PROTEINS: Structure, Function, and Genetics 48:227–241 (2002)

Scoring Residue Conservation

William S.J. Valdar*
Biomolecular Structure and Modelling Unit, Department of Biochemistry and Molecular Biology, University College London, London, United Kingdom

ABSTRACT **The importance of a residue for maintaining the structure and function of a protein can usually be inferred from how conserved it appears in a multiple sequence alignment of that protein and its homologues. A reliable metric for quantifying residue conservation is desirable. Over the last two decades many such scores have been proposed, but none has emerged as a generally accepted standard. This work surveys the range of scores that biologists, biochemists, and, more recently, bioinformatics workers have developed, and reviews the intrinsic problems associated with developing and evaluating such a score. A general formula is proposed that may be used to compare the properties of different particular conservation scores or as a measure of conservation in its own right. Proteins 2002;48:227–241.** © 2002 Wiley-Liss, Inc.

Key words: protein sequence analysis; amino acid; variability; evolutionary conservation; multiple sequence alignment

INTRODUCTION

A multiple-sequence alignment is a historical record. The patterns of amino acid variability in its columns tell a story of evolutionary pressure, mutation, recombination, and genetic drift that often spans many millions of years. This story can be read in different ways. According to the neutral model of molecular evolution, once a protein has evolved to a useful level of functionality, most new mutations are either deleterious, in which case they are removed by negative selection, or neutral, in which case they are kept. Therefore, most of the substitutions observed in an alignment are neutral; rather than representing improvements in a protein, they indicate how tolerant the protein is to change at that position. In an already optimized protein, the rate of substitution will be inversely correlated with the functional constraints acting on that protein. The most functionally important residues of hemoglobin, those that secure the heme group, show a much lower rate of substitution than do others in the protein. The selectionist model of molecular evolution, although agreeing that most mutations are deleterious and removed, argues that accepted mutations usually confer a selective advantage, whereas neutral mutations are rare (Ref. 1 and refs. therein). Although both models have their place, this review takes the perspective of the neutral model only. That model accords better with the idea of conservation among functionally equivalent sequences and is arguably the more evident in alignments from structural biology.

If the degree of functional constraint dictates how conserved a position is, then the converse must also be true, that is, the degree of conservation must indicate the functional importance of that position. Thus, identifying conserved regions of a protein is tremendously useful. In the past, patterns of conservation in multiple alignments were identified by inspection alone. However, the rapid increase of available sequences and published analyses has emphasized the need for objective, automated methods, and in the last decade or so, this has been the subject of considerable research. Much of that work has focused on extracting global patterns and motifs from multiple alignments, often with a view to exploring the relationships between homologues and developing diagnostic tests for functions of newly discovered sequences. For instance, statistically robust profile methods, such as PSI-BLAST[2] and those based on hidden Markov models,[3] have become increasingly popular.

Despite these advances, there have been few recent insights into the derivation of a quantitative conservation measure for a single aligned position, and there certainly is no standard method. Ask a life scientist how similar two sequences are and he will probably quote a percentage identity or an E-value. Ask him how conserved a position is in a family and the reply is most likely to be qualitative. This review discusses what a quantitative measure of conservation should actually measure and, by surveying almost 20 scores, examines some of the problems inherent in developing such a score.

Exercises for a Conservation Score

There is no rigorous mathematical test for judging a conservation measure; if there were, one would use the test and not bother with an additional score. Rather than accuracy then, a conservation score may be judged on its verisimilitude: its ability to depict realism and its concordance with biochemical intuition. Figure 1 helps make these abstract notions more concrete. It shows columns of amino acids taken from hypothetical multiple-sequence alignments of functionally equivalent orthologues. For simplicity, we assume each sequence contributes equally

*Correspondence to: William S.J. Valdar, Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. E-mail: William.Valdar@well.ox.ac.uk

Received 12 October 2001; Accepted 22 February 2002

Sequences	Columns (a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)
1	D	D	D	D	D	D	I	P	D	L	L
2	D	D	D	D	D	D	I	P	V	L	L
3	D	D	D	D	D	D	I	P	Y	L	L
4	D	D	D	D	D	D	I	P	A	L	L
5	D	D	D	D	D	D	L	W	T		–
6	D	D	E	D	E	E	L	W	K		–
7	D	D	E	D	E	E	L	W	P		–
8	D	D	E	D	E	E	L	W	C		–
9	D	D	E	D	E	F	V	S	R		–
10	D	E	E	F	F	F	V	S	H		–

Fig. 1. Some example columns from different multiple alignments. Each labeled column represents a residue position in a multiple-sequence alignment. The rows denote the sequence number of a particular amino acid. Amino acids are identified by their one-letter code, and gaps by a dash ("—"). Note that column (k) comes from an alignment of 10 sequences, whereas column (j) comes from an alignment of only 4.

to the information in the alignment and that none require extra weighting. Applying basic biochemical knowledge to this collection of columns reveals some concrete qualitative comparisons. Specifically, from most conserved to least conserved, the following orders seems reasonable: (a) > (b) > (c) > (d) > (e) > (f), then (g) > (h) > (i), and (j) > (k).

Column (a) contains only D and therefore is the most obviously conserved. Column (b) also contains E, so (b) is more variable than (a). Column (c) contains D and E but is less dominated by any one than (b), so (c) is more variable than (b). Column (d) contains nine D and one F; it is clearly more variable than column (a), but is it more variable than column (b)? Phenylalanine is large and nonpolar, whereas aspartate and glutamate are both smaller and polar. Because the amount of stereochemical variability in column (d) is greater than in column (b), it seems likely a mutation from D to E glutamate would be more tolerable than one from D to F (a conclusion supported by the exchange probabilities in a mutation data matrix; see later). Column (e) implies both conservative substitutions (between D and E) and nonconservative ones (between the acids and phenylalanine). Thus, column (e) is the least conserved so far. Column (f) contains the same amino acid types as (e). However, because it is less skewed toward an abundance of D and E, that is, more evenly mixed, (f) is more variable.

Columns (g) and (h) are equivalent in the number and frequency of their amino acids. However, because (g) contains only branch chained amino acids whereas (h) encompasses a broader mix of stereochemical characteristics, (g) is more suggestive of conservative substitutions in response to negative selective pressure. Column (i) is the most variable column encountered so far, as judged by biochemistry or amino acid frequency.

Columns (j) and (k) illustrate the importance of gaps. Column (j) is taken from an alignment of four sequences. In each sequence, a leucine is present at that position. Column (k) also contains four leucines, but because it comes from an alignment of 10 sequences, it also contains six gaps. For column (k), then, there is strong evidence that leucine is not functionally constrained. After all, this amino acid has been shed from six other orthologues with apparent impunity. There is no such evidence for column (j), the conservation of which remains untarnished. The comparison between columns (j) and (k) also highlights the dangers of having too small an alignment. The alignment of (j) could be the same as that of (k) but with six sequences missing; it is an example of lack of data producing completely different conclusions about the same site.

Figure 1 will be used here as a testing ground for some of the scores surveyed later.

Requirements of a Conservation Score

A score that quantifies the degree of conservation at an aligned position should fulfill the following criteria:

1. Mathematical properties. The score should be a function that maps a set of arguments (the input space), which includes the aligned column and possibly other information, to a number (the output space). Convenient scores will have an output space that is continuous and bounded.
2. Amino acid frequency. The score should take account of the relative frequencies of amino acids in a column. For instance, by using the columns from Figure 1 it should reproduce the ranking (a) > (b) > (c) > (e) > (f).
3. Stereochemical properties. The score should recognize conservative replacements and that some substitutions incur more chemical and physical change than others. For instance, it should score column (g) as more conserved than column (h).
4. Gaps. Numerous gaps suggest a position can be deleted without significant loss of protein function. Therefore, the score should penalize such positions and should rank column (j) as more conserved than column (k). An ideal score might also recognize that, in protein structure, the difference between a small residue (e.g., glycine) and a gap is less than between a large residue (e.g., tryptophan) and a gap.
5. Sequence weighting. A typical alignment often includes some sequences that are very closely related to each other. These clusters of highly similar sequences may reflect bias in the sequence databases or result from nature's irregular sampling of the space of acceptable mutations. Either way, such clusters can monopolize alignments, masking important information about allowed variability from more sparsely represented sequences. A good conservation score should find some way to normalize against redundancy and bias in the alignment without loss of evolutionary information.
6. Simplicity. Most scoring methods, from E-values that describe sequence similarity to school exam grades, have their limitations. Understanding the shortcomings of these methods is key to using them wisely and interpreting their results meaningfully. Therefore, on the reasonable assumption that no method is perfect, a good conservation score should be no more complex than it needs to be so its deficits can be understood.

A SURVEY OF CONSERVATION SCORES

Over the last 30 years a number of methods have been proposed to score residue conservation. The scores surveyed here are presented in approximately increasing order of sophistication in what they try to achieve. For clarity, the names given to each score by its authors are ignored in favor of the following convention. Scores whose values increase with increasing conservation are denoted C_{name}, where the subscript identifies the author. Scores that do the converse are denoted V_{name}.

