Published in final edited form as:

Phys Biol. ; 13(2): 025003. doi:10.1088/1478-3975/13/2/025003.

Correlated rigid modes in protein families

D A Striegel¹, D Wojtowicz², T M Przytycka², and V Perival¹

¹Laboratory of Biological Modeling, NIDDK, USA

²Computational Biology Branch, NCBI, NLM, The National Institutes of Health, Bethesda MD 20892, USA

Abstract

A great deal of evolutionarily conserved information is contained in genomes and proteins. Enormous effort has been put into understanding protein structure and developing computational tools for protein folding, and many sophisticated approaches take structure and sequence homology into account. Several groups have applied statistical physics approaches to extracting information about proteins from sequences alone. Here, we develop a new method for sequence analysis based on first principles, in information theory, in statistical physics and in Bayesian analysis. We provide a complete derivation of our approach and we apply it to a variety of systems, to demonstrate its utility and its limitations. We show in some examples that phylogenetic alignments of amino-acid sequences of families of proteins imply the existence of a small number of modes that appear to be associated with correlated global variation. These modes are uncovered efficiently in our approach by computing a non-perturbative effective potential directly from the alignment. We show that this effective potential approaches a limiting form inversely with the logarithm of the number of sequences. Mapping symbol entropy flows along modes to underlying physical structures shows that these modes arise due to correlated compensatory adjustments. In the protein examples, these occur around functional binding pockets.

Keywords

protein families; normal modes; pattern recognition

1. Introduction

Information propagation is a fruitful way of thinking about biological processes. The propagation of genetic information could even be considered a definition of life. The linear character of genetic information leads to the linear residue chains that fold and interact to form functional proteins. Given obvious evolutionary similarities between organisms, it is natural to attempt to extract commonalities between homologous proteins in an effort to gain insight into their structural and functional properties. This area of research has seen activity for several decades [1–5].

vipulp@mail.nih.gov.

Supplementary material for this article is available [online](#)

Structurally and functionally similar proteins can have remarkably different amino-acid sequences. Nevertheless, there are physical constraints that are encoded in families of sequences that contain information about the physical properties of the entire family of proteins. All-atom molecular dynamics (MD) simulations of protein activity have made tremendous progress [6–8], and many approaches take structure and sequence homology into account. Besides the references above, models of protein residue interactions motivated by statistical physics and phylogenetic relationships have been studied in various ways [9–28]. On the other hand, there clearly is a need for multiple viewpoints on this difficult problem [29, 30].

In more depth, the well-established approaches in the analysis of protein structure, allostery, and function, that take sequence features into account [1–5] comprise two main veins in the literature, those that introduce an explicit parametrized model of residue interactions [13–15, 23, 24] with priors on parameter values, and those motivated by analogies with random matrices and finance [16, 21, 22]. A distinct approach is in [17], where the results are compared to the approach of [3, 16]. The former direct-coupling analysis approach [13, 14, 23, 24] has achieved significant results in predicting native contacts and requires considerable computational effort. The latter approach, statistical coupling analysis, motivated by finance and random matrix theory, has also made significant predictions for allostery [16, 21] that have been experimentally tested with success [22]. A two-point correlation matrix, albeit in a binary approximation [21, 22], has entries multiplied by weights computed from a multinomial distribution based on the number of sequences, in this approach, and eigenvectors and eigenvalues of this matrix are of central interest. The connection of the weighted correlation matrix with random matrix theory is not evident to us.

Given that information theory is important, and Shannon entropy variants are even explicitly used in defining the statistical ‘free energy’ of these formalisms [3, 17], we recently approached the analysis of a family of symbol sequences from the point of view of empirical Bayesian analysis [31], but carried out directly with the Shannon entropy of the sequence family as the foundation. Our approach worked well for predicting phenotypes from associated sequence data [32]. Here we show that our conceptually simple and computationally tractable method can lead to interesting deductions even when the only phenotype is, literally, existence, i.e., the sequences available are ones that survive in the evolutionary record. To be specific, much of what we carry out is geared to the case of a minuscule fraction of sequences compared to the number of possible sequences. Thus intuitions for statistical physics and thermodynamic limits need to be tempered with the reality of extremely sparse sampling.

A few points of interest in comparing our work with some earlier work are worth noting at the outset. We do not introduce an *a priori* statistical model with coupling parameters in our approach. In other words, we do not set up the problem in a parametric form, with explicitly assumed interaction forms, usually linear and quadratic for residue-residue interactions in proteins (see e.g. [13, 15, 23]), accompanied by a parameter determination step with (maximum entropy: MaxEnt or other) Bayesian priors. The parameter determination is computationally difficult in some of these cases, as there are on the order of (length-of-

sequence $\times$ number-of-symbols)² parameters to be determined for a model with quadratic interaction terms. Rather, we aim at a non-parametric functional inference using standard statistical physics tools to compute the function of interest directly, but coupled with a null Bayesian prior, one that explicitly assumes that there are no interactions between residues in assigning probabilities to models. As we showed in [32], and re-derive in the present context, in our approach the number of parameters is independent of the order of interactions, be they quadratic or cubic or higher in residues. As befits a non-parametric approach, the question becomes one of extracting answers of interest from the inferred function. This is the point where we resort to an expansion in fluctuations about a fixed point found in the functional inference.

Perhaps the most relevant counterpoint to our work is that of [18–20], where the normal modes of networks of interactions deduced from known threedimensional structures are used to inform an analysis of the sequence characteristics that are relevant for these modes. Our approach is in some sense the opposite direction of inference from the elastic network approach used in [18–20], as we start from multiple sequence alignments and infer sequence normal modes which we then overlay on known structures as an approximate visual demonstration of contiguity, while these works [18–20] determined structure based normal modes and then showed that these modes were evolutionarily robust under sequence variation.

We exhibit three novel physical phenomena: (1) there may be a small number of isolated low-lying modes in families of proteins which scale logarithmically with the number of proteins in the sequence dataset, as we show in some examples. (2) These modes are associated with compensatory covariation of amino-acids at different positions in the sequence and predict correlated fluctuations in individual proteins. (3) Discrete symmetry-breaking is uncovered when structurally distinct yet compositionally similar sequences are analyzed.

