Approval for publication Signed ______________________ Date ____________ Number of amended pages returned ____________

Appl Bioinformatics 2004; 3 (2): 1
1175-5636/04/0099-0001/$31.00/0

Inferring Property Selection Pressure from Positional Residue Conservation

Rose Hoberman,[1] *Judith Klein-Seetharaman*[1,2] and *Roni Rosenfeld*[1]

1 School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
2 University of Pittsburgh Medical School, Pittsburgh, Pennsylvania, USA

Abstract

This study attempts to understand and explain positional selection pressure in terms of underlying physical and chemical properties. We propose a set of constraining assumptions about how these pressures behave, then describe a procedure for analysing and explaining the distribution of residues at a particular position in a multiple sequence alignment. In contrast to previous approaches, our model takes into account both amino acid frequencies and a large number of physical-chemical properties. By analysing each property separately, it is possible to identify positions where distinct conservation patterns are present. In addition, the model can easily incorporate sequence weights that adjust for bias in the sample sequences. Finally, a measure of statistical significance is provided for our conservation measure((**Author: please suggest another word to replace "measure", which is used twice in this sentence**)). The applicability of this method is demonstrated on two HIV-1 proteins: Nef and Env. The tools, data and results presented in this article are available at http://flan.blm.cs.cmu.edu.((**Author: unable to access this website. Please check URL**))

Introduction

Motivation

Multiple sequence alignments (MSAs) contain a wealth of information about the structure, function and evolution of proteins. In particular, residue conservation has long been associated with functional and structural significance. MSAs therefore provide important guidance in designing laboratory experiments such as studying the effects of site-directed mutants at the conserved sites. Due to the explosion in the amount of protein data that are available, quantitative measures of positional conservation have been devised. Measures such as entropy are used to systematically evaluate the distribution of amino acids in each column in an MSA, assigning a numerical value to each position to quantify its degree of conservation. These quantitative measures have been used to identify functionally or structurally important residues and to predict the structure and function of the entire protein.[1-3] Other uses for these measures include building motif-based models of protein families or domains.[4] Measures of conservation have also been used to detect functional surfaces, such as active sites, ligand-binding sites or protein–protein interaction domains.[5-7] Finally, conservation measures can provide a way to evaluate and refine MSAs.[8]

While many measures have been proposed for quantifying positional conservation (see Valdar[9] for a review), identifying positional conservation is only a first step. A more difficult task is to understand the *specific* selective pressures that have influenced the distribution of residues at each position. Different positions have different functional and structural roles, and thus the amino acid in each position will be subject to different constraints. Biologists often manually analyse an MSA by using software packages that colour positions in an alignment according to the types of amino acids present. These packages use simple heuristics to determine the colour of each position, such as grouping amino acids by one or a few properties, and colouring the column according to the property that is most often represented. However, these methods cannot be used in any systematic way. They consider only a few possible conservation patterns and do not provide any statistical methodology for testing whether the apparent conservation of a particular property in a position is statistically significant.

The goal of this study is to develop a methodology for identifying and quantifying the selection pressures acting on amino acid properties in each position in an MSA. We propose a model that analyses the residue distribution at a particular position in an MSA and explains sequence conservation in terms of selective pressures on a specific set of underlying physical and chemical properties. Our method requires only an MSA as input and determines the most highly conserved properties in each position, as well as their statistical significance.

Related Work

The most common measure of positional conservation is entropy (Sander and Schneider 1991(**(Author: unable to locate this ref in the ref list. Please advise)**)).[10] However, like other methods based purely on symbol frequencies, entropy assumes that each amino acid is an independent symbol with no relation to any of the other symbols. This assumption is clearly invalid for amino acids. Furthermore, an explicit evaluation of an entropy score showed little correlation with structural conservation.[11]

A second class of approaches considers only the physical-chemical properties of each amino acid. These methods[12] (Zvelebil 1987(**(Author: unable to locate this ref in the ref list. Please advise)**)) classify amino acids based on their position in a Venn diagram of overlapping sets representing different physical and chemical properties. For each aligned position, such a method finds the smallest set of properties that explains the observed amino acids at that site. Since the number of combinations of properties is large, only a limited number of combinations is considered. The *ad hoc* selection of a subset of combinations is necessarily subjective, and limiting. In addition, these methods only work with binary properties and, perhaps more importantly, fail to account for amino acid frequency. For example, they do not distinguish between a position with one leucine and 100 arginine from a position with 100 leucine and 100 arginine.

A third hybrid class of approaches tries to derive entropy scores that are sensitive to the physical and chemical relationships between amino acids.[9] These methods divide the 20 amino acids into a small number of physically or chemically related clusters (about 5–10) and calculate entropy on this reduced symbol set. Again, this clustering is *ad hoc*, and by choosing a handful of clusters, the number of properties modelled is necessarily limited. In addition, each amino acid in a cluster is considered to be uniformly distant from the amino acids in every other cluster. Consequently, even the properties that are used for clustering are modelled at only a very coarse level.

More sophisticated approaches use mutation-based substitution matrices to quantify the physical-chemical similarity between amino acids.[9] The values in these matrices are derived by averaging over many positions and types of proteins, which can at best capture the selection pressures corresponding to the most dominant properties. While this sort of averaging yields a reasonable measure of *overall* conservation, it will blur specific pressures, making it difficult to distinguish the effect of one property over another. So, although calculating a similarity measure for each pair of amino acids is more informative than a simple clustering approach, it will inevitably ignore similarities based on particular properties that are only infrequently selected for. Furthermore, amino acids may be similar with respect to one property, but not with respect to another. Since different positions select for different properties, certain amino acids are more likely to be aligned at one position than another. However, any approach based on a single similarity matrix cannot represent such position-dependent distinctions.

Position-specific patterns of amino acid conservation have been estimated for use in building better profile hidden Markov models (HMMs) of protein families(**(Author: please confirm that rewording is ok)**). One step in building a profile HMM is to learn the probability of observing each amino acid at a given site in the protein. When the number of training sequences is small however, the observed frequency of each amino acid is a poor estimate of the true probability. To improve these estimates, a regularisation method has been developed based on mixtures of Dirichlets.[13] A small number of Dirichlet distributions and mixing weights are estimated from a large set of protein alignments. These distributions often appear to be correlated with physical and chemical properties such as size, hydrophobicity and charge. Nevertheless, the mixture of Dirichlets is used merely as a prior to smooth the posterior probability estimates, not as a means of explaining the observed amino acid counts. Given a set of Dirichlets, one can calculate the posterior probability that each of them generated the position-specific conservation pattern from which the observed amino acids were sampled. Although the Dirichlet assigned the highest probability can be said to provide the best explanation of the data, no statistical test is provided to determine if that Dirichlet provides a significant fit to the data. In addition, the number of components must be preselected, and even once the number is fixed there is no guarantee that the components learned are optimal.

