Extracting Molecular Diversity Between Populations Through Sequence Alignments

Steinar Thorvaldsen[1], Tor Flå[1], and Nils P. Willassen[2]

[1] Dept of Mathematics and Statistics, Faculty of Science
{steinart,tor}@math.uit.no
[2] Department of Molecular Biotechnology, Faculty of Medicine,
University of Tromsø, 9037 Tromsø, Norway
nilspw@fagmed.uit.no

Abstract. The use of sequence alignments for establishing protein homology relationships has an extensive tradition in the field of bioinformatics, and there is an increasing desire for more statistical methods in the data analysis. We present statistical methods and algorithms that are useful when the protein alignments can be divided into two or more populations based on known features or traits. The algorithms are considered valuable for discovering differences between populations at a molecular level. The approach is illustrated with examples from real biological data sets, and we present experimental results in applying our work on bacterial populations of *Vibrio*, where the populations are defined by optimal growth temperature, T_{opt}.

Keywords: sequence analysis; structural analysis; physicochemical properties; extremophiles; Fisher's exact test; Wilcoxon test.

1 Biological Motivation

Extreme environments are those that fall outside the limited range in which we, and most other eukaryotes can survive, and are inhabited by the *extremophiles*. Among extremophiles, which include thermophiles, psychrophiles, acidophiles, alkalophiles, halophiles, barophiles and xerophiles, those who live and prefer low temperatures are the largest and least studied group. Psycrophilic organisms are living at temperatures close to the freezing point of water. It is of great interest to understand how these organisms can function at "the limits of life" [1].

Living at extreme temperatures requires a multiplicity of crucial adaptations including preservation of membrane stability and maintenance of enzymatic activities at appropriate levels. At these temperatures a number of physiological factors are changed; the solubility of gases is not the same, the viscosity of water increases several folds as temperature is changed towards the extreme areas, for example.

The number of characterized cold or heat adapted proteins, reported sequences and high resolution structures is growing. The *Vibrios* are of the species with the greatest amount of published genomes, reaching five completed genomes this year, and seven ongoing whole genome sequencing projects including the cold adapted *Vibrio salmonicida*.

Alignment-free analysis has been used previously to compare amino acid compositions in whole genome and proteome datasets [2][3]. In this study, we focus on a set of homolog protein data from a relatively narrow range of closely related species

J.L. Oliveira et al. (Eds.): ISBMDA 2005, LNBI 3745, pp. 317–328, 2005.

belonging to the group *Vibrios* of gamma proteobacteria - a strategy also adopted by [4]. In our comparative study, we employ alignment based methods for examination of similarities and chemical differences at the molecular level by comparing amino acids and their *physicochemical properties* in the proteins. Different new methods of univariate analysis have been developed and used in this analysis.

The definition and analysis of chemical similarity has long been an active area of study in theoretical and computational chemistry [5-7]. Currently, there seems to be no generally agreed quantitative, or even qualitative, definition of chemical diversity. In formulating any description of quantitative chemical distance, one is obliged to make approximations and to use heuristically derived solutions. Many different ways have been used to represent chemical structures leading to many different approaches to assessing their similarity. These include methods based on three-dimensional representations. Two-dimensional approaches are, perhaps, even more numerous. The term 2D is a convention, as it is in general the properties of the molecular graph which are of interest, and not its pictorial representation in the plane. There have also been attempts to consider the measured biological properties of compounds as the basis for diversity analysis. The descriptors may take the form of measured or computed physical properties such as topological or constitutional indices. There are several approaches based on some count of shared features. Such features include atom or element types, bonds, topological torsions, etc.