Symbol Frequency Scores

Scores in this category consider amino acids as symbols in a uniformly diverse alphabet. They focus on their relative frequency of these symbols and do not account for sequence redundancy in the alignment. Because, by definition, none model stereochemical properties (criterion 3) or weight their sequences (criterion 5), the discussion instead concentrates on how well they fulfill the remaining criteria.

In 1970, Wu and Kabat[4] introduced the first widely accepted measure of conservation. Their score, which they used to identify the variable regions on antibodies, was defined as

$$V_{\text{Kabat}} = \frac{k}{n_1} \times N, \tag{1}$$

where k is the number of amino types present at the aligned position, n_1 is the number of times the most commonly occurring amino acid appears there, and N is the number of sequences in the alignment. The variable N acts as a scaling factor and is constant for a given alignment. For clarity, this survey will tend to set such constants apart from the main equation.

Applying the score to Figure 1, V_{Kabat} correctly reproduces the ranks (a) > (b) > (c) > (e) but fails to distinguish (e) from (f). This is because it cares only about the frequency of the most commonly occurring symbol and ignores the frequencies of the rest. V_{Kabat} has other problems; for one, it is discontinuous along its output space. A strictly conserved column, such as column (a), always scores 1. A column that is strictly conserved except for one aberrant amino acid is $2\,\frac{N}{N-1} > 2$, regardless of how many sequences are in the alignment. This discontinuity is biologically meaningless.[5] The score also fails to consider gaps and so fulfils only the criterion of simplicity.

Jores et al.[6] recognized the inability of V_{Kabat} to distinguish (e) from (f) and in response proposed a modified version:

$$V_{\text{Jores}} = \frac{k_{\text{pair}}}{n_{\text{pair}_1}} \times \frac{1}{2} N(N-1), \tag{2}$$

where $1/2\ N(N-1)$ is the number of possible pairs of amino acids in the column, k_{pair} is the number of distinct pairs, and n_{pair_1} is the number times the most frequently distinct pair occurs. By considering pairs rather than singlets, this score improves on V_{Kabat}. However, all other deficits remain. It is still discontinuous: although complete conservation scores one, the next most conserved value possible is $2\,\frac{N}{N-2} > 2$. It does not account for gaps. Even its simplicity is questionable: V_{Jores} does the same job but is significantly more awkward to compute than the symbol entropy scores discussed later.

Lockless and Ranganathan[7] propose a different type of symbol frequency score. They measure the conservation at an aligned position as the extent to which amino acid frequencies at that position deviate from frequencies over the whole alignment. To model this deviation, they use binomial probabilities. If an amino acid a occurs in the sequence databases at fractional frequency q_a, then the probability of a occurring n_a times in a column of N residues is $P(X = n_a)$ where $X \sim Bin(N, q_a)$. This probability is compared with the probability for the overall frequency of a in the alignment to give a measure of deviation

$$d(n_a, \bar{n}_a) = \ln\left(\frac{P(X = n_a)}{P(X = \bar{n}_a)}\right), \tag{3}$$

where n_a is the average frequency of a in the whole alignment. The distance d describes how much the frequency a at the position differs from that of a across the alignment. When these frequencies are the same, $d = 0$; when they are different, d may be positive or negative. The conservation for the column, C_{Lockless},* is taken as the root-mean-square deviation (RMSD) over all 20 amino acids, that is,

$$C_{\text{Lockless}} = \sqrt{\sum_a d(n_a, \bar{n}_a)^2}. \tag{4}$$

If a single column can be represented by a point in 20-dimensional space of binomial probabilities, then C_{Lockless} measures the Euclidean distance between that point and the point representing the "average" column. In a typically diverse alignment, Lockless and Ranganathan's score identifies columns dominated by only a few amino acids, because the binomial probabilities of these columns would be small. Some strictly conserved columns score higher than others. For instance, if cysteine occurs least frequently in the alignment, a strictly conserved column of cysteine will score higher than a strictly conserved column of histidine. Although this has some intuitive appeal—strictly conserved columns of rare amino acids are visually more striking in an alignment—the authors do not argue its case. But this arbitrariness is symptomatic of a deeper malaise: that C_{Lockless} is complex. Its purpose is to measure how different a column is from the rest of the alignment. However, d could be calculated far more simply, say, as the Euclidean distance between the two sets of amino acid frequencies. Instead C_{Lockless} uses a binomial model that brings in further data, namely, the frequencies of amino acids from a sequence database. Considering that (arguably) more important information about stereochemistry, gaps, and the like is omitted, this added complexity seems hard to justify.

*In their original article, Lockless and Ranganathan presented their score as $\Delta G_i^{\text{stat}} = \text{kT}^* \sqrt{\Sigma_a (\ln P(X = n_a)/P(X = \bar{n}_a))^2}$. For clarity, references to thermodynamics have been removed to C_{Lockless}.

Symbol Entropy Scores

Symbol entropy scores are a specialization of symbol frequency scores. Scores in this category all account for the relative frequencies of symbols using Shannon's entropy or variations thereof.

Background

Shannon's information theoretic entropy[8] (hereafter referred to as "Shannon's entropy") is an often-used measure of diversity.[9,10] It can be derived from two roots: one combinatoric and one information theoretic. The combinatoric derivation goes as follows. Given N objects that fall into K types, the number of distinct ways they can be permuted W is given by the multinomial coefficient,

$$W = \frac{N!}{\prod_i^K n_i}, \tag{5}$$

where n_i is the frequency of the ith type. As N becomes large, $N!$ can be calculated by using Sterling's approximation, $\ln N! = N \ln N - N$, such that

$$\ln W = -N \sum_i^K p_i \ln p_i, \tag{6}$$

where $p_i = n_i/N$, the fractional frequency of type i. Transforming linearly gives Shannon's entropy:

$$S = -\sum_i^K p_i \log_2 p_i, \tag{7}$$

The quantities S and W monotonically increase with each other. S ranges from zero, when objects of only one type are present, to $S_{\max} = \log_2 K \geq 0$, when all types are present in equal proportion. It has been shown that Shannon's entropy belongs to a general class of diversity index,[11]

$$D(\alpha,\beta) = \sum_i^K p_i^{\alpha}(-\log p_i)^{\beta}. \tag{8}$$

Of this class, both Shannon's entropy ($D(1,1)$) and Simpson's index ($D(2,0) = \Sigma_i^K p_i^2$) have been used in ecology for measuring species diversity (Ref. 9 and refs. therein). Note that the base of the logarithm affects only the unit of measurement, not the score itself because, for any a and b, $\log_a x \propto \log_b x$.

The original use of Shannon's entropy was in information theory (Refs. 10 and 12 and refs. therein). In many older telecommunication systems, such as radio, the signal constructs the output. In more modern digital systems, the range of possible outputs is small and known in advance, allowing the more economical approach of encoding, in which the signal selects output from a finite list. The selective information content of an encoded signal depends not on the size or complexity of the output as such, but on the number of alternative forms it might have taken, and on the relative likelihood of each. The total selective information content of a signal in bits, I, is defined as the amount of uncertainty it resolves and can be deduced from the difference between Shannon's entropy of the alternative forms before and after the signal is sent, that is,

$$I = S_{\text{before}} - S_{\text{after}}. \tag{9}$$

Scores

In 1991, two groups proposed residue conservation scores based on Shannon's entropy. Until then, entropy had been used for scoring positional conservation, but only in nucleotide sequences (Ref. 13 and refs. therein).

Sander and Schneider[14] defined their score as a normalized Shannon's entropy:

$$V_{\text{Schneider}} = -\sum_i^K p_i \ln p_i \times \frac{1}{\ln K}, \tag{10}$$

where $K = 20$, representing the 20 amino acid types. Shenkin et al.[5] proposed the related score

$$V_{\text{Shenkin}} = 2^s \times 6, \tag{11}$$

where S is Shannon's entropy (7) and K in that equation is also 20. These scores are transformations of each other and so are trivially different. Both purport to have conveniently bounded ranges: $0 \leq V_{\text{Schneider}} \leq 1$ and $6 \leq V_{\text{Shenkin}} \leq 120$. However, neither would score column (i) in Figure 1 as maximally variable. This is more a minor artifact than a serious deficit. Maximal diversity occurs when all types are represented evenly. But if there are more amino acid types than there are sequences to represent them, this limit can never be reached. Similarly, $V_{\text{Schneider}}$ and V_{Shenkin} can reach their top value only when there are at least 20 sequences in the alignment.

Gerstein and Altman[15] present another variation on this theme. To compare sequence conservation with structural conservation in a multiple alignment of protein structures, they define V_{Gerstein}, which measures the entropy of a position relative to that if the sequences were aligned randomly:

$$V_{\text{Gerstein}} = \sum_i^K \bar{p}_i \ln \bar{p}_i - \sum_i^K p_i \ln p_i, \tag{12}$$

where $\bar{p}_i$ is the average frequency of amino acid i in the alignment and $K = 20$. This score, which is in the same form as Eq. 9, measures the information content of the position in bits. This does not bestow any particular advantage; like the other entropy scores, V_{Gerstein} delivers nothing grander than a conveniently expressed multinomial coefficient (Eq. 5).

