BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 19 2005, pages 3704–3710
doi:10.1093/bioinformatics/bti616

Sequence analysis

Pairwise alignment incorporating dipeptide covariation

Gavin E. Crooks[1,*,†], Richard E. Green[1,2,†] and Steven E. Brenner[1,2]

[1]Department of Plant and Microbial Biology, 111 Koshland Hall #3102 University of California, Berkeley, CA, USA and [2]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720-3102, USA

Received on March 16, 2005; revised on July 28, 2005; accepted on August 4, 2005

Advance Access publication August 25, 2005

ABSTRACT

Motivation: Standard algorithms for pairwise protein sequence alignment make the simplifying assumption that amino acid substitutions at neighboring sites are uncorrelated. This assumption allows implementation of fast algorithms for pairwise sequence alignment, but it ignores information that could conceivably increase the power of remote homolog detection. We examine the validity of this assumption by constructing extended substitution matrices that encapsulate the observed correlations between neighboring sites, by developing an efficient and rigorous algorithm for pairwise protein sequence alignment that incorporates these local substitution correlations and by assessing the ability of this algorithm to detect remote homologies.

Results: Our analysis indicates that local correlations between substitutions are not strong on the average. Furthermore, incorporating local substitution correlations into pairwise alignment did not lead to a statistically significant improvement in remote homology detection. Therefore, the standard assumption that individual residues within protein sequences evolve independently of neighboring positions appears to be an efficient and appropriate approximation.

Availability: Sequence data, software and matrices are freely available from http://compbio.berkeley.edu/

Contact: gec@compbio.berkeley.edu

Supplementary information: Supplementary data for this paper is available at *Bioinformatics* online.

1 INTRODUCTION

Among the most commonly used tools in computational biology are the pairwise protein sequence alignment methods, such as SSEARCH, FASTA and BLAST (Smith and Waterman, 1981; Pearson and Lipman, 1988; Altschul *et al*., 1990; Durbin *et al*., 1998). These algorithms are elegant, efficient and effective methods of detecting similarity between closely related protein sequences. However, the ability of fast pairwise methods to detect homology deteriorates as the divergence between the sequences increases. Past the 'twilight zone' (20–30% pairwise sequence identity), only a small fraction of related proteins can be found (Sander and Schneider, 1991; Doolittle, 1992; Brenner *et al*., 1998; Green and Brenner, 2002). Therefore, in order to make better use of the vast and increasing amount of available biological sequence data, there is an immediate need for more sensitive, fast database search methods.

For the sake of computational efficacy, current pairwise alignment methods make several simplifying assumptions. First, amino acid substitutions are assumed to be homogeneous between protein families. The most commonly used substitution matrices [BLOSUM (Henikoff and Henikoff, 1992) and PAM (Dayhoff *et al*., 1978)] are thus generic models of protein sequence evolution across all protein sequence families at various evolutionary distances. Second, substitutions at a given site are assumed to be uncorrelated with those on neighboring sites, i.e. the likelihood of substituting an amino acid X for amino acid Y is assumed to be independent of the sequence context of X. It is known that both of these simplifying assumptions introduce errors into homology searching. Relaxing the assumption of homogeneous substitution across protein families can significantly improve the performance of pairwise alignment methods (Yu *et al*., 2003). Furthermore, alignment methods that remove the assumption of homogeneity among different positions in the sequence, and instead model the heterogeneity of the given protein family, have been found to be dramatically superior for remote homology detection (Park *et al*., 1998, R. E. Green and S. E. Brenner, Unpublished data). Unfortunately, these profile methods [e.g. PSI-BLAST (Altschul *et al*., 1997), HMMER (http://hmmer.wustl.edu/) (Eddy, 2001), SAM (Karplus *et al*., 1998)] are not tractable for all query sequences. They require the presence, identification and correct alignment of homologous sequences in order to generate a model of the query sequence's family. Therefore, the fast and universally applicable pairwise methods remain widely used for database searching, despite their lower sensitivity.

One proposed strategy for increasing the sensitivity of pairwise alignment is to use a more sophisticated scoring function for amino acid substitutions, namely one that is sensitive to the sequence context in which the residue resides. For example, amino acid sequences are correlated with secondary structural features, such as helixes and loops, which can directly lead to structure-dependent substitution patterns (Thorne *et al*., 1996; Topham *et al*., 1997; Goldman *et al*., 1998). Similarly, one might intuitively expect structurally and functionally important residues, such as cysteines and prolines, to be more or less conserved depending on their local sequence environment and the prevalence of particular motifs.