2 The Algorithms

2.1 Residue Frequencies

We wanted to examine amino acid occurrences in relation to background distributions, and for this purpose we analysed the composition of amino acids in two different sequence populations. The distribution of the categorical variable (amino acid type) in the sequence samples can be modelled and compared statistically. The basic chain-like structure of proteins, allows an abstract view of these as strings, or sequences, over a finite alphabet. Protein sequences could in principle, as a first approximation, be considered as random samples taken from some distribution. Let X be a discrete random variable with a finite set of possible values, or categories {A,R,N,D,Q,E,G,H,I,L,K,M,F,P,S,T,W,Y,V}. For convenience the amino acids are often encoded by ordinal numbers x = 1, 2,…, 20. If sites were independent, the *multinomial* distribution would follow. This is the multivariate generalisation of the common binomial distribution.

However, gene sequences are not completely random, but display various kinds of structure. E.g. a family of homologous proteins is likely to have similar amino acid residues in “equivalent” positions, and the amino acid frequencies are expected to be different from gene to gene. Let two populations be defined by some trait (like different temperature preference). For every gene family g = 1, 2,…, k; we may for each of the 20 amino acids compute the frequency change vector between the mean frequencies in the two populations:

$$\Delta \boldsymbol{f}_g = (\Delta f_1, \Delta f_2, \ldots, \Delta f_{20}), \quad g = 1,2,\ldots, k$$

This is a variable that can be studied statistically across the dataset concerning significant changes.

Still, the statistics above have serious limitations. They simply indicate the degree of evidence for an over- or under-representation of the variables, and are not adequate for answering other more interesting questions about the data. One should also study the nature and effects of these differences.

2.2 Residue Substitution Pairs

A statistical approach to molecular sequence analysis also involves the stochastic modelling of the substitution, insertion and deletion processes. We also present an analysis of amino acid substitution data matrices from an independent set of paired homolog protein sequences. The method is based on trusted alignments, where observed amino acid replacements are tallied in a raw residue replacement matrix. We also present an analysis of the amino acid substitution pattern.

The modelling of amino acid replacement by a Markov chain has been introduced by Dayhoff et al [8]. Our strategy is much of the same as the method Dayhoff' used to estimate the well known initial PAM1 (Percent Accepted Mutation) transition matrix based on just 1572 substitutions. But our technique is different and our aim is to study the internal distance between populations, not to extrapolate to higher PAM distances.

Brenner et al [9] were the first to point out the problem in Dayhoff's method of estimating *one* consistent model from an *inhomogeneous* pool of aligned sequence data. Bias in the sequence selection may influence the frequencies of substitutions. Though biases cannot be eliminated entirely when data are sparse, one has to minimize biases in the data selection. Without clustering, some closely related sequences may be over-represented, and special care must be taken to locate representative samples as a safeguard against obtaining bias comparisons and ditto results in the investigations. Jones et al [10] presented an alternative method where the set of sequences are clustered at the 85 % identity level. The *closest* relating pairs of sequences are aligned, and observed amino acid exchanges tallied in a matrix.

We count by the Jones-method to minimize biases. Two pairs in an aligned site can be classified as invariant (where the same amino acid are conserved in the two populations) or variant (where there is a difference). Our main goal is not to measure conservation, but its opposite, deviation. A matched *Substitution Pair* (SP) is defined as the ordered combination of two amino acids observed in an alignment position, and a SP-matrix is the accumulation of all such pairs by summing over all positions. The accumulated array contains the frequency of all position specific pairing of residues. Thus, for any number of aligned amino acid sequences, the number of possible SP in each position is between 1 and 400, but only a small fraction of these SPs are observed, and the majority of sequence positions are covered by less than 10 SPs for closely related homologous proteins. This method of mapping and expressing alignment data by SP-matrices trim down problems caused by statistical dependences between the sequences and the uncertain phylogeny involved in the PAM procedure. Note that no counting is done when a residue is aligned to a gap. The accepted SPs are counted by the algorithm shown in Figure 1.

We may both calculate SP-matrices *between two* populations, and *within one* population. SPs may be used to address the following question: Which and how many SPs account for the major significant variations between the populations?