Instead of deriving a similarity matrix indirectly from mutation data, there have been attempts to make direct use of a large set of continuous, experimentally determined properties. These include methods that build a similarity matrix directly from physical-chemical properties and methods that use dimensionality reduction to reduce the large number of properties to a small set of prototypical ‘aggregate’ properties. Because of their aggregation of multiple properties, these direct approaches suffer from similar weaknesses to those based on substitution matrices. One specific model

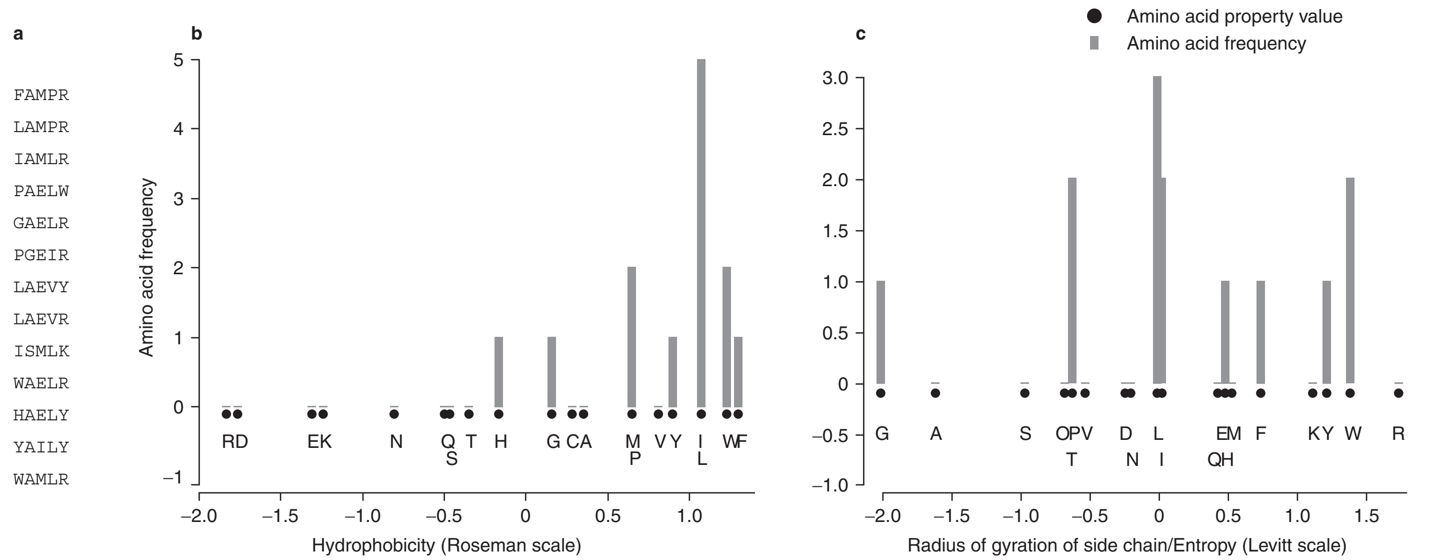

Fig. 1. Frequencies of amino acids in a specific position in the alignment. (**a**) A small sample alignment. Distribution of residues in the first position of the sample alignment are plotted against two different scales: (**b**) hydrophobicity; and (**c**) radius of gyration of side chains.**((Author: please confirm that rewording on the figure and in figure caption is ok))**

of this type is the exponential fitness model of Koshi, Mindell and Goldstein.[14] This method not only aggregates multiple columns but uses only two hybrid properties, corresponding roughly to hydrophobicity and size.

All of the previous approaches((**Author: do you mean the approaches described above?**)) are based solely on the identities of the amino acids observed at each position in the alignment. However, the likelihood of seeing a particular amino acid at a particular position depends not only on the fitness of that amino acid, but on its mutational distance from the ancestral amino acid. As a result, any method based solely on observed amino acid frequencies will unintentionally confound mutation effects with fitness. Yampolsky and Stoltzfus (1998)((**Author: unable to locate this ref in the ref list. Please advise**)) have attempted to decouple these two effects. A measure of amino acid exchangeability has been derived based solely on fitness measurements obtained by experimental replacement of amino acids. This article, however, is based on learning a global exchangeability matrix and, thus, cannot make predictions in a position-specific manner.

Method

In contrast with previous methods, our model takes into account both amino acid frequencies and a large number of experimentally derived physical-chemical properties. Our approach is position- and property-specific: the degree of conservation of each property is tested independently at each position. We do not aggregate over many positions, and therefore our method can detect property constraints specific to single positions. We do not aggregate multiple properties into a similarity matrix or a small set of prototypical properties; our method can therefore identify conservation of even those properties that only occasionally affect fitness.

To maximise the sensitivity of our conservation test, we use continuous instead of binary properties. Each property scale assigns a numerical value to each of the 20 amino acids. Frequencies of amino acids in a specific position in the alignment can be mapped directly into frequencies along the property scale. For example, the amino acid frequencies in the first position of the alignment shown in figure 1a are also shown plotted against two different properties (figure 1b and figure 1c). In figure 1b these frequencies are mapped onto the (normalised) Roseman hydrophobicity scale. Figure 1c shows the same frequencies mapped onto the Levitt scale, which measures the radius of gyration of side chains. It is not difficult to see from these graphs that hydrophobicity is more conserved in this position than size.

Two Simple Measures of Property Conservation

A simple way to identify conserved properties is to look for properties with low variance in particular positions in a protein family. More specifically, let N((**Author: could this be changed to lower case "n"? [if so, we will also change it in the equations]**)) be the number of sequences in the alignment and $v_{k,i}$ be the numerical value assigned to amino acid i by the k-th property scale. If $n_{i,j}$ is the number of times amino acid i occurs in column j in the alignment, then the mean and variance of property k in the j-th position are:

$$\mu_{k,j} = \frac{1}{N}\sum_{i=1}^{20} n_{j,i} * v_{k,i}$$

$$\sigma^2_{k,j} = \frac{1}{N-1}\sum_{i=1}^{20} n_{j,i} * \left(v_{k,i} - \mu_{k,j}\right)^2$$

Since property scales are measured in a variety of different units, the expected variance for a random sample of amino acids can differ substantially for different properties. An observed variance of 0.5 along a scale that ranges from 0 to 100 is not equivalent to the same variance measured on a scale that ranges from −1 to 1. To ensure that variances of different scales are comparable, each scale is normalised to have unit variance.