*To whom correspondence should be addressed at Physical Biosciences Division, Lawrence Berkeley Natl Lab., Berkeley, CA 94720, USA

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

The first large-scale exploration of the effect of sequence context on amino acid evolution was performed by Gonnet *et al*. (1994), who examined the frequencies of dipeptide substitutions, and compared them with the dipeptide substitution frequencies expected assuming no sequence dependent correlations. Despite the fact that nearly half of the elements of the 400 × 400 observed dipeptide matrix were vacant (owing to the sparsity of data) several interesting patterns were evident. The chief trend was that amino acids are generally more likely to be conserved if they are adjacent to positions that are also conserved.

More recently Jung and Lee (2000) have taken advantage of the large increase in available data to reexamine trends in dipeptide evolution. They used the observed patterns of substitution within a large set of structure-based alignments to generate dipeptide substitution matrices. Furthermore, they developed an extension to the standard Smith–Waterman alignment algorithm that incorporates a term from these dipeptide matrices. By using sequence and structure context information, they show some improvement in homolog detection in a limited test set. However, their method could not be extensively tested or practically utilized because an efficient dynamic programming method for finding the optimal alignment was not known to the authors. Instead, they adopted a heuristic search that is not guaranteed to find optimal alignments.

In this study, we have extended the work described above by examining the strength of local, dipeptide substitution correlations using the massive amount of alignment data within the BLOCKS database. We have also extended the standard Smith–Waterman algorithm to include local dipeptide correlation information over a user-defined distance. Similar to Smith–Waterman, this new polynomial time algorithm, `doublet`, finds the optimal alignment under the scoring scheme described. Using a standard remote homolog detection evaluation strategy, we have tested `doublet` against the Smith–Waterman algorithm to measure the impact of including this extra information. Perhaps surprisingly, we found that incorporating doublet substitution correlations leads to a statistically insignificant difference in homology detection.

2 METHODS

2.1 Quantifying substitution correlations

Consider two aligned, ungapped sequences, $x = x_1, x_2, \cdots, x_n$ and $y = y_1, y_2, \cdots, y_n$, both of length n, where each element represents one of the 20 canonical amino acids. We wish to use the patterns of conservation and variation between these sequences to estimate the log odds that the sequences are homologous (i.e. that both sequences have descended from a common ancestor).

$$\mathcal{S} = \log \frac{q(x;y)}{p(x)p(y)} = \log \frac{q(x_1,x_2,\cdots,x_n;y_1,y_2,\cdots,y_n)}{p(x_1,x_2,\cdots,x_n)p(y_1,y_2,\cdots,y_n)}. \quad (1)$$

Here, $p(x)$ is the background probability of the given amino acid segment and $q(x;y)$ is the target probability of observing the pair of segments in diverged homologous sequences.

Except for very short segments, the background and target probability distributions are large and cannot be directly measured. Therefore, Equation (1) is typically simplified by assuming that substitutions probabilities are homogeneous (independent of the location in the fragment) and that both the substitutions and the sequences themselves are uncorrelated from one position to the next. Consequentially, the total similarity score is now a sum of independent parts,

$$\mathcal{S} \approx \sum_k s(x_k;y_k), \quad s(i;j) = \log \frac{q(i;j)}{p(i)p(j)} \quad (2)$$

The log odds of residue replacement, $s(i,j)$, is an element of a standard singlet substitution matrix, of the type widely used in pairwise sequence alignment (Altschul, 1991).

This approximation of the full similarity by a sum of singlet substitution scores requires that we neglect all intersite correlations. We can perform a more controlled approximation by noting that a homogeneous multivariate probability can be expanded into a product of single component distributions, pairwise correlations, triplets correlations and so on.

$$P(z_1,z_2,\cdots,z_n) = \prod_i P(z_i) \times \prod_{i<j} \frac{P(z_i,z_j)}{P(z_i)P(z_j)} \times \prod_{i<j<k} \frac{P(z_i,z_j,z_k)P(z_i)P(z_k)P(z_j)}{P(z_i,z_j)P(z_i,z_k)P(z_j,z_k)} \cdots \quad (3)$$

If we assume that substitution probabilities are independent of the location within the fragment, then we can apply this expansion to the segment homology score [Equation (1)].