Algorithm CountSubstitutionPairs

Input	k gene families each with m(k) aligned protein sequences $s_1,s_2,\ldots s_{m(k)}$ from two populations.
Process	find in all genes all closest sequence pairs between the populations, and count substitutions over all positions
1	for all genes
2	for every sequence pair (s_{Pop1},s_{Pop2}) with max similarity
3	for every residue position j = 1,…,n
4	find all residue pairs SP_j = {sp(x, x'): x ϵ Pop1 and x' ϵ Pop2}
5	SP := SP + SP_j
Output	substitution pair matrix SP between the two populations

Fig. 1. An algorithm to compute the SP-matrix. The amino acids are denoted by x = 1,2,…,20

Instead of doing an overall test of the big SP-matrix, we partition the matrix in a meaningful manner and focus on more targeted tests. We want to analyse the over- or under representation of single SPs compared to a random model. The occurrence of SP might be expected to differ in some measure due solely to chance factors of sampling, and for other reasons which might be attributed to random causes. And what we shall need to find out is whether or not the observed differences are too large to be credited to such causes.

An enduring problem in statistics is the analysis of 2x2 contingency tables, and there has been a lot of research and debates [11, 12]. The main debate has at least two components. The first is to select either an exact test (e.g. Fisher's exact test) or an asymptotic test (e.g. Pearson's chisquared test). The second is which test procedure should be employed among many candidates in each group. There has been an effort to determine the best exact test among the Fisher's exact test, the exact chi-squared test and the exact likelihood ratio test in 2x2 tables in both large and small samples. Kang and Kim [13] compared the three conditional tests and showed that the Fisher's exact test turns out to be the best choice in most cases. Consequently, because of the practical values in our data, we decided to use *Fisher's exact test* to find statistically whether there is any non-random relation between any two categorical variables with two observed levels found from the SP-matrix.

2.3 Residue Properties

An alignment of homolog sequences is a set of matched *pairs* where there is a meaningful one-to-one correspondence between the data points in one group and those in the other. This gives us the possibility to investigate the mean property differences (like hydrophobicity) in the sequences by a probabilistic framework.

For two amino acids, x and x', we denote their linear chemical *difference measure*:

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

This difference yields real values when we assume that we have a table of quantitative chemical values, q, for each amino acid. The measure is an expression of the diversity between the amino acids, and the choice of measures to be used depends on the test we want to perform.

Let s_m, $m = 1, 2, \ldots M$, be M aligned amino acid sequences, and let the amino acid at position 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 position j by averaging the measurements at position j within each population:

$$d(p_{1,j}, p_{2,j}) = \bar{q}(x_{Pop2,j}) - \bar{q}(x_{Pop1,j})$$

By this we measure n differential effects between population 1 and 2, where n is the length of the gapless alignment. We find the mean chemical difference, D, between the two populations by:

$$D(p_1, p_2) = \frac{1}{n}\sum_{j=1}^{n} d(p_{1j}, p_{2j})$$

Assuming that the distribution of the differences in each position, j, is identical, we obtain the expected value Δ:

$$\delta = E(D(p_1, p_2)) = \frac{1}{n}\sum_{j=1}^{n} E(d(p_{1j}, p_{2j})) = E(d(p_{1j}, p_{2j})), j = 1,2,\ldots n$$

A standard parametric test would be to approximate D with the normal distribution, and apply the paired t-test. However, since it is not always clear that this is appropriate with the protein properties; other statistical test should be considered. In the statistical analysis, it is also important that the significant difference found between means (or not found) be due to the different conditions of the populations, and not due to the organisation and conservation of the particular enzyme in the study.

A relevant alternative to the t-test is the *Wilcoxon signed-rank test* which is a non-parametric test that also can be used on continuous type of paired data, both when the underlying population is normal and when not [14]. This test automatically discards all differences equal to zero from the analysis (conserved sites). It can in some cases be better than the paired t-test for non-normal populations, although non-parametric procedures in general need larger sample size than t-tests.