Even after this normalisation however, low variance in itself does not provide sufficient evidence that a property exerts pressure on a particular position. If the entropy is low (i.e. only a few different amino acids are present in this position), then *many* properties will have low variance. In the extreme case of only a single amino acid present, *all* properties will have zero variance. In addition, a variance-based measure is inadequate because it fails to penalise for residues that have not been observed. Figure 2 illustrates this point for an example protein. Shown is the distribution of amino acids at position 85 of the HIV-1 negative factor protein (Nef), plotted on a scale that quantifies the normalised frequency of a bend. Four amino acids occur in this position (I, V, L and F). All four of these amino acids have a similar frequency of occurring in a bend region of a protein, with the result that this property has low variance in this position. Nonetheless, the striking absence of cysteine raises a red flag. This 'hole' causes us to doubt whether this property is actually being selected for. Thus, we make an additional assumption: when selection is based on a single property, fitness is unimodal (i.e. holes constitute negative evidence).

This new assumption suggests another very simple strategy: look for properties where all the residues at a particular position are adjacent in property space. Adjacency requires that if two residue types both occur in a particular position of the MSA, then any residue that has a property value lying between their property values must also occur in that position. Thus, properties for which all the amino acids in a specific position are adjacent will be considered significantly conserved. Clearly, for positions with only a few amino acids, this method will have a high false discovery rate. A position with only a single residue will appear equally conserved with respect to almost all properties. Even two-residue positions will appear to be equally conserved with respect to many properties. The false discovery rate for a position with only i amino acids can be quantified precisely – it is just the probability that these i amino acids are adjacent according to a random property:

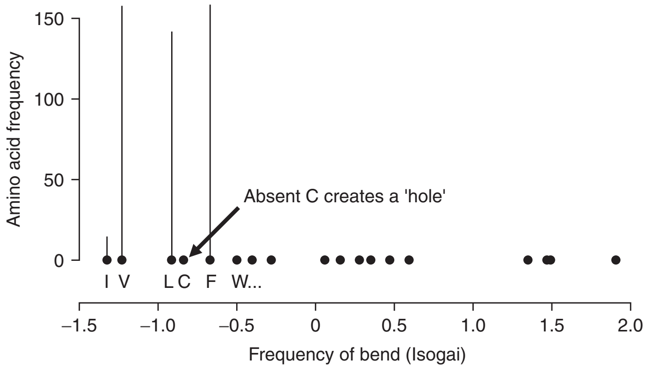

Fig. 2. Distribution of residues in position 85 of HIV-1 Nef protein plotted against a scale that quantifies the normalised frequency of a bend.**((Author: what do the filled dots and thin vertical bars in the figure represent? Also, what is "Isogai" in the x axis label?))**

$$\frac{(21-i)}{\binom{20}{i}}$$

Thus, the false discovery rate is 10% for a position with only two amino acids present, 1.6% for three amino acids, 0.35% for four and 0.1% for five.

Gaussian Fitting

Unlike variance, the adjacency measure does not explicitly evaluate the degree of dispersion of the property values of residues at a specific position. Furthermore, it only considers whether an amino acid is present or absent. This leads to a strict definition of a hole and a subsequent failure to reject property distributions that are clearly not unimodal. For example, in figure 2, if cysteine was not completely absent in this position, but occurred in only one or two sequences, then the adjacency method would view this amino acid as 'present' and fail to detect the very clear presence of a hole.

Adjacency measures also ignore the distances between the property values of different amino acids. For instance, consider the two property plots of the HIV-1 Nef protein at position 85 (figure 3). The two graphs show the same data plotted against two different property scales: bulkiness (figure 3a) and normalised hydrophobicity (figure 3b). In both cases, all the observed residues are adjacent (in figure 3a, I and L have *identical* property values). Nevertheless, in figure 3a, V is closer to W than it is to any of the other amino acids occurring at this position (F, I and L). Given the relative distances between these amino acids, the total absence of W in this alignment position makes it doubtful that bulkiness is truly conserved, but does not affect our confidence in the conservation of hydrophobicity.

We would like our model to take these relative distances into account, and so we propose using a Gaussian distribution to model property conservation. A Gaussian is a good choice because it is

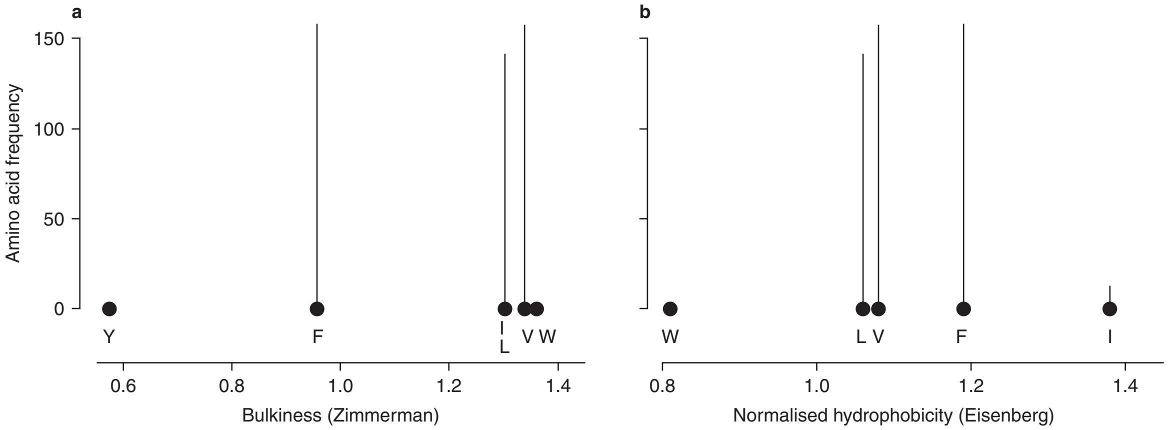

Fig. 3. Distribution of residues in position 85 of HIV-1 Nef protein plotted against two different property scales: (**a**) bulkiness and (**b**) normalised hydrophobicity.**((Author: please confirm that rewording in figure caption is ok. Also, what do the filled dots and thin vertical bars in the figure represent? What are "Zimmerman" and "Eisenberg" in x axis labels referring to))?**

unimodal and will thus penalise heavily for missing amino acids with property values near the mean value. In addition, distributions such as that in figure 3a will score badly because the Gaussian density decreases steeply only when the variance is small.

Given the amino acid frequencies for a position in the MSA, for each property we could select the parameters of the Gaussian that give maximum likelihood to the data. However, maximum likelihood estimation is not robust: a single amino acid with a property value far from the mean will result in fitting a Gaussian with a large variance. In addition, since the space of possible observations is not continuous, but discrete (with at most 20 values), the sample mean and covariance are not actually the maximum likelihood estimates. Indeed, when the space of possible observations is discrete and fixed, the maximum likelihood estimates for the parameters cannot be found analytically. Consequently, we smooth the likelihood function by interpolating it with a uniform background distribution over the 20 amino acids, then use a simple heuristic search procedure to efficiently select the parameters of the Gaussian to fit the observed data.