$$\mathcal{S} = \sum_{k=1}^{n} s(x_k;y_k) + \sum_{l=1}^{L}\sum_{k=1}^{n-L} d_l(x_k,x_{k+l};y_k,y_{k+l}) + \cdots \quad (4)$$

The first term of this expansion represents single residue replacements, as in Equation (2). The next term defines the doublet substitution scores,

$$d_l(i,i';j,j') = \log \frac{q_l(i,i';j,j')}{p_l(i,i')p_l(j,j')} - s(i;j) - s(i';j'). \quad (5)$$

Here, i and i' are residues separated by a distance l along one amino acid chain, whereas j and j' are the corresponding aligned residues on the putative homologous sequence; $q_l(i,i';j,j')$ is the target probability of observing this aligned quartet, and $p_l(i,i')$ is the background probability of this residue pair in protein sequences. These doublet scores represent the additional similarity owing to correlations between substitutions.

By truncating the expansion of the full similarity score at doublet terms [Equation (4)], we are assuming that triplet and higher order correlations between substitutions are relatively uninformative. For reasons discussed below, this is probably a reasonable approximation. Furthermore, the most important intersite correlations are between residues neighboring on the chain (Fig. 3). Therefore, we can restrict the maximum distance over which doublet interactions are scored without serious error.

The average similarity score is the interhomolog mutual information I (Cover and Thomas, 1991), a measure of the interequence correlations. A high mutual information value indicates strong correlation, whereas a mutual information value of zero indicates uncorrelated variables. Mutual information has various advantages as a correlation measure: it is firmly grounded in information theory, it is additive for independent contributions and it has consistent, intuitive units (bits).

$$I(x;y) = \sum q(x,y) \log_2 \frac{q(x,y)}{p(x)p(y)}. \quad (6)$$

The average singlet score is the interhomolog mutual information per residue, under the assumption that replacements are uncorrelated. This is frequently reported as the 'relative entropy' of the substitution matrix. The average doublet score is the first order correction to the intersequence mutual information owing to intersite correlations. Consequentially, we may evaluate the comparative importance of singlet and doublet contributions to the sequence similarity by examining the average contributions of these different components to the full interhomolog mutual information.

The preceding analysis applies to contiguously aligned sequence segments. However, in addition to substitutions, protein sequences are modified by the insertion and deletion of residues. Since it is not obvious how to capture the existence of indels in doublet scores, in the following discussion we assume that dipeptide correlations do not extend across gaps, and we adopt the simple and standard affine model of gap lengths. This approximation should have little impact, since aligned detectably homologous sequences tend to have relatively few indels, particulary in regions that are significantly similar.

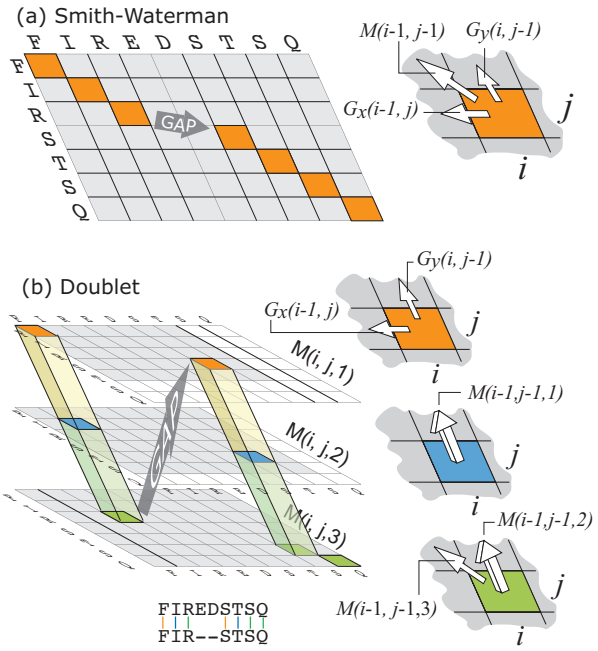

Fig. 1. A comparison of Smith–Waterman and `doublet` sequence alignment. (**a**) A Smith–Waterman match table, with the optimal alignment highlighted. The value of each cell is the maximum of (1) The singlet match score (this is the start of an alignment), (2) The singlet score plus the match score from the previous cell along the diagonal (this extends an aligned region) or (3) the singlet score plus the optimal score from a gap score table (the previous residue was not aligned) (**b**) For `doublet`, multiple match tables are used [Equations (9–11)]. The number of match tables is one plus the distance over which dipeptide correlation information is considered (in this example, two). Again, the optimal alignment is highlighted. The top table corresponds to the starts of aligned regions, the middle table corresponds to aligned regions of at least two consecutive residues and the bottom table corresponds aligned regions of at least three consecutive residues. The alignment path through these tables falls through to lower tables in regions of consecutive aligned residues and begins again in the top table following gaps. To extend dipeptide context scoring over longer distances requires additional match tables.

