PROPERTY-DEPENDENT ANALYSIS OF ALIGNED PROTEINS FROM TWO OR MORE POPULATIONS

STEINAR THORVALDSEN, ELINOR YTTERSTAD, TOR FLÅ

Dept of Mathematics and Statistics, Faculty of Science, University of Tromsø, 9037 Tromsø - Norway.

Multiple sequence alignments can provide information for comparative analyses of proteins and protein populations. We present some statistical trend-tests that can be used when an aligned data set can be divided into two or more populations based on phenotypic traits such as preference of temperature, pH, salt concentration or pressure. The approach is based on estimation and analysis of the variation between the values of physicochemical parameters at positions of the sequence alignment. Monotonic trends are detected by applying a cumulative Mann-Kendall test. The method is found to be useful to identify significant physicochemical mechanisms behind adaptation to extreme environments and uncover molecular differences between mesophile and extremophile organisms. A filtering technique is also presented to visualize the underlying structure in the data. All the comparative statistical methods are available in the toolbox *DeltaProt*.

1 Introduction

Comparative analysis of proteins and proteomes derived from their genome data has already proven powerful in gene identification and in prediction of structure and function. Multiple sequence alignment can also provide information for comparative physicochemical analysis of proteins. Analysis of variations in amino acids at the different alignment positions allows inferences to be made about the pair- and multiple wise relationships between sequences or populations of sequences. The main approach in this paper is based on estimation and analysis of the variation between the values of a physicochemical parameter at positions of a sequence alignment.

The description and definition of chemical similarity and dissimilarity of molecules has long been an active area of study in experimental, theoretical and computational chemistry [1]. The descriptors may take the form of measured or computed physical properties such as topological or constitutional indices. There are several approaches based on counting shared features. Such features include atom or element types, bonds, topological torsions, etc. In formulating other descriptions of quantitative physicochemical distance, one is obliged to make approximations and to use heuristically derived solutions. There have also been several attempts to consider the measured biological properties of the amino acids as a basis for the diversity.

Microorganisms are found almost everywhere on earth. Some are able to tolerate extreme conditions such as low and high temperatures, low and high pH, high salt concentrations, high pressure and high radiation levels. These organisms are commonly referred to as *extremophiles*. Habitats with high temperature harbour the *thermophilic* organisms, which favour temperatures between 45-100º Celsius, and even higher [2]. At

the other end of the temperature scale are the *psychrophiles*, or cold-loving organisms, that are able to live even below the freezing point of water [3].

Clearly, to survive at low temperatures, the organisms have to face different challenges. One of the most important tasks is that the metabolism and the enzymes must be able to maintain adequate activity, which otherwise would slow down dramatically at low temperatures. Higher viscosity of water reduces the diffusion rates of substrates and products. Many enzymes from cold sources are shown to be more thermolabile and have a higher catalytic efficiency when compared to orthologous enzymes from warmer sources [4]. A commonly stated hypothesis is that this increased activity is caused by a more flexible structure at low temperature. Several articles concerning sequence comparisons, which aim to find common denominators for cold adaptation, have been published in recent years [4]. But the mechanism still remains unclear. Cold-adapted enzymes have been reported to have some general collective characteristics such as fewer salt bridges, reduction in the R/(R +K) ratio and the number of R (considered stabilizing because it has a long side chain which is both hydrophobic and can form salt bridges and hydrogen bonds), and a lower fraction of larger aliphatic residues expressed by the (I+L)/(I+L+V) ratio indicating a reduced core packing. But at the compositional level the differences in cold-adapted populations seem to be marginal, with overlapping standard deviation intervals.

To study more closely the mechanisms involved in protein cold adaptation on a molecular level, the enzyme *Uracil-DNA N-glycosylase* (UNG; E.C.3.2.2.3) has been chosen from related mesophile and psychrophile bacteria. UNG is an important intracellular, monomeric enzyme which recognizes and removes uracil occurring in DNA [5,6]. A disturbance of this reparation system results in occurrence of a number of diseases, including various kinds of a cancer. The enzyme consists of a classic single domain alpha/beta fold with a central four stranded beta-sheet surrounded by ten alpha-helices. By using a bioinformatics approach, we attempt to reveal the trends in temperature adaptations of this enzyme.