The structure of our paper is as follows: in the theory section we provide a complete and self-contained derivation of our method. After the materials and methods section, in the results section we apply our method to some concrete applications, including a famous dataset consisting of handwritten digits. While this dataset may appear far removed from biology, there are two reasons for including it. The first is that, in fact, the handwriting samples are concrete examples of biological variation, produced by distinct organisms, for a specific purpose, communication. Homologous proteins in different organisms could be characterized with exactly the same terminology. The second reason is that the results we find in this example can be intuitively checked by a visual inspection. Real proteins are much more complex and claiming an intuition for protein structure based on an analysis of sequences is above our pay-grade. Thus the digit dataset results serve as a sanity check and, as such, this seemed to us to be an apt dataset for demonstrating our method before applying it to proteins. We conclude with a discussion.

2. Theory

For orientation, we start with a sketch of the fundamental idea behind our approach. Many datasets in biology can be formally represented as a family of sequences of symbols, which can represent amino-acids along a protein chain or nucleotides in a DNA or RNA sequence. For example, an enzyme such as dihydrofolate reductase (DHFR) is found in many organisms with the same function but with varying amino-acid composition. A set of such related enzyme sequences can be aligned so that amino-acids at the same position in different sequences can be compared. For example, there are often times conserved residues that are crucial for function and exhibit no variation at all in the family of sequences. The alignment will usually place these residues at the same position along the sequences. In our approach, we will embed these sequences in a frequency model space, and think of these as observed configurations of a statistical mechanical system. In a physics analogy, any configuration of a set of Ising spins can be embedded into a space of possible local magnetizations. Aligned protein (DNA) sequences have $G = 20$ ($G = 4$) (or more, depending on gaps in the alignment) possible residues (nucleotides). Thus, mapping such sequences, each of length R , to a frequency model space means mapping them to binary matrices of size $R \times G$, where the columns correspond to possible amino-acids and rows corresponding to positions along the sequence. For any specific sequence, the entry in row i and column A is 1 if the sequence has amino-acid σ_A at position i and 0 for all other columns in that row. Thus each matrix, which we denote by σ_{iA} , associated with a sequence has the constraint that the entries in every row add up to 1.

If the family of sequences, S , has n_{seq} sequences, each sequence is by this mapping embedded into a continuous frequency model space, which is therefore $[0, 1]^{GR}$, a product of unit intervals of dimension $R \times G$, but constrained by the R constraints $\sum_A f_{iA} = 1$, where f_{iA} is the expectation value of σ_{iA} for each position i in the alignment of the sequences. We refer to any point in this continuous space as a frequency model, $F \equiv \{f_{iA}\}$.

2.1. Partition function definition

A simple way to compute all possible correlations between amino-acids at different positions is provided by the formalism of statistical physics [33, 34]. We think of each sequence as a configuration of ‘magnetic spins’ that point in specific amino-acid directions. Introducing J_{iA} as a local magnetic field, we define a generating function for all connected correlation functions, $W(J_{iA})$, as the logarithm of the partition function through

$$\exp(MW(J_{iA})) = \sum_{\alpha \in S} \exp\left(M \sum_{i,A} J_{iA} \sigma_{iA\alpha}\right). \quad (1)$$

$\sigma_{iA\alpha}$ is the frequency model binary matrix for the α sequence in the set of sequences, S . M is a parameter we introduce for later notational convenience. It follows from this definition that the normalized expectation value of a residue A at position i is given by

$$\langle \sigma_{iA} \rangle_J = \frac{\partial W}{\partial J_{iA}} = f_{iA}(J). \quad (2)$$

Thus the introduction of J leads to interpreting the frequency model F as a local magnetization. As the frequencies are the J -weighted expectation values of σ_{iA} , altering J_{iA} leads to changes in the frequency model, as sequences with spins ‘pointing’ in the direction favored by J get higher weighting in the statistical sum over all sequences.

2.2. Legendre transform

The sequence data gives us frequencies of occurrence, while the generating function W is a function of the magnetic field. However, there is a standard reformulation of the partition function as a function of J as a free energy or effective potential as a function of magnetization, F . The Legendre transform makes the magnetic field the dependent variable and the magnetization the independent variable. In our case, the Legendre transform of W is the effective potential E_{eff} :

$$W(J) + E_{\text{eff}}(F) = J \cdot F = \sum J_{iA} f_{iA}. \quad (3)$$

The effective potential defined in this manner is a function of f_{iA} , and the symmetry of the Legendre transform gives

$$\frac{\partial E_{\text{eff}}(F)}{\partial f_{iA}} = J_{iA}(F). \quad (4)$$

For later use, remark that the Legendre transform has the important mathematical property that the Legendre transformed function, in our case $E_{\text{eff}}(F)$, is a convex function by construction. This is true even if W is not a convex function, though as defined in (1), Jensen’s inequality implies that W is convex. Now, $\sum_A f_{iA} = 1$ so J at each position is only determined up to an additive constant, which can be subtracted, allowing us to normalize J by demanding $\sum_A J_{iA} = 0$. The physical interpretation is that $\exp(-ME_{\text{eff}}(F))$ gives the free energy of the frequency model F achieved by an appropriate choice of weights $\exp(MJ(F) \cdot \sigma)$ for sequences.

2.3. Effective potential + Shannon entropy = 0

However, as we have defined it, $ME_{\text{eff}}(F)$ has a much more interesting interpretation. It is the negative of the Shannon entropy for this family of sequences with the assignment of likelihoods $\exp(MJ \cdot \sigma_a)$ for each sequence a , and normalizing to get probabilities. One can see this by noting

$$1 = \sum_{\alpha} \exp(MJ \cdot (\sigma_{\alpha} - F) + ME_{\text{eff}}(F)) \equiv \sum_{\alpha} p_{\alpha}, \quad (5)$$

and therefore

$$S \equiv - \sum_{\alpha} p_{\alpha} \ln p_{\alpha} = M \sum_{\alpha} (-J \cdot (\sigma_{\alpha} - F) - E_{\text{eff}}(F)) p_{\alpha} \quad (6)$$

but $\langle \sigma_{\alpha} \rangle = F$, so the Shannon entropy $S = -ME_{\text{eff}}(F)$. Thus minimizing $E_{\text{eff}}(F)$ with respect to F is equivalent to maximizing the Shannon entropy of the weighted sequences by choosing an appropriate local magnetic field $J_{iA}(F)$. That is, however, not what we want to do.

