Evolutionary Parameters in Sequence Families
Cold Adaptation of Enzymes

Said Hassan Ahmed and Tor Flå

Dept of Mathematics and Statistics, University of Tromsø, 9037 Tromsø, Norway

Abstract. In an attempt to incorporate environmental effects like cold-adaptation into models of sequence evolution on a phylogenetic tree, we present a viable way of representing descriptive statistics of sequence observables under reversible Markov models of sequence evolution. Local variation in amino acid distribution along and across the sequence family can be connected to enzymatic adaptation to different temperatures. Here, we estimate a few amino acid properties and how the variations of these properties both with respect excess mean values (EMVs) and covariance classify the protein family into clusters. Application of a multiscale and multivariate method to an aligned family of distinct trypsin and elastase sequences shows drift of centroid mean sequences of cold adapted enzymes compared to their warm-active counterparts.

1 Introduction

Phylogenetic tree-building methods presume particular evolutionary models [2]. Current evolutionary models of amino acid sequence evolution generally depend on mathematical models based on empirical observations using either comparisons of the observed amino acid sequences or their physical-chemical properties [1, 2]. These models estimate evolutionary distances in terms of expected number of substitutions per site by assuming evolution with independent sites and that the sequences in each site are assumed to evolve according to a single stochastic process, and that this process is fixed across all sites. For instance, in Markov models of amino acid replacement, the Markov process is assumed to be stationary, homogeneous and reversible so that the amino acid distribution and the rate of replacement are assumed to be fixed in time and positions, and that the forward and reverse substitution rates are assumed to be the same [15, 16].

We will be interested in the possibility to parameterize environmental effects like cold adaptation into Markov transition and the corresponding rate matrices. In particular, we are interested in amino acid distribution profile in an aligned family with cold-adapted representatives. Cold-adapted enzymes are characterized by clusters of glycine residues, a reduced number of proline residues in loop regions, a general reduction in charged residues on the surface and exposure of hydrophobic residues to solvent [19, 20]. All these features are thought to give rise to the increased structural flexibility observed in some of the regions of the enzyme. Flexibility seems to be a strategy for cold-adapted enzyme to maintain high catalytic activity at low temeperatures [18, 19]. Often a few conserved

V. Kadirkamanathan et al. (Eds.): PRIB 2009, LNBI 5780, pp. 1–12, 2009.

residues within each temperature class sequence positions are detemining factors of their strategies to adapt to cold/warm temperature.

Here, we study how approximations on Markov models for standard phylogeny will give us an opportunity to obtain first hand insight into the statistics of observables related to index and counting variables. Based on parameterized sequence features, we will carry out a multiscale and multivariate data analysis on an aligned family of distinct trypsin and elastase sequences. The basis of this multivariate analysis is that covariation of residue sites in evolution is related mainly to structural or functional site fitness as parameterized by models of mean amino acid distributions and certain amino acid properties. These correlated residues based on amino acid properties (property sequences) show deviation from a common position dependent mean value. Such mean deviations, which we refer to as *excess mean values (EMVs)*, are due to species dependent variations in local and global fitness without affecting the overall 3D fold and fitness with respect to protein (enzyme) function. Our goal is to extract these EMVs from evolutionary noise both along and across the sequence family.

On application, the method revealed drift of centroids due to features of cold adaptation. Such deviations could be used as measures of evolutionary adapted fitness landscapes corresponding both to the folding rate as parameterized by the global relative energy gap/energy standard deviation ratio (funneling picture) and the local fitness adaptations at the active site measuring the binding activity effects.

2 Parameterization of Sequence Features

2.1 Statistics Based on Amino Acid Unit Count Vectors

We assume an aligned family of L homologous protein sequences of length N. Let $\alpha(l,s)$ be the residue at position s and species l, where $l \in \{1,\ldots,L\}$, $s \in \{1,\ldots,N\}$. Then we describe $\alpha(l,s)$ numerically in the vector space of amino acid *unit counts*, denoted as $\mathbf{Y}_{l,s} = \mathbf{Y}_{\alpha(l,s)}$, where $\mathbf{Y}_{\alpha(l,s)} = (\delta_{\alpha,\,\alpha(l,s)}) \in \mathbb{R}^{20}$. Here, $\alpha \in \mathcal{A}$, is one of the 20 amino acid categories, $\delta_{\alpha,\,\alpha(l,s)}$ is 1 if one of the amino acids α equals the amino acid at (l,s), $\alpha(l,s)$, or 0 otherwise (the Kronecker delta function). With this representation, we have that the average over the observed present time leaf distribution of the protein amino acids at (l,s) is given by the amino acid distribution

$$< \mathbf{Y}_{l,s} >= \mathbf{p}^{l,s} = (p_{\alpha}^{l,s}) \ , \tag{1}$$

where $\langle . \rangle$ is the expectation operator (with respect to phylogenetic distributions). For completness, we have taken into account that $\mathbf{p}^{l,s}$ will vary both on subset of species and positions due to different species clusters and functional (or structural) constraints, in our case, clusters and residue determinants of cold-adapted enzymes.