2 Materials and Methods

2.1 Sequence Data

Orthologues of UNG protein sequences from gamma-proteobacteria were collected from various genome sequence projects around the world, and a total of 32 amino acid sequences were found. The optimum growth temperature, T_{opt}, was determined by studying the literature, and by searching The Prokaryotic Growth Temperature Database, PGTdb [7]. Sequence identity is in the range from 38 to 98%. They were divided in *three populations* defined by temperature adaptation: *mesophilic* (T_{opt}=31-40°C, 20 sequences), *psychrotolerant* (T_{opt}=21-30°C, 9 seq.), and *psychrophilic* (T_{opt}=5-20°C, 3 seq.). Temperature is the main environmental trait that separates these populations, not factors such as highly concentrated salts or toxics.

Structural alignments of the data sets were created using the crystal structure from *E.coli* UNG (1FLZ.pdb) as guide. A structural alignment is a set of matched pairs or blocks where there is a meaningful correspondence between the data points in one population and those in the other. This gives us the possibility of investigating the physicochemical measurements in the sequences by statistical methods.

Furthermore, for the sake of a more specific analysis, the amino acid data was decomposed into structural elements in two different ways, and sectioned along the sequence according to these criteria:

- 2D structure region (alpha, beta, loop)
- 3D structure location (surface, twilight zone, core)

For this purpose the secondary (2D) structure was downloaded from the DSSP server, and solvent Accessible Surface Area (ASA) was calculated using the program GETAREA [8] with the crystal structure of *E.coli* as template and default settings. The spatial (3D) location was attached according to solvent accessible surface area of the side chain, where *surface*: 100-50% exposed side chain, *twilight zone*: 50-10%, and *core*: 10-0%. This method makes it possible to analyze the data relative to both its 2D and some of its 3D structural constraints.

2.2 Comparing Sequence Properties by Statistical Methods

Protein sequences are, with rare exceptions (e.g. long fibrous proteins like collagen or silk), quasi-random strings of amino acids with scant evidence of order or periodicity. As a first approximation we will consider these strings as random and independent sampling from a pool of amino acids with a specified probability distribution. The dataset in general will be unbalanced, in the sense that there are different numbers of sequences in each population.

In all the statistical analyses, it is important that the significant differences found (or not found) are due to the different conditions of the populations and not due to the organisation and conservation of the particular enzyme in the study. Therefore, the statistical tests automatically discard sites with no differences (the conserved sites in the alignment) from the analysis.

Each amino acid has many different indices, ranging from molecular mass to helix formation parameter. Sixty physicochemical, steric and other numerical properties of the amino acids were downloaded from the database AAindex release 6 [9] or collected from the literature [10]. The statistical tests will be performed for each of these properties, and the properties are assumed to be additive in the protein structure. For some properties (like molecular mass) this is self-evident, for others (like water-accessible surface area), it may only be an approximation for mean values, and for some (like heat capacities) it may be decided by experiments [11]. We applied *three* univariate statistical tests in the analysis. The main question of interest is whether there are any significant differences between the populations defined by temperature adaptation.

One-way ANOVA

In a previous study [10] properties of amino acids were averaged in unaligned sequences from different temperature populations, but no other statistical analyses than regression studies were performed. When the number of sequences in each temperature population is different, a statistical comparison may also be made by unbalanced one-way analysis of variance (parametric ANOVA).

Let the amino acids in a sequence *s* be x_j, j = 1,2,….n; where n is the number of non-conserved sites, and let the measurement of a particular property of the amino acids of the same sequence be $q(x_j)$, j = 1,2,….n. This yields real values when we assume that we have a table of quantitative chemical values, *q*, for each amino acid. There are a total of M sequences from three temperature populations. The test is based on mean values from each sequence *i*:

$$\bar{q}_i = \tfrac{1}{\mathrm{n}} \sum_{j=1}^{\mathrm{n}} q(x_j), \quad i = 1,2,\ldots,\mathrm{M}$$

Since the sample size n is moderately large, and by assuming independence between the sites, the normality assumption of the ANOVA is fulfilled. An ANOVA will test whether one or more of the population means are unequal, see Figure 2.

Matched-pair tests for step change

We want to improve this strategy above by focusing on *positions* of aligned sequences, rather than averaging over unaligned sequences. Our data (amino acids) can be grouped according to both population and position in the aligned sequence they belong to. The second grouping factor, position, is included in the model to take into account possible differences between positions along the aligned sequence.

For amino acids *x* and *x*', we may define a chemical *difference measure* for the two amino acids:

$$d(x, x') = q(x') - q(x)$$

This measure is an expression of the diversity between the amino acids, and the choice of measures to be used depends on the property we want to test.

