlation between these potentials and the OGTs of organisms produced the proteins (Table 3). A significant improvement of the correlation coefficient by using the difference or the ratio of potentials was observed. Figure 4 displays the plot of DFIRE* against OGTs. The highest correlation coefficient reached 0.782, suggested that the potentials have strong correlation with OGTs and thermostability since there is direct correlation between them.

## CONCLUSION

In this paper we have reported several distance-dependent statistical potentials for discriminating thermophilic and mesophilic proteins. The overall distribution of residue pairs used to derive these potentials was analyzed and typical regions associated with protein secondary structures were observed. We found that the size of distance shells has limited effects on final discriminatory performance. More importantly, the discriminating ability of each potential enhanced dramatically when the difference or the ratio of the potentials based on thermophilic and mesophilic proteins was used. We expect that to construct knowledge-based potentials based on pre-grouped large amount of proteins may help further investigations to find characteristic patterns responsible for protein thermostability.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We thank Dr. Robert Hanzlik of the University of Kansas for his valuable scientific input. This work was partially supported by National Institutes of Health grant P01 AG12993 (PI: E. Machaelis).

## References

1. Sterner R, Liebl W. Thermophilic adaptation of proteins. Critical Reviews in Biochemistry and Molecular Biology. 2001; 36:39–106. [PubMed: 11256505]
2. Dahiyat BI. In silico design for protein stabilization. Current Opinion in Biotechnology. 1999; 10:387–390. [PubMed: 10449321]
3. Korkegian A, Black ME, Baker D, Stoddard BL. Computational thermostabilization of an enzyme. Science. 2005; 308:857–860. [PubMed: 15879217]
4. Lazar GA, Marshall SA, Plecs JJ, Mayo SL, Desjarlais JR. Designing proteins for therapeutic applications. Curr Opin Struct Biol. 2003; 13:513–518. [PubMed: 12948782]
5. Dill KA. Dominant forces in protein folding. Biochemistry. 1990; 29:7133–7155. [PubMed: 2207096]
6. Reese MG, Lund O, Bohr J, Bohr H, Hansen JE, Brunak S. Distance distributions in proteins: a six-parameter representation. Protein Eng. 1996; 9:733–740. [PubMed: 8888138]
7. Sippl MJ. Who solved the protein folding problem? Structure. 1999; 7:R81–83. [PubMed: 10196132]
8. 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 Sci. 2002; 11:2714–2726. [PubMed: 12381853]
9. Hoppe C, Schomburg D. Prediction of protein thermostability with a direction- and distance-dependent knowledge-based potential. Protein Sci. 2005; 14:2682–2692. [PubMed: 16155198]
10. Parthiban V, Gromiha MM, Hoppe C, Schomburg D. Structural analysis and prediction of protein mutant stability using distance and torsion potentials: role of secondary structure and solvent accessibility. Proteins. 2007; 66:41–52. [PubMed: 17068801]
11. Clark AT, Smith K, Muhandiram R, Edmondson SP, Shriver JW. Carboxyl pK(a) values, ion pairs, hydrogen bonding, and the pH-dependence of folding the hyperthermophile proteins Sac7d and Sso7d. J Mol Biol. 2007; 372:992–1008. [PubMed: 17692336]
12. Eijsink VG, Gaseidnes S, Borchert TV, van den Burg B. Directed evolution of enzyme stability. Biomol Eng. 2005; 22:21–30. [PubMed: 15857780]
13. Ghosh K, Dill KA. Computing protein stabilities from their chain lengths. Proceedings of the National Academy of Sciences. 2009; 106:10649–10654.
14. Li Y, Middaugh CR, Fang J. A novel scoring function for discriminating hyperthermophilic and mesophilic proteins with application to predicting relative thermostability of protein mutants. BMC Bioinformatics. 2010; 11:62. [PubMed: 20109199]
15. Max KE, Wunderlich M, Roske Y, Schmid FX, Heinemann U. Optimized variants of the cold shock protein from in vitro selection: structural basis of their high thermostability. J Mol Biol. 2007; 369:1087–1097. [PubMed: 17481655]
16. Radestock S, Gohlke H. Exploiting the Link between Protein Rigidity and Thermostability for Data-Driven Protein Engineering. Engineering in Life Sciences. 2008; 8:507–522.
17. Robinson-Rechavi M, Alibes A, Godzik A. Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: lessons from structural genomics of Thermotoga maritima. J Mol Biol. 2006; 356:547–557. [PubMed: 16375925]
18. Sagarik K, Chaiyapongs S. Structures and stability of salt-bridge in aqueous solution. Biophys Chem. 2005; 117:119–140. [PubMed: 15935545]