To measure property conservation we then apply a goodness-of-fit test to the learned model. We can calculate the number of times we expect to observe amino acid i in column j:

$$E_j[i] = (1-\lambda)\frac{f_i}{\sum_{k=1}^{20} f_k} + \lambda\left(\frac{1}{20}\right)$$

where

$$f_i = \exp\left(\frac{(v_{k,i}-\mu)^2}{-2\sigma^2}\right)$$

where μ and σ are the parameters of the Gaussian, λ is the interpolation weight of the background distribution[1] and $v_{k,i}$ is the numerical value assigned to amino acid i by the k-th property scale. These expected counts are compared with the observed counts with a χ^2 goodness-of-fit test. We don't expect an absolute fit of the data to a Gaussian, but we can use the relative fit to rank different hypotheses.

Significance Testing

Clearly, it is not difficult to fit a Gaussian to the amino acid frequencies if only a few of them are non-zero. Consider, for example, a position with almost perfect conservation. The best Gaussian for each property will have a very small variance and a mean centred at the dominant amino acid. As a result, almost every property will fit the data equally well. In these cases, we will have a high false discovery rate. We therefore designed a Monte Carlo-type significance test that determines which goodness-of-fit values are statistically significant. The procedure to determine the significance of a property in a particular position is described as follows.

A large set of 'pseudo properties' is generated by randomly permuting the numerical vectors corresponding to the physical-chemical properties. A Gaussian is then fitted to each of the randomised properties and its χ^2 goodness-of-fit statistic determined. Cumulatively, these χ^2 values thus provide an estimate of the empirical distribution of the χ^2 values for each position under the null (= randomness(**(Author: could this say just "(randomness)" or do you mean "null hypothesis = randomness"?)**)) assumption. This distribution is then used to convert the χ^2 statistic of the original physical-chemical property into a significance score, which plays the role of an empirically estimated p-value.

[1] The weight of the uniform background component was set to 0.05 for these experiments.

This score approximates the probability of obtaining a χ^2 value smaller than the observed value by chance – small scores represent a highly significant property. Unlike the original χ^2 statistics, these significance scores are comparable across positions with significantly different entropies.

Since we are conducting a large number of tests at each position in the MSA, any p-value threshold should be corrected to account for the number of comparisons being performed. For instance, if we want the experiment-wide chance of a type I error to be <5%, then a Bonferroni correction for multiple testing (of 240 properties) yields a significance threshold of $\alpha = 0.05/240 = 0.0002$. The Bonferroni correction is an overly conservative adjustment, especially since there are significant correlations between many of the properties we are testing.[15]

Indeed, it is often the case that, for a single site, the p-values of multiple property scales are significant. In this case, we report all of them as likely alternative explanations of the observed positional conservation. In most cases, these scales will be highly correlated, as they are often just different experimental measurements of the same underlying property. When this is not the case and the scales truly measure different properties, our results can give guidance in designing mutation experiments to fully determine the physical or chemical constraints imposed on the amino acid at this position in the protein.

Adjusting for Sequence Bias

Sequences in an MSA are not independent. In most cases, they have diverged from a common ancestor. In other cases, the sampling of sequences is biased toward certain subfamilies. As a result, a typical alignment contains clusters of similar sequences. If these related clusters are large, they can significantly bias our estimates of amino acid frequency, effectively obscuring the variability exhibited by the less common subfamilies.

Many different methods have been proposed for reducing this bias, the majority of which assign a numerical weight to each sequence, with the goal of down-weighting closely related sequences. Although there are many algorithms for determining sequence weights (see Durbin et al.[16] for a review), most measures of positional conservation provide no logical way to incorporate these weights. Since our methods are based solely on the frequency of each amino acid in each position, it is quite simple to incorporate sequence weights. We merely weight each position by its corresponding weight before counting frequencies. If $W_i(a_j)$ is the cumulative weight of all sequences that have amino acid a_j in position i, the weighted frequency of the j-th amino acid is:

$$f_i\left(a_j\right)=\frac{W_i\left(a_j\right)}{\sum_{k=1}^{20} W_i\left(a_k\right)}$$

The tree-weighting algorithm of Gerstein/Sonnhammer/Chothia (as implemented in ClustalW[17]) was used in this analysis. This algorithm first constructs a guide tree using pairwise distances between sequences and then assigns weights according to the branch lengths and local density of the tree. The method essentially works by down-weighting those sequences that appear to have diverged from each other recently.

Sequence weighting methods are generally simple, fast to compute, and effectively adjust for sample bias. They are not, however, able to fully account for the influence of phylogenetic relationships among the sequences.[18] Fully correcting for phylogenetic correlations requires a more sophisticated approach, and this is discussed further in the section Conclusions and Future Work.

Experiments

Data Sources

A set of 240 property scales was obtained from the online databases PDbase (http://www.scsb.utmb.edu/compbiol.html/venkat/prop.html **((Author: unable to access this URL. Please advise)**)) and ProtScale (http://us.expasy.org/cgi-bin/protscale.pl). This set includes both experimentally derived scales, such as average accessible surface area, and scales derived computationally from data, such as relative frequency in an α-helix. Each scale assigns a numerical value to each of the 20 amino acids.

The dataset we used to develop and test our method comprises proteins from HIV, for the following reasons. HIV, like other RNA viruses, has a high rate of mutation. This leads to a large and variable collection of protein sequences that provides a wealth of information about the selective constraints acting at particular positions in the HIV proteins. MSAs for two HIV proteins were downloaded from the HIV Molecular Immunology Database (http://www.hiv.lanl.gov/content/immunology). These alignments include many isolates of each HIV-1 subtype and inter-subtype recombinants, but include only one sequence from any one individual. The alignments were built by scientists at the Los Alamos National Laboratory. They used HMMs and hand editing to construct very accurate alignments. Furthermore, since the functions of the proteins in the alignment are identical, and the sequences are derived from the same organism, the amino acid substitutions in the MSAs are not affected by possible contributions from organism-specific amino acid usage preferences, nor substitutions that slightly modulate function (such as would be the case in protein families where individual members bind different ligands, for example). The envelope protein (Env) alignment was constructed from 388 protein sequences. The alignment for the Nef protein is slightly larger, encompassing 484 sequences. The Nef sequences are about 200 residues long and the Env proteins are over 800 residues long.

A large number of gaps often indicates that a position is not important to the structure or function of the protein. In addition, there is no way to assign property values to gaps. Hence, we analyse only those positions where a majority of sequences do not contain gaps. For efficiency reasons, we only test positions with entropy >0.5, since our method will not be able to identify the cause of conservation in such highly conserved positions. After filtering, there were 145 columns in the Nef alignment and 625 in the Env alignment.

In the following sections**((Author: do you mean for the remainder of this paper?))**, unless otherwise stated, all position numbers refer to the HIV reference proteins with Swiss-Prot IDs NEF_HV112 and ENV_HV1H2.