Let s_m, m = 1, 2, …M, be the M aligned amino acid sequences, and let the amino acid at site j in s_m be denoted by $x_{m,j}$. We define the difference between the sequences from population $p_{(1)}$ and $p_{(2)}$ at site j by averaging the measurements at site j within each population:

$$d_j(p_{(1)}, p_{(2)}) = \bar{q}(x_{Pop\,2,j}) - \bar{q}(x_{Pop\,1,j}), \quad j = 1,2,\cdots,\mathrm{n}$$

By this random variable *d* we measure n assumed independent and identically distributed differential effects between population 1 and 2, where n is the length of the non-conserved alignment. The underlying distributions of the variables *d* are very seldom known, and for many physicochemical properties there are distinct indications that the distributions are non-normal. In such situations the use of the standard parametric

methods assuming normality may be criticized regarding validity and optimality. However, nonparametric methods based on ranks are valid for a broad family of underlying distributions. This gives rise to the possibility of using a data-adaptive test. A standard goodness of fit test was used for to screen the chemical data for deviations from normality (Kolmogorov-Smirnov test) with a significance level of 0.25. By this some physicochemical properties are found to be normal, and some are not (like Kyte-Doolittle hydrophobicity with more than one peak). In general, non-parametric tests will need larger sample size than the corresponding tests based on normality.

In the case of *normality* we approximate *d* with the normal distribution, and apply the *paired t-test*. In the non-normality case we may apply the paired *Wilcoxon signed-rank* test [12].

These two tests, based on paired data, define a useful and reliable statistical method when we are investigating a variable along the sequence in two population groups. In practice we select the most extreme temperature populations for this test (psychrophiles and mesophiles). The tests can be used on all continuous type of paired data with symmetric distributions. The variances of each of the mean values may be unequal, which will be the case for unbalanced data sets. Results of the Wilcoxon test will still be valid if the distributions have different symmetrical shapes and a common mean.

Cumulative Mann-Kendall test for monotonic trend

A statistical test for trend in the population levels was also performed. Our data consists of temperature-ordered populations, and this ordering may be utilised in the statistical approach by applying a general trend test. The step change presented above is a special case of a more general type of trend often called *monotonic* trends. A monotonic trend indicates that the properties shift monotonically with temperature, but does not specify if this occurs continuously, linearly, in one or more discrete steps, or in any other explicit pattern.

Mann-Kendall tests are a group of nonparametric tests for detection of monotonic relationship between two variables [13]. The collected data are separated into n different sites. Within each site *j* in the alignment we have M observations of property value *q* and population group *p* obtained in pairs:

$$(q_i, p_i), \quad i = 1,2,\ldots,\mathrm{M}$$

The test uses only the relative magnitude of the data rather than their measured values. By the Mann-Kendall test we rank both the property data and the grouping data at each site, and base the test statistic on these ranks. In case of tied data (equal-values), we use the average ranking. The Mann-Kendall statistic K_j at site *j* is computed by comparing each of the M(M-1)/2 possible pairs of observations, and examine if the two variables are ranked in the same order or in reverse order:

$$K_j = \sum_{m=1}^{M-1} \sum_{i=m+1}^{M} sign\left[(q_i - q_m)(p_i - p_m)\right]$$

where

$$sign(x) = \begin{cases} +1, & if\ x > 0 \\ 0, & if\ x = 0 \\ -1, & if\ x < 0 \end{cases}$$

This is the number of positive differences minus the number of negative differences. The distribution of K_j has been studied by Kendall, and its variance is known [12,13].

Next we need a way to consider all sites in the alignment simultaneously. In a cumulative test the Mann-Kendall test is applied to each site separately, and then the results are combined for an overall test:

$$K = \sum_{j=1}^{n} K_j$$

Under the null hypothesis (no trend), K is asymptotically normally distributed with mean zero and variance:

$$Var(K) = \sum_{j=1}^{n} Var(K_j)$$

Each site by itself may show a positive trend, none of which is significant, but the overall cumulative Mann-Kendall statistic may still be significant.

Visualization

The differences between the populations can be compared graphically as well as statistically, and we used a smoothing technique to be able to recover and visualize underlying structure in the data set [14]. We use a 2D rectangular box-filter where the vertical filter size is all the amino acids in the aligned sequence position of the population, and the horizontal window size can be varied. This 2D filter can be used to plot smoothed line of amino acid properties, such as comparative plots shown in Fig. 1.