19. Scandurra R, Consalvi V, Chiaraluce R, Politi L, Engel PC. Protein thermostability in extremophiles. Biochimie. 1998; 80:933–941. [PubMed: 9893953]

20. Wu LC, Lee JX, Huang HD, Liu BJ, Horng JT. An expert system to predict protein thermostability using decision tree. Expert Systems with Applications. 2009; 36:9007–9014.

21. Zeldovich KB, Berezovsky IN, Shakhnovich EI. Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol. 2007; 3:e5. [PubMed: 17222055]

22. Glyakina AV, Garbuzynskiy SO, Lobanov MY, Galzitskaya OV. Different packing of external residues can explain differences in the thermostability of proteins from thermophilic and mesophilic organisms. Bioinformatics. 2007; 23:2231–2238. [PubMed: 17599925]

23. 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. [PubMed: 2359125]

24. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature. 1992; 358:86–89. [PubMed: 1614539]

25. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol. 1998; 275:895–916. [PubMed: 9480776]

26. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins. 2001; 44:223–232. [PubMed: 11455595]

27. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci. 2006; 15:2507–2524. [PubMed: 17075131]

28. Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. Proteins. 2000; 38:3–16. [PubMed: 10651034]

29. Pavelka A, Chovancova E, Damborsky J. HotSpot Wizard: a web server for identification of hot spots in protein engineering. Nucleic Acids Res. 2009; 37:W376–383. [PubMed: 19465397]

30. Huang SL, Wu LC, Liang HK, Pan KT, Horng JT, Ko MT. PGTdb: a database providing growth temperatures of prokaryotes. Bioinformatics. 2004; 20:276–278. [PubMed: 14734322]

31. Zeldovich KB, Berezovsky IN, Shakhnovich EI. Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol. 2007; 3:62–72.

32. Puigbo P, Pasamontes A, Garcia-Vallve S. Gaining and losing the thermophilic adaptation in prokaryotes. Trends Genet. 2008; 24:10–14. [PubMed: 18054113]

33. Heinzelman P, Snow CD, Wu I, Nguyen C, Villalobos A, Govindarajan S, Minshull J, Arnold FH. A family of thermostable fungal cellulases created by structure-guided recombination. Proc Natl Acad Sci U S A. 2009; 106:5610–5615. [PubMed: 19307582]

34. Trivedi S, Gehlot HS, Rao SR. Protein thermostability in Archaea and Eubacteria. Genet Mol Res. 2006; 5:816–827. [PubMed: 17183489]

35. Stetter KO. History of discovery of the first hyperthermophiles. Extremophiles. 2006; 10:357–362. [PubMed: 16941067]

36. Laksanalamai P, Robb FT. Small heat shock proteins from extremophiles: a review. Extremophiles. 2004; 8:1–11. [PubMed: 15064984]

37. Sterner R, Liebl W. Thermophilic adaptation of proteins. Crit Rev Biochem Mol Biol. 2001; 36:39–106. [PubMed: 11256505]

38. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. Bioinformatics. 2003; 19:1589–1591. [PubMed: 12912846]

39. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995; 247:536–540. [PubMed: 7723011]

40. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215:403–410. [PubMed: 2231712]

41. Zhang C, Liu S, Zhou Y. Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. Protein Sci. 2004; 13:391–399. [PubMed: 14739324]