Like V_{Jores}, Shannon-based scores rank (a), (b), (c), (e), and (f) correctly. Unlike V_{Jores}, they are continuous. In a column strictly conserved but for one aberrant residue, the entropy decreases to the score's minimum with an increasing number of sequences. Shannon's entropy is also much simpler to calculate. Although V_{Jores} requires information

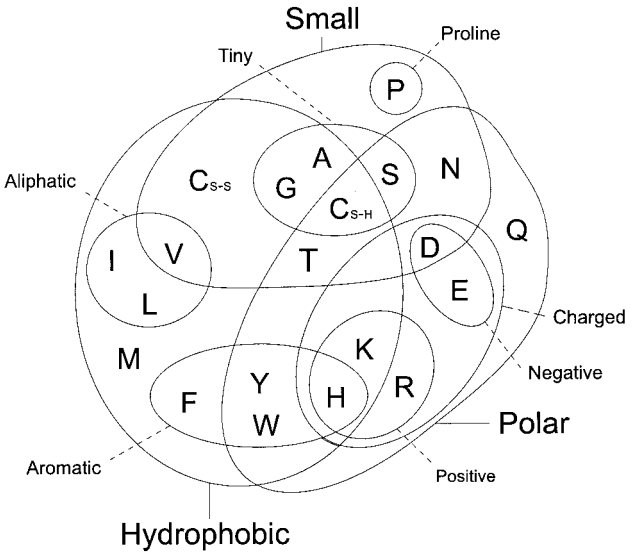

Fig. 2. Taylor's Venn diagram of amino acid properties. Taylor argues Cys should appear twice because although the reduced form ($C_{S\text{-}H}$) has similar properties to Ser, the oxidized form ($C_{S\text{-}S}$) is more like Val. Adapted from Refs. 18 and 21.

about pair frequencies, which itself requires combinatoric calculations, entropy requires only fractional frequencies of the symbol types. But as well as being simple, these scores are simplistic. Amino acids are not uniformly different, no matter how mathematically convenient it is to think otherwise. None of these scores could distinguish column (g) from (h) in Figure 1. When Gerstein and Altman compare structural conservation, using an atom coordinates-based scheme, with sequence conservation, using V_{Gerstein}, they find the two have little in common.[15] Perhaps a sequence conservation score that considered stereochemistry would have led them to a different conclusion.[†]

More worryingly, none of these scores account for gaps. This is a problem. In the Shannon scheme, it is most natural to consider a gap as another symbol type, the "21st" amino acid. Doing this, however, has absurd consequences. For instance, column (k), which is predominantly gapped, would score as more conserved than columns (c) or (g).

Stereochemical Property Scores

Scores in this category consider only the stereochemical properties of the amino acids in a column.

In 1986, Taylor[18] classified amino acid types according to their stereochemical properties and their patterns of conservation in the Dayhoff mutation data matrix.[19] He embodied them in a Venn diagram (Fig. 2), in which each overlapping set represents a distinct physical or chemical property. Taylor then devised a set theoretic method based on this diagram to score positional conservation. His method finds the smallest set or subset that describes the amino acid types observed at an aligned position. The variability of the column is taken as the total number of residue types belonging to that set. The number of possible subsets of the Venn diagram is large, and many of these sets have little physical meaning. To reduce the possibility of high conservation being ascribed to meaningless subsets, Taylor compiled a list of 70 sets and subsets that might reasonably be conserved and suggested only these "valid" sets should be considered. Taylor's score can be expressed as

$$V_{\text{Taylor}} = \min(n(\{X: \mathit{Aligned} \subseteq X, X \in \mathit{Valid}\})), \quad (13)$$

where *Aligned* is the set of amino acids at the aligned position, *Valid* is Taylor's set of 70 valid sets and $n(X)$ is the number of elements in set X. V_{Taylor} ranges from 1 to 20.

Taylor's score accomplishes some things the symbol scores could not. It recognizes that column (b) from Figure 1 is more conserved than (d) and that (g) is more conserved than (h). It does not explicitly model gaps, but there is a natural way these could be incorporated into the scheme: a gap could belong only to the largest superset. But Taylor's score is clumsy. The ad hoc clause of reducing the number of valid sets to 70 makes the score more computationally tractable but diminishes its simplicity and elegance. To interpret the score properly, one must accept that some subsets in the Venn diagram are forbidden. This introduces a degree of subjectivity on top of that supplied by the Venn diagram itself.

Taylor's score has more conspicuous problems. First, the score of strictly conserved columns depends on the amino acid: a column of P's scores 1, H's score 3, and W's score 4. Similarly, column (g) in Figure 1 would score the same as a strictly conserved column of I. Second, it fails to account for amino acid frequencies and cannot distinguish column (b) from (c) or (e) from (f). Clearly, a Venn diagram is picturesque but unwieldy. Fortunately, Zvelibil et al.[20] abridge it to something more convenient: a truth table of amino acids with 10 property descriptors (Fig. 3). They define their score as

$$V_{\text{Zvelibil}} = n_{\text{const}} \times \frac{1}{10}, \quad (14)$$

where n_{const} is the number of properties whose state (i.e., truth or falsehood) is constant for all amino acids in the column. For example, column (b) in Figure 1 contains D and E, which share nine properties and scores 0.9. Although it has a less erratic output space, V_{Zvelibil} retains V_{Taylor}'s failure to account for amino acid frequency. In their program Analysis of Multiply Aligned Sequences (AMAS), Livingstone and Barton[21] turn this weakness into a strength. AMAS uses V_{Zvelibil} to split sequences into subgroups and thus infer an evolutionary or functional hierarchy from the alignment. The success of their method shows V_{Zvelibil} is better suited to this kind of selectionist analysis.

[†]Interestingly, that an almost identical criticism was recently leveled by Mirny and Shakhnovich[16] at a comparison of structure and sequence conservation by Plaxco et al.[17] Plaxco et al. used a score much like V_{Gerstein} and formed similarly heretical conclusions.

Fig. 3. Truth table profile of amino acid properties. Amino acids (across) are each described in terms of 10 properties (down). A filled circle means the amino acid above it possesses that property. The symbol "Δ" represents a gap, which is considered to have all properties. Adapted from Ref. 21.

Mutation Data Scores

Scores in this category use mutation data from a substitution matrix to quantify stereochemical variability in an aligned column. No scores in this category normalize against sequence redundancy in the alignment.

Background

Substitution matrices provide a quantitative and reasonably objective assessment of likely amino acid substitutions. Through analogy, they have also been used to quantify amino acid similarity. The nondiagonal pairwise scores indicate how likely one amino acid is to be substituted by another in a homologous protein. The analogy here is strong: functionally important positions between orthologues accept replacements more readily if those replacements are conservative. The diagonal scores, which pitch an amino acid against itself, indicate how likely an amino acid is to substituted at all, that is, its "mutability." This is where the analogy between a substitution matrix and a similarity matrix breaks down.

The diagonal of a substitution matrix helps alignment algorithms decide whether two amino acids should be aligned. Its use is normative concerning the alignment. A similarity matrix used by a conservation score must assess the similarity of the amino acids in a column. Its use is descriptive concerning the alignment. The conservation score does not seek to question the validity of the alignment; rather, it assumes the alignment is correct and seeks to describe its features. These two motives are fundamentally different. For instance, the diagonal in a substitution matrix tells you that Trp is rarely substituted, whereas Arg is substituted more readily. This makes sense; Trp is unique among amino acids, whereas Arg has more obvious replacements. Therefore, a conservation score that used the substitution matrix to measure similarity would rate a column containing only Trp as more conserved than one containing only Arg. This is would be wrong. Given that we trust the alignment, strict conservation of a more replaceable amino acid suggests a greater evolutionary constraint on that position. The functional constraint could be such that, although other amino acids are similar, because they differ even slightly in their geometry and chemistry, being similar is not enough. Therefore, a measure of replaceability is not just different from a measure of similarity; it is actually at odds with a descriptive measure of conservation.

How can this be resolved? The simplest answer is to redefine the diagonal, hence explicitly converting the substitution matrix into a similarity matrix, ideally in a way that minimally disturbs the off-diagonal values. For example, all diagonal values could be constant, set to the highest diagonal or off-diagonal value. Alternatively, the entire matrix could be normalized to take into account diagonal values—although this would count as perturbation of perfectly good off-diagonal values. If the similarity matrix is explicitly for measuring conservation, there is even a case for scaling diagonal values inversely to their values in the substitution matrix (i.e., the more replaceable an amino acid is, the more significant an event is its conservation and so the higher its self-similarity score).