It is not easy to compare the two test procedures in a general theoretical way. One widely used measure in the literature is *asymptotic relative efficiency* (ARE, [14]). The ARE of one test relative to another is the limiting ratio of the sample size necessary to obtain identical error probabilities for the two procedures. For normal populations the ARE of the Wilcoxon test relative to the t-test is $3/\pi \approx 0.95$, and for non-normal populations the ARE is ≥ 0.86 which means that it in some cases will exceed 1. Although these results are for large samples, and do not necessary tell us anything for small samples, one may generally conclude that the Wilcoxon signed-rank test will never be much worse than the t-test, and in many cases where the population is non-normal it may be better. Our experience with using the two tests on the protein data is that they do not make much difference, with the Wilcoxon test as the most conservative, except for properties with more than one peak distribution (like Kyte-Doolittle hydrophobicity).

This defines a useful and reliable statistical model when we are investigating a variable along the sequence in two population groups. The formal statement of the hypothesis of interest is

$$H_0 : \delta = 0, (\text{zero mean difference})$$

$$H_1 : \delta \neq 0$$

We may test the significance of property alterations, e.g. a decreased surface hydration free energy, in psychrophilic sequence populations. This gives an efficient and more reliable comparison of protein populations than earlier studies.

The differences between two sequence populations can be compared graphically as well as statistically, and we developed a smoothing technique to be able to recover and visualize underlying structure in the data set [15]. We use a 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 filter can be used to plot smoothed lines of amino acid properties, such as comparative plots shown in Figure 5. All analyses reported in this work were implemented in Matlab.

2.4 False Discovery Rate

A common objection against the testing algorithms described above will be the multiple comparisons problem. Benjamini and Hochberg have suggested that the false discovery rate (FDR) may be the suitable error rate to control such multiple testing problems [16]. FDR is the expected proportion of false rejections among all rejections and is a new measure of error rate. A simple procedure was given by them as an FDR controlling procedure for independent test statistics and was shown to be much more powerful than comparable procedures like Bonferroni correction which may be much too conservative. The original formulation FDR presumes independence among the different amino acid properties, which is far from correct in our case. But in a recent paper [17], the FDR criterion has also been extended to multiple testing under dependency.

However, a more straightforward way to overcome this difficulty is just to analyse more than one dataset by the same procedure, and only report features that are significant across many independent protein groups.

3 Experimental Design and Results

3.1 Methodology and Datasets

Sequenced proteins from the *Vibrios* were downloaded from standard databases in order to identify homologues. With a minimum cut-off score of 70% sequence identity the corresponding amino acid sequences of 7-14 vibrio species (4-10 in the *mesophile* population and 2-4 in the *psychrophile*) were extracted and 25 alignments of intracellular sequences were made with T-coffee. Physicochemical, steric and other properties were downloaded from the database AAindex release 6 [18], which contain 494 quantitative properties of the amino acids, and collected from the literature [19].

We applied the methods of comparative analysis of protein sequences by focusing on discriminative features extracted from rationally selected parts of the data sets. In this approach we made extensive use of the sequence based predictors developed in the latest years, and our alignment-based data sets were decomposed and clustered in relevant subclasses. We used these tools: sub-cellular location (CELLO, Yu 2004), surface (SABLE, Adamczak 2004) and secondary structure (PSIPRED, McGuffin 2000). The secondary structure was predicted with default settings, and the solvent Accessible Surface Area (ASA) was predicted with thresholds 0-1: completely buried, 2-3: twilight zone, and 4-9: surface. All predictions were based on the sequences from

Vibrio cholerae. The alignments and annotations were automatically done with a program script, and made it possible to analyse the alignment data relative to its cellular, 2D and some of its 3D structural constraints:

- cellular location (intracellular, membrane, extracellular)
- 2D secondary structure region (alpha, beta, loop)
- 3D structure location (surface, twilight zone, core)

3.2 Experimental Results

Compositional Analysis

Different subsets of the protein data were used in the analysis; we examined the distribution in the compositional space. An example showing the variation in amino acid compositions are shown in Figure 2. In general it seems difficult to resolve general elements of cold adaptation at the basic compositional level. The standard deviations between the populations are overlapping. The greatest differences are found to be at the surface and the smallest in the interior of the molecule.