Results

Applying the Gaussian fitting method to the two HIV protein alignments described in the Data Sources section, we observe that for certain positions the method is clearly identifying properties that fit the data much better than would be expected by chance. The use of randomised properties for significance testing of property conservation is illustrated by figure 4, which shows the empirical cumulative distribution of χ^2 goodness-of-fit statistics for three different positions in the HIV-1 Nef protein.

Figure 4a shows a position for which a small number of properties fit the amino acid distribution much more closely than would be expected by chance. The enlarged detail of this graph in figure 4c illustrates how the distribution of random χ^2 values can be used to establish a significance cutoff. If we want the experiment-wide chance of a type I error to be <5%, then a Bonferroni correction yields a significance threshold of $\alpha = 0.05/240 = 0.0002$, which for this cumulative distribution function yields a critical value of $\chi^2 < 102$. Any property with a goodness-of-fit that is better than this cutoff is highly unlikely to have fitted the data well purely by chance. In this case, there are two (almost identical) property scales that pass this threshold: ‘Percentage of buried residues/Janin’ and ‘Free energy of transfer from inside to outside of a globular protein/Janin’.

Figure 4b and its enlarged detail (figure 4d), on the other hand, show a position where none of the 240 properties achieve a significant fit. There is little difference between the distribution of the physiochemical**((Author: is “physiochemical” the same as “physical-chemical” as is used elsewhere in the manuscript?))** properties and that of random properties. Either this position has very little conservation or the selective pressures are acting on more than one property, with the result that our method fails to detect a single property that is conserved.

In figure 4e, an almost perfectly conserved position is shown. Since almost all properties (including random properties) can be fitted well by a Gaussian, none of the 240 properties achieve a fit that is significantly better than random. Thus, our method is not applicable for highly conserved positions where sequence conservation is a better descriptor for evolutionary pressure than property conservation.

Table I lists some of the most highly conserved properties at specific positions in the Nef and Env proteins. For each of the reported positions, either the property listed in the table is the only property that passes the given significance threshold or it is highly correlated with all other scales found to be significantly conserved at that position.

Discussion

For a qualitative evaluation and interpretation of the results obtained with our method, we analysed in detail the MSAs of the HIV-1 Nef and Env proteins in relation to the existing knowledge about their structure and function. HIV-1 is a lentivirus, which as a retrovirus encodes the prototypic *Gag*, *Pol* and *Env* genes, plus a number of accessory genes including *Nef*.**((Author: please confirm that rewording of last sentence is ok and that the names in italics are indeed genes [house style is to italicise gene names]))**

Property Conservation in Nef Proteins

Nef is a 27-kDa myristoylated protein, observed in the cytoplasm of cells infected by lentiviruses.[19] Nef is involved in viral signalling processes and has been shown to be associated with cellular membranes and the cytoskeleton, depending on myristoylation of the N-terminal glycine (Franchini et 1986; Niederman et al. 1993)**((Author: unable to locate these refs in the ref list. Please advise))**).[20] Cleavage between W57 and L58 separates the protein into two domains:[21] the C-terminal well folded core domain (amino acids 58 to 206), with structural similarities to DNA-binding proteins containing a helix-turn-helix motif,[22,23] and the flexible N-terminal sequence that is proposed to function as a membrane anchor.[21,24] The N-terminal fragment is largely unstructured in solution, but in the presence of myristoylation there are two relatively well defined helices separated by a turn.[25] The structures of fragments corresponding to the two domains obtained by nuclear magnetic resonance (NMR) spectroscopy are shown in figure 5a and figure 5b, for the N-terminal and C-terminal fragments, respectively. The structures are shown in the