Before describing matrix-based scores, it is interesting to contrast substitution matrices with their cousins, sequence profiles. A position under functional constraint will tolerate some substitutions better than others. But different positions have different functional and structural roles, and each particular role will inevitably place slightly different demands on its actor. For instance, sometimes it is important for a position to be polar, whereas other times bulkiness is more desirable. A transition from, say, W (large and polar) to S (small and polar) will, therefore, be more probable at the first type of position than at the second. So transition probabilities are not absolute but depend on context; this is what a sequence profile tries to model. But by modeling context dependence, a profile invites further questions: "How does one model these different contexts?" and "When data are sparse, that is, when there are few sequences, how can one reliably infer the underlying context without overfitting?"

The two most popular profile methods, PSI-BLAST[2] and hidden Markov models (HMMer,[3] SAM[22]) address these questions differently. PSI-BLAST assumes a single context, based on the BLOSUM62 substitution matrix. To avoid overfitting, it adds pseudocounts based on the Robinson amino acid frequencies. In contrast, HMMer and SAM assume there are nine prototypical contexts and that the context of any particular position can be described as a chimera of these nine. For example, a given position may be 30% like prototypical context one, 4% like context two, and so on. The contexts are modeled statistically by using Dirichlet mixtures,[23] which handle overfitting implicitly.

Sequence profiles are more sensitive at detecting remote homologues than are ordinary sequence comparisons that use a one-size-fits-all substitution matrix,[24] and at first glance, this might seem to suggest they offer a more elegant framework for scoring conservation. Not necessarily. Conservation scores and sequence profiles are guided by different motives, ask different questions, and make different assumptions. Without repeating the discussion above of diagonal elements in substitution matrices, it suffices to say that although a good sequence profile should

provide a flexible probabilistic framework for describing a complex observation, a good conservation score reduces this complexity to a single, decisive statistic.

Scores

Karlin and Brocchieri[25] propose the following score, which they use to study conserved positions in DNA-binding proteins:

$$C_{\text{Karlin}}(x) = \sum_{i}^{N} \sum_{j>i}^{N} M(s_i(x),s_j(x)) \times \frac{2}{N(N-1)}, \qquad (15)$$

where $s_i(x)$ is the amino acid at column x in the ith sequence, and $M(a,b)$ is the similarity between amino acids a and b. The similarity matrix M is defined such that

$$M(a,b) = \frac{m(a,b)}{\sqrt{m(a,a)m(b,b)}}, \qquad (16)$$

where m is BLOSUM62[26] or a similar substitution matrix. The normalization above ensures that $M(a,a) = 1$ always and that, provided m has a typical range, $-1 < M(a,b) \leq 1$. This in turn means C_{Karlin} ranges from -1 to 1. C_{Karlin} is a so-called "sum-of-pairs" (SP) score. It describes conservation by calculating the sum of all possible pairwise similarities between residues in an aligned column.

One criticism leveled at SP scores is that they do not make sense in what the statistic $m(a,b)$ means.[10] Scoring column (e) of Figure 1 involves summing the similarities of 29 pairs of amino acids. Yet, it is implausible that the diversity in that column arises from 29 amino acid substitutions among 10 homologues. However, if one treats the substitution matrix as no more than a quantitative guide to pairwise amino acid similarity, the SP score is no less than a convenient way to consolidate this two-dimensional information into a single number. Besides, the perceived overcounting of amino acid substitutions in the SP scores is assuaged somewhat by its overcounting of self-similarity terms. In fact, SP scores can be seen as a tug of war between self-similarity and substitution, that is,

$$C_{\text{Karlin}} \propto \sum_{a}^{K} \sum_{b\geq a}^{K} \begin{cases} a = b \rightarrow \dfrac{n_a(n_a-1)}{2} M(a,a) \\ a \neq b \rightarrow n_a n_b M(a,b) \end{cases}, \qquad (17)$$

where K is the number of amino acid types and n_a is the number of occurrence of amino acid type a. The upper term ($a = b$) bestows high scoring $M(a,a)$ values, whereas the lower term (a ≠ b) provides predominantly low scoring $M(a,b)$ values. But even if it escapes this criticism, C_{Karlin} deserves a further reproach: it does not account for gaps.

The score of Armon et al.[27] does account for gaps. Armon et al. present "ConSurf," an implementation and extension of the evolutionary trace method of Lichtarge et al.[28] ConSurf measures conservation by using a variation on the SP theme, defining its score as

$$V_{\text{Armon}} = \sum_{a>b}^{20} f_{ab} D(a,b), \qquad (18)$$

where

$$f_{ab} = \begin{cases} 1 & \text{if amino acids } a \text{ and } b \text{ present} \\ 0 & \text{otherwise} \end{cases}, \qquad (19)$$

and D is a dissimilarity matrix.[‡] In the tug-of-war notation of Eq. 17, this can be expressed as

$$V_{\text{Armon}} = \sum_{a}^{K} \sum_{b\geq a}^{K} \begin{cases} a = b & \rightarrow & \begin{cases} n_a > 1 & \rightarrow & n_a D(a,a) \\ n_a \leq 1 & \rightarrow & 0 \end{cases} \\ a \neq b & \rightarrow & (n_a + n_b) D(a,b) \end{cases}, \qquad (20)$$

where the upper terms contribute conserved scores and the lower terms contribute variability. Rather than basing their similarities on a substitution matrix, Armon et al. use a physicochemical distance matrix[29] (although the effect is much the same). D is zero for all $D(a, a)$, as much as 4.88 [= D(D,W)] for comparisons between real amino acids, 6 for $D(a,-)$ and 0.5 for $D(-,-)$. Distances for gaps are set heuristically.

Gaps aside, V_{Armon} and C_{Karlin} are only subtly different. C_{Karlin} emphasizes self-similarity more than V_{Armon}. This is evident from their self-similar terms in Eqs. 17 and 20: in C_{Karlin}, the coefficient of $M(a,a)$ grows quadratically with respect to n_a; for V_{Armon}, the coefficient of $D(a,a)$ grows linearly. The scores are also different in that V_{Armon} uses a physicochemical distance matrix. But to what advantage? Armon et al. argue that polarity and volume are the most important factors governing conservation of amino acid type. The most obvious way to check this would be to see if these factors dominate a substitution matrix. If volume and polarity do dominate, Armon et al. might as well have used a substitution matrix instead. If volume and polarity do not dominate, this challenges their assertion and needs to be explained.

Thompson et al.[30] do not use an SP score; they prefer instead a vectorial measure. Their program CLUSTALX, a graphical user interface to the CLUSTALW multiple alignment package,[31] plots a graph of positional conservation beneath a visual display of a multiple alignment. In the column at position x in the alignment, Thompson et al. consider the residue of the ith sequence, $s_i(x)$, to be a point $\mathbf{X}_i$ in K-dimensional space:

$$\mathbf{X}_i = \begin{bmatrix} M(a_1,s_i(x)) \\ M(a_2,s_i(x)) \\ \vdots \\ M(a_K,s_i(x)) \end{bmatrix}, \qquad (21)$$

where a_n is the nth symbol in an alphabet of K possible amino acids and $M(a,b)$ is similarity as judged by a substitution matrix. The consensus amino acid, which for columns that are not strictly conserved will be a hypothetical construct, is the center of gravity of all points from the column, $\bar{\mathbf{X}}$, that is,

[‡]Technically, f_{ab} is the number of times a and b are seen to exchange in a phylogenetic tree of the sequences. However, the definition above is a fair approximation.

$$\bar{\mathbf{X}} = \frac{1}{N}\sum_{i}^{N}\mathbf{X}_i, \tag{22}$$

The degree of conservation among these points is then related to the average Euclidean distance of all points from the consensus point:

$$C_{\text{Thompson}} = p_{\text{amino}} \times \frac{1}{N}\sum_{i}^{N}|\bar{\mathbf{X}} - \mathbf{X}_i|, \tag{23}$$

where p_{amino} is the fraction of symbols that are not gaps.

All three of C_{Karlin}, V_{Armon}, and C_{Thompson} correctly order columns (a)–(f) and (g)–(i) in Figure 1 and have mathematically continuous output spaces. C_{Thompson} has the aesthetic advantage of defining a consensus point in amino acid space. Although this may not correspond to a particular amino acid, the closest amino acid could easily be found.

Pilpel and Lancet[32] use a mutation data score to help analyze amino acid variability in olfactory receptor sequences. They define their score as

$$V_{\text{Lancet}} = \sum_{a}^{K}\sum_{b}^{K}\frac{p_a p_b}{M(a,b)}, \tag{24}$$

where p_a is the fractional frequency of amino acid a in the aligned column, the alphabet of amino acids is $K = 20$ and $M(a,b)$ is a BLOSUM62 or similar substitution matrix. Although this score is not directly comparable to those above, the following alteration makes it extremely similar to the lower term in the tug-of-war definition of C_{Karlin} (Eq. 17):

$$V_{\text{NotLancet}} = \sum_{a}^{K}\sum_{b}^{K} p_a p_b M(a,b), \tag{25}$$

This intermediate score now has properties almost identical to that of C_{Karlin}, so further discussion is best focused on how $C_{\text{NotLancet}}$ differs from V_{Lancet}, that is, the placing of the term $M(a,b)$. Having $M(a,b)$ as a denominator blights V_{Lancet} with an idiosyncratic output space. For instance, if a column contains a and b such that $M(a,b) = 0$, then V_{Lancet} for this column will be infinity. It is also difficult to imagine a reasonable matrix normalization that would avoid this problem. Thus, V_{Lancet} fails on at least two counts: its mathematical properties make it awkward to use and it fails to account for gaps.