3 Applications and Results

3.1 Difference in Physiochemical Properties

We have compared amino acid sequences from three populations by statistical tests and found that it provides interesting results. A summary of the results are shown in Table 1. Results are only reported in the table if the P-values are found to be significant ($P<0.05$) in at least two of the three tests described above. We observe interesting differences especially in the surface and twilight zone, and in the loop regions of the molecule.

We observe decreasing trends with cold adaptation for hydrophobicity, isoelectric point, and long range non-bounded energy. The exterior of the molecules is more negatively charged and will increase their solubility in water. The loop and exterior of the psychrophilic enzymes has a more negative potential (low isoelectric point)

Table 1. List of main differences from mesophile to psychrophile population. P-values are obtained by the ANOVA test (P_{ANOVA}), the paired tests ($P_{t/W}$), the cumulative Mann-Kendall test (P_{cumMK}), and are reported in the given order. Details in the references to amino acid properties can be found in [9, 10].

Property	Trend	2D region P-values	3D region P-values	Entire sequence P-values	Ref.
Hydrophobicity	↓		Surface: 0.05/0.31/0.003	0.04/0.04/0.003	Ponnus-wamy 1993
Isoelectric point	↓	Loop: 0.001/0.03/0.002	Twilight: 0.03/0.04/0.20	0.0001/0.01/0.03	Zimmerman 1968
Molecular weight	↓↑		Twilight: ↓ 0.01/0.04/0.01 Surface: ↑ 0.10/0.04/0.01	0.93/0.82/0.40	Fasman 1976
Volume	↓↑		Twilight: ↓ 0.02/0.03/0.005 Surface: ↑ 0.10/0.003/0.007	0.91/0.40/0.58	Gunsteren-Mark, 1992
Energy long range	↓	Loop: 0.007/0.30/0.0001	Surface: 0.007/0.23/0.0004	0.02/0.12/0.005	Oobatake-Ooi, 1977
Heat capacity	↑		Surface: 0.14/0.04/0.003	0.28/0.30/0.01	Hutchens, 1970
Compressibility	↓	Loop: 0.09/0.02/0.03	Surface: 0.03/0.0007/0.0005	0.18/0.30/0.01	IqbalVerrall 1988
ΔH (unfolding enthalpy change	↑		Twilight: 0.02/0.05/0.10	0.14/0.39/0.20	Oobatake-Ooi, 1993
-TΔS (unfolding entropy change)	↓		Twilight: 0.01/0.06/0.05	0.07/0.36/0.19	Oobatake-Ooi, 1993
ΔG (Gibbs free energy change)	↑		Twilight: 0.004/0.09/0.004	0.57/0.53/0.49	Oobatake-Ooi, 1993
Shape (pos. of branch point)	↓	Loop: 0.01/0.08/0.02	Twilight: 0.0002/0.002/0.0007	0.02/0.06/0.02	Gunsteren-Mark, 1992

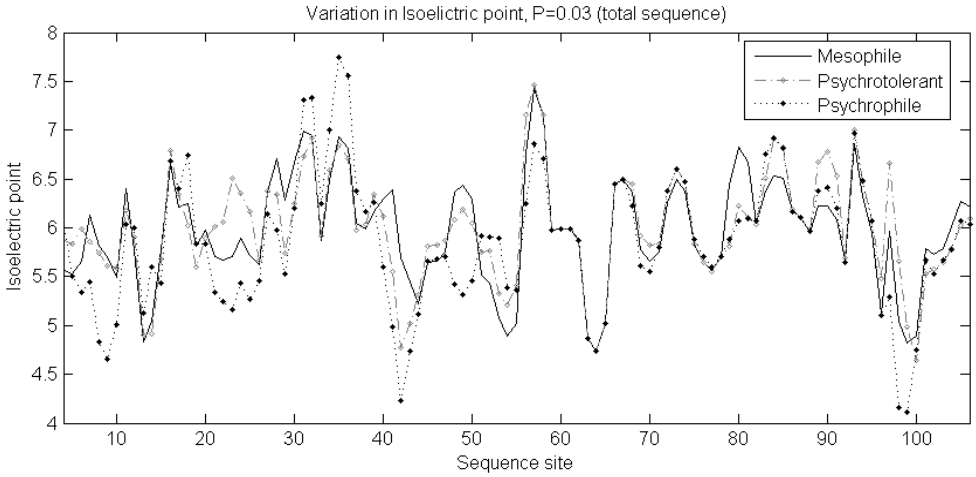

Figure 1. Variation of the isoelectric point property in the first part of the alignment of UNG. We use a box filter of size m x 3 to recover the underlying structure in the data, where m is the number of sequences in the population. The psychrophilic population has 3 sequences, psychrotolerant 9, and mesophilic 20. In average the psychrophilic population appears to have lower values than the mesophilic counterparts.