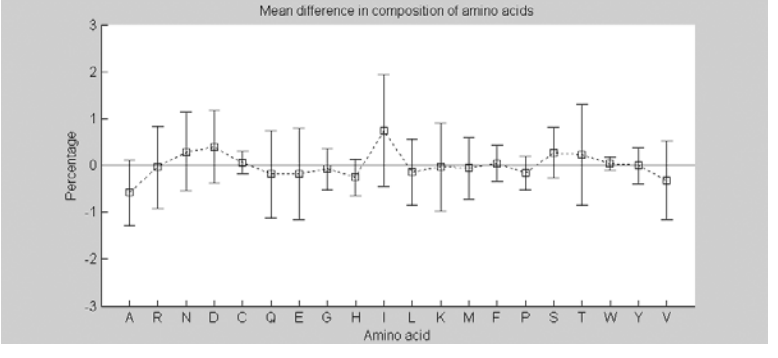

Fig. 2. Average amino acid difference in compositions from the mesophilic to the psychrophilic temperature domain of 25 cytoplasmic proteins from the bacteria *Vibro*. Bars are the empirical standard deviations

Analysis of SP Patterns

In the matrix shown in Figure 3, we present the SPs found between the mesophilic and psycrophilic populations in one of our data analysis. They were all tested for statistical significance by using Fisher's exact test (Figure 4).

The data may also be decomposed and studied in more detail, such as *secondary structure* and *external/internal* surface position. Analyses of residue properties reveal that many frequent SPs consist of the most hydrophilic/ hydrophobic residues, or residues with high propensities to form secondary structures. In general, residues forming the most significant SPs have extreme values of one or more essential Physicochemical property.

(V,I) is the most frequent SP. These residues have in common a maximal propensity to form beta-sheets, and the highest minimum width of the side chain. Among all pairs of residues, their steric similarity is the highest, as expressed by the partial specific volume and bulkiness. They are considered the most hydrophobic and conse-

quently the most often buried residues which transfer with the highest free energy from the exterior to the core of the protein. The (L,I) has similar characteristics to (V,I) and is also frequent, but I is more similar to V than to the differently branching L.

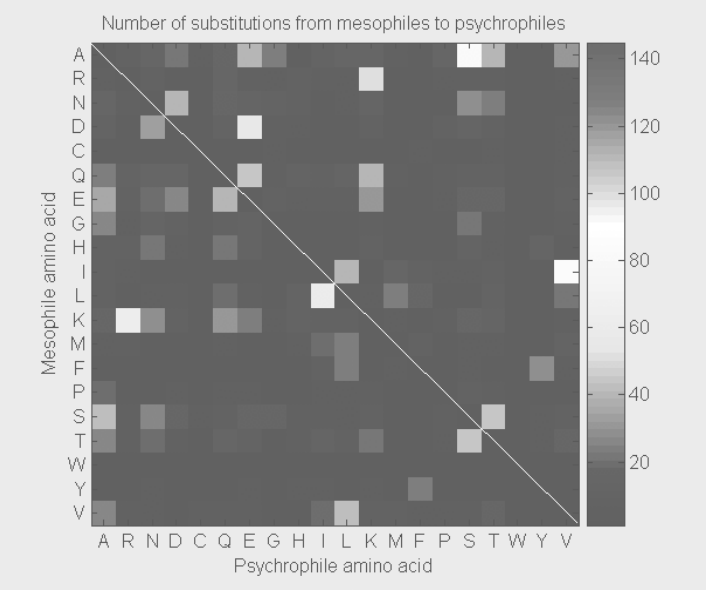

Fig. 3. Visualisation of all the pairwise substitutions from the mesophilic (left) to the psychrophilic (bottom) group. Two clear trends can be seen: Val (V) is replaced by Ile (I), P-value=0.007; and Glu (E) is replaced by Asp (D), P-value=0.21. Many of the other minor changes are found to be statistically significant (P<0.01)