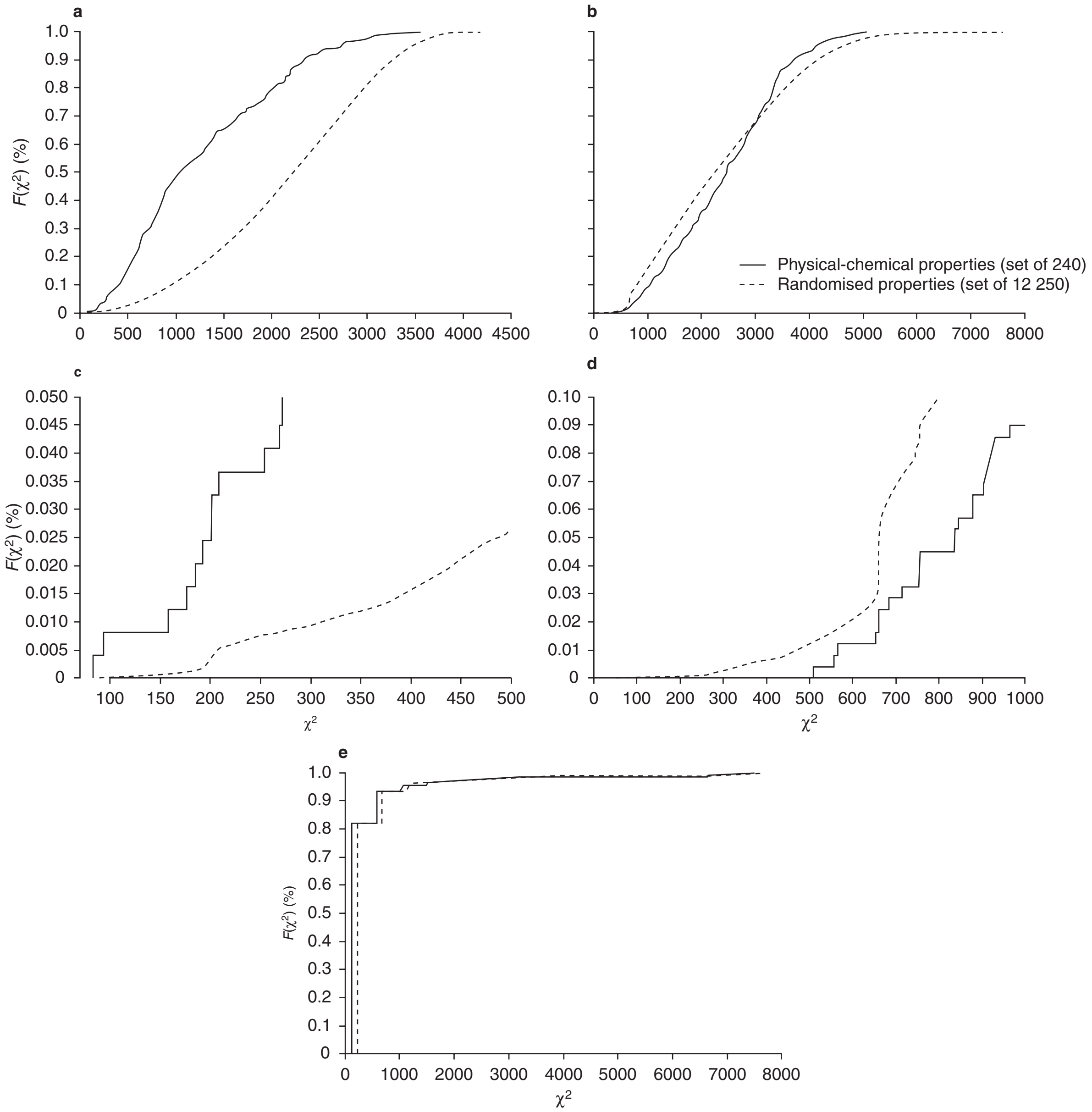

Fig. 4. Empirical cumulative distribution function, *F*, of the χ^2 goodness-of-fit statistic for physical-chemical properties compared with randomised properties for three different positions in the HIV-1 Nef protein. (**a**) Nef position 85; (**b**) Nef position 54; (**c**) Nef position 85 (detail); (**d**) Nef position 54 (detail); (**e**) Nef position 52. The x axis shows the value of the χ^2 statistic, where a value of zero represents a perfect fit to a Gaussian. The y axis shows the percentage of properties whose goodness-of-fit statistic is less than or equal to that value. The subfigures illustrate the significant differences between the distribution of χ^2 values at different positions in the Nef alignment. Subfigure (a) and its detail (c) illustrate a more variable site, with a small number of significantly conserved properties. Subfigure (b) and its detail (d) also illustrate a variable site, but in this position there are no physiochemical**((Author: is "physiochemical" the same as "physical-chemical" as is used earlier in the caption?))** properties with a χ^2 value significantly smaller than would be expected by chance. Subfigure (e) shows a highly conserved site, so the average χ^2 statistic is close to zero, and there are no properties with a χ^2 value significantly smaller than expected under the null model.

predicted correct relative orientation between the two domains in the full-length protein.[26]**((Author: please confirm that rewording of last sentence is ok))**

Table I lists the properties found to be the most significantly conserved at specific positions in the Nef proteins, where the significance level is α = 0.05, and thus the corrected p-value threshold is 0.0002. These positions are highlighted in the structures shown in figure 5 (labelled in black with the amino acid identity and position of the HIV-1 Nef reference protein). Previous structure-function studies have correlated surface accessibility

Table I. Properties found to be most significantly conserved at specific positions in the Nef and Env proteins, where the significance level is $\alpha = 0.05$. The threshold after a Bonferroni correction for multiple testing thus requires a p-value < 0.0002

Protein	Alignment position	Reference position	Conserved property (p-value < 0.0002)**((Author: what do the names in the column below mean? If they are a reference, the full details of each study will need to be included in the reference list, and the reference number added to the names below))**
Nef	17	16	Percentage of exposed residues/Janin
Nef	22	21	Hydropathy index/Kyte-Doolittle
Nef	62	45	Relative mutability/Jones
Nef	63	46	Retention coefficient in HPLC, pH7.4/Meek
Nef	106	85	Percentage of buried residues/Janin
Nef	110	89	Average flexibility indices/Bhaskaran-Ponnuswamy
Nef	119	98	Net charge/Klein
Nef	162	135	Hydrophobicity from HPLC peptide retention times/Wilson
Nef	177	149	Hydrophobicity at pH 3.4 determined by HPLC/Cowan
Nef	214	184	Partition energy/Guy
Nef	223	192	Optimised transfer energy parameter/Oobatake
Env	48	32	Normalised flexibility/B-values
Env	81	63	Flexibility parameter for no rigid neighbours/Karplus-Schulz
Env	203	165	Hydrophobicity/Prabhakaran
Env	212	174	Absolute entropy/Hutchens
Env	356	291	Side chain torsion angle φ/Levitt
Env	427	352	Recognition factors
Env	442	363	Long-range nonbonded energy per atom/Oobatake-Ooi
Env	595	494	Polarity/Grantham
Env	739	620	Hydrophobic parameter π/Fauchere-Pliska

Env = envelope protein; **HPLC** = high performance liquid chromatography; **Nef** = negative factor protein.

with secondary structure elements and sequence conservation.[27] It was found that conserved motifs involved in Nef-mediated CD4**((Author: should CD4 here, and all instances of CD4 in this article, be CD4+?))** and MHC**((Author: please define))** I downregulation are located in flexible regions of the Nef protein, suggesting that the formation of the transient trafficking complexes depends on the recognition of primary sequences. In contrast, the interaction sites for signalling molecules that contain SH3 domains or the p21-activated kinases are associated with the well folded core domain, suggesting the recognition of highly structured protein surfaces.[27] Critical for the folding of the C-terminal domain are residues W141 and W183, which form the scaffold of the protein and assemble the two core helices to the β-pleated sheet (figure 5).[26] As can be seen, several of the amino acids with conserved properties are located in close proximity to these two amino acids, which are highlighted in red in figure 5b.

An interesting observation arising from our analysis is that a single property, the retention coefficient in NaH_2PO_4, appears to be conserved in seven amino acids**((Author: please confirm that rewording is ok))**^{2} that are all located in a not well conserved, variable duplication region in the loop following the first helix of the N-terminal domain (shown in red in figure 5a**((Author: could this say "shown with a red arrow in figure 5a, as the region itself doesn't appear to be I in red"?))**). A likely reason for the conservation of the retention coefficient in NaH_2PO_4 (a measure of hydrophobicity) at these sites is suggested by the main function of the N-terminal Nef domain in membrane anchoring. The insertion of amino acids with the same property would therefore be predicted to enhance the interactions of Nef with the membrane. This is a hypothesis that can be tested directly through wet-lab experiments. Thus, this example highlights the type of hypothesis that can be generated using property conservation, revealing evolutionary pressure that is not captured by sequence conservation.