42. Bueno M, Camacho CJ, Sancho J. SIMPLE estimate of the free energy change due to aliphatic mutations: superior predictions based on first principles. Proteins. 2007; 68:850–862. [PubMed: 17523191]

43. Tan YH, Luo R. Protein stability prediction: a Poisson-Boltzmann approach. J Phys Chem B. 2008; 112:1875–1883. [PubMed: 18211063]

44. Lonquety M, Lacroix Z, Papandreou N, Chomilier J. SPROUTS: a database for the evaluation of protein stability upon point mutation. Nucleic Acids Res. 2009; 37:D374–379. [PubMed: 18945702]

45. Kang S, Chen G, Xiao G. Robust prediction of mutation-induced protein stability change by property encoding of amino acids. Protein Eng Des Sel. 2009; 22:75–83. [PubMed: 19054789]

46. Chan CH, Liang HK, Hsiao NW, Ko MT, Lyu PC, Hwang JK. Relationship between local structural entropy and protein thermostability. Proteins. 2004; 57:684–691. [PubMed: 15532068]

47. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. J Phys Chem B. 1998; 102:3586–3616.

48. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983; 22:2577–2637. [PubMed: 6667333]

49. Li Y, Zhang Y. REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. Proteins. 2009; 76:665–676. [PubMed: 19274737]

50. Kumar S, Nussinov R. Relationship between Ion Pair Geometries and Electrostatic Strengths in Proteins. Biophysical Journal. 2002; 83:1595–1612. [PubMed: 12202384]

51. Kumar S, Nussinov R. Salt bridge stability in monomeric proteins. J Mol Biol. 1999; 293:1241–1255. [PubMed: 10547298]

52. Mallick P, Boutz DR, Eisenberg D, Yeates TO. Genomic evidence that the intracellular proteins of archaeal microbes contain disulfide bonds. Proc Natl Acad Sci U S A. 2002; 99:9679–9684. [PubMed: 12107280]

53. Folch B, Rooman M, Dehouck Y. Thermostability of salt bridges versus hydrophobic interactions in proteins probed by statistical potentials. J Chem Inf Model. 2008; 48:119–127. [PubMed: 18161956]

