Estimation of Evolutionary Average Hydrophobicity Profile from a Family of Protein Sequences

Said H. Ahmed and Tor Flå

Dept of Mathematics and Statistics, University of Tromsø, 9037 Tromsø- Norway
{said.hassan.ahmed,tor.fla}@matnat.uit.no
http://www.uit.no

Abstract. Hydrophobicity has long been considered as one of the primary driving forces in the folding of proteins. We discuss here the evolutionary average of the hydrophobicity profile in an aligned family of proteins and found a patchy mean hydrophobicity profile. This is in contrast to Bastolla et al (2005b) results for the large superfamily of globular proteins. The idea is to use singular value decomposition and cavity filtering in order to remove the eigensequences burried in the evolutionary noise

1 Introduction

It is well known that hydrophobicity is a major determinant of protein stability and evolution. With respect to sequence-structure correlation, the evolutionary average of hydrophobicity profiles of sequences with the same fold correlates with principal eigenvector of fold's contact matrix (PE) much strongly than the hydrophobicity profile (HP) of its single sequence [1]. In the Structurally Constrained Neutral (SCN) model of protein evolution [2,3,4] the correlation is perfect (almost one), and yields

$$h^s_{evol} \equiv \sum_{a=1}^{20} \pi^s_a h_a = \sqrt{\frac{\langle h^2_{evol}\rangle - \langle h_{evol}\rangle^2}{(\langle c^2\rangle - \langle c\rangle^2)}}\,(c_s - \langle c\rangle) + \langle h_{evol}\rangle \ , \tag{1}$$

where h_{evol} is the position specific evolutionary average of the HP, π^s_a is the position specific amino acid distribution at site s resulting from the evolutionary process (a indicates one of the 20 amino acid types) and c_s is the PE component of the contact matrix of the family. Assuming this equation is the only relevant condition, the amino acid distribution at site s is predicted to be the distribution of maximal entropy [11] with mean given above , i.e.

$$\pi^s_a = \frac{\exp[-\beta_s h_a]}{\sum_{a'=1}^{20} \exp[-\beta_s h_{a'}]} \ . \tag{2}$$

The site specific Boltzmann parameters ('inverse temperature') β_s determine the width of the amino acid distribution. The width parameter varies from site

J.C. Rajapakse, B. Schmidt, and G. Volkert (Eds.): PRIB 2007, LNBI 4774, pp. 158–165, 2007.

to site and measures the tolerance of site s to accept mutations over very long evolutionary time. In principle it can catch external parameter dependence of the distribution due to say temperature, regulatory effects, e.t.c.

In this paper we estimate the evolutionary average hydrophobicity sequence from a set of aligned protein sequences from elastase family. The idea is to use eigensequences related to the inter-species hydrophobicity sequence correlation matrix to remove the evolutionary noise from the sequences and hence avoid inspection of large database to compute the mean hydrophobicity. For example, Bastolla et al. (2005a) used thousands of globular sequences from the PFAM, the FSSP, and the SCN databases in order to compute the evolutionary mean hydrophobicity profile. Since the aligned sequences are represented through hydrophobic profiles by quantifying each of the amino acids in the sequences using for example Kyte and Doolittle hydropathy scale it can be viewed as multi-dimensional heterogenous hydrophobicity sequences. We then use Singular Value Decomposition (SVD) and cavity filtering in order to decorrelate and remove the eigensequences burried in evolutionary noise. The average hydrophobicity profile is then computed from the first few useful eigensequences corresponding to the largest eigenvalues of the cross species hydrophobicity covariance matrix.

2 Dataset and Methods

2.1 Dataset

The dataset consists of $L = 32$ aligned sequences of length $N = 247$ (including gaps) from elastase family. The sequences were located from a search in the NCBI and SWISSPROT protein data banks. Elastase is a member of the large family of serine proteinases which includes trypsin and chemotrypsin, and is synthesized initially in the pancreas as an inactive precursor. The 3D structure of these molecules has also abeen modeled at the department of chemistry, university of Tromsø. The dataset can be obtained on request from us.