2.2 Alignment algorithm

We have extended the standard Smith–Waterman optimal local sequence alignment algorithm (Smith and Waterman, 1981) to incorporate doublet substitution scores (Fig. 1). The time complexity of Smith–Waterman is $O(nm)$, where n and m are the lengths of the two sequences. Adding doublet scores increases the complexity to $O(nmL)$, where L is the distance over which substitution correlations are scored. This efficient dynamic programming alignment is possible because, although we are scoring correlations between residues that are not directly aligned, these correlations are local along the chain. The space complexity of our implementation is also $O(nmL)$, which could be improved using standard techniques (Durbin *et al.*, 1998).

The additional similarity score associated with adding the final match pair x_i, y_j to the alignment contains singlet (S) doublet (D) substitution scores;

$$S(i,j) = s(x_i, y_j), \tag{7}$$

$$D(i,j,r) = \sum_{l=1}^{r} d_l(x_{i-l}, x_i; y_{j-l}, y_j). \tag{8}$$

Here, r is the length of the preceding contiguous segment of aligned residues, or the maximum sequence separation over which doublet correlations are scored, whichever is less. Deletions of length k are weighted with the affine penalty $-(g_{\text{open}} + (k-1)g_{\text{ext}})$, where g_{open} and g_{ext} are positive constants. This standard affine gap length model is both computationally efficient and surprisingly effective. (Smith and Waterman, 1981; Altschul and Erickson, 1986; Zachariah *et al.*, 2005).

The optimal, highest scoring alignment between two sequences ($x = x_1, x_2, \cdots, x_n$ and $y = y_1, y_2, \cdots, y_m$) is found by populating a series of score tables, also known as dynamic programming matrices. The entries of the match table, $M(i,j,r)$, are the maximum alignment score for an alignment that terminates with an ungapped segment of length r, ending at the i-th position of x and the j-th position of y. Similarly, the gap tables $G_x(i,j)$ and $G_y(i,j)$ contain the maximum alignment similarity given that the alignment ends with x_i or y_j gapped. The entries of these tables can be efficiently computed starting from the following boundary conditions: $M(i,0,l), M(0,j,l), G_{x/y}(i,0), G_{x/y}(0,j) = -\infty$. A single aligned amino acid pair may signal the beginning of a new local alignment, or it may occur immediately after any alignment gap.

$$M(i,j,1) = \max \begin{cases} S(i,j) \\ S(i,j) + G_x(i-1,j) \\ S(i,j) + G_y(i,j-1) \end{cases} \tag{9}$$

In standard Smith–Waterman this is the only necessary match score table. However, in `doublet` we require additional match tables so that we may keep track of match scores over extended, contiguously aligned regions. Of necessity, longer ungapped segments occur only after shorter segments. We restrict the maximum distance L over which doublet correlations are scored, since we expect that the useful information that can be extracted from doublet correlations will decay rapidly with sequence separation (Fig. 3). Consequentially, we do not need to explicitly consider ungapped segments of length greater than $L+1$.

$$\begin{aligned} M(i,j,2 \le r \le L) &= S(i,j) + D(i,j,r-1) \\ &\quad + M(i-1,j-1,r-1). \\ M(i,j,L+1) &= S(i,j) + D(i,j,L) \\ &\quad + \max \begin{cases} M(i-1,j-1,L), \\ M(i-1,j-1,L+1). \end{cases} \end{aligned} \tag{10}$$

Gaps in the alignment are either preceded by a match or they extend an existing gap.

$$\begin{aligned} G_x(i,j) &= \max_{r=1,L} \begin{cases} M(i-1,j-1,r) - g_{\text{open}}, \\ G_x(i-1,j) - g_{\text{ext}}. \end{cases} \\ G_y(i,j) &= \max_{r=1,L} \begin{cases} M(i-1,j-1,r) - g_{\text{open}}, \\ G_y(i,j-1) - g_{\text{ext}}. \end{cases} \end{aligned} \tag{11}$$

The largest score within the match table marks the last aligned position of the optimal alignment. The full alignment can be found by backtracking through the table, according to the choices previously made during the scoring step.