surrounding the active sites where DNA binds, than the other enzymes. The substrate for UNG is the negatively charged DNA. However, the active site of the UNG enzymes are dominated by conserved positively charged residues where DNA binds. A highly negative charged surface may lead to slow binding of the enzyme to the substrate. But on the other hand, this may contribute to a faster catalysis, as the negative potential forces the product quickly away from the enzyme. Hence, a lower potential will promote weaker electrostatic interactions between the DNA and the enzyme, but in addition it may show a tendency to optimize the electrostatics around the active site.

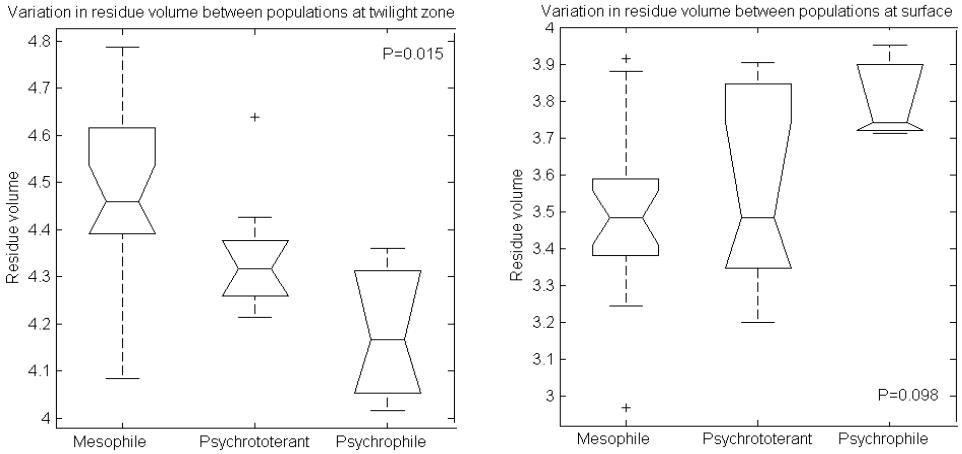

Figure 2. Boxplots showing variations in the volume within and between the populations at the twilight zone (left) and at the surface (right). The analyses show an interchange between the two regions. The one-way ANOVA test is used to compute each P-value.

The heat capacity is one of the fundamental parameters describing thermodynamic properties of a system, and for the amino acids the heat capacity is known to decrease with temperature. In our data the heat capacity is increased from the mesophilic to the psychrophilic population at the surface, but also slightly in general. We interpret this modification to be compensatory to the inherent effects of temperature change. The compressibility parameter also follows an inverse change compared to heat capacity. Most studies of the stability of proteins are concentrated on evaluation of the Gibbs free energy of unfolding, a parameter that provides a measure of thermodynamic stability of the protein molecule. We found no significant overall difference for this parameter between the populations, only a local increase in the twilight zone. However, the Gibbs energy, ΔG, consists of two terms describing the enthalpic, ΔH, and entropic, $-T\Delta S$, contribution, i.e. $\Delta G=\Delta H-T\Delta S$. The enthalpic and entropic contributions for a given system appear to have a close relationship, the so-called enthalpy/entropy compensation. In some cases the enthalpy/entropy compensation is significantly close to obscure the occurrence of the changes in a system, if the analysis is done only in terms of Gibbs energy. The differences between the separate changes in the enthalpy and entropy may be

quite significant, as we found it to be in our data analysis, where ΔH goes up and $-T\Delta S$ down (Table 1). Both ΔH and ΔS are dependent on the heat capacity of the involved amino acids and will decrease with temperature, so the changes found in enthalpy and entropy are fully consistent with the observed increase in heat capacity of residues found in the psychrophilic population.

In Table 1, it is also of importance to note that the shape property, defined as position of branch point in the side chain (e.g. ranging from 0 in Alanin to 5 in Argenin), decreases with adaptation to cold. This observation is a direct extension in agreement with earlier results found for thermophilic proteins [10], and may indicate a more flexible exterior because of early or no branching. Surprisingly, we observe no significant changes in the alpha or beta structures.