2.4. Data-driven prior on null models

If there is no pattern at all in the data, and symbols appear at positions based solely on the frequency with which they appear in the entire dataset, we would expect the probability of finding F to be distributed according to the multinomial distribution

$$P(F|M, Q) \propto \prod_i \left[\frac{\Gamma(M+1)}{\prod_A (\Gamma(Mf_{iA}+1)) \prod_A q_A^{Mf_{iA}}} \right], \quad (7)$$

where $Q = \{q_A\}$ are the frequencies of the symbols occurring at any position in the data and M is usually interpreted as the number of trials. In Bayesian inference, $P(F|M, Q)$ is the prior probability of the set of frequencies, F , and $-ME_{\text{eff}}(F)$ is the log likelihood of the data being consistent with the ‘model’, F . Thus each choice of F is a distinct null model in the sense that none of these sets of frequencies assume any correlations between the symbols that appear at different positions in the sequences. In any Bayesian analysis where no symmetry principles are invoked, the choice of model priors is exactly that, a choice, and we argue that the choice we make here is the one most aligned with the statistical concept of a null hypothesis. If correlations or patterns appear with this choice of model prior, it is likely that these patterns are not random. In particular, the determination of Q is data-driven as in empirical Bayes approaches [31].

2.5. Maximizing the posterior: iterative solution

Define $K(F|M, Q)$ by $\log P(F|M, Q) \equiv MK(F|M, Q)$. We can ask for the maximum posterior null model, the model that is most likely given the data and our uncorrelated multinomial prior. Using the basic Bayes’ theorem, we want therefore to maximize $K - E_{\text{eff}}$ to find this maximum posterior model given the sequences. As the space of models parametrized by F is a continuous space, we may find the minimum by computing a derivative. We have

$$\frac{\partial K}{\partial f_{iA}} - \frac{\partial E_{\text{eff}}}{\partial f_{iA}} = 0. \quad (8)$$

Combining (4) and (8) relating J , E_{eff} and K , we find

$$J_{iA}(F) = \frac{\partial K}{\partial f_{iA}} \quad (9)$$

at the most likely model. This is the crucial step because the solution of (8) can now be written as an explicitly computable direct fixed point iteration:

$$f_{iA}^{(n+1)} = \langle \sigma_{iA} \rangle_{J = \partial K / \partial f |_{F = F^{(n)}}} \quad (10)$$

Once we find the fixed point values F_E , we know the weights of the sequences through knowing $J(F_E)$. This lets us use standard results on the Legendre transform [34] to compute two and higher point connected correlation functions directly from the weighted data with no approximations. Thus, suppressing indices for clarity

$$\langle \sigma \sigma \rangle_{J(F_E)} = M^{-1} \frac{\partial^2 W}{\partial J^2} = \left(\frac{\partial^2 M E_{\text{eff}}}{\partial f^2} \right)^{-1}. \quad (11)$$

This evaluation of the second (and higher) derivatives of E_{eff} lets us expand E_{eff} in a Taylor series for any model f_{iA} , with $\delta F \equiv f_{iA} - f_{iA,E}$ and $\delta E_{\text{eff}}(F) \equiv E_{\text{eff}}(F) - E_{\text{eff}}(F_E)$:

$$\delta E_{\text{eff}}(F) = \delta F \cdot J(F_E) + \frac{1}{2} \delta F \frac{\partial^2 E_{\text{eff}}}{\partial F^2}(F_E) \delta F + \dots \quad (12)$$

This Taylor expansion will most likely only converge for F such that δF lies in a subspace of small eigenvalues of $D_2 \equiv \partial^2 M E_{\text{eff}} / \partial f_{iA} \partial f_{jB}(F_E)$. When the eigenvalues of D_2 are small enough, it can be used to compute the entropy of a given sequence, since we have an embedding of sequences into the frequency model space, as described at the beginning of the theory section.

Referring back to the frequency model space, the quadratic term in the Taylor expansion computed about F_E can help us in understanding the spectrum of likely fluctuations in the model space. As remarked above, actual sequences can be embedded into this frequency model space as frequency models where the frequencies are all 0 or 1, and therefore any amino-acid changes in a sequence can be associated with Shannon entropy changes. In the

statistical mechanics of physical systems, equipartition of energy implies that the low-frequency fluctuations are associated with large-amplitude collective motions while high-energy fluctuations correspond to localized small-amplitude vibrations. The spatial physical intuition for a chain of beads or springs is absent from the present context and we do not have exchangeability so equipartition is not valid. However, the mathematical association of collective motions associated with the small eigenvalues of D_2 is more general and the low eigenmodes of D_2 correspond to the highly correlated modes of interest as deformation directions for the sequence family that decrease the entropy the least. Therefore, we use the term rigid for them. (There are no continuous symmetries in this case that would give rise to truly flat directions such as rigid body rotations or translations in physics.) In other words, these modes correspond to coordinated changes in the amino-acids at many positions in the sequence, just as in the motion of a string of beads the low-frequency modes require neighboring beads to move in a synchronized manner. It should also be noted that if $M=0$, these correlated modes are the principal components. For small numbers of sequences, as is the case in biological applications, the re-weighting of the data sequences when $M=0$ is critical for better predictions in the quantitative phenotype case [32]. If one had all or even a substantial fraction of all possible sequences, our analysis would constitute superfluous learnedness.

2.6. Summary of algorithm

In this section, we summarize the steps in our overall algorithm.

Step 0: Alignment of data—Our algorithm requires sequences that can be compared, position by position. Arranging the sequences in rows, this assumes that sequences are of equal length, so that calculations like $\sum_{i,A} J_{iA} \sigma_{iA\alpha} \equiv J \cdot \sigma_\alpha$ make sense for each sequence (labeled by α). If the dataset has sequences of unequal length, for biological applications, protein residue sequences or DNA sequences can be aligned by multiple sequence alignment tools, and our algorithm can then be used, albeit that gaps (usually denoted by ‘-’ in alignments) mean that the symbol ‘-’ is treated as an additional symbol. Another possibility, used extensively in [32] but not required for the protein sequences considered in the present work, is to pick a set of short sequences, called motifs, that occur in the given sequences, and convert the original sequences to sequences of symbols A (absent) or P (present) for each motif. If the motif set is diverse enough that every pair of distinct original sequences leads to distinct A/P sequences, then we have converted the dataset into a new aligned sequence problem, with the original positions in the sequence replaced by motif number in the set of chosen motifs. The motifs chosen are the same for each sequence in the dataset, and hence the associated A/P sequences are equal length and aligned. Thus the algorithm is quite general, but does require either multiple sequence alignments or conversion to aligned motif sequences.