Stereochemically Sensitive Entropy Scores

The entropy scores discussed above quantified symbol diversity in an elegant and intuitive way. Their problem was they failed to account for stereochemistry. Scores in this section represent attempts to build stereochemical sensitivity into the entropy model.

Entropy measures the diversity of N symbols from an alphabet κ comprising K types. The difference between a symbol of one type and that of another is ineluctably uniform. What can be changed is how κ partitions amino acid space. Recognizing the deficiencies of symbol entropy scores, Mirny and Shakhnovich[33] use the following stereochemically sensitive entropy score to analyze conservation at protein structure cores:

$$V_{\text{Mirny}} = \sum_{i}^{K} p_i \ln p_i, \tag{26}$$

that is, Shannon's entropy, where $K = 6$ and κ is the set (eligible amino acids in square brackets): aliphatic [AVLIMC], aromatic [FWYH], polar [STNQ], positive [KR], negative [DE], and special conformations [GP]. Williamson[34] provides a similar score, which he uses to look at sequence variability in transporter proteins:

$$V_{\text{Williamson}} = \sum_{i}^{K} p_i \ln\left(\frac{p_i}{\bar{p}_i}\right), \tag{27}$$

where $\bar{p}_i$ is the fractional frequency of type i in the whole alignment, $K = 9$, and κ is the set: [VLIM], [FWY], [ST], [NQ], [HKR], [DE], [AG], [P], and [C]. An improvement over the scores discussed in the section on pure entropy scores, V_{Mirny} and $V_{\text{Williamson}}$ correctly order columns (a), (b), and (c) from Figure 1 as more conserved than columns (d), (e), and (f). They also order columns (g)– (i) correctly. However, unlike the scores in that section, neither can distinguish among (a), (b), and (c) or among (d), (e), and (f). So grouping residues in this way has its price. Moreover, neither score accounts for gaps. In their analysis, Mirny and Shakhnovich acquit themselves of this charge by choosing to ignore columns that contain gaps. But the problem of how to model gaps in the entropy score remains. One solution that has not been implemented might be to factor in gaps at the end using a multiplier such as in Eq. 23.

Incorporating stereochemistry into an entropy score involves compromise. But does the choice have to be so stark between, on the one hand, a robust but stereochemically insensitive description of relative amino acid frequencies (e.g., $V_{\text{Schneider}}$) and on the other, a clumsy partitioning of the 20 amino acids that accounts for some stereochemistry but ignores relative frequencies within a partition (e.g., V_{Mirny})? There is a third way. In their pattern-induced multi-sequence alignment (PIMA) algorithm, Smith and Smith[35] use a hierarchical clustering of amino acids to extract sequence profiles from multiple alignments (Fig. 4).

Given the set amino acid types from an aligned column, PIMA finds the smallest possible "covering class" in the hierarchy that includes them all. One possible conservation score might use Shannon's entropy to assess the diversity of symbols in a column and then factor in the exclusivity of the smallest subset to which those symbols belong, for example,

$$V_{\text{PIMAInspired}} = f\left(\sum_{i}^{K} p_i \ln p_i, \gamma\right), \tag{28}$$

where $K = 21$ (i.e., 20 amino acids plus one gap symbol), γ is the cardinality of the smallest covering class (see Fig. 4),

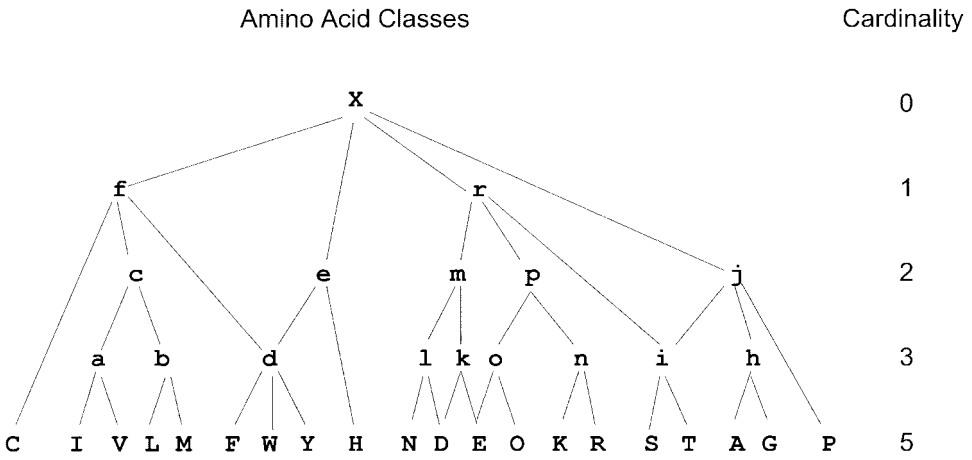

Fig. 4. Amino acid class hierarchy used in PIMA. Upper case characters are amino acids; lower case characters are amino acid classes. X is a wild-card character of any type, including a gap. In its original use, which was pairwise alignment, the match score between two aligned amino acids is the cardinality of the smallest class that includes both elements. This use is extended by $V_{\text{PIMAinspired}}$ (see text). Adapted from Ref. 35.

and f is some combining function. Gaps are penalized because they belong to only the largest superset and so have low cardinality ($\gamma = 0$). This is much like a synthesis of V_{Taylor} and $V_{\text{Schneider}}$. But because the PIMA hierarchy is ad hoc, and f is likely to be so, any statistical rigor potentially conferred by the use of entropy is lost.

Weighted Scores

Scores in this category attempt to normalize against sequence redundancy in the alignment.

Background

Normalizing against redundancy is a concern not only for scoring conservation but also for building sequence profiles. Thus, sequence weighting has received attention from exponents of both fields, and there are a large number of methods to choose from. A selection of the simpler methods is reviewed here.

The weight of a sequence is inversely related to its genetic distance from other sequences in the alignment. The simplest formulation is that given by Vingron and Argos,[36] where the weight of a sequence is equal to its average distance from all other sequences, that is,

$$w_i = \frac{1}{N-1}\sum_{j\neq i}^{N} d(s_i,s_j), \tag{29}$$

where w_i is the weight of the ith sequence, s_i, and $d(s_i,s_j)$ is the genetic distance between the ith and jth sequences, measured as their percentage identity or some more sophisticated measure. Sander and Schneider incorporate a variation of this into their HSSP database.[14] They define the weight of a sequence in not only of sequence distance but also of the weights of all other sequences:

$$\lambda w_i = \sum_{j\neq i}^{N} w_j d(s_i,s_j), \tag{30}$$

where λ is a scaling constant. Expressed in the above form, this apparently circular definition can be solved as an eigenvalue problem. Such self-consistency may be aesthetically appealing but, because it makes the weight calculation more complex, and because Sander and Schneider do not justify it, is an unnecessary mathematical flourish. More sophisticated schemes in this mold weight sequences according to how well they match a profile of the alignment (e.g., see Karchin and Hughey[37]).

Another formulation attempts to maximize the spread of data in aligned columns by using a metric related to symbol entropy.[38] This method first weights sequences at individual positions in an alignment and then combines position weights to give sequence weights. The weight of the ith sequence at position x is

$$w_{i_x} = \frac{1}{k_x n_{x_i}}, \tag{31}$$

where k_x is the number of amino acid types present in column x and n_{x_i} is the frequency of the ith sequence's amino acid at that position. By averaging along all positions in an alignment, each sequence then has weight

$$w_i = \frac{1}{L}\sum_{x}^{L} w_{i_x}, \tag{32}$$

where L is the length of the alignment. A more elaborate method that considers each aligned column separately is the position-specific independent counts method of Sunyaev et al.[39] More sophisticated weighting schemes based on entropy are described by May[40] and in Durbin et al.[10]

Sibbald and Argos apply Voronoi diagrams to the problem of sequence weighting. They consider sequences in an alignment to be a cloud of points in high dimensional space and apply the Voronoi procedure (Fig. 5), defining polyhedra around each point, and take the weight of a sequence as the volume of its surrounding polyhedron. The more