We are interested in the amino acid distribution given in (1), in terms of two sequence ensembles $(\mathbf{Y}_{l,s}, \mathbf{Y}_{l',s'})$, namely, the first-and second-order marginals, $\mathbf{p}^{(1)l,s}$ (we will supress the superscript (1)) and $\mathbf{P}^{(2)l,l';s,s'}$, respectively. Since the

first-order marginal, $\mathbf{p}^{l,s} = \prod_{l=1}^{L} p_{\alpha}^{l,s}$, is a product of single-site multinomial type probability with no information about the sequence pair probabilites necessary to describe standard phylogenetic tree parameters, we consider correlation of unit count vectors. For simplicity, we look at the two-point covariation, which is given by

$$\begin{aligned}\left\langle (Y_{\alpha(l,s)} - p_{\alpha}^{l,s})(Y_{\beta(l',s')} - p_{\beta}^{l',s'}) \right\rangle =& P_{\alpha,\beta}^{(2)l,l';s,s'} - p_{\alpha}^{l,s} p_{\beta}^{l',s'} \ , \\ =& \rho_{\alpha\beta}^{(2)l,l';s,s'} \, p_{\alpha}^{l,s} p_{\beta}^{l',s'} \ ,\end{aligned} \tag{2}$$

where $\rho_{\alpha\beta}^{(2)l,l',s}$ is the pair (α at (l,s) and β at (l',s')) dependent correction. Based on a reversible Markov model of amino acid replacement, with the instantaneous rate of replacement of amino acid α by amino acid β defined by the rate matrix $\mathbf{Q} = (Q_{\alpha\beta})$, as described in Sect. 1, the two point correlation could, for relatively short evolutionary times compared to the mutation rate $Q_{\alpha\alpha}$ and within the same cluster $c \in \{1, 2, \ldots, K\}$, such that the mean amino acid distribution is fixed, $\mathbf{p}^{l,s} = \mathbf{p}^{c,s}$, $c(l) = c$, be modelled as[1]

$$\rho_{\alpha\beta}^{(2)l,l';s} \, p_{\alpha}^{s} p_{\beta}^{s} \cong p_{\alpha}^{s}\delta_{\alpha,\beta} - p_{\alpha}^{s} p_{\beta}^{s} + (T_l + T_{l'}) \, \Lambda_{\alpha\beta}^{s} \, p_{\alpha}^{s} p_{\beta}^{s} \ , \tag{3}$$

Here T_l is the edge length corresponding to species l, and $\mathbf{\Lambda}_{\alpha\beta}$ is constrained so that the row sums are all zero: $Q_{\alpha\alpha} = \Lambda_{\alpha\alpha}^{s} p_{\alpha}^{s} = -\sum_{\beta \neq \alpha} = \Lambda_{\beta\alpha}^{s}$, $\alpha \neq \beta$ (symmetry), which ensures reversibility of the Markov process. Notice that for long evolutionary times between leaf nodes l and l', the process will effectively be independent and $\rho_{\alpha,\beta}^{(2)l,l';s} \cong 0$. This effect will tend to divide our protein sequences into clusters of close neighbors in evolutionary time $T_l + T_{l'}$ where the above model will be used within each cluster. We can extend the above model of the two-point correlation for all sites $s = 1, 2, \ldots, N$ and find covariance of two unit count vectors between two protein sequences $\langle (\mathbf{Y}_{l,s} - \mathbf{p}^{l,s})(\mathbf{Y}_{l,s} - \mathbf{p}^{l',s'}) \rangle$, for short evolutionary times as

$$\left(\rho_{\alpha\beta}^{(2)l,l';s,s'} \, p_{\alpha}^{s} p_{\beta}^{s'}\right) \cong p_{\alpha}^{s}\delta_{\alpha,\beta}\delta_{s,s'} - p_{\alpha}^{s} p_{\beta}^{s'} + (T_l + T_{l'}) \Lambda_{\alpha\beta}^{s} p_{\alpha}^{s} J_{\alpha\beta}^{ss'} \, p_{\beta}^{s'}, \, \forall \alpha\beta \ , \tag{4}$$

where $J_{\alpha\beta}^{s,s'}$ is the pair dependent correlation that ensures symmetric reversible substitution matrix $\forall(\alpha,\beta)$ pairs, $\Lambda_{\alpha\alpha}^{s} p_{\alpha}^{s} - \sum_{\beta \neq \alpha} \Lambda_{alpha\beta}^{s} \sum s' \neq s J_{\alpha\beta}^{s,s'} \, p_{\beta}^{s'}$.

We are interested in physical-chemical observables and how they are reflected in the amino acid distribution of the family. As they are linearly dependent on the unit count vectors, parameterized features (covariance) based on physical-chemical observables can be derived by unit count vector projections as described below. Thus, we consider the unit count vectors as our basic observables.