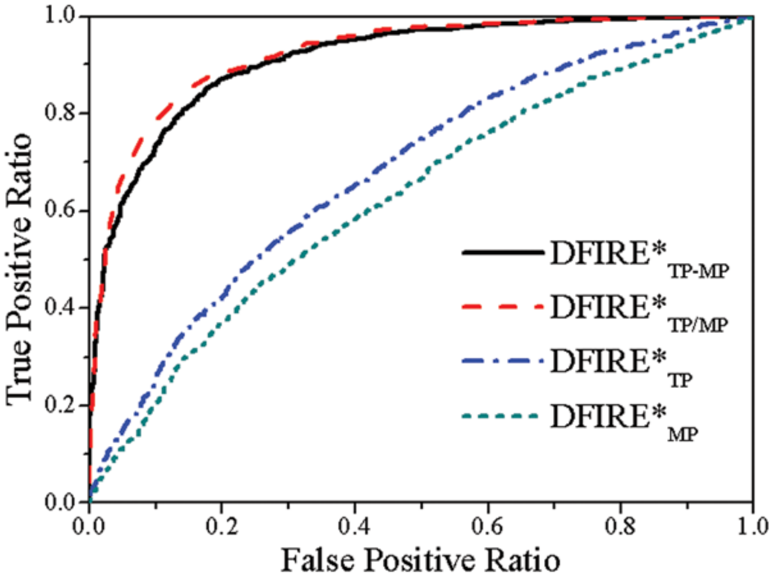

**Figure 1.**
The ROC curves of DFIRE*.

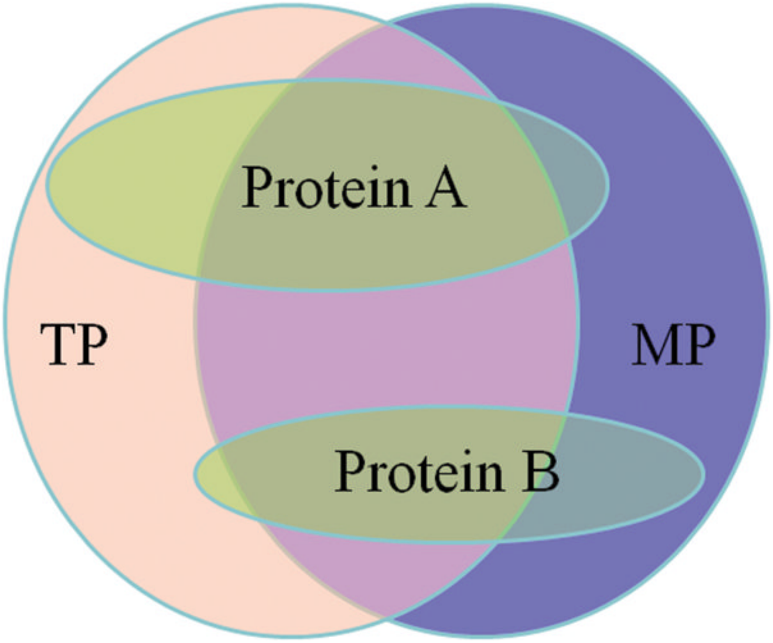

**Figure 2.**

Schematic plots to illustrate why the potential difference or the potential ratio can enhance discriminatory ability.

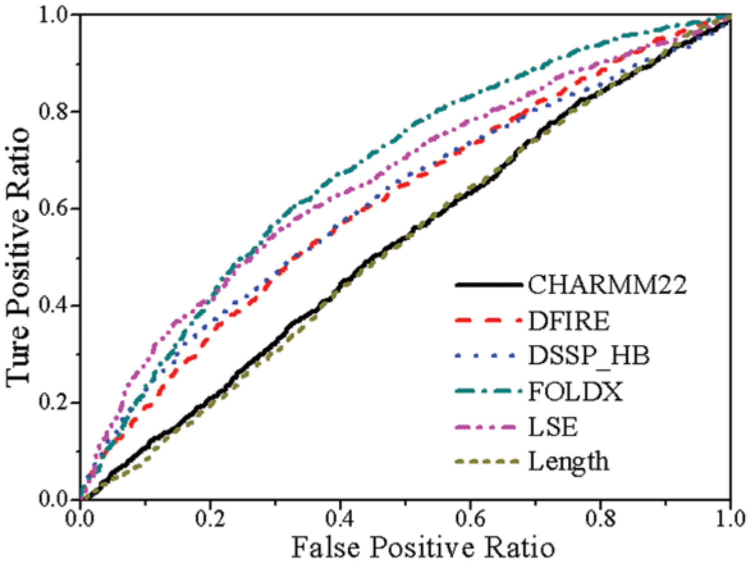

**Figure 3.**

The ROC curves of other classical potentials/properties in discriminating TPs and MPs.