Example: Substitution pair: V -> I

Population	Yes	No
Mesophilic vs. Psychrophilic	146	141
Psychrophilic vs. Mesophilic	83	132

Fig. 4. Example showing the use of Fisher's exacte test. P-value= 0.007

(E,D) is the second most frequent SP. These hydrophilic and accessible residues have high transfer free energies from water to organic solvents. They have the lowest propensity for bata-structures and the highest helix termination parameter.

(K,R) is also a frequent SP. These are very hydrophilic and accessible residues and have the most side chain heteroatoms. The gyration radius and the side chain interaction parameters are very high because of the long side chains of these amino acids.

In terms of involvement in replacement pairs, A is a very changeable residue. The high mutability of A is probably due to its role as a default residue, with positive contributions to alpha-helix propensity. Lack of gamma-carbon also allows substitutions with small steric obstructions. But replacements to the somewhat rigid A induce smaller changes in the fold than substitutions to the very flexible G.

(A,S) also frequently replace each other, mostly in surface and loop areas. Both residues have low free energy of hydration. (T,S) are also repeated substitutions.

In the literature, there are also published some paired indexes of amino acid *changes* [20]. This model is described by a 20x20 matrix *C*, and may also be used to define the chemical difference *d(x,x')* between every pair *x*, *x'* of amino acids in a similar manner as we did above.

Physicochemical Properties

We analysed different sequence populations defined by origin and temperature, and conducted a large scale statistical test to identify systematic change and significant differences between the mesophilic and psychrophilic populations (described in part 2.3 of the paper). 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 significantly low (P<0.01). We observe that there are many interesting differences especially at the surface and in the alpha helices of the molecules.

The Gibbs free energy is a fundamental parameter that provides a measure of thermodynamic stability of the protein molecule. Most studies of the stability of proteins are concentrated on evaluation of the Gibbs free energy of unfolding. We found no significant difference for this parameter between the populations. 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 very significant, as we found it to be in our data analysis, where ΔH goes down and $-T\Delta S$ up (Table 1).

Table 1. List of main differences from mesophile to psychrophile populations. P-values are obtained by the Wilcoxon test of the total data set of 25 alignments from *Vibrio* bacteria. Particular 2D and 3D regions are also marked if they have significant hits (P<0.01): A=Alpha helix, B=Beta sheet, L=Loop, C=Core, T=Twilight zone and S=Surface. Details in the references to amino acid properties can be found in [18, 19]. Other properties gave no significant hits across the data set

Property	Change	Entire seq. P-value	Significant 2D/3D region	Ref. property index
Hydrophobicity	↓	$<10^{-5}$	A,L,C,T	Kyte-Doolittle, 1982
Buriedness	↓	$<10^{-5}$	A,L,C,T	Chothia, 1976
Molecular weight	↑	$<10^{-5}$	A,L,C,T	Fasman, 1976
Volume	↑	$<10^{-5}$	A,L,T. (C↓)	Zamyatin, 1972
Metabolic costs	↓	$<10^{-5}$	A,L,C,T	Akashi, 2002
Alpha-helix, frequency of	↓	$4\ 10^{-3}$	A	Chou-Fasman, 1978
Beta-sheet, frequency of	↓	0.6	B	Chou-Fasman, 1978
Side-chain contrib. to stab.	↓	$6\ 10^{-3}$	B,C	Takano-Yutani, 2001
Average flexibility	↑	$3\ 10^{-5}$	A,B,C	Bhaskaran et al 1988
ΔH (unfolding enthalpy)	↓	$<10^{-5}$	A,B,L,C,T	Oobatake-Ooi, 1993
$-T\Delta S$ (unfolding entropy)	↑	$5\ 10^{-5}$	A,B,C,T	Oobatake-Ooi, 1993