4 Some Conclusions

We have applied and expanded the methods of comparative analysis of proteins. The improved strategy is partly extensions of traditionally used statistics [10], e.g., residue frequencies, residue properties, but applied to positions of aligned sequences rather than averaged over unaligned sequences. In this paper a unified framework for context-sensitive and property-dependent analysis of alignments is developed, including data representation and efficient computations using statistical methods. We have demonstrated how alignment data can be incorporated into a Mann-Kendall trend test. In this *cumulative* Mann-Kendall test the alignment sites are tested individually, and then combined into one overall test result. We extracted significant differences into several distinct physicochemical factors.

In the present study of UNG, we found that the properties shape (defining length to the first branch point of side chain), and isoelectric point are generally the most important properties for adaptation to cold. The exterior of the molecules is more negatively charged and will increase their solubility in water and provide weaker electrostatic interactions with the negative substrate (DNA). But small areas around the active site have a positive potential, which possibly acts to improve the interactions.

Furthermore, the shape parameter is decreasing mainly in loop regions and at the twilight zone, indicating weaker medium range interactions, and an increased flexibility at the surface and between the secondary structure elements. This may indicate that the cold adapted protein are characterised by an improved flexibility of the structural components involved in the catalytic cycle. In addition the heat capacity, unfolding enthalpy, and unfolding entropy are found to be different in a direction that compensates the inherent chemical effects of temperature change.

Hence, the ability to be active at temperatures that are close to the freezing point of water requires an array of minor adaptations to compensate for the temperature loss and maintain the enzymatic function. However, these results are based on bioinformatics results only and need to be verified by more detailed analyses in the laboratory.

Some of the features observed may be specific to groups of proteins, and different enzymes may have different strategies for cold adaptation. But also the same enzyme may have different strategies depending on its working environment, and more sequence families should be analyzed to detect both general and special molecular determinants of cold adaptation. A multivariate extension of the present analysis may also be of interest.

All analyses reported in this work were implemented in Matlab. Our toolbox *DeltaProt* can be downloaded from our web-site at: http://www.math.uit.no/bi/deltaprot/

Acknowledgments

The sequence alignment of UNG was kindly provided by Nils P. Willassen.

References

1. N. Nikolova, J. Jaworska. Approaches to measure chemical similarity - A review. *QSAR & Combinatorial Science,* (9-10): 1006-1026, 2004.
2. K. Kashefi and D. R. Lovley. Extending the upper temperature limit for life. *Science*, 301(5635): 934, 2003.
3. K. Junge, H. Eicken, et al. Bacterial Activity at -2 to -20 degrees C in Arctic wintertime sea ice. *Appl Environ Microbiol*., 70(1): 550-7, 2004.
4. G. Feller and C. Gerday. Psychrophilic enzymes: hot topics in cold adaptation. *Nat Rev Microbiol* 1: 200-208, 2003.
5. S.S. Parikh, C.D Mol, D.J Hosfield and J.A Tainer. Envisioning the molecular choreography of DNA base exicions repair. *Curr Opin Struct Biol* 9(1):37-47, 1999.
6. I. Leiros, E. Moe, A.O. Smalås and S. McSweeney. Structure of the uracil-DNA N-glycosylase from Deinococcus radiodurans. *Acta Cryst*., D61:1049-1056, 2005.
7. L.C.W Huang, K.H. Laing, K.T. Pan and J.T. Horng. PGTdb: a database providing growth temperatures of prokaryotes. *Bioinformatics* 20(2): 276-278, 2004.
8. R. Fraczkiewicz and W. Braun. Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J. Comp. Chem*., 19: 319–333, 1998.
9. S. Kawashima, H. Ogata and M. Kanehisa. AAindex: amino acid index database. *Nucleic Acids Res.*, 27, 368-369, 1999.
10. M.M. Gromiha, M. Oobatake and A. Sarai. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. *Biophysical Chemsitry*, 82, 51-67, 1999.
11. M. Hackel, H. J. Hinz and G. R. Hedwig. Additivity of the partial molar heat capacities of the amino acid side-chains of small peptides: Implications for unfolded proteins. *Physical Chemistry Chemical Physics*, 2 (23): 5463-5468, 2000.
12. M. Hollander and D. A. Wolfe. *Nonparametric Statistical Methods*, 2.ed. Wiley, 1999.
13. M.G. Kendall and J.D Gibbons. *Rank Correlation Methods,* 5th ed. Edward Arnold, 1990.
14. A. V. Oppenheim and R.W. Schafer. *Discrete-Time Signal Processing,* Englewood Cliffs, NJ: Prentice-Hall, 1989.