We represented the sequences through hydrophobic profiles by quantifying each of the amino acids in the sequences using Kyte and Doolittle hydropathy scale [7]. That is the hydrophobicity of residue a at position s in a sequence is given by

$$H_{a(s)} = \mathbf{Y}_{a(s)}^T \mathbf{f} \tag{3}$$

where $\mathbf{Y}_{a(s)} = (0, 0, \ldots, 1_{a(s)}, 0, \ldots, 0) \in \mathbb{R}^{21}$ is a count vector for residue $a^1 = \{1, 2, \ldots, 21\}$ at site s and $\mathbf{f}$ is the hydrophobic index in Kyte and Doolittle. For all the consecutive amino acids in sequence l we have

$$H_l = \mathbf{Y}_l^T \mathbf{f} \tag{4}$$

where $\mathbf{Y}_l = \{\mathbf{Y}_{l,a(s)}\}_{s=1}^N \in \mathbb{R}^{21 \times N}$ is a dummy matrix that consists of N unit count vectors. Hence $\mathbf{H}$ is $L \times N$ elastase sequences represented through hydrophobic profiles. Plot of hydrophobicity level of ela-pig (PDB:1qnj) is shown in Figure 1.

[1] Gaps were treated as if they were a 21st amino acid type.

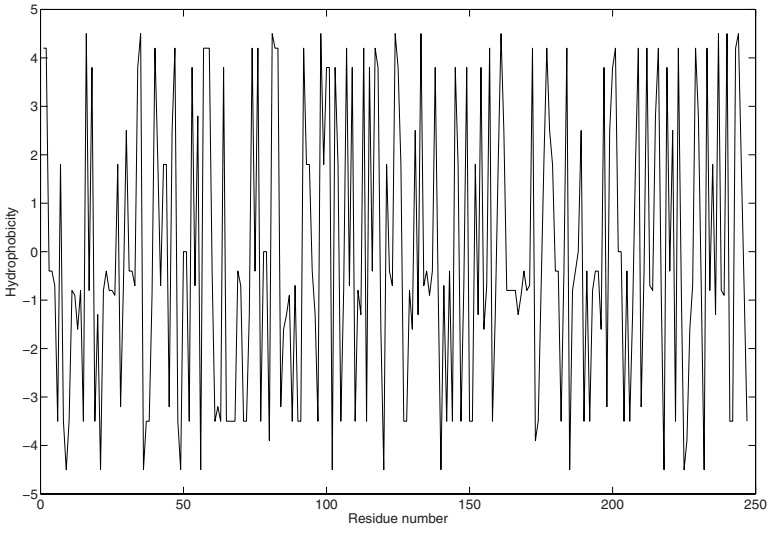

Fig. 1. Hydrophobicity profile of one of the elastase sequences,1QNJ, generated by quantifying each of the amino acids in the sequence using Kyte and Doolittle hydropathy scale

2.2 Estimating Average Hydrophobicity Profile (HP)

The problem of computing the average HPs from $\mathbf{H}$ can be considered as extracting mean hdrophbocity sequence from a noisy one[2]. We assume two types of noise contributions in our data - one along the sequence chain (due to for example the stochasticity of the folded protein chain) and the other across the sequences (due to evolutionary noise). In order to decorrelate and remove the eigensequences burried in evolutionary noise we eigen decompose (SVD - Singular Value Decomposition) the estimate of the noise covariance matrix $\hat{\mathbf{\Sigma}}$,

$$\hat{\mathbf{\Sigma}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathbf{T}} \tag{5}$$

where $\mathbf{U} \in \mathbf{R}^{L\times L}$ is an orthogonal matrix (i.e.,$\mathbf{U}^{\mathbf{T}}\mathbf{U} = \mathbf{I}$), the columns of $\mathbf{U}$ form an orthonormal basis for the HPs of the sequences and $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_L)$ is a diagonal matrix with entries λ_l, eigenvalues in decreasing order. The noise sequences are approximated by subtracting a smoothed (denoised) mean HP from each of the observed HPs. We choose a deterministic gaussian 'cavity filtering' procedure [10] due to local amino acid interactions along the protein sequence. It has also the non-enhancement property of local extrema: values of local maxima cannot increase and respective values of local minima cannot decrease [8]. The

[2] We are presently developing a Boltzmann lattice approximation for discrete evolutionary sequence noise and protein observables in an aligned phylogenetic protein family.