Step 1: Initialization—Compute the $Q \equiv \{q_A\}$ vector which counts the overall frequency of each symbol A in the sequence dataset. Initialize a random J matrix, with rows corresponding to position and columns corresponding to symbol. Normalize the J matrix so that each row sum vanishes. This choice fixes the arbitrariness in J that arises from the fact

that each position in the sequence has a unique choice of symbol, so the sum of frequencies of symbols at each position adds up to unity.

Step 2: Iteration—Compute the normalized weighted expectation value of each symbol at each position by summing over all sequences with weight $\exp(MJ \cdot \sigma_\alpha)$,

$$f_{kC}(J) \equiv \frac{\sum_{\alpha \in S} \sigma_{kC\alpha} \exp(M \sum_{i,A} J_{iA} \sigma_{iA\alpha})}{\sum_{\alpha \in S} \exp(M \sum_{i,A} J_{iA} \sigma_{iA\alpha})} \quad (13)$$

and compute the gradient of $K(F|M, Q)$ at these f_{kC} values. This gradient is the updated value of J after normalizing to vanishing row sums, and we iterate these steps until convergence. Denote the fixed point expectation value matrix as F_E and the associated J matrix as J_E .

Step 3: Weighted correlation matrix—Compute the weighted connected correlation matrix, D_2^{-1} :

$$D_{2kC,jB}^{-1} \equiv \frac{\sum_{\alpha \in S} \sigma_{kC\alpha} \sigma_{jB\alpha} \exp(M \sum_{i,A} J_{iA} \sigma_{iA\alpha})}{\sum_{\alpha \in S} \exp(M \sum_{i,A} J_{iA} \sigma_{iA\alpha})} - f_{kC,E} f_{jB,E}. \quad (14)$$

The eigenvalues and eigenvectors of D_2 are the modes of interest, sorted from lowest to highest. As in classical mechanics, the lowest eigenvalues correspond to low-energy excitations, and in statistical physics one would expect these to correspond to larger amplitudes by equipartition. The amplitudes in the present context correspond to correlated sequence variations, as we have embedded the discrete symbol sequences into a continuum expectation value (in other words, frequency model) space, much as the magnetization of an Ising model takes continuous values while the actual spin configurations are discrete. However, one must be careful with interpretations as the present context is neither statistical physics with exchangeability nor classical mechanics.

If not for the crucial weighting of sequences by $J(F_E)$ that enables us to relate ME_{eff} to the Shannon entropy, the eigenvectors of D_2^{-1} would be just the principal components. Even with the weighting, by construction as the eigenvectors of a real symmetric matrix, these eigenvectors are orthonormal when the eigenvalues are non-zero. The eigenvalues are nonnegative due to the properties of the matrix in any event.

2.7. Mutual information (MI)

As our approach is based directly on the Shannon entropy, it is natural to quantify correlations between positions in terms of MI. Define the MI between two positions i and j in the sequence as the MI computed from the entropy of deviating from the fixed point model by altering the frequencies of symbols only at these two positions. To be explicit

$$MI(i, J) \equiv \sum_{A, B} \Pr(iA, jB) \ln \frac{\Pr(iA, jB)}{\Pr(iA, j^*) \Pr(i^*, jB)}, \quad (15)$$

where

$$\ln \Pr(iA, jB) \propto \epsilon MJ(F_E) \cdot (\delta_{iA} + \delta_{jB}) + \epsilon^2 \delta_{iA} \delta_{jB} \frac{\partial^2 ME_{\text{eff}}}{\partial f_{iA} \partial f_{jB}}(F_E) \quad (16)$$

and δ_{iA} is a unit change in the probability of symbols at position i , increasing the probability of symbol A and decreasing the probability of all other symbols at that position appropriately to maintain unit probability. ϵ is chosen to be the square-root of the smallest eigenvalue of the two-point correlation matrix that is retained in the inversion following the SVD. We normalize the probability $\Pr$ before computing MI .

The eigenvectors of the second derivative of E_{eff} , D_2 , contain information about normal modes in amino-acid occurrence in the set of aligned sequences. We compute this in terms of changes of the Shannon symbol entropy. Let $V \equiv \{v_{iA}\}$ be any vector in the space of frequency models. The Shannon entropy of a position i in the aligned sequences is defined as $I_i(F) \equiv -\sum_A f_{iA} \log f_{iA}$. If we change F to $F + tV$, the entropy changes to

$$I_i(F + tV) \equiv -\sum_A [(f_{iA} + tv_{iA})/N] \times \log[(f_{iA} + tv_{iA})/N], \quad (17)$$

where $N \equiv \sum_A (f_{iA} + tv_{iA})$ normalizes the changed frequencies. Thus the change in entropy at each position along the direction V can be computed as

$$\begin{aligned} \mathcal{L}_V I_i(F) &= \left. \frac{dI_i(F + tV)}{dt} \right|_{t=0} \\ &= -\sum_A v_{iA} (\log f_{iA} + I_i(F)). \end{aligned} \quad (18)$$

To ascertain the relative changes in information content at different positions, we subtract a position-independent constant from $\mathcal{L}_V I_i(F)$ so that $\sum_i (\mathcal{L}_V I_i(F) - \text{const.}) = 0$. We term this the entropy flow $\widehat{\mathcal{L}}_V$ in the direction V .

3. Materials and methods

We obtained 708 amino-acid sequences of DHFR enzymes similar to *E. coli* DHFR from UniRef50 [35]. We obtained 146 aligned sequences of the intimin/invasin virulence factor found in Gram-negative bacteria, solved recently [36]. We obtained the sequences for the NTPase *Maf_HamI* domain which is shared by two subfamilies of proteins with a

structurally similar nucleotide binding cleft but altered conserved residue location and nature [37].