Property Conservation in Env Proteins

The HIV-1 Env glycoprotein**((Author: elsewhere in this article, ENV is referred to just as a protein. Should it be glycoprotein?))** mediates the fusion of viral and cellular membranes during

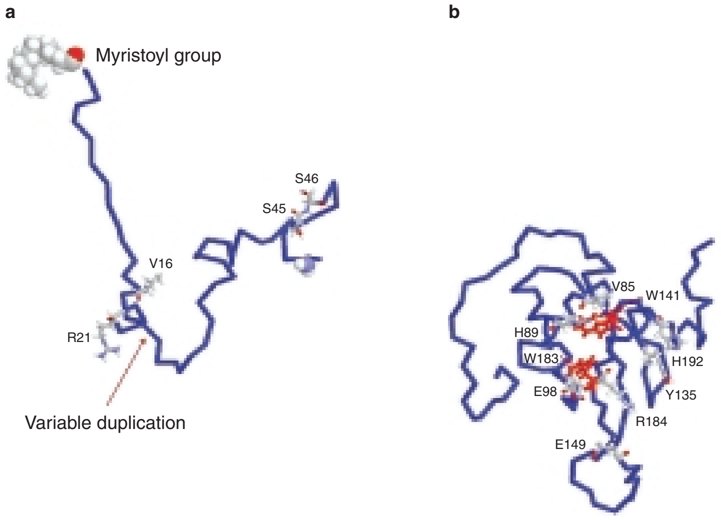

Fig. 5. Structures of two fragments of the HIV-1 Nef protein obtained from nuclear magnetic resonance (NMR) spectroscopy. (**a**) N-terminal Nef anchor fragment, myristate G2 to W57, Protein Data Bank (PDB) code 1QA5.[25]**((Author: is the figure image reproduced from this reference?))** (**b**) C-terminal Nef core fragment, 40–206 with a deletion from 159 to 173 and C206A point mutation, PDB code 2NES.[22]**((Author: is the figure image reproduced from this reference?))** Residues with conserved properties (see) are highlighted in stick representation using CPK**((Author: please define))** colour coding. The amino acid identities and positions are labelled in black. Two other amino acids are highlighted in red: W141 and W183. The locations of the myristoyl group and the variable duplication region are shown. The amino acids from 40 to 56 are unstructured and are not shown. The structure was drawn using the noncovalent bond finder (MDL**((Author: please define))**).**((Author: please confirm that rewording of figure caption is ok))**

HIV-1 infection of cells. Env is synthesised as the precursor protein gp160, which is proteolytically cleaved into two subunits, the surface subunit gp120 and the transmembrane subunit gp41. Both fragments remain noncovalently associated with each other after cleavage in a trimeric structure. HIV-1 infection is initiated by the binding of gp120 primarily to CD4 with co-receptors of the chemokine receptor family on the host surface. Subsequent conformational changes in Env result in exposure of a hydrophobic N-terminal fusion peptide within the gp41 domain, thus initiating membrane fusion.

Only one of the amino acids with highly conserved properties is located in gp41, in the disulphide-bonded loop region. This region is thought to act as a hinge in the formation of the trimer-of-hairpins structure of gp41 that brings the cellular and viral membranes in((**Author: should "in" be "into"?**)) proximity and to play a key role in the interaction between gp120 and gp41.[28] Thus, the conservation of hydrophobicity at((**Author: should "at" be "of the"?**)) amino acid at position 739 in the alignment may play an important role in this interaction. All other amino acids are located in the gp120 fragment. The structure of the gp120 fragment has been solved by x-ray crystallography in complex((**Author: should this be "in a complex"?**)) with CD4 and a neutralising antibody.[29] Only three of the amino acids are located in the fragment that has been crystallised: alignment positions 356, 427 and 442, corresponding to S291, Q363 and Q352 in the HIV-1 reference protein. These three amino acids are highlighted in figure 6. All are located on different parts of the surface of the molecule. Different properties are conserved at all three sites; however, all of them are conceivably connected to the possible recognition of protein partners. All other amino acids with conserved properties are located in regions that are flexible and thus not amenable to x-ray crystallographic analysis. This includes the N-terminus (positions 48 and 81), for which both properties are

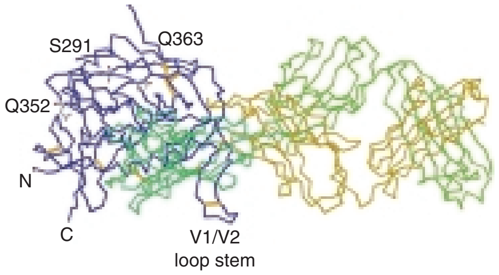

Fig. 6. Crystal structure model of gp120 (blue) in complex with CD4 and neutralising antibody (green and yellow**((Author: is CD4 shown in green OR is the neutralising antibody both green and yellow?))**). N- and C-terminals are labelled. Three amino acids with conserved properties are modelled in this structure, shown as stick representations. The structures were drawn using the noncovalent bond finder (MDL**((Author: please define))**).**((Author: please confirm that rewording of figure caption is ok))**

[2] These seven amino acids are at positions 29, 31, 32, 33, 34, 37 and 38 in the alignment. The first amino acid corresponds to residue 28 in the reference protein.

related to its flexibility, supporting the hypothesis that these amino acids contribute to the observed flexibility. The other two amino acids, at positions 203 and 212, are located in the V1/V2 loop, a sequence variable region that was deleted in the gp120 crystal structure to allow crystallisation. This loop has been shown to rescue changes in the important V3 loop on the opposite side of the molecule and is a potential interaction site between individual molecules of the three units of the trimer. The lack of sequence conservation in this region, but high conservation of absolute entropy (position 212) and hydrophobicity (position 203), suggests that these residues may be involved in the inter-monomer interactions, an experimentally testable hypothesis.

Conclusions and Future Work

We have described a framework for systematically identifying the property-based selection pressures affecting each position in an MSA. We first calculate a distance from the observed distribution to an idealised distribution (currently a Gaussian), assuming a particular property is conserved at this position. We then use a Monte Carlo procedure to estimate the probability of achieving such a good fit by chance. Our method requires only an MSA as input and determines the most highly conserved properties in each position, as well as their statistical significance. In our biological evaluation, we discuss the significance of our results with respect to current understanding of the structure and function of two proteins from HIV-1, Nef and Env. We make predictions regarding previously unknown sites of property conservation, providing guidance for future site-directed mutagenesis studies by experimental biologists.

Our current test evaluates whether a particular property scale explains the observed data better than we would expect by chance. However, it is often the case that more than one property scale is determined to provide a significant fit to the data. This usually occurs because there are many highly correlated scales that are just different experimental measurements of the same underlying property (such as hydrophobicity or size). In this case, we do not expect to be able to discriminate between these highly correlated scales. In the less common case, where two properties that are not highly correlated are both scored as significant, we would like to test whether the different properties explain the data equally well or whether one property is significantly better than the other.

Currently, we try to limit the confounding effects that are attributable to the lack of independence between sequences by using sequence weights. However, it has been shown by Bruno[18] that since sequence weights assume that the effects of phylogenetic correlation are identical at every position in the alignment, they cannot fully correct for phylogenetic dependencies. A more thorough, but also more computationally intensive, solution to this problem would either use Bruno's RIND program to estimate site-specific residue frequencies or explicitly incorporate the phylogenetic tree into a generative model of protein evolution that includes selection of physical and chemical properties.