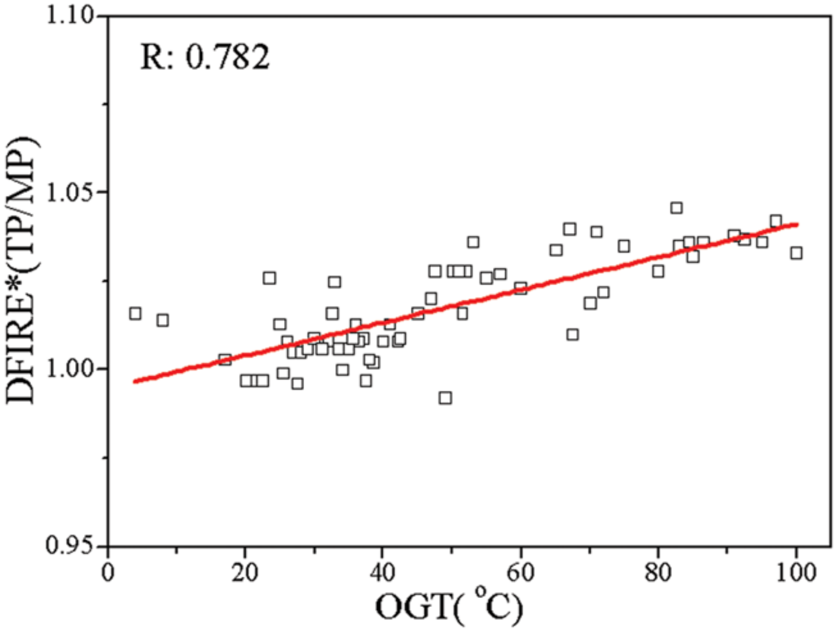

**Figure 4.**
Linear regression of DFIRE* potential ratios against OGTs.

## Table 1

The AUC and the accuracy (ACC) at 80% specificity of the potentials for discriminating thermophilic and mesophilic proteins.

| Settings | | Bin1(Backbone + C$_\beta$) | | | | Bin1(C$\alpha$) | | | | Bin2 (C$\alpha$) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **potential** | | **TP** | **MP** | **TP-MP** | **TP/MP** | **TP** | **MP** | **TP-MP** | **TP/MP** | **TP** | **MP** | **TP-MP** | **TP/MP** |
| | DFIRE* | 0.681 | 0.628 | 0.912 | 0.924 | 0.683 | 0.632 | 0.865 | 0.866 | - | - | - | - |
| AUC | E1 | 0.526 | 0.541 | 0.874 | 0.828 | 0.513 | 0.536 | 0.873 | 0.868 | 0.500 | 0.531 | 0.860 | 0.842 |
| | E2 | 0.533 | 0.545 | 0.874 | 0.868 | 0.527 | 0.544 | 0.872 | 0.867 | 0.521 | 0.533 | 0.872 | 0.866 |
| | DFIRE* | 0.736 | 0.722 | 0.812 | 0.817 | 0.700 | 0.729 | 0.797 | 0.799 | - | - | - | - |
| ACC | E1 | 0.698 | 0.703 | 0.800 | 0.800 | 0.739 | 0.702 | 0.800 | 0.800 | 0.705 | 0.698 | 0.795 | 0.789 |
| | E2 | 0.699 | 0.705 | 0.800 | 0.802 | 0.703 | 0.709 | 0.800 | 0.800 | 0.695 | 0.698 | 0.799 | 0.798 |

## Table 2

The AUC of the potentials for the testing dataset.

| Settings | Bin1(Backbone + $C_\beta$) | | | | Bin1 (Cα) | | | | Bin2 (Cα) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Potential** | **TP** | **MP** | **TP-MP** | **TP/MP** | **TP** | **MP** | **TP-MP** | **TP/MP** | **TP** | **MP** | **TP-MP** | **TP/MP** |
| DFIRE* | 0.654 | 0.634 | 0.759 | 0.762 | 0.683 | 0.662 | 0.739 | 0.735 | - | - | - | - |
| E1 | 0.500 | 0.513 | 0.835 | 0.828 | 0.513 | 0.509 | 0.834 | 0.827 | 0.519 | 0.512 | 0.831 | 0.805 |
| E2 | 0.506 | 0.517 | 0.835 | 0.824 | 0.502 | 0.514 | 0.834 | 0.834 | 0.502 | 0.514 | 0.823 | 0.837 |

## Table 3

The correlation coefficients of potentials against the OGTs of organisms produced the proteins.

| Settings | Bin1 (Backbone + $C_\beta$) | | | | Bin2(Cα) | | | |
|---|---|---|---|---|---|---|---|---|
| **Potential** | **TP** | **MP** | **TP-MP** | **TP/MP** | **TP** | **MP** | **TP-MP** | **TP/MP** |
| DFIRE* | 0.445 | 0.346 | 0.774 | 0.782 | - | - | - | - |
| E1 | 0.143 | 0.101 | 0.682 | 0.687 | 0.245 | 0.162 | 0.709 | 0.695 |
| E2 | 0.124 | 0.090 | 0.682 | 0.672 | 0.187 | 0.156 | 0.684 | 0.571 |