We obtained the MNIST hand-written digit image database [38] and digitized it as follows. This database consists of 60 000 gray-scale images of numbers 0 through 9. We reduced these 28×28 pixel format images to an unordered sequence of symbols by associating symbols with each of the 16 possible combinations of black and white pixels in a 2×2 square (figure 1(a)). Each digit was thus associated with a family of about 6000 strings of symbols, each string of length 729 symbols (figures 1(b)–(f)). We reduced these images to unordered strings of length 729 symbols (figure 1). In all cases, proteins and digits, we kept only positions in the sequences that showed less than 95% conservation. We checked that changing 95%–99% did not change any qualitative results (supp. figure S1).

For varying the number of sequences, we picked five random selections of subsets of sequences of the appropriate cardinality, and took the mean values of the eigenvalues found for these random subsets. We repeated this for each choice of the number of sequences. Obviously, as one approaches the total available number of sequences, the variation in the eigenvalues between randomly selected subsets vanishes.

The code was written in Python and C++. Let G be the number of possible symbols so there are G^R possible sequences. The two-point connected correlation matrix is singular as it has $G \times R$ rows and columns but there are only $n_{\text{seq}} < G \times R$ sequences in all the cases considered in this paper. We performed a singular value decomposition (SVD) of the second derivative matrix of $ME_{\text{eff}}(F)$ expanded about F_E , and only kept only n_{seq} (or fewer) eigenvalues. The case of fewer eigenvalues arises when some of the sequences are almost degenerate, leading to numerical instability in the SVD. The results did not depend on the threshold with which we detected this instability, which was to limit the range of values taken by the entries in the inverted matrix to machine precision on a 64 bit processor. $M = 0.001 - 0.007$ in this paper, but there are no qualitative changes if we vary M to similar higher or lower values (supp. figures S1, S3).

4. Results

4.1. Dihydrofolate reductase

The DHFR enzyme has been studied for its conformational changes related to catalytic activity [39]. For DHFR sequences, we computed the effective potential describing these sequences up to quadratic order for varying numbers of sequences following [32], using the fixed-point of the iteration in (10). We found that the eigenvalues of D_2 scaled inversely with the logarithm of the number of sequences, $1/\log n_{\text{seq}}$ (figure 2(a)). Thus E_{eff} may converge to an n_{seq} -independent form, though this is finite size-scaling as the number of possible sequences is bounded, albeit extremely large.

To test the dependence of this logarithmic scaling on the correlations between amino-acids at distinct positions, we permuted the amino-acids column-wise independently for each position in the aligned sequences. This removed correlations between distinct positions but preserved position dependence of amino-acid occurrence. The asymptotic value of the

eigenvalues increased, implying that the effective potential became steeper (figure 2(b)). We then permuted each sequence independently and then applied column-wise permutation as before, effectively removing both correlations between positions and the position dependence of amino-acid occurrence. This led to an even steeper effective potential (figure 2(c)). Together, these phenomena suggested that both amino-acid frequencies at specific positions and compensating variations between amino-acids at distinct positions were important for the existence of deformations. Most importantly, for DHFR, only three isolated eigenvalues occurred in the spectrum, and there were no such isolated eigenvalues for either type of permuted sequences.

4.2. Entropy flow: sanity check with handwritten digits

Is this a general property of analyzing symbol sequences in this manner or is it specific to protein sequences? For a completely different set of sequences, we applied these concepts to the MNIST hand-written digit sequences [38]. Starting from the second derivative matrix of $ME_{\text{eff}}(F)$ expanded about F_E , the eigenvalues of this matrix (which is the inverse of the weighted version of the connected two-point correlation function computed in our formalism for the digit symbol sequences) scale logarithmically as well (supp. figure S4).

We can see in figure 3 that these low eigenvalue directions are global in character in the sense that they have a geometric meaning on the two-dimensional image, even though the analysis made no use of the ordering of the sequence or of two-dimensional continuity. Changing the fixed point model F_E along the directions given by the orthonormal eigenvectors associated with low-lying eigenvalues induces entropy changes at each position in the sequence. By visualizing this entropy change using $\widehat{\mathcal{L}}_V I_i(F_E)$ (theory section) along the direction V , we observe that some positions in the sequence become more specialized and lose entropy and others gain entropy as the probabilities change along a flat direction. The three lowest eigenvalue directions for each digit are shown in figure 3, and have two-dimensional geometric meanings as inter-individual handwriting size and slant variations, that are absent from randomized digit sequences (supp. figure S5). For example, one can deduce that most people are right-handed by noting the slant towards the right.

Thus the low eigenvalue eigenvectors, deduced using only discretized sequence information (see methods) with no knowledge of two-dimensions, nevertheless capture correlated changes in symbol frequencies that are consistent with the functional constraints of representing the two-dimensional digit. This is the main point of this MNIST digit sanity check, to note that this is similar to the situation we expect in real proteins, where sequence proximity has little bearing on tertiary proximity in the three-dimensional structure.

What do the eigenvectors associated with these eigenvalues reveal? We mapped $\widehat{\mathcal{L}}_V$ for V associated with the first three eigenvalues to DHFR structures (figure 4). The isolated character of the associated eigenvalues, their contiguous appearance when superposed on the DHFR structure, and their evanescence upon permutation (supp. figure S6), suggests that they are correlated collective modes enabled by contact interactions. If they were not rigid, we would expect them to be part of a continuum of deformations, much as the difference between coordinates describing the configuration of a rod or of a string. As we used no

structural connectivity information in our analysis of DHFR, it must be that the sequences in toto contain this information.

4.3. Intimin/invasin barrels

The intimin/invasin family of virulence factors are produced by Gram-negative bacteria. As a check on our method with far fewer sequences, we applied it to the intimin/invasin virulence factor [36]. Computing the entropy flow, $\widehat{\mathcal{L}}_V$, induced by the flattest mode (figure 5), revealed that this mode corresponds to reducing variation in either the side chains projecting towards the outside of the barrel and allowing more variation in those projecting towards the inside, or vice versa [40]. These features are erased by column permutations (supp. figure S7).