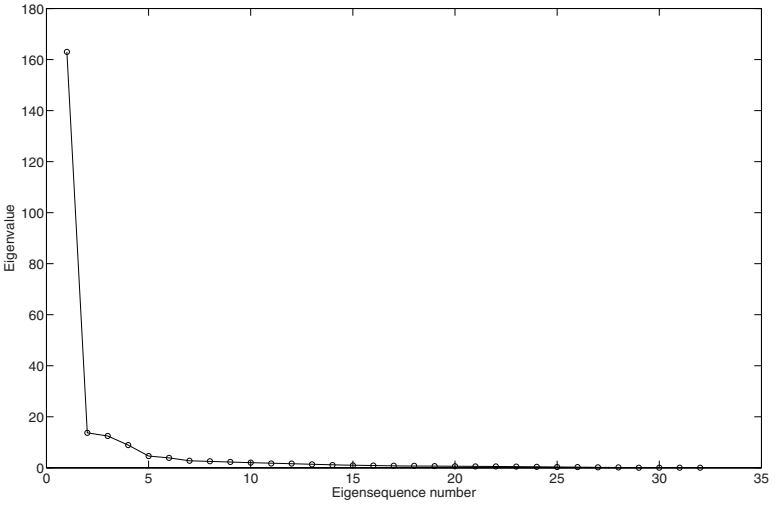

Fig. 2. Scree plot: A plot of eigenvalues λ_l, in decreasing order. The plot is used to decide the number of eigensequences that are useful (eigensequences to the left of the elbow or bend).

hydrophobicity profiles of sequences, $\mathbf{H}$ are then projected into new coordinates to obtain the eigensequences

$$\mathbf{Q} = \mathbf{H^T U} \ . \tag{6}$$

The eigensequences due to evolutionary noise are then filtered out by using the first $K = 3$ eigensequences. K is determined by the point at which the remaining eigenvalues are relatively small and all about the same size. One way to determine K, the number of eigensequences $\mathbf{Q} = [\mathbf{q_1 q_2} \ldots \mathbf{q_K}]$ to retain is by use of a scree plot [6], a plot of λ_l (the eigenvalues in deccreasing order) versus l. A scree plot for the HPs of elastases, $\mathbf{H}$ is shown in Figure 2. To determine K, we look for an 'elbow' (bend) in the scree plot. The eigensequences whose eigenvalues plot to the right of such 'elbow' are ignored since they are defined here to be due to evolutionary noise. Thus the information in the scree plot indicates that we extract the first three eigensequences.

The denoised eigensequences $\hat{\mathbf{Q}}$ are inverse projected to obtain a denoised version $\hat{\mathbf{H}}$ of $\mathbf{H}$:

$$\hat{\mathbf{H}}^T = \hat{\mathbf{Q}}^T \mathbf{U}^T \ . \tag{7}$$

The site specific average hydrophobicity profile of the aligned elastases is calculated by taking the mean of the denoised HPs of the sequences:

$$\overline{h}_s = \frac{1}{L} \sum_{l=1}^{L} \hat{H}_{l,s} = (\overline{h}_1, \overline{h}_2, \ldots, \overline{h}_N) \ . \tag{8}$$

Finally we perform cavity field on the average hydrophobicity profile. The cavity fields is defined as [10]

$$\overline{h}_s = \sum_{t \neq s} J_{st} \overline{h}_t \tag{9}$$

where the couplings J_{st} are taken to be translational invariant gaussian. We choose a deterministic cavity field since our J_{st} parameters are assumed to have small variance compared to their mean. The cavity field describes the local internal filed which the amino acid 'sees'.

The algorithm to estimate the site specific average hydrophobicity profile can then be divided into seven main steps:

1. Estimate the noise hydrophobicity sequences by subtracting a cavity filtered cross species mean HP from all the HPs of the sequences.
2. Compute the estimated noise covariance matrix $\hat{\mathbf{\Sigma}}$.
3. Diagonalize $\hat{\mathbf{\Sigma}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathbf{T}}$, where $\mathbf{U} \in \mathbf{R}^{L \times L}$ is an orthogonal matrix (singular vectors), $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_L)$ are the eigenvalues. Decorrelate the HPs of the sequences by projecting them into new coordinates to obtain the eigensequences, i.e., $\mathbf{H}^T\mathbf{U}$