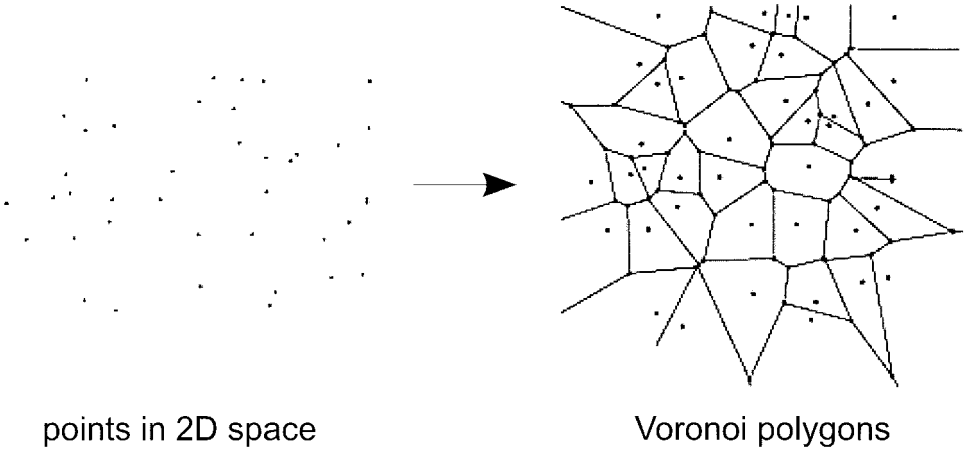

Fig. 5. Voronoi diagram. Neighboring points in a two-dimensional space are separated by a network of planes. Each plane is defined by the perpendicular to the bisector of two neighboring points. Sibbald and Argos liken each sequence to a point and the volume of the surrounding Voronoi polygon to the weight of that sequence.

isolated a sequence is, the larger its polyhedron and the greater its weight. Sibbald and Argos estimate volumes of the polyhedra by filling the high dimensional space with random sequences. They show their method calculates more intuitive weights than that of Vingron and Argos. However, the margin of difference is small and the Voronoi method is inconsistent, producing inexplicably different weights for equally redundant sequences.

Many weighting methods rely on a phylogenetic tree of the multiple alignment. One of the earliest is described by Altschul and Lipman (Ref. 41 and refs. therein). This allocates weights according to Kirchhoff's laws, which describe how charge and voltage are distributed in an electrical circuit. The tree is viewed as a system of wires and nodes, and a voltage is applied to the root. Kirchhoff's Current Law enforces conservation of charge: moving from the root to the leaves, it distributes current at each node so that the amount of charge entering at the root equals the amount exiting from the leaves. The current exiting at a leaf is taken as the weight of the corresponding sequence. So far so good. But the current entering a node is not necessarily distributed equitably among the outputs (branches). Rather, this distribution is governed by Kirchhoff's Voltage Law, which apportions greater current to the branch with more leaves. This inequitable distribution may be sensible for electric currents or water systems, but it acts contrary to the motives of sequence weighting. Given a node that bifurcates into a highly populated subfamily and a sparsely populated one, the highly populated subfamily will receive the larger share of current and thus be upweighted.

Gerstein et al.[42] approach the tree from the opposite direction: they start with the leaves and work up to the root, each sequence accumulating a share of the branch length as its weight. More sophisticated tree-based methods are described by Altschul et al.[43] and Eddy et al.[44] and are discussed at length in Durbin et al.[10] Tree-based weighting schemes require more assumptions than those based on only the alignment. After all, many plausible trees can describe a single alignment. Choosing one, even if it is the most probable, introduces additional uncertainty and thus hidden complexity.

Scores

Sequence weighting can be more easily incorporated into some scoring models than others. The sum-of-pairs model accumulates contributions on a per sequence-pair basis. This provides an obvious placing for sequence weights. For entropy scores, which bundle amino acids according to type and disregard for which sequence each came from, the placing is less obvious. Perhaps for this reason, weighted scores have tended to follow the SP model.

Landgraf et al.[45] use the following score to extend the evolutionary trace method of Lichtarge et al.,[28]

$$V_{\text{Landgraf}}(x) = \frac{1}{N}\sum_{i}^{N}\sum_{j>i}^{N}\left(w_i D(s_i(x),s_j(x)) + w_j D(s_j(x),s_i(x))\right), \tag{33}$$

where $s_i(x)$ is the amino acid at position x of the ith sequence and w_i is the weight of sequence s_i as calculated by the Voronoi scheme of Sibbald and Argos (see above). $D(a,b)$ measures the dissimilarity of the amino acids a and b and is calculated as

$$D(a,b) = \frac{m(a,a) - m(a,b)}{m(a,a)}, \tag{34}$$

where m is the Gonnet substitution matrix.[46] One of the first things to notice about D is its asymmetry. Intuitively, the difference between two amino acids is commutative such that $D(a,b) = D(b,a)$. However, because, as in most matrices, the diagonal scores in the Gonnet matrix differ depending on the amino acid, there are many cases when

$m(a,a) \neq m(b,b)$ and therefore $D(a,b) \neq D(b,a)$. Landgraf et al. recognize this inconsistency and hedge their bets in Eq. 33, with a sum of the form $w_iD(a,b) + w_jD(b,a)$. However, because this may give a different result from $w_jD(a,b) + w_iD(b,a)$, their handling of D's asymmetry is somewhat arbitrary.

Sander and Schneider do not use sequence weights as such.[14] Rather, they moderate comparisons by the genetic distance between the sequences being compared:

$$C_{\text{Sander}}(x) = \lambda \sum_{i}^{N} \sum_{j>i}^{N} d(s_i,s_j)m(s_i(x),s_j(x)), \quad (35)$$

where $d(s_i(x),s_j(x))$ is the distance between sequence s_i and s_j measured as 100% minus their percentage identity in the alignment, m is the Dayhoff substitution matrix,[19] and λ scales C_{Sander} to range [0,1], that is,

$$\lambda = \left(\sum_{i}^{N} \sum_{j>i}^{N} d(s_i,s_j) \right)^{-1}. \quad (36)$$

Valdar and Thornton[47] use sequence weights in the Vingron and Argos mold for their SP score:

$$C_{\text{Valdar}}(x) = \lambda \sum_{i}^{N} \sum_{j>i}^{N} w_i w_j M(s_i(x),s_j(x)), \quad (37)$$

where λ scales C_{Valdar} to range [0,1], that is,

$$\lambda = \left(\sum_{i}^{N} \sum_{j>i}^{N} w_i w_j \right)^{-1}. \quad (38)$$

Their comparison matrix M is a linear transformation of the substitution matrix m such that M takes values in the range [0,1] and all exchanges involving a gap score 0, that is,

$$M(a,b) = \begin{cases} \dfrac{m(a,b) - \min(m)}{\max(m) - \min(m)} & \text{if } a \neq \text{gap and } b \neq \text{gap} \\ 0 & \text{otherwise} \end{cases}. \quad (39)$$

m itself is a modified version of the pairwise exchange table (PET),[48] differing from the original in that all diagonal elements are set constant and equal to the rounded average of diagonal elements in the unmodified matrix.

V_{Landgraf}, C_{Sander}, and C_{Valdar} are much alike and all fulfill or partially fulfill many of the criteria laid out at the beginning of the review. All three correctly order columns (a)–(f) and (g)–(i) in Figure 1 (remember that all sequences in that alignment have equal weights). Their output space is continuous and bounded; they account for amino acid frequency; they quantify stereochemical diversity with a full substitution matrix; and they normalize against redundancy in the alignment.

There are also differences. Only C_{Valdar} penalizes gaps and so only that score correctly orders columns (j) and (k). In C_{Sander} and C_{Valdar}, weights are derived in a way that is simpler and more consistent than in V_{Landgraf} but which may give marginally inferior results. V_{Landgraf}, C_{Sander}, and C_{Valdar} illustrate three distinct ways weights can be incorporated into an SP score. Surprisingly, as the remainder of this section shows, the method of incorporation may be more important than the choice of weighting metric.

Consider an alignment of three sequences that are all equally different from one another. The column at position x in the alignment contains three different amino acids. The first sequence, s_1, has a W at this position, sequence s_2 has a Q, and s_3 an R. Because the sequences are uniformly different, this alignment is "ideal" and requires no sequence weighting. Applying a simple unweighted sum-of-pairs score gives the result

$$C_{\text{simple}}(x_{\text{ideal}}) = \sum_{i}^{N} \sum_{i>j}^{N} M(s_i(x),s_j(x)) = M(\text{W, Q}) + M(\text{W, R}) + M(\text{Q, R}), \quad (40)$$

where M is a symmetric similarity measure. Now add duplicates of sequence s_3 to the alignment to make it redundant and in need of sequence weighting. Figure 6 shows column x, which now contains one W, one Q, and n R's, corresponding to the n duplicates of s_3. Applying C_{simple} to the new alignment gives the result

$$C_{\text{simple}}(x) = M(\text{W, Q}) + n(M(\text{W, R}) + M(\text{Q,R})) + \frac{n(n-1)}{2} M(\text{R, R}), \quad (41)$$

Clearly, as n increases, two undesirable things happen. First, the M(W,Q) term vanishes out of existence. Second, the spurious M(R,R) term dominates. A good weighted SP score applied to the redundant alignment should at best reproduce the result in Eq. 40, at least moderate the affects of increasing n, and at worst reproduce the result in Eq. 41.