2.2 Statistics Based on Physico-Chemical Properties

Given a vector of amino acid properties $\mathbf{C} \in \mathbb{R}^{|\mathcal{A}|}$, we find that for a family of sequences with unit count vectors $\mathbf{Y}_{l,s}$, there is a family of *property sequences* given as

[1] For simplicity, we assume one cluster and skip the cluster index c in this section.

$$\mathbf{C}_l^{(N)} = (C_{i(l,s)})_{s\in\{1,\ldots,N\}} = \left(\mathbf{C}^T\mathbf{Y}_{l,s}\right)_{s\in\{1,\ldots,N\}}, \quad l = 1,\ldots,L \ , \tag{5}$$

where the superscript (N) indicates the length of the sequence. Since (5) a linear mapping of the unit count vectors, the *mean of the property sequences* can be expressed as

$$\bar{\mathbf{C}}_l^{(N)} = (\bar{C}_{c(l),s}) = \left(\mathbf{C}^T\mathbf{p}^{l,s}\right)_{s\in\{1,\ldots,N\}} \ , \tag{6}$$

where $\bar{C}_{c(l)s}$ is the mean property in cluster $c(l) = c$ and we assume a fixed amino acid distribution for each cluster $\mathbf{p}^{l,s} = \mathbf{p}^{c,s}$. Here, $\bar{\mathbf{C}}_l^{(N)}$ could be the mean property of the whole family or as above of a cluster $c(l) = c$ within the protein family. Let $\widetilde{\mathbf{C}}_l^{(N)} = \mathbf{C} - \bar{\mathbf{C}}_l^{(N)}$ be the mean subtracted property sequences (this subtraction to be explained below). Then a similar model of covariation as in (4) based on property sequences can be derived by projecting mean subtracted property vector on the parameterized covariance of the unit count vectors in fixed cluster c:

$$\begin{aligned}
\mathbf{\Sigma}_{ll'}^c = \frac{1}{N-1}\langle\widetilde{\mathbf{C}}_l^{(N)}, \widetilde{\mathbf{C}}_{l'}^{(N)}\rangle_{(c(l),c(l'))=c} &\approxeq \frac{1}{N-1}\sum_s(\tau_c^s + (T_l + T_{l'})\bar{\mathcal{S}}_c^s \ , \qquad (7)\\
\tau_c^s &= \sum_\alpha (C_\alpha - \bar{C}_c^s)^2 p_\alpha^{c,s} = \sum_\alpha C_\alpha^2 p_\alpha^{c,s} - (\bar{C}_c^s)^2,\\
\bar{\mathcal{S}}_c^s &= \sum_{\alpha,\beta}(C_\alpha - \bar{C}_c)p_\alpha^{c,s}\Lambda_{\alpha\beta}^{c,s}p_\beta^{c,s}(C_\beta - \bar{C}_c) \ ,
\end{aligned}$$

where $\bar{C}_c^s = \sum_{\{\alpha\}}\bar{C}_{c(l),\alpha}^s = \sum\{\alpha\}C_\alpha p_\alpha^{c(l),s}$, is the average amino acid property[2] for proteins in species l in cluster $c = c(1) \in \{1,2,\ldots,K\}$ since we assume that proteins come in say K groups with more or less the same properties within a group and are independent between groups.[3]. The logic of the mean subtraction prior to parameterization, which is also the basis for our data analysis, is the relation for the substitution matrix, $\sum_\alpha \Lambda_{\alpha\beta}^{c,s}p_\beta^{c,s} = 0$, which is valid for equilibrium amino acid distributions and symmetric substitution matrix. This would lead to that correlation within a subfamily will be described by a simple variance. If some cold-adapted representatives are present within the family or subfamily, it will imply that the mean amino acid distribution will change. Consequently, both the center of the cluster as described by the mean properties and covariance matrix will move relative to that of standard mesophilic temperature class.

Excess mean values (EMVs). When there are more than one cluster in the sequence family or subfamily, each cluster might have a different mean amino acid distribution $\mathbf{p}^{c,s}$. Often the size of the data contained in each cluster is not sufficient enough to estimate the sequence position dependent mean necessary to observe cluster deviations. In this case, one would have to be satisfied with

[2] τ_c^s is the local property variance obtained for short evolutionary times compared to the local mutation rate, i.e. $T\mu^s \ll 1$, $\mu^s = \sum_\alpha Q_{\alpha\alpha}^s$.

[3] The formula for covariance clusters is simply obtained by summing the covriance of each cluster with respect to the cluster prior probability $p(c)$, $c = 1,2,\ldots,K$.

a common mean $\mathbf{p}^s$. This analysis lead to artificial, extrinsic correlations which we attribute to the EMVs $\delta\mathbf{p}_e^{c,s} = \mathbf{p}^s$ and $\delta C_e^{c,s} = C^{c,s} - \bar{C}^s$. Additional average might come due to intrinsic correlations, interactions /and dependencies along the protein sequence as discussed above and this would lead to cavity filter (the coefficients $J^{s,s'}$ in 4)) averaged fields $\mathbf{p}_{cav}^s$. Both extrinsic and intrinsic excess mean values lead to extra correlations to linear and quadratic order (or higher) in our measurements and theoretical model of covariance might indeed give a substantial contribution to the clustering we study below. The theory and detailed discussion of these effects is beyond the scope of this paper. Still we will refer to EMVs in our discussion of the results of the data analysis below.

3 Multivariate Data Analysis

The goal of the data analysis to find excess mean values which maximize covariations between clusters of cold-and warm adapted enzymes in an aliged family of homologs of differently adapted enzymes.