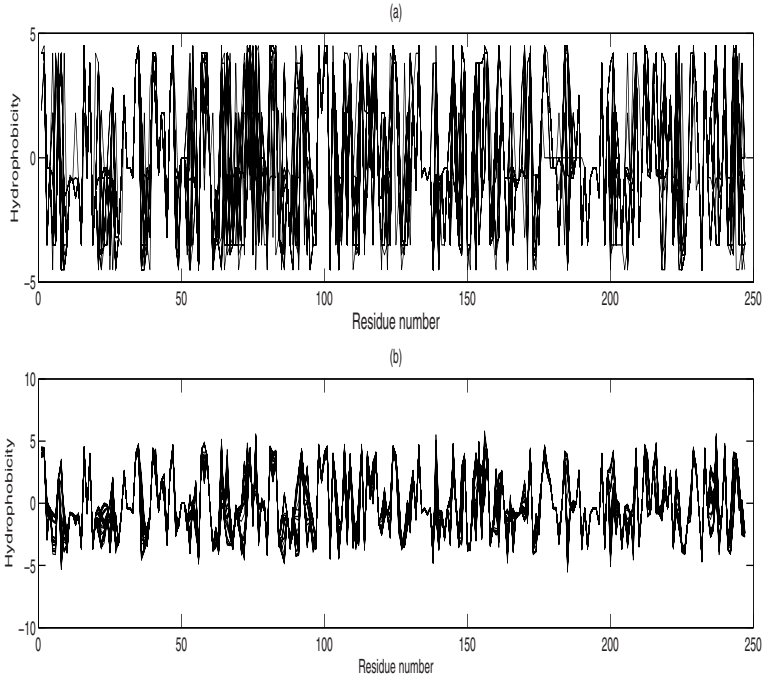

Fig. 3. (a) Hydrophobicity profiles of all the aligned elastase sequences. The hydrophobicity profiles were generated by assigning a hydrophobicity value to each of the amino acids in each sequence using Kyte and Doolittle hydropathy scale. (b) SVD denoised version of HPs of the sequences. The HPs were reconstructed using only the first three useful eigensequences that account 82.4% of the total variance.

4. Remove the eigensequences due to evolutionary noise by choosing the first K useful eigensequences (use for example a scree plot to decide the number of eigensequences to retain).
5. Reconstruct the hydrophobicity sequences, $\hat{\mathbf{H}}$ from the denoised eigensequences (multiply by $\mathbf{U}^T$).
6. Calculate the site specific average hydrophobicity profile from the denoised HPs of the sequences using (8).
7. Perform cavity filtering on the average hydrophobicity profile using (9).

3 Results and Discussion

We demonstrated our method using the aligned protein sequences from elastase family represented through their HPs (see Materials and Methods). Figure 3 shows plot of HPs of all the aligned elastase sequences and their SVD denoised version. From the figure we see a lot of variations (evolutionary noise) in the original sequences while in the second plot much of the evolutionary noise is removed. Only the first three eigensequences that account 82.4% of the total variace were used in the reconstruction. The site specific average hydrophobicity profile was then estimated from the reconstructed denoised eigensequences (first

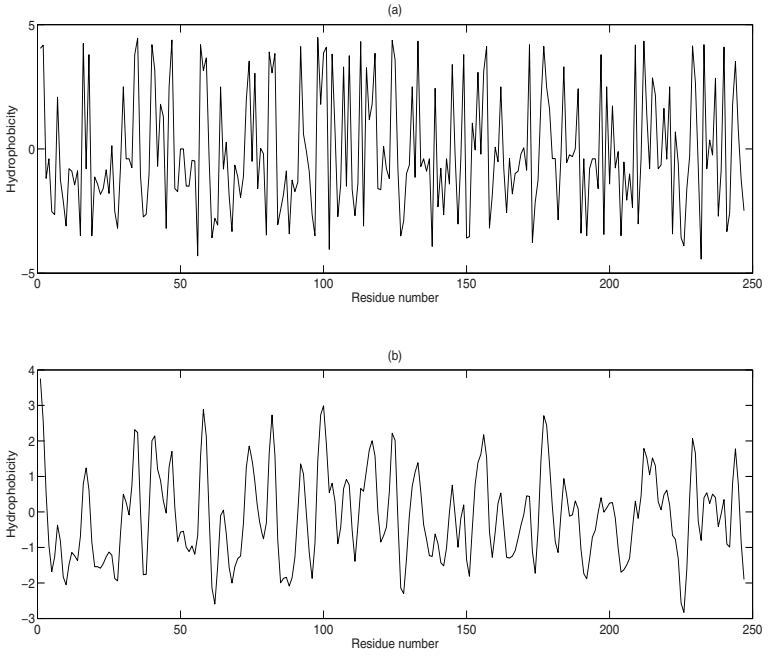

Fig. 4. (a) Site specific average HP estimated from the reconstructed denoised eigensequences. (b) Result of cavity 'filtering', short range interaction - three amino acid local interactions along the mean HP sequence.

three eigensequences) using equation 3. Finally cavity filtering was applied on the average HP. Short range amino acid interactions (three local amino acid interactions) along the sequence profile was used. Figure 4 shows plot of average HP and its cavity filtered version. From Figure 4(a) we see that the estimated average hydrophobicity profile is still patchy. This might be due to variation along the sequences. We therefore used cavity filtering to smooth this variation (see Figure 4(b)).