Let the distance between sequences $d(s_i(x), s_j(x))$ be 0 if $s_i = s_j$ and 1 otherwise. A C_{Sander}-like modification to C_{simple} gives

$$C_{\text{distance}}(x) = \sum_{i}^{N} \sum_{j>i}^{N} d(s_i,s_j)M(s_i(x),s_j(x)), \quad (42)$$

which, when applied to the redundant position x gives

$$C_{\text{distance}}(x) = M(\text{W, Q}) + n(M(\text{W, R}) + M(\text{Q, R})). \quad (43)$$

This is certainly an improvement on C_{simple} because M(R,R) has been factored out. However, it still has the problem that as n increases, M(W,Q) disappears, and the only effective comparisons are those involving s_3. A V_{Landgraf} -like modification (ignoring inconsistencies) to C_{simple} gives

$$C_{\text{sum}}(x) = \sum_{i}^{N} \sum_{j>i}^{N} (w_i + w_j)M(s_i(x),s_j(x)). \quad (44)$$

Sequence	Column x
s_1	W
s_2	Q
s_{3_1}	R
⋮	⋮
s_{3_n}	R

Fig. 6. Column from a redundant-sequence alignment. Sequence s_{3_i} is the ith copy of sequence s_3. See text for details.

For simplicity, we calculate w_i similarly to Vingron and Argos (Eq. 29) as

$$w_i = \frac{1}{N-1}\sum_{j>i}^{N} d(s_i,s_j). \quad (45)$$

(This is reasonable because comparisons of this method with other weighting methods by Gerstein et al.[42] and Sibbald and Argos$_{49}$ showed it was only slightly inferior). According to this scheme, $w_1 = w_2 = n + 1$, whereas the weight of a single duplicate of s_3 is $w_3 = 2$. Applying C_{sum} to the redundant column gives

$$C_{\text{sum}} = (2n+2)M(\text{W, Q}) + (n^2+3n)(M(\text{W, R}) + M(\text{Q, R})) + (2n^2-2n)M(\text{R, R}). \quad (46)$$

This result is better than in Eq. 41 but worse than in Eq. 43. M(W,Q) will still disappear because it increases with linearly with n, whereas the other terms increase geometrically. M(R,R) is also present. The C_{Valdar} strategy is

$$C_{\text{product}}(x) = \sum_{i}^{N}\sum_{j>i}^{N} w_i w_j M(s_i(x),s_j(x)), \quad (47)$$

which, when applied to the column in Figure 6 gives

$$C_{\text{product}} = (n^2+2n+1)M(\text{W, Q}) + (2n^2+2n) \times (M(\text{W, R}) + M(\text{Q, R})) + (2n^2-2n)M(\text{R, R}). \quad (48)$$

All terms are now on an equal footing with respect to n. Result 48 is clearly better than result 46 or 41. It is arguably more desirable than result 43 in that, although the spurious M(R,R) features, no term disappears with increasing n.

A GENERALIZED FORMULA FOR SCORING CONSERVATION

No score is perfect, but some scores are less perfect than others. Scores discussed later in the survey tended to satisfy more of the criteria outlined at the start of the review than those discussed earlier. Shannon's entropy offered an elegant way to measure diversity among uniformly different symbols but faltered when accounting for stereochemistry. Property-based scores (e.g., V_{Taylor}) respected stereochemistry but failed to register symbol diversity. The most successful compromises were seen in the sum-of-pairs scores, although they exposed some limitations of using substitution matrices.

So far this review has mainly discussed scores following either the entropy or the substitution matrix model. But is this dichotomy inevitable? One could devise a score that plays entropy and mutation data to their relative strengths by keeping the assessment of relative symbol frequencies and the assessment of stereochemistry separate. An example of such a score is considered here.

Positional variability may be seen to have three elements:

- symbol diversity, normalized to take account of sequence redundancy
- stereochemical diversity
- gaps

For a given position, each element can be assigned a score that measures the extent to which it describes that column. Let t be the normalized symbol diversity (diversity), let r be the stereochemical diversity (stereochemistry), and let g be the gap cost (gap). For convenience, all measures are continuous and bounded in the range 0–1, where 0 means that element is not present and 1 means that element is at its maximum. For instance, $r = 0$ means there is no stereochemical diversity at the position, whereas $r = 1$ means the position could not be any more stereochemically diverse. Conservation is a function of these three variables. More intuitively, an assessment of conservation can be seen as a three-pronged attack: a position is criticized on its symbol diversity, its stereochemical diversity, and its gappiness. For a position x, we can write

$$C_{\text{trident}}(x) = (1-t(x))^{\alpha}(1-r(x))^{\beta}(1-g(x))^{\gamma}. \quad (49)$$

The exponents α β, and γ weight the importance of each element. Suppose they are all equal to one. If position x is strictly conserved, then $C_{\text{trident}} = (1–0) \times (1–0) \times (1–0) = 1$. As position x becomes more afflicted with gaps, stereochemical diversity or symbol diversity, C_{trident} drops toward zero. The relative impacts of these three elements on the conservation score were rigidly prescribed in C_{Valdar}. In C_{trident}, however, the sharpness of each prong may be adjusted freely to suit the purpose of the user. For example, if C_{trident} with $\alpha = \beta = \gamma = 1$ is too lenient on gaps and too strict on stereochemistry for a particular application, one could instead try $\alpha = 1$, $\beta = 1/2$, and $\gamma = 2$.

C_{trident} is so far more a convenient division of labor than a score, because it is open how any particular prong is

defined. To make the score more concrete, one can start by specifying t as Shannon's entropy:

$$t(x) = \lambda_t \sum_a^K p_a \log_2 p_a, \tag{50}$$

where K, the alphabet size, is 21 (20 amino acids plus one gap symbol), and p_a is the probability of observing the ath symbol type. λ_t scales the entropy to range [0,1] and is defined as

$$\lambda_t = [\log_2(\min(N,K))]^{-1}, \tag{51}$$

where N is the number of sequences in the alignment, so that $t(x)$ can reach its maximum of 1 even when there are fewer that K amino acids in the column. Sequence weighting can be incorporated into Shannon's entropy by normalizing each p_a thus

$$p_a = \sum_{i \in \{i : s_i(x) = a\}} w_i \tag{52}$$

where w_i is the weight of the ith sequence and $s_i(x)$ is the symbol type at position x in that sequence. In words, the probability of observing symbol type a is the summed weight of sequences manifesting a. Ideally, the sum of all weights should be 1. Therefore, an apposite weighting scheme, which is related to an entropy model, is that of Henikoff and Henikoff[38]:

$$w_i = \frac{1}{L} \sum_x^L \frac{1}{K_x n_{x_i}}, \tag{53}$$

where L is the length of the alignment, k_x is the number of symbol types present at the xth position and n_{x_i} is the number of times the symbol type manifested by the ith sequence occurs at that position.

The second prong of C_{trident} measures stereochemical diversity but does not need to take account of symbol frequency or gaps. One candidate for this is V_{Zvelibil}. The one used here, which is less ad hoc, uses a substitution matrix and is related to the scoring model used in C_{Thompson}. Let amino acid a be represented by a point $\mathbf{X}_a$ in 20-dimensional space such that

$$\mathbf{X}_a = \begin{bmatrix} M(a,a_1) \\ M(a,a_2) \\ \vdots \\ M(a,a_{20}) \end{bmatrix}, \tag{54}$$

where a_i is the ith amino acid type. For example, the position of Cys in this space is defined by its mutational proximity to all other amino acids. $M(a,b)$ is the similarity between amino acids a and b judged by a normalized substitution matrix. One consistent normalization would be that of Karlin and Brocchieri[25] (Eq. 16). For a position x, the consensus amino acid type is calculated as point $\bar{\mathbf{X}}(x)$:

$$\bar{\mathbf{X}}(x) = \frac{1}{k_x} \sum_a^{k_x} \mathbf{X}_a, \tag{55}$$

where k_x is the number of amino acid types present in the column. The stereochemical diversity may be calculated as the average distance of observed amino acids from the consensus point:

$$r(x) = \lambda_r \frac{1}{k_x} \sum_a^{k_x} |\bar{\mathbf{X}}(x) - \mathbf{X}_a|, \tag{56}$$

where the scalar $\lambda_r = [\sqrt{20(\max(M) - \min(M))^2}]^{-1}$ ensures $r = 1$.

The third prong of C_{trident}, the gap cost, is more straightforward. The more gaps, the less selective pressure is assumed to have acted at the position. Thus, $g(x)$ can be defined simply as the fraction of symbols in column x that are gaps.