4.4. Convexity and NTPase subfamily separation

E_{eff} is convex by construction (figure 6(b)). If there are distinct structures in the set of sequences with no connecting continuous mode, will this approach detect structural differences? This would correspond to discrete symmetry breaking. The NTPase *Maf*-*Ham1* domain is shared by two subfamilies of proteins with a structurally similar nucleotide binding cleft but altered conserved residue location and nature [37]. The NTPase *Maf*-*Ham1* sequences are a good test. Figure 6(a) shows that the lowest few modes completely separated the two subfamilies. The classification of these proteins may actually correspond to three subfamilies. Entropy flows within each subfamily are associated with deformations of the binding pocket (figure 7 and supp. figure S8). When we analyzed digits with sequences for '1' and '7' together we obtained similar results (figure 6(c)), showing clear separation.

Do the pairwise interactions in all of D_2 contain information beyond the eigenvectors? We computed the MI between different symbol positions from E_{eff} . Mapping the residues with high MI values to the dimerized NTPase structure showed that these residues were concentrated around the ligand binding pocket that forms at the interface between the two chains (figure 8). Certain isolated high MI residues were not at the dimer interface (supp. table S1) but the sequences available are not a complete evolutionary record. It may be that these residues were part of a structural interaction that is no longer maintained in the sample of present-day sequences we used.

5. Discussion

We showed here that our method, developed originally for predicting quantitative phenotypes [32], provides a computationally efficient way to deduce some general properties of biologically interesting sequence alignments from the typically sparse set of possible configurations available, using the fundamental principles of statistical physics and Bayesian analysis in a completely data-driven manner. The only choice we made was the multinomial null prior but we showed that the single parameter M introduced in our analysis did not affect the results significantly.

The key points we emphasize are that biological applications are varied, the mechanisms underlying observed variation are seldom completely known, and biological sequence

datasets are usually extremely sparse in the context of all possible sequences of the same length. In this context, using a data-driven and parameter-free approach based on fundamental concepts such as the maximization of Shannon entropy, the convexity of the Legendre transform, and Bayesian probability has merit in two senses: (1) there are essentially no theoretical approximations so when the analysis extrapolated to test sequences gives inconsistent results, it is more likely that the experiments have latent unobserved variables than that the theoretical setup used approximations. One could use this as a simple-minded discovery tool for latent variables. (2) The null hypothesis built into our prior is, again, data-driven [31] and is key to both computational efficiency in finding the iterative fixed-point and to overcoming sparse sampling.

The conceptual symmetry of weighting data sequences to determine the likelihood of the data given a model through the Shannon entropy of the weighted data may appeal (only) to theorists. The derivative expansion we used was computationally convenient, but we could use convex optimization techniques to define the Legendre transform using the full Legendre-Fenchel definition [41]:

$$E_{\text{eff}}(F) = \sup_J (J \cdot F - W(J)). \quad (19)$$

It is crucial to note here that convexity is a guaranteed property in our approach, not a matter of choice, and it is even useful as we saw in the separation of the two subfamilies of the *Maf_Ham1* proteins and digit sequences (figure 6).

To compare and contrast our approach with common Bayesian approaches, expectation maximization (EM) [42] is an iterative algorithm for finding the maximum likelihood parameters of model or latent variables that maximize the probability of the dataset. EM requires no model prior for its interleaving of expectation and maximization steps. MaxEnt model priors attempt to be maximally unbiased or uninformative, though biologically appropriate MaxEnt priors are harder to define than priors in situations where symmetry principles are available. We have no maximization step in our iteration and cannot work without the null model prior that explicitly precludes correlations, thus going beyond MaxEnt. Another interesting approach, variational bayes [43], works with possible latent variables and analytically computable approximate priors, and uses convexity in some incarnations. We introduce no latent variables as our approach is completely data-driven.

A recent overview of applications of statistical physics to protein structure is given in Cocco *et al* [15]. Here, we concentrated on deducing general properties of the protein family. The Taylor expansion of the effective potential may converge only for deformations in the low-energy modes directions so our approach may not apply directly for deducing contact maps for specific proteins but could, however, be used as a basis for generating heuristic starting points for parameter-dependent optimization of specific residue-residue interactions. Our approach is very general and we could even apply it to sequences such as in hydrophobic–hydrophilic simplifications of actual protein residues, or to variants that group amino-acids according to other properties of interest. This may be potentially of interest in cases where few sequences are available.

There are differences between our approach and the approaches in the literature that pertain to the analysis of protein structure, allostery, and function, taking sequence features into account. Comparing to the parameterized approach [13–15, 23, 24] with priors on parameter values, we did not introduce any explicit parameters for interactions between residues. Instead, we introduced a weighting factor $\exp(J \cdot \sigma)$ and a Bayesian prior on expectation values of residue frequencies. The interaction parameters between residues appear in our approach as expansion coefficients in a derivative expansion of the non-parametrically inferred $E_{\text{eff}}(F)$, and are therefore not limited to just quadratic residue interactions, but could be computed for arbitrarily high-order residue interactions. Such higher order interactions have been shown to be important for protein structure [44–46], but are impossible to compute in a parameter-optimization approach as the number of parameters grows too rapidly. We are continuing work on our approach to be able to make similar predictions on native contacts. In contrast to approaches motivated by analogies with random matrices and finance [16, 21, 22], we weighted sequences with weights determined according to a null Bayesian model prior, and computed the effective action at the maximum of the posterior. The quadratic term in a derivative expansion of this $E_{\text{eff}}(F)$ is, in fact, the inverse of a two-point correlation matrix computed from the weighted sequences. As such, the eigenmodes associated with this matrix have a direct interpretation as directions in sequence space where the Shannon entropy decreases the least. It is important to compare our results with [21, 22] but we have been unable to implement an analog of the binary approximation in our approach in order to compare with their experimentally tested results.

As we noted in our example of the MNIST digit images, the sequence analysis recovered features of two-dimensional contiguity that were not evident in the sequences. A direction for future research is to use our sequence analysis applied to proteins to help determine three-dimensional structures or the role of these sequence-space deformations in protein allosteric function, as shown in [18–20]. A longer-term goal would be to determine an elastic network model from the low-lying frequency model E_{eff} fluctuation expansion. In principle, the results of [18–20] suggest that this is not impossible.