Amino acids have different propensities to form helical structures, and the composition in helical regions may affect both the helix stability (Figure 5) and the overall

stability of the protein. Our calculations of the alpha-helical sequences show a better stabilisation for mesophiles compared with psychrophiles (Table 1).

We observe increasing trends with cold adaptation for molecular weight, volume, and average flexibility. Hydrophobicity, side-chain contribution to stability and metabolic cost are decreasing properties.

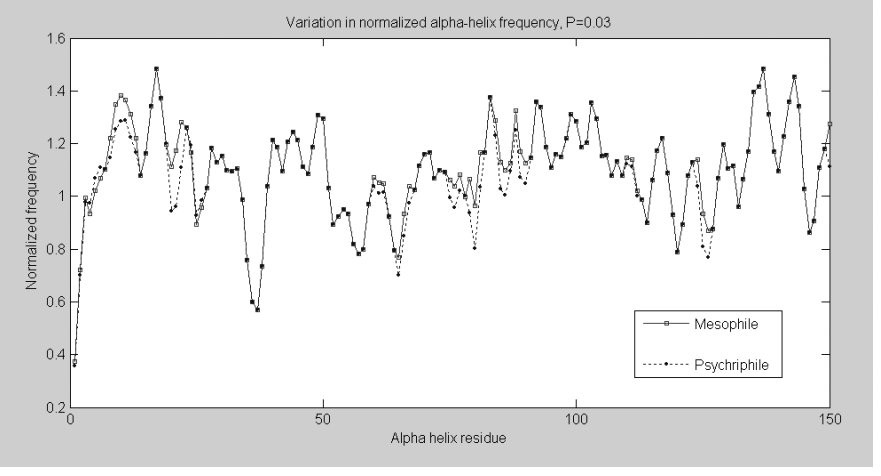

Fig. 5. Comparison of the mean helix formation parameter in the cytoplasmic protein *Isocitrate lyase* of 5 *Vibrio* gamma protobacteria (3 mesophilic and 2 psychrophilic). We used a box-filter of size m x 3 as smoothing technique to recover the underlying structure in the data, where m is the number of sequences in the population. For this particular scale (Chou-Fasman, 1978) the helix-favourable values are at the positive y-axis. In average the mesophilic sequences appear to have more favourable values than the psychrophile counterparts

4 Conclusion

We performed comparative analysis of genetic variability using protein sequences from bacterial populations of *Vibrio* with different temperature preferences. The use of data from the same taxonomic groups reduced problems associated with physiology and phylogenetic noise that have been a problem in other studies.

We have applied and expanded the methods of comparative analysis of proteins. The improved strategy is partly extensions of traditionally used statistics [19], e.g., residue frequencies, residue properties, but applied to *positions* of aligned sequence pairs rather than averaged over unaligned sequences. Statistics also include an amino acid replacement matrix approach to identify residue substitution pairs that differs between populations. The approach of using aligned sequence pairs yield better comparisons, and in this paper an appropriate probabilistic model of context-sensitive and property-dependent analysis of alignments is developed, including efficient algorithms for constructing them. We extracted compositional differences into several distinct physicochemical factors.

In the present *Vibrio* study we found that decreasing hydrophobicity and buriednes are generally (and especially for core residues) the most important properties for adaptation to cold in cytoplasmic proteins. Moreover, unfolding enthalpy and unfolding

entropy are found to be different in a direction that compensates concerning the Gibbs free energy.