3.1 Data Sets

As a benchmark data, we used trypsin (a well studies enzyme w.r.t. cold adaptation) sequences studied by Nils Peder Willassen and coworkers [20]. The sequences are divided into 3 groups: trypsins from the higher vertebrate, the cold-adapted fish and the other fish. Additionally, we used 29 elastase sequences (though could not show all results due to small space), composed of the three types of elastases, namely, elastase type-I (with 3 cold-adapted representatives), II (with 5 cold-adapted elastases) and III. The elastases were collected from homologous search by blast at the data bases available at NCBI (http://www.ncbi.nml.nih.gov/blast) and SiB (http://au-expasy-org/tools/blast). Multiple alignment was performed using Geneious, version 3.7.1 (Biomatters Ltd.). In this paper, the property sequences are based on hydrophobicity (Kyte-Doolittle, 1983) and polarity (Zimmermann, 1968). All analyses subsequently described were implemented in MATLAB 7.0.

3.2 Discrete Wavelet Transform (DWT)

We begin with a brief review of orthogonal forward discrete wavelet transform (DWT). An important concept in wavelets is the multiresolution analysis (MA) [8], which decomposes the property sequences as coefficients A_n^{j+1} at reference level 0 (unit scale) in orthonormal basis functions $\{\phi_k^{j+1}(t)\}$ in the space V_{j+1} into approximation and detail coefficients, $A_k^{(j)}$ and $D_k^{(j)}$, at level 1 in orthonormal basis functions $\{\phi_k^j(t), w_k^j(t)\}$ in the nested spaces V_j and W_j ($V_{j+1} = V_j \oplus W_j$), respectively:

$$f^{j+1}(t) = \sum_n A_n^{(j+1)} \psi_n^{j+1}(t) = \sum_k A_k^{(j)} \psi_k^j(t) + \sum_k D_k^{(j)} \psi_k^j(t) \ , \qquad (8)$$

where, $q_k^j(t) = 2^{j/2}q(2^jt-k)$. That is, the scaling and the wavelet basis functions, $\psi_k^j(t)$ and $w_k^j(t)$, are dyadic dilations ($(1/2)^j$) and integer translations ($(1/2)^j k$) of the father and mother functions $\psi(t)$ and $w(t)$, which connect DWT to subband filtering (see the footnote below). As the basis functions are orthonormal at each level j, the corresponding coefficients can be obtained by taking the inner products $\langle f^{j+1}(t), \psi_k^j(t)\rangle$ and $\langle f^{j+1}(t), w_k^j(t)\rangle$ to yield[4]

$$A_k^{(j)} = \sum_n A_n^{(0)} g(n-2k), \qquad D_k^{(j)} = \sum_n A_n^{(0)} h(n-2k) \ . \tag{9}$$

This is filtering and downsampling operations in the analysis filter bank [10]. It is convolution with time reversed lowpass ($g(-n)$) and highpass filters ($h(-n)$).

We performed a 4-level wavelet decomposition of each $\mathbf{C}_l^{(N)}$ using Symlet (sym4), a near symmetric, orthogonal wavelet with 4 vanishing moments. In DWT [8] starting with the approximation coefficient at reference level $j = 0$ (unit scale), (8) was recursively applied on $A_k^{(j)}$ at coaser levels, i.e. levels $j = 1, \ldots, J$, up to the desired level $J = 4$. The detail (differences) coefficients at each level and the approximation (averages) coefficients at final level J were extracted. For $l = 1, 2, \ldots, L$, the set of L detail coefficients were arranged as $L \times N_j$ matrices, denoted by $\mathbf{D}^{(j)} = (D_{lk}^{(j)})$, and the approximation coefficients at the final level as $L \times N_J$ matrix, denoted by $\mathbf{A}^{(J)} = (A_{lk}^{(j)})$, where $N_j \approx N(1/2)^j$ is the number of coefficients at level j.

The goal with using *orthogonal* DWT is that it produces uncorrelated ensembles (due to orthonormality of the basis functions) of $\mathbf{C}_l^{(N)}$ along the sequence family based on $\mathbf{D}_l^{(j)}$ across k, creating a sparse representation[5]. An important feature of this representation is that most of the energy of the ensembles is concentrated on a few number of large coefficients $D_{lk}^{(j)}$ that contain correlated features, partly due to intrinsic and mostly due to extrinsic correlations (see Sect. 2.2) across the species labels, that could be associated with "EMV variation" (fitness w.r.t. environmental effects like cold adaptation). In other words, most of the coefficients at finer levels are attributed to evolutionary noise (background noise) across the sequence family, with small energy spread out equally across the scales. Additionally, orthogonal DWT represents these ensembles at local-space $(1/2)^j k$ and at scales $(1/2)^j$, hence giving an accurate local description and separations of the high-frequency-features (small j) at different resolutions. Figure 1 shows histograms of $\mathbf{D}^{(j)}$ at levels 1 to 3 (4 not shown but displays a similar form). From the figure, it is clear that the multivariate distributions of $\mathbf{D}_l^{(j)}$, $\forall l$ has small variance with mean close to zero. Thus, a Gaussian distribution is a reasonable probabilistic model for the multivariate distributions at these

[4] Use the refinement equations of $q_k^j(t)$, i.e. shift the dilation and wavelet equations $\psi(t) = \sqrt{2}\sum_m g(m)\psi(2t-m)$ and $w(t) = \sqrt{2}\sum_m h(m)\psi(2t-m)$ by k and set $m = n - 2k$.