We chose to use a Gaussian to model property conservation effects because it is unimodal and has other desirable properties (discussed in the section Gaussian Fitting). However, we do not necessarily believe that selective pressures are truly Gaussian in nature. Another option would be to not assume any particular unimodal distribution, but just test the hypothesis that the distribution of property values in a position is unimodal. A dip test is one way to determine whether a distribution is unimodal.[30]

The current method is only designed to find evidence for positional conservation of a *single* property, but we know that in many cases selection exerts pressure on multiple properties simultaneously. In particular, there are many cases where we find a property that is more conserved than would be expected by chance, but this property still has unavoidable 'holes', which could be explained by the intersection of two property pressures. We would like to extend our approach to use multidimensional fitting of properties in these cases where a single property is insufficient to explain the observed data. Although testing for combinations of properties could be carried out by a simple multivariate extension of the Gaussian-fitting test, the larger search space makes multiple testing a more problematic issue. To avoid false positives, we will have to reduce the number of tests conducted, possibly by considering only combinations of properties that were conserved (to some extent) individually.

((Author: unable to locate reference [31] in the text – please advise))

Acknowledgements

This research was supported by National Science Foundation Information Technology Research grant NSF0225656 and by the American Society for Engineering Education (ASEE) through a National Defense Science & Engineering Graduate (NDSEG) fellowship. We are grateful to Bill Bruno for his detailed and helpful comments and to Larry Wasserman for numerous helpful discussions.

Author: please provide information, for publication in the acknowledgements section of the manuscript, on any sources of funding that were used to assist in conducting the study; and on any potential conflicts of interest that the authors may have that are directly relevant to the contents of this manuscript. If this information is not provided we will include the following statement: 'The authors have provided no information on sources of funding or on conflicts of interest directly relevant to the content of this review/study.'

References

1. Oliveira L, Paiva P, Paiva A, et al. Identification of functionally conserved residues with the use of entropy-variability plots. Proteins 2003; 52: 544¯52
2. Oliveira L, Paiva P, Paiva A, et al. Sequence analysis reveals how G protein-coupled receptors transduce the signal to the G protein. Proteins 2003; 52: 553¯60
3. Grigoriev I, Kim S. Detection of protein fold similarity based on correlation of amino acid properties. Proc Natl Acad Sci U S A 1999; 96: 14318¯23
4. Mathura V, Schein C, Braun W. Identifying property based sequence motifs in protein families and super-families: application to DNase-1 related endonucleases. Bioinformatics 2003; 19: 1381¯90
5. **((Author: please supply volume number))**Ouzounis C, Perez-Irratxeta C, Sander C, et al. Are binding residues conserved? Pac Symp Biocomput 1998; : 401¯12
6. Villar H, Kauvar L. Amino acid preferences at protein binding sites. FEBS Lett 1994; 349: 125¯30
7. Lichtarge O, Bourne H, Cohen F. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996; 257: 342¯58
8. Pei J, Grishin N. AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 2001; 17: 700¯12
9. Valdar W. Scoring residue conservation. Proteins 2002; 48: 227¯41
10. Shenkin P, Erman B, Mastrandrea L. Information-theoretical entropy as a measure of sequence variability. Proteins 1991; 11: 297¯313
11. Gerstein M, Altman R. Average core structures and variability measures for protein families: application to the immunoglobulins. J Mol Biol 1995; 251: 161¯75
12. Taylor W. The classification of amino acid conservation. J Theor Biol 1986; 119: 205¯18
13. Sjolander K, Karplus K, Brown M, et al. Dirichlet mixtures: a method for improving detection of weak but significant protein sequence homology. Comput Appl Biosci 1996; 12: 327¯45
14. Koshi J, Mindell D, Goldstein R. Beyond mutation matrices: physical-chemistry based evolutionary models. Genome Inform Ser Workshop Genome Inform 1997; 7: 80¯89
15. Wasserman L. All of statistics. New York: Springer, 2004
16. Durbin R, Eddy S, Krogh A, et al. Biological sequence analysis. Cambridge: Cambridge University Press, 1998
17. Thompson J, Higgins D, Gibson T. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994; 22: 4673¯4680
18. Bruno W. Modeling residue usage in aligned protein sequences via maximum likelihood. Mol Biol Evol 1996; 13: 1368¯74
19. Harris M. From negative factor to a critical role in virus pathogenesis: the changing fortunes of Nef. J Gen Virol 1996; 77 (Pt 10): 2379¯92
20. Fackler OT, Kienzle N, Kremmer E, et al. Association of human immunodeficiency virus Nef protein with actin is myristoylation dependent and influences its subcellular localization. Eur J Biochem 1997; 247: 843¯51
21. Freund J, Kellner R, Houthaeve T, et al. Stability and proteolytic domains of Nef protein from human immunodeficiency virus (HIV) type 1. Eur J Biochem 1994; 221: 811¯19
22. Grzesiek S, Bax A, Clore GM, et al. The solution structure of HIV-1 Nef reveals an unexpected fold and permits delineation of the binding surface for the SH3 domain of Hck tyrosine protein kinase. Nat Struct Biol 1996; 3: 340¯5
23. Lee CH, Saksela K, Mirza UA, et al. Crystal structure of the conserved core of HIV-1 Nef complexed with a Src family SH3 domain. Cell 1996; 85: 931¯42
24. Freund J, Kellner R, Konvalinka J, et al. A possible regulation of negative factor (Nef) activity of human immunodeficiency virus type 1 by the viral protease. Eur J Biochem 1994; 223: 589-93
25. Geyer M, Munte CE, Schorr J, et al. Structure of the anchor-domain of myristoylated and non-myristoylated HIV-1 Nef protein. J Mol Biol 1999; 289: 123¯38
26. Geyer M, Peterlin BM. Domain assembly, surface accessibility and sequence conservation in full length HIV-1 Nef. FEBS Lett 2001; 496: 91¯5
27. Geyer M, Fackler OT, Peterlin BM. Structure¯function relationships in HIV-1 Nef. EMBO Rep 2001; 2: 580¯5
28. Wang S, York J, Shu W, et al. Interhelical interactions in the gp41 core: implications for activation of HIV-1 membrane fusion. Biochemistry 2002; 41: 7283¯92
29. Kwong PD, Wyatt R, Robinson J, et al. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. Nature 1998; 393: 648¯59
30. **((Author: please confirm journal title))**Hartigan JA, Hartigan PM. The dip test of unimodality. Ann Stat 1985; 13: 70¯84
31. **((Author: (1) unable to locate this reference in the text – please advise. (2) Has this manuscript been accepted for publication by a journal [if so, please update details]? (3) If it is simply an unpublished manuscript, which of the 2 authors is the 'originating' author?))**Yampolsky L, Stoltzfus A. Amino acid exchangeability from experimental data. 2004. Unpublished manuscript

Correspondence and offprints: Dr *Rose Hoberman*, **((Author: please supply a corresponding address))**, **((Author: please confirm that you are the corresponding author and are happy for us to publish your e-mail addresses))**.

E-mail: roseh@cs.cmu.edu