Our method works for a wide spectrum of dataset sizes. We had 6000 sequences for each digit, but for the intimin/invasin and the two subfamilies of the *Maf*-*HamI* proteins we computed potentials with only 146, 67 and 70 sequences, respectively. We showed that the eigenvalues of the second derivative matrix of the effective potential, D_2 , approach constants as the inverse of the logarithm of the number of sequences analyzed. The corresponding geometric entropy flows for the isolated lowest eigenmodes of D_2 identified isolated structural deformations made possible by covariation of amino-acids at distinct positions in the sequence. The large n_{seq} limit of the correlations implied that the greatest spread in the asymptotic probability distribution held for the actual sequences. While surprising at first glance, this is a natural consequence of functional restrictions that lead to compensatory covariation, whether in handwriting or in protein function. The off-diagonal interactions in the effective potential allow more functional variation for actual sequences than is possible for the diagonal interactions of sequences that have been permuted to remove covariation. The appearance of flat directions in an effective potential is associated with degenerate ground states, symmetry breaking and Goldstone modes in statistical physics. Here, the flatter directions of the effective potential are associated with protein sequences that are

structurally almost degenerate, and are related by covariation along a few specific directions in the space of sequences. The isolated eigenvalues are indicative of rigid modes, as a string-like configuration space would have a continuum of eigenvalues. No single sequence position can capture the ‘8-ness’ of a sequence of symbols associated with the digit ‘8’, yet symbol entropy flows along modes capture the interactions around protein binding pockets as well as the slant differences in handwriting.

MD simulations of protein activity are hampered by the computational difficulty of sampling transient high-energy states. These motions are predicted to occur in micro-second to milli-second ranges, can be functionally important, and are difficult to model with atomic-scale MD. We used no structural information so the discovery of isolated structural elements in protein structures from sequence alignments using our method may help overcome some of these difficulties in MD by suggesting elements capable of independent motion [47]. Pan *et al* [48] computed a residue-specific connectivity matrix (RSC) for *E. coli* DHFR and provided evidence that these pair-wise couplings measure the magnitude and extent of mutational effect propagation. MI found in our approach for DHFR is remarkably similar to RSC. The characteristics of E_{eff} determined by our cross-sectional analysis of a protein family that are relevant for the folding kinetics of proteins in that family is a central question that we hope to address in future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the Intramural Research Program of the National Institutes of Health, NIDDK and NCBI. We are grateful to C Chow and A Sherman for helpful discussions.

References

- [1]. Lesk AM and Chothia C 1980 J. Mol. Biol. 136 225–70 [PubMed: 7373651]
- [2]. Altschuh D, Lesk A, Bloomer A and Klug A 1987 J. Mol. Biol. 193 693–707 [PubMed: 3612789]
- [3]. Lockless SW and Ranganathan R 1999 Science 286 295–9 [PubMed: 10514373]
- [4]. Kern D and Zuiderweg ER 2003 Curr. Opin. Struct. Biol. 13 748–57 [PubMed: 14675554]
- [5]. Dima RI and Thirumalai D 2004 Bioinformatics 20 2345–54 [PubMed: 15073020]
- [6]. Chang HW, Bacallado S, Pande VS and Carlsson GE 2013 PLoS One 8 e58699 [PubMed: 23565139]
- [7]. Lane T, Shukla D, Beauchamp K and Pande V 2013 Curr. Opin. Struct. Biol. 23 58–65 [PubMed: 23237705]
- [8]. Henry E, Best R and Eaton W 2013 Proc. Natl Acad. Sci USA 110 17880–5 [PubMed: 24128764]
- [9]. Kass I and Horovitz A 2002 Proteins: Struct. Funct. Bioinformatics 48 611–7
- [10]. Goodey NM and Benkovic SJ 2008 Nat. Chem. Biol 4 474–82 [PubMed: 18641628]
- [11]. Daily MD and Gray JJ 2009 PLoS Comput. Biol 5 e1000293 [PubMed: 19229311]
- [12]. Del Sol A, Tsai CJ, Ma B and Nussinov R 2009 Structure 17 1042–50 [PubMed: 19679084]
- [13]. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R and Sander C 2011 PLoS One 6 e28766 [PubMed: 22163331]
- [14]. Marks DS, Hopf TA and Sander C 2012 Nat. Biotechnol 30 1072–80 [PubMed: 23138306]
- [15]. Cocco S, Monasson R and Weigt M 2013 PLoS Comput. Biol 9 e1003176 [PubMed: 23990764]