[5] Since $\mathbf{A}^{(j)}$ contain few data points of low-frequency features, we concetrate on the details (variations), that is, $\mathbf{D}^{(j)}$.

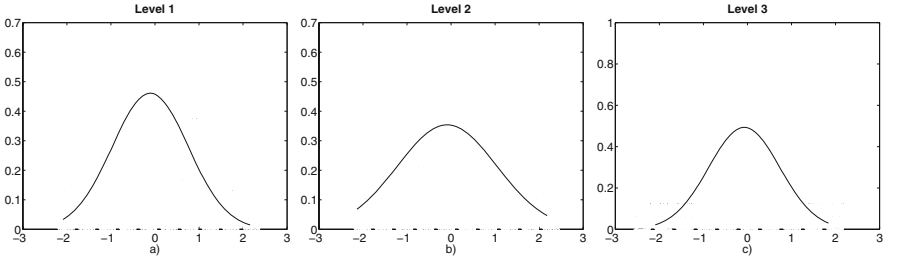

Fig. 1. Histograms showing multivariate distribution of the set of L detail coefficients corresonding to the 27 hydrophobicity sequences based on trypsin sequence data at levels 1 (a), 2 (b), and 3 (c). The superimposed curves correspond to theoretical probability density functions.

levels (see also the transposed curves). Consequently, we can use diagonalizing transformations to elimate the effects of correlation [7].

3.3 Diagonalization

Let the energy (correlation) of the ensembles of the property sequences based on $\mathbf{D}^{(j)}$ at level j be given by

$$\mathbf{R} = \left\langle (\mathbf{D}^{(j)})(\mathbf{D}^{(j)})^T \right\rangle = \left(R_{ll'} = \frac{1}{N_j - 1} (\mathbf{D}_l^{(j)})(\mathbf{D}_{l'}^{(j)})^T \right) \ , \tag{10}$$

where $R_{ll'}$, $l, l' \in \{1, 2, \ldots, L\}$ is the $l - l'$-th element of the symmetric matrix $\mathbf{R}$, a measure of correlation between $\mathbf{D}_l^{(j)}$ and $\mathbf{D}_{l'}^{(j)}$. Since $\mathbf{R}$ is symmetric and positive definite (in our case), there exists an $L \times L$ orthogonal matrix $\mathbf{U}$ such that

$$\mathbf{R} = \mathbf{U}\mathbf{D}(\sigma_i)\mathbf{U}^T \Longrightarrow \mathbf{U}^T\mathbf{R}\mathbf{U} = \mathbb{D}(\sigma_i) \ , \tag{11}$$

where the columns of $\mathbf{U}$ is given by a set of L orthonormal eigenvectors $(\mathbf{u}_i)$ and (σ_i) are the corresponding eigenvariances, ordered from high-to-low species variations, $(\sigma_1 > \sigma_2 > \ldots \sigma_i > \ldots \sigma_L)$, and $\mathbb{D}(.)$ is a diagonal matrix with eigenvariances as elements in the bracket. Then projection of $\mathbf{D}_l^{(j)}$ along the L orthogonal directions $(\mathbf{u}_i)$ will create a set of L uncorrelated coefficients in i $\widetilde{\mathbf{D}}_i^{(j)}$ that are normally distributed with mean $\widetilde{\mathbf{D}}_1^{(j)}$ (the subscript 1 indicates the 1st row of the transformed sequences):

$$\widetilde{\mathbf{D}}^{(j)} = \mathbf{U}^T\mathbf{D}^{(j)} \ \sim \ N\left[\widetilde{\mathbf{D}}_1^{(j)}, \mathbb{D}(\sigma_i)\right] \ , \tag{12}$$

Where the rows of $\widetilde{\mathbf{D}}_i^{(j)}$ indexed by i are arranged from high-to-low species variation, as described by (σ_i). The effect of the diagonalizing transformation in (11) is that two highly correlated transformed detail coefficients $\widetilde{\mathbf{D}}_i^{(j)}$ and $\tilde{\mathbf{D}}_{i'}^{(j)}$ will contribute less that two nearly correlated transformed sequence of detail coefficients $(\mathbb{D}(\sigma_i))$, thus eliminating the effect of such correlation. Figure 2 shows