C_{trident} is not so much one score but a framework in which many different conservation scores can be imitated. For instance, if $\alpha = 1$, $\beta = 0$, and $\gamma = 0$, C_{trident} resembles $C_{\text{Schneider}}$ (except for the sequence weighting). However, because C_{trident} raises more questions than it answers, it is better deployed in the analysis of conservation scores than in scoring conservation as such. After all, such a versatile framework can imitate uninformative scores as well as useful ones. An intriguing question is whether C_{trident} can imitate one of the weighted SP scores. To investigate this, scores from C_{Valdar} were compared with scores from C_{trident} for different values of α, β, and γ. Specifically, C_{Valdar} and C_{trident} were used to score all positions in six multiple-sequence alignments. The similarity of the two scores was measured as Pearson's correlation coefficient of the two outputs. This was performed for 1000 different sets of α, β, and γ. To avoid unnecessary confounding variables, the similarity matrix used for C_{trident} was the same as that used for C_{Valdar}. The alignments used were those belonging to the six homodimer families examined by Valdar and Thornton.[47] This data set comprised 1595 residue positions in all and contained data from 195 sequences. In this cursory investigation, the maximum correlation reached was 0.98 when $\alpha = 1$, $\beta = 0.5$, and $\gamma = 3$. This correlation seems high. However, because the data set used is small and may not uniformly exercise all aspects of C_{Valdar}, this result should be considered only as a rough estimate. In particular, the analysis showed that when α and β are optimal, varying γ has little effect on the correlation. This reflects the small number of gaps in the six carefully compiled alignments. A gappier set of alignments might raise the profile of this third parameter. Acknowledging these caveats, it is, nevertheless, interesting to contrast the parameter set necessary to simulate C_{Valdar} with that required to simulate $C_{\text{Schneider}}$. For both scores, $\alpha = 1$. This reflects the fact that C_{Valdar} and $C_{\text{Schneider}}$ both account for the relative frequencies of amino acids. The two scores differ on β. For C_{Valdar}, β is nonzero, indicating this score is sensitive toward stereochemistry, whereas for $C_{\text{Schneider}}$, $\beta = 0$, indicating stereochemistry is ignored. C_{Valdar} penalizes gaps, whereas $C_{\text{Schneider}}$ does not. Similarly, $\gamma = 3$ for C_{Valdar}, indicating this score's acknowledgment of gaps, whereas for $C_{\text{Schneider}}$, which does not penalize gaps, $\gamma = 0$.

C_{trident} combines the strengths of previously disparate approaches. Although its flexibility undermines any authority it has as a concrete score, it does provides a framework for dissecting the character of other scores. This kind of meta-analysis is interesting from an abstract theoretical point of view. It may also be useful in a more practical sense. Given a data set of multiple alignments with "correct" scores—these scores might be inferred from orthogonal information relating to the importance of particular residues in structure or function—the parameters of α, β, and γ could be optimized so that C_{trident} imitates these scores. At present, however, such data sets are not available and even somewhat difficult to conceive.

CONCLUSIONS

This review has described 18 scores (not including C_{trident}) and several distinct approaches for quantifying evolutionary conservation at an aligned position. No score achieved both biological and statistical rigor. The most meaningful scores were relatively ad hoc. However, given the success of probabilistic sequence profiles,[3,50] which are a different but related emprise, it seems likely that a statistically robust score is possible and will emerge. Until then, there are a number of powerful measures available (described above), which, despite their shortcomings, have proved useful in the analysis of conservation.

ACKNOWLEDGMENTS

William Valdar thanks Roman Laskowski, Andrew Martin, Richard Mott, and Janet Thornton for helpful discussions, and Geoff Barton for giving permission to reproduce two figures from his article.[21] This work was supported by a BBSRC Special studentship.

REFERENCES

1. Page RDM, Holmes EC. Molecular evolution: a phylogenetic approach. 2nd ed. Oxford: Blackwell Science; 1998.
2. Altschul SF, Madden TL, Schäffer, AA, Zhang J, Zhang Z, Miller W, Lipman D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402.
3. Eddy SR. Hidden markov models. Curr Opin Struct Biol 1996;6: 361– 365.
4. Wu TT, Kabat EA. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. J Exp Med 1970;132:211–249.
5. Shenkin PS, Erman B, Mastrandrea LD. Information-theoretical entropy as a measure of sequence variability. Proteins 1991;11:297–313.
6. Jores R, Alzari PM, Meo T. Resolution of hypervariable regions in T-cell receptor β chains by a modified Wu-Kabat index of amino acid diversity. Proc Natl Acad Sci USA 1990;87:9138–9142.
7. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 1999;286:295–299.
8. Shannon CE. A mathematical theory of communication. The Bell System Technical J 1948;27:379–423, 623–656.
9. Baczkowski AJ, Joanes DN, Shamia GM. Properties of a generalized diversity index. J Theor Biol 1997;188:207–213.
10. Durbin R, Eddy S, Krogh A, Mitchison G. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge (UK): Cambridge University Press; 1998.
11. Good IJ. The population frequencies of species and the estimation of population parameters. Biometrika 1953;40:237–264.
12. Gregory RL, Zangwill OL. The Oxford companion to the mind. Oxford (UK): Oxford University Press; 1987.
13. Schneider TD. Information content of individual genetic sequences. J Theor Biol 1997;189:427–441.
14. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
15. Gerstein M, Altman RB. Average core structures and variability measures for protein families: application to the immunoglobulins. J Mol Biol 1995;251:161–175.
16. Mirny L, Shakhnovich E. Evolutionary conservation of the folding nucleus. J Mol Biol 2001;308:123–129.
17. Plaxco K, Larson S, Ruczinski I, Riddle D, Buchwitz B, Davidson A, Baker D. Evolutionary conservation in protein folding kinetics. J Mol Biol 2000;298:303–312.
18. Taylor WR. The classification of amino acid conservation. J Theor Biol 1986;119:205–218.
19. Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins: matrices for detecting distant relationships. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Washington (DC): National Biomedical Research Foundation; 1978. p 345–358.
20. Zvelibil MJ, Barton GJ, Taylor WR, Sternberg MJ. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J Mol Biol 1987;195:957–961.
21. Livingstone CD, Barton GJ. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput Appl Biosci 1993;9:745–756.
22. Hughy R, Krogh A. Hidden Markov models for sequence analysis: extension and analysis of the basic method. Comput Appl Biosci 1996;12:95–107.
23. Sjölander K, Karplus K, Brown M, Hughey R, Krogh A, Mian S, Haussler D. Dirichlet mixtures: a method for improving detection of weak but significant protein sequence homology. Comput Appl Biosci 1996;12:327–345.
24. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 1998;284:1201–1210.
25. Karlin S, Brocchieri L. Evolutionary conservation of RecA genes in relation to protein structure and function. J Bacteriol 1996;178: 1881–1894.
26. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
27. Armon A, Graur D, Ben-Tal N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J Mol Biol 2001;307:447–463.
28. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 1996;257:342–358.
29. Miyata T, Miyazawa S, Yashunaga T. Two types of amino acid substitutions in protein evolution. J Mol Evol 1979;12:219–236.
30. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTALX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997;25:4876–4882.
31. Higgins DG, Thompson JD, Gibson TJ. Using CLUSTAL for multiple sequence alignments. Methods Enzymol 1996;266:383–402.
32. Pilpel Y, Lancet D. The variable and conserved interfaces of modeled olfactory receptor proteins. Protein Sci 1999;8:969–977.
33. Mirny LA, Shakhnovich EI. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J Mol Biol 1999;291:177–196.
34. Williamson RM. Information theory analysis of the relationship between primary sequence structure and ligand recognition among a class of facilitated transporters. J Theor Biol 1995;174:179–188.
35. Smith RF, Smith TF. Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modeling. Protein Eng 1992;5:35–41.
36. Vingron M, Argos P. A fast and sensitive multiple sequence alignment algorithm. Comput Appl Biosci 1989;5:115–121.
37. Karchin R, Hughey R. Weighting hidden Markov models for maximum discrimination. Bioinformatics 1998;14:772–782.

38. Henikoff S, Henikoff JG. Position-based sequence weights. J Mol Biol 1994;243:574–578.
39. Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber BE, Tumanyan VG, Kuznetsov EN. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng 1999;12:387–394.
40. May AC. Optimal classification of protein sequences and selection of representative sets from multiple alignments: application to homologous families and lessons for structural genomics. Protein Eng 2001;14:209–217.
41. Altschul SF, Lipman DJ. Equal animals. Nature 1990;348:493–494.
42. Gerstein M, Sonnhammer ELL, Chothia C. Volume changes in protein evolution. J Mol Biol 1994;236:1067–1078.
43. Altschul SF, Carroll RJ, Lipman DJ. Weights for data related by a tree. J Mol Biol 1989;207:647–653.
44. Eddy SR, Mitchison G, Durbin R. Maximum discrimination hidden Markov models of sequence consensus. J Comp Biol 1995;2:9–23.
45. Landgraf R, Fischer D, Eisenberg D. Analysis of heregulin symmetry by weighted evolutionary tracing. Protein Eng 1999;12:943–951.
46. Benner SA, Cohen MA, Gonnet GH. Amino acid substitution during functionally constrained divergent evolution of protein sequences. Protein Eng 1994;7:1323–1332.
47. Valdar WSJ, Thornton JM. Protein-protein interfaces: analysis of amino acid conservation in homodimers. Proteins 2001;42:108–124.
48. Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 1992;8:275–282.
49. Sibbald PR, Argos P. Weighting aligned protein or nucleic acid sequences to correct for unequal representation. J Mol Biol 1990;216:813–818.
50. Mott R. Accurate formula for P-values of gapped local sequence and profile alignments. J Mol Biol 2000;300:649–659.