- [16]. Stüel GM, Lockless SW, Wall MA and Ranganathan R 2003 *Nat. Struct. Mol. Biol.* 10 59–69
- [17]. Dima RI and Thirumalai D 2006 *Protein Sci.* 15 258–68 [PubMed: 16434743]
- [18]. Zheng W, Brooks BR and Thirumalai D 2006 *Proc. Natl Acad. Sci.* 103 7664–9 [PubMed: 16682636]
- [19]. Zheng W, Brooks BR and Thirumalai D 2007 *Biophys. J.* 93 2289–99 [PubMed: 17557788]
- [20]. Zheng W, Brooks BR and Thirumalai D 2009 *Curr. Protein Peptide Sci* 10 128 [PubMed: 19355980]
- [21]. Halabi N, Rivoire O, Leibler S and Ranganathan R 2009 *Cell* 138 774–86 [PubMed: 19703402]
- [22]. Reynolds KA, McLaughlin RN and Ranganathan R 2011 *Cell* 147 1564–75 [PubMed: 22196731]
- [23]. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T and Weigt M 2011 *Proc. Natl Acad. Sci. USA* 108 E1293–301 [PubMed: 22106262]
- [24]. Morcos F, Jana B, Hwa T and Onuchic JN 2013 *Proc. Natl Acad. Sci. USA* 110 20533–8 [PubMed: 24297889]
- [25]. Tsai CJ, Del Sol A and Nussinov R 2009 *Mol. Biosystems* 5 207–16
- [26]. Nussinov R, Tsai CJ and Ma B 2013 *Annu. Rev. Biophys* 42 169–89 [PubMed: 23451894]
- [27]. Aow J, Dore K and Malinow R 2015 *Proc. Natl Acad. Sci USA* 112 14711–6 [PubMed: 26553983]
- [28]. Stein RR, Marks DS and Sander C 2015 *PLoS Comput. Biol* 11 e1004182 [PubMed: 26225866]
- [29]. Kiss G, ÇelçbiÖlcüm N, Moretti R, Baker D and Houk K 2013 *Angew. Chem. Int. Ed.* 52 5700–25
- [30]. Micheletti C 2013 *Phys. Life Rev* 10 1–26 [PubMed: 23199577]
- [31]. Efron B and Morris C 1973 *J. Am. Stat. Assoc.* 68 117–30
- [32]. Shreif Z, Striegel DA and Periwal V 2015 *J. Theor. Biol.* 380 399–413 [PubMed: 26092377]
- [33]. Landau L and Lifschitz E 1980 *Statistical Physics* (Oxford: Pergamon)
- [34]. Zinn-Justin J 2002 *Quantum Field Theory and Critical Phenomena* (New York: Oxford University Press)
- [35]. Suzek BE, Huang H, McGarvey P, Mazumder R and Wu CH 2007 *Bioinformatics* 23 1282–8 [PubMed: 17379688]
- [36]. Fairman JW, Dautin N, Wojtowicz D, Liu W, Noinaj N, Barnard TJ, Udho E, Przytycka TM, Cherezov V and Buchanan SK 2012 *Structure* 20 1233–43 [PubMed: 22658748]
- [37]. Chakrabarti S, Bryant SH and Panchenko AR 2007 *J. Mol. Biol.* 373 801–10 [PubMed: 17868687]
- [38]. Lecun Y, Bottou L, Bengio Y and Haffner P 1998 *Proc. IEEE* 86 2278–324
- [39]. Page MJ and Di Cera E 2008 *Cell. Mol. Life Sci.* 65 1220–36 [PubMed: 18259688]
- [40]. Oberhettinger P, Schütz M, Leo JC, Heinz N, Berger J, Autenrieth IB and Linke D 2012 *PLoS One* 7 e47069 [PubMed: 23056583]
- [41]. Lucet Y 1996 *Comput. Optim. Appl* 6 27–57
- [42]. Dempster A, Laird N and Rubin D 1977 *J. R. Stat. Soc. B* 39 1–38
- [43]. Friston K, Mattout J, Trujillo-Barreto N, Ashburner J and Penny W 2007 *NeuroImage* 34 220–34 [PubMed: 17055746]
- [44]. Godzik A, Kolinski A and Skolnick J 1992 *J. Mol. Biol.* 227 227–38 [PubMed: 1522587]
- [45]. Krishnamoorthy B and Tropsha A 2003 *Bioinformatics* 19 1540–8 [PubMed: 12912835]
- [46]. Deutsch C and Krishnamoorthy B 2007 *Bioinformatics* 23 3009–15 [PubMed: 17921497]
- [47]. Ho BK and Agard DA 2009 *PLoS Comput. Biol* 5 e1000343 [PubMed: 19343225]
- [48]. Pan H, Lee JC and Hilser VJ 2000 *Proc. Natl Acad. Sci. USA* 97 12020–5 [PubMed: 11035796]

Figure 1.

Construction of a sequence from an image. (a) Every black/white combination of a 2×2 square. Each individual combination is associated with a letter, *A–P* going across. (b)–(d) An example of a 3×3 -pixel image that is reconstructed vis-*vis* the alphabet shown in (a). There are four 2×2 -pixel subunits (c) centered around the four red dots in (b). Each subunit is translated to a letter from (a) and is shown in (d). (e) and (f) Reconstruction of the digit 5 with respect to the alphabet in (a).

Figure 2.

Eigenvalue versus $1/\log n_{\text{seq}}$ for DHFR sequences (a), position-wise permutations of DHFR sequences (b), and sequence and position permuted DHFR sequences (c). The number of sequences was varied by picking five random selections of subsets of sequences of the appropriate cardinality, and taking the mean of the eigenvalues computed from these subsets.

Figure 3.

Entropy flows along the first three eigenvector directions (columns 1–3, respectively) and image samples (column 4) for each digit. Note: (a) most digits (except for ‘0’) show a tilt deformation to the right in column 1; (b) the digit ‘1’ tilts (with much smaller amplitude) left (column 3); (c) the width of most digits (except for ‘0’ where columns 1 and 2 show the interchanged variations) is the variation that appears in column 2; (d) emphasized endings on ‘8’, ‘5’, ‘2’ are evident in column 3.

Figure 4.

Entropy flows along the first three eigenvector directions of DHFR (projected onto DHFR chicken structure, PDB 1DR1) identify structural changes in the folate binding pocket (a), the G-H loop (b), and the junction between the α_4 and α_6 helices (c). Spheres indicated the ten highest, positive (red) and ten lowest, negative (blue) values. Notice the structural differences found in *E. coli*, PDB 1DRE (top insert), chicken (middle insert) and human, PDB 1MVS (bottom insert) DHFR for each mode.

Figure 5. Entropy flow along the first eigenvector direction of the intimin/invasin family of virulence factors. Front (a), cut-through (b) and back (c) views of the intimin/invasin β barrel, PDB 4E1S. Red positions are losing symbol entropy and blue positions are gaining entropy, relative to the most likely model (or vice versa due to linearity). The symbol entropy changes are consistent with β -barrel shear.

Figure 6.

Convexity of E_{eff} and clustering of dissimilar sequences, (a) Each *Maf_Ham1* sequence projected onto the first three modes. The separation between subfamily 1 (blue circles) and 2 (red triangles) is visible, (b) Digit sequence data projected onto the first two modes (green) and the convex effective potential function (blue) approximating the data, (c) For Images '1' and '7' each sequence was projected onto the first two modes of the calculated effective potential. Color represents the percentage of sequences for every projection combination. Blue and red histograms represent the distributions for the first and second modes, respectively.

Figure 7.

The effective potentials of *Maf_Ham1* subfamilies projected onto protein structures. The effective potential was calculated for subfamily 1 + 2, subfamily 1, and subfamily 2. $\widehat{\mathcal{L}}_V$ for each effective potential was projected onto human inosine triphosphotase (PDB 2J4E), a member of subfamily 1 (a)–(c), and *B. subtilis* Maf protein (PDB 1EXC), a member of subfamily 2 (d)–(f). Notice how the first mode corresponds to the binding pocket in (b) and (f).

Figure 8.

Ball-and-stick representations on the dimerized NTPase structure indicate residues belonging to a high MI value pair. These pairs and their inter- and intra-chain distances are given in table S1.