Furthermore, decreased stability parameters correlates both in alpha helices and in beta-strands. All these results suggest that the maintenance of proper balance between stability and flexibility is critical for proteins to function at their environmental temperatures.

Some of the features observed may be specific to intracellular proteins or to the *Vibrio* species, and more sequence families should be analyzed to detect both general and special determinants of cold adaptation.

Acknowledgements

We thank Elinor Ytterstad for suggestions regarding the statistical analysis. Some of the data used in this work were sequenced at the University of Tromsø by Erik Hjerde.

Upon publication, our Matlab code can be downloaded from our web-site at: http://www.math.uit.no/bi/deltaprot/

References

1. Russell, N.J.: Toward a molecular understanding of cold activity of enzymes from psychrophiles. Extremophiles 4: 83-90. 2000.
2. Karlin S, Brocchieri L, Trent J, Blaisdell BE, Mrazek J. Heterogeneity of genome and proteome content in bacteria, archaea, and eukaryotes. Theor Popul Biol. 61:367–390. 2002
3. Pe'er I, Felder CE, Man O, Silman I, Sussman JL, Beckmann JS: Proteomic signatures: Amino acid and oligopeptide compositions differentiate among phyla. Proteins-Structure Function and Genetics 54 (1): 20-40. 2004
4. Saunders NFW, Thomas T, Curmi PMG, et al.: Mechanisms of thermal adaptation revealed from the genomes of the Antarctic Archaea Methanogenium frigidum and Methanococcoides burtonii. Genome Res 13 (7): 1580-1588. 2003.
5. Nikolova N, Jaworska J.: Approaches to measure chemical similarity - A review. QSAR & Combinatorial Science (9-10): 1006-1026, 2004.
6. Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physicochemical Property Descriptors. J. Chem. Inf. Comput. Sci. 1996, 36, 118-127.
7. Basak, S. C.; Grunwald, G. D. Molecular Similarity And Estimation of Molecular-Properties. J. Chem. Inf. Comput. Sci. 1995, 35, 366-372.
8. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). A model of evolutionary change in proteins. Atlas of Protein Sequences and Structure, 5 suppl 3, 345–352.
9. Benner, S. A., Cohen, M. A. & Gonnet, G. H. (1994) Amino acid substitution during functionally constrained divergent evolution of protein sequences. Protein Eng., 7 (11), 1323–1332.
10. Jones DT, Taylor WR, Thornton JM (1992): The rapid generation of mutation data matrices from protein sequences. Computer Applications in the Biosciences 8 (3): 275-282.
11. Agresti A. (2001): Exact inference for categorical data: recent advances and continuing controversies. Statistics in Medicine. 20 (17-18): 2709-2722.
12. Agresti, A. (2002). Categorical Data Analysis. 2. ed. John Wiley & Sons.

13. Kang S.H., Kim S.J. (2004). A comparison of the three conditional exact tests in two-way contingency tables using the unconditional exact power. Biometrical Journal 46(3): 320-330.
14. Conover, W.J. (1999). Practical Nonparametric Statistics, 3.ed. John Wiley & Sons.
15. Oppenheim, A. V. & Schafer, R.W. (1999). Discrete-Time Signal Processing, 2.ed. Prentice-Hall.
16. Benjamini, Y & Hochberg, Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B, 57:289-300.
17. Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Ann Stat 29 (4): 1165-1188.Rabus R, Ruepp A, Frickey T, et al.:
18. Kawashima, S., Ogata, H., and Kanehisa, M.; AAindex: amino acid index database. Nucleic Acids Res. 27, 368-369 (1999).
19. Gromiha, M.M., Oobatake, M. & Sarai, A. (1999). Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins. Biophysical Chemsitry 82, 51-67.
20. Hua Tang, Gerald J. Wyckoff, Jian Lu, and Chung-I Wu A Universal Evolutionary Index for Amino Acid Changes. Mol Biol Evol 2004 21: 1548-1556