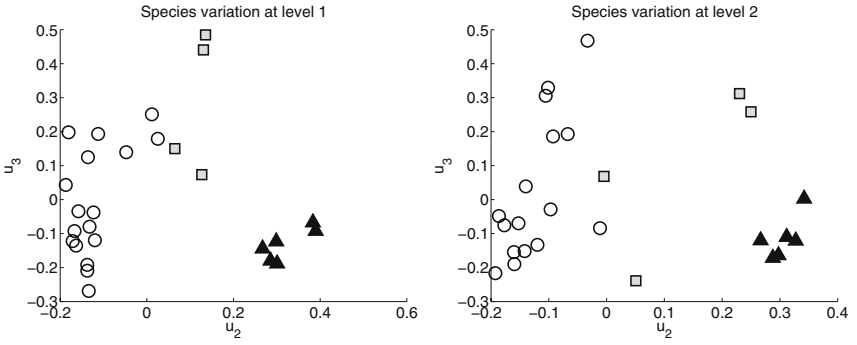

Fig. 2. Species variation at decomposition levels (from left) 1 and 2 based on the transformed detail coefficients along the 2nd and 3rd eigenvectors obtained from energy of the ensembles across the sequence family, based on detail coefficients at these levels

species variation along 2nd and 3rd eigenvectors ($\mathbf{u}_2, \mathbf{u}_3$) based on 27 hydrophobicity sequences (trypsin). We see that the extrinsic variations, especially due to cold-and warm adapted trypsins, are represented by the detail coefficients at level 1 and 2, while correlated variations at level 3 and 4 (not shown) are mostly due to within cluster variations. We could use the information at the two finest level to extract extrinsic correlations associated to environmental effects like cold adaptation, instead we chose to remove redundancy from the L set of transformed sequences. Since each of the L transformed detail coefficients are uncorrelated, we performed one-dimensional Hard-thresholding ("keep or kill" approach) using a universal threshold [11] based on eigenvariances σ_i derived from diagonalization of $\mathbf{R}$ at level 1. That is, $\varepsilon_i = \sqrt{2\sigma_i logN}$ for the i-th transformed sequence of detail coefficients, $\widetilde{\mathbf{D}}_i^{(j)}$, $i = 1, \ldots, L$, $j = 1, \ldots, 4$. We chose the eigenvariances derived at level 1 for two main reasons: (1) a better estimate of σ_i can be obtained due to high noise level, (2) σ_i determined from coarser level with more large coefficients can eliminate significant coefficients. Finally, we performed diagonalization on the covariance matrix based on approximation coefficients at the final level $\mathbf{A}^{(4)}$. In this case, we removed redundancy by keeping the first two significant components of $\widetilde{\mathbf{A}}_i^{(j)}$, $i = 1, 2$ (using a scree-plot). The output of the inverse transformed coefficients (by transpose $\mathbf{U}^T$ and transpose of analysis filter bank due to orthogonality) corresponds to a smoothed version of the original property sequences $\mathbf{C}_l^{(N)}$.

For visualization and extraction of extrinsic variations both across and along the sequence family, we performed mean subtraction, that is , $(\hat{\mathbf{C}}_l^{(N)} - \bar{\hat{C}}_l^{(N)})$ before computing the covariance. Components of the first two eigenvectors and the corresponding eigensequences (projections on the first two orthogonal directions), obtained from diagonalization of the covariance matrix based on the mean subtracted smoothed property sequences, were used to visualize species variation and the underlying residue positions responsible for this variation.

4 Results and Discussion

We presented a viable way of representing an aligned family of protein sequences through evolutionary parameterization of features. Based on these parameterized

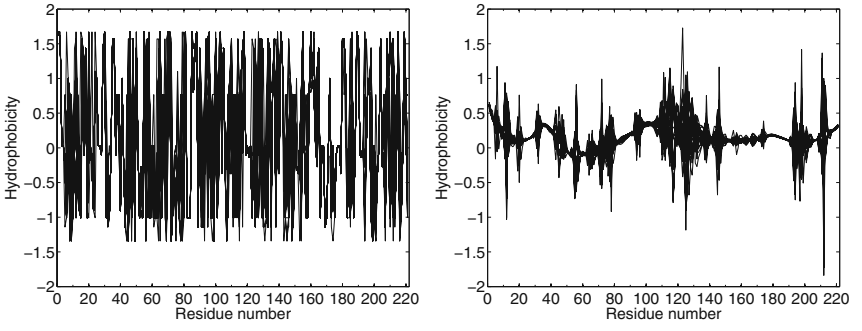

Fig. 3. Pattern of original hydrophobicity values along the 27 sequences and their smoothed version based on trypsin sequence data