We have analyzed (not yet published) the correlation between this estimated average hydrophobicity and average surface exposure of our proteins and found that the correlation is stronger than when the average hydrophobicity is computed by just averaging the HPs of the sequences or estimated using wavelet based smoothing methods.

4 Conclusions and Further Work

In this paper, we developed a method to estimate average hydrophobicity sequence from a set of aligned sequences from one protein family. We tested this method on aligned sequences from elastase family. The method has removed the evolutionary noise effectively. We are still working further to test how effective the method is by analyzing the correlation between mean hydrophobicity and surface-exposure or principal eigenvector of fold's contact matrix for varioius families. This mean profile can be improved if we for example use more physico-chemical properties like charge, electrostatic interactions, e.t.c.

So far we have computed an estimate of a mean HP profile but our future aim is to estimate the site specific Boltzmann paramters, β_s from this mean HP. This width parameter in principle can catch external parameter dependence of the distribution due to for example temperature. So we think that this parameters can be used to classify proteins within a family, for example identify significant diferences between mesophilic and psychrophilic populations [13].

References

1. Bastolla, U., Porto, M., Roman, H.E., Vendruscolo, M.: The principal eigenvector of contact matrices and hydrophobicity profiles in proteins. Proteins 58, 22–30 (2005a)
2. Bastolla, U., Porto, M., Roman, H.E., Vendruscolo, M.: Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution. J. Mol. Evol. 56, 243–254 (2003)
3. Bastolla, U., Porto, M., Roman, H.E., Vendruscolo, M.: Lack of self-averaging in neutral evolution of proteins. Phys. Rev. Lett. 89, 208101/1–208101/4 (2002)
4. Bastolla, U., Porto, M., Roman, H.E., Vendruscolo, M.: Statistical properties of neutral evolution. J. Mol. Evol. 57, S103–S119 (2003)
5. Branden, C., Tooze, J.: Introduction to Protein Structure, 2nd edn. Garland publishing, New York (1999)
6. Johnstone, R.A., Wichern, D.W.: Applied Multivariate Statistical Analysis, 5th edn. Prentice Hall, Englewood Cliffs (2002)

7. Kyte, J., Doolittle, R.F.: A Simple Method for Displaying the Hydropathic character of a Protein. J. Mol. Biol. 157, 105–132 (1982)
8. Lindeberg, T.: Scale-Space Theory in Computer Vision. The Kluwer International Series in Engineering and Computer Science. Kluwer Academic Publishers, Dordrecht (1994)
9. Miyazawa, S., Jernigan, R.L.: Self-consistent estimation of inter-residue protein contact energies based on an equillibrium mixture approximation of residues. Proteins: Structure and Molecular Principles 34, 49–68 (1999)
10. Opper, M., Winther, O.: From Naive Mean Field Theory to the TAP Equations. The MIT Press, Cambridge, Massachusetts London, England (2002)
11. Porto, M., Roman, H.E., Vendruscolo, M., Bastolla, U.: Prediction of site-specific amino acid distributions and limits of divergent evolutionary changes in protein sequences. Mol. Biol. Evol. 22, 630–638 (2005)
12. Fornasari, M.S., Parisi, G., Echave, J.: Site-specific amino acid replacement matrices from structurally constrained protein evolutioin. Mol. Biol. 19, 352–356 (2002)
13. Thorvaldsen, S., Flå, T., Willassen, N.P.: Extracting molecular diversity between populations through sequence alignments. In: Oliveira, J.L., Maojo, V., Martín-Sánchez, F., Pereira, A.S. (eds.) ISBMDA 2005. LNCS (LNBI), vol. 3745, pp. 317–328. Springer, Heidelberg (2005)
14. Wall, M.E., Rechtsteiner, A., Rocha, L.M.: Singular Value Decomposition and Principal Component Analysis. In: Berrar, D.P., Dubitzky, W., Granzow, M. (eds.) A Practical Approach to Microarray Data Analysis, pp. 91–109. Kluwer, Norwell, MA (2003)
15. Tang, C.: Simple Models of the Protein Folding problem. Physica A 31, 288 (2000)
16. Moelbert, S., Emberly, E., Tang, C.: Correlation between sequence hydrophobicity and surface-exposure pattern of database proteins. Protein Science 13, 752–762 (2004)