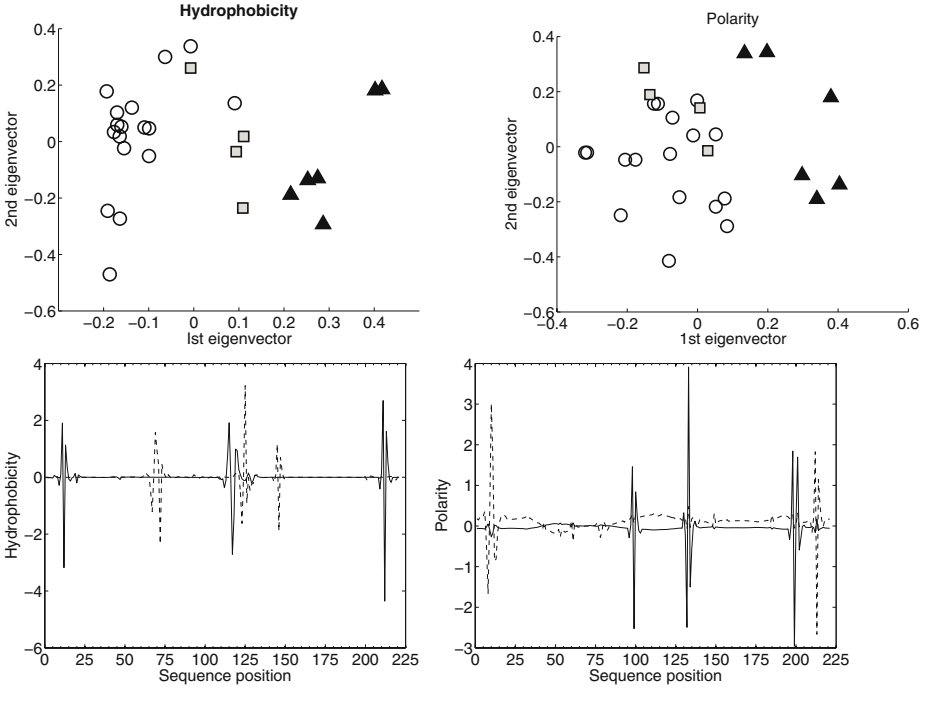

Fig. 4. Species variation based on trypsin sequence data in the space spanned by the first two eigenvectors (1st row) and the corresponding excess eigensequences, 1st (solid line) and 2nd (dashed line). The sequenes are based on hydrophobicity (on the left side) and poliarity (right).

features under reversible Markov models, we carried out a multiscale and multivariable data analysis on distinct alignment of L trypsin and elastase sequences of length N (L is 27/29, N is 221/240 for trypsin/elastase) based on hydrophobicity and polarity sequences $\mathbf{C}_l^{(N)}$. Since the sequences in both groups of enzymes are closely related, for simplicity, we removed the few columns containing gaps. The basis of the data analysis is that covariation of residue sites in evolution is related to mainly structural and functional site fitness as parameterized by models of amino acid distributions and certain amino acid properties. These correlated residues based on property sequences show deviation from a common position dependent mean value. In principle, this requires a sequence data of sufficient size and diversity (at each site) to compute such position dependent mean values for each cluster. Therefore, we used multivariate method to remove background noise and extract extrinsic correlations due to environmental effects like cold adaptation. Description of the method is given in details with some illustrations in Sect. 3. The idea is to use orthogonal wavelets to obtain a sparse representation of the property sequences based on detail coefficients and perform diagonalizing transformation in the wavelet domain to decorrelate the small number of large detail coefficients (representing the sequence ensembles) with high energy. One-dimensional thresholding, in this case, Hard thresholding can then be applied on the uncorrelated wavelet coefficients in order to separate out the larger coefficients that are associated with variations due to environmental effects like cold cold adaptation. The resulting backward transformed, denoised property sequences are smoothed version of the original property sequences as shown in Fig. 3. In this figure (to the right), the thick horizontal curve represents the

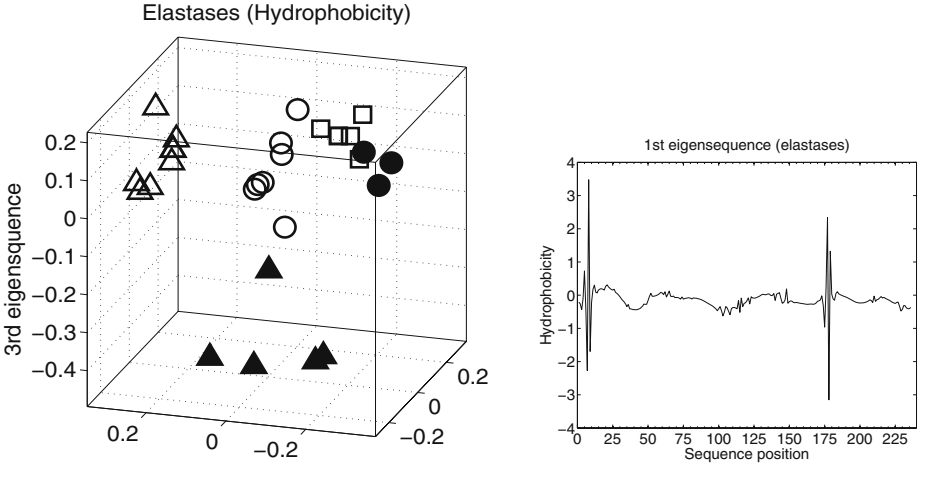

Fig. 5. Drift of centroids, as described by the species covariance matrix (figure on the left) and centroid sequence (figure on the right) based on hydrophobicity. The solid triangles and circles represent cold-adapted elastases of type I and II, respectively. The corresponing open triangles and circles represent their warm-active counterparts. The squares represent elastase type-III with no cold representatives.

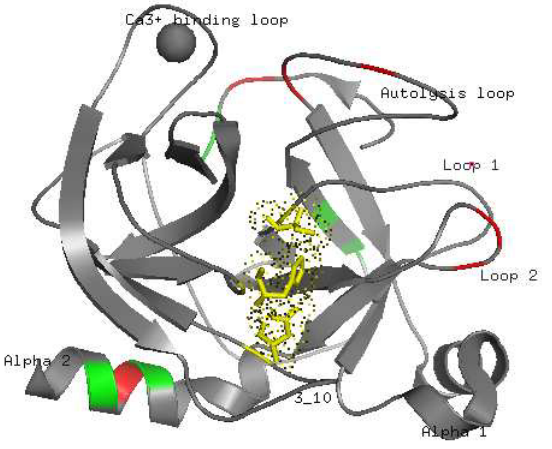

Fig. 6. 3D structure of trysin from Bovine (PDB:3PTB) showing surpport of the components of the excess eigensequences based on hydrophobicity (green) and polarity (red). Region of active site is shown in yellow.

centroid sequence of the trypsin family based on 27 hydrophobicity sequences. The larger spikes are due to extrinsic variations that are associated with residue (hydrophobic) determinants of cold-adapted trypsins. The smaller spikes are due to intrinsic variations, partly due to that there are several clusters, in the trypsin case, the warm-active higher vertebrates and other fishw, and partly due to assymmetries in covariance induced by evolutionary time since two leaf nodes were merged. Such drift of centroids, in terms of species and position variations can be clearly observed in the subspace spanned by the first two eigenvectors with largest variances, derived from diagonalization based on the smoothed property sequences after subtracting the mean profile in Fig 3, and projecting the excess variations from th mean profile along the two eigenvectors. Drift of centroids in terms of covariation and mean sequences are shown in Fig. 4 and Fig. 5. Fig. 6 shows support of the excess eigensequences, namely, in the N-and C terminals for distability and around active site [20].

References

[1] Goldman, N., Yang, Z.: A codon based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11(5), 725–736 (1994)
[2] Felsenstein, J.: Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17, 368–376 (1981)
[3] Pollock, D., Taylor, W., Goldman, N.: Coevolving protein residues: Maximum likelihood identification and relationship to structure. J. Mol. Biol. 287, 187–198 (1999)
[4] Jones, D.T., Taylor, W.R., Thornton, J.M.: The rapid generation of muatation data matrices from protein sequences. Comp. Appl. Biosci. 8, 275–282 (1992)

[5] Kishino, H., Miyata, T., Hasegawa, M.: Maximum likelihood inference of protein phylogenies and the origin of chloroplasts. J. Mol. Evol. 31, 151–160 (1990)
[6] Hasegawa, M., Fujiwara, M.: Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor joining methods for estimating protein phylogeny. Mol. Phylog. and Evol. 2, 1–5 (1993)
[7] Johnson, R.A., Wichern, D.W.: Applied Multivariate Statistical Analysis, 5th edn. Prentice Hall, Upper Saddler River (2002)
[8] Mallat, S.G.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. 11(7), 674–693 (1989)
[9] Suzuki, T., Srivastava, A., Kurokawa, T.: cDNA cloning and phylogenetic analysis of pancreatic serine proteases from Japanese flounder, Paralichthys olivaceus. Comp. Biochem. and Physiol. Part B 131, 63–70 (2001)
[10] Strang, G., Nguyen, T.: Wavelets and Filter Banks. Wellesley-Cambridge Press (1997)
[11] Donoho, D., Johnstone, I.: Ideal spatial adaptation by wavelet shrinkage. Biometrika 81(3), 425–455 (1994)
[12] Koshi, J.M., Mindell, D.P., Goldstein, R.A.: Beyond mutation matrices: physical-chemistry based evolutionary models. In: Miyano, S., Takagi, T. (eds.) Genome informatics, pp. 80–89. Universal Academy Press, Tokyo (1997)
[13] Casari, G., Sander, C., Valencia, A.: A method to predict functional residues in proteins. Nat. Struc. Biol. 2(2), 171–178 (1995)
[14] Durbin, R., Eddy, S., Krogh, A., Mitchison, G.: Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge (1998)
[15] Whelan, S., Goldman, N.: A general empirical model of protein evolution derived from multiple protein famililies using a maximum-likelihood approach. Mol. Biol. and Evol. 18, 691–699 (2001)
[16] Goldman, N., Whelan, S.: A novel use of equillibrium frequencies in models of sequence evolution. Mol. Evol. 11, 1821–1831 (2002)
[17] Ahmed, S.H., Flå, T.: Estimation of evolutionary average hydrophobicity profile from a family of protein sequences. In: Rajapakse, J.C., Schmidt, B., Volkert, L.G. (eds.) PRIB 2007. LNCS (LNBI), vol. 4774, pp. 158–165. Springer, Heidelberg (2007)
[18] Feller, G., Gerday, C.: Psychrophilic enzymes: molecular basis of cold-adaptation. Cell Mol. Life Sci. 53, 830–841 (1997)
[19] Georlette, D., Blaise, V., Collins, T., D'Amico, S., Gratia, E., Hoyoux, A., Marx, J.C., Sonan, G., Feller, G., Gerday, C.: Some like it cold: biocatalysis at low temperatures. FEMS icrobiol. Rev. 28, 25–52 (2004)
[20] Schrøder, H.-K., Willassen, N.P., Smalås, A.O.: Residue determinants and sequence analysis of cold-adapted trypsins. Extremophiles (2), 5–219 (1999)