We used the method of Bailey and Gribskov (2002) to fit an extreme value distribution to the results of aligning a query sequence against a database of possible homologs. The maximum-likelihood parameters are then used to assign E-values to each alignment.

2.3 Doublet BLOcks SUbstitution matrix

A doublet substitution matrix [Equation (5)] contains $20^4 = 160\,000$ entries, of which $20^2 \times (20^2+1)/2 = 80\,200$ are unique as a result of the underlying symmetry, $d_l(i,i';j,j') = d_l(j,j';i,i')$. To accurately estimate these scores we require a very large collection of reliably aligned protein sequences. The BLOCKS database is one such resource (Henikoff and Henikoff, 1992; Henikoff *et al.*, 2000). Each database block consists of a reasonably reliable, ungapped multiple sequence alignment of a core protein region. BLOCKS version 13+ contains 11 853 blocks, containing, on average, 56 segments of average length 26 residues. Overall, about 10^9 pairwise amino acid comparisons are available for study.

BLOSUM65 (from BLOCKS 13+)

Singlet Substitutions

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	6	-1	-2	-2	-2	-1	-1	0	-2	-2	-2	-2	0	-2	-1	1	0	-3	-3	0	-2	-1	-1
R	-1	7	0	-1	-4	2	1	-3	0	-3	-3	3	-2	-3	-2	-1	-1	-2	-2	-3	0	1	-1
N	-2	0	8	2	-3	1	0	0	1	-3	-3	0	-2	-3	-1	1	0	-2	-2	-2	5	0	-1
D	-2	-1	2	9	-4	1	3	-1	-1	-4	-4	0	-3	-4	-1	0	-1	-3	-3	-3	6	2	-1
C	-2	-4	-3	-4	16	-4	-4	-4	-3	-3	-3	-4	-2	-2	-4	-2	-3	-3	-3	-2	-3	-4	-3
Q	-1	2	1	1	-4	6	3	-2	0	-3	-2	2	-1	-3	-1	0	0	-2	-2	-2	1	4	0
E	-1	1	0	3	-4	3	7	-3	-1	-3	-3	1	-2	-4	0	-1	-1	-3	-3	-3	2	5	-1
G	0	-3	0	-1	-4	-2	-3	9	-3	-5	-5	-2	-3	-4	-2	0	-2	-3	-4	-4	-1	-2	-2
H	-2	0	1	-1	-3	0	-1	-3	13	-4	-4	0	-3	-2	-2	-1	-2	-2	1	-3	0	0	-1
I	-2	-3	-3	-4	-3	-3	-3	-5	-4	6	3	-3	2	1	-3	-3	-1	-2	-2	4	-4	-3	-1
L	-2	-3	-3	-4	-3	-2	-3	-5	-4	3	6	-3	3	2	-3	-3	-2	0	-1	1	-4	-3	-1
K	-2	3	0	0	-4	2	1	-2	0	-3	-3	7	-2	-3	-1	-1	-1	-2	-2	-3	0	2	-1
M	0	-2	-2	-3	-2	-1	-2	-3	-3	2	3	-2	7	1	-3	-2	-1	0	-1	1	-3	-2	-1
F	-2	-3	-3	-4	-2	-3	-4	-4	-2	1	2	-3	1	9	-3	-3	-2	3	4	0	-3	-3	-1
P	-1	-2	-1	-1	-4	-1	0	-2	-2	-3	-3	-1	-3	-3	11	0	-1	-2	-3	-2	-1	-1	-1
S	1	-1	1	0	-2	0	-1	0	-1	-3	-3	-1	-2	-3	0	5	2	-2	-2	-2	0	0	0
T	0	-1	0	-1	-3	0	-1	-2	-2	-1	-2	-1	-1	-2	-1	2	6	-1	-2	0	0	0	0
W	-3	-2	-2	-3	-3	-2	-3	-3	-2	-2	0	-2	0	3	-2	-2	-1	16	4	-2	-3	-2	-1
Y	-3	-2	-2	-3	-3	-2	-3	-4	1	-2	-1	-2	-1	4	-3	-2	-2	4	11	-2	-2	-2	-1
V	0	-3	-2	-3	-2	-2	-3	-4	-3	4	1	-3	1	0	-2	-2	0	-2	-2	6	-3	-2	-1
B	-2	0	5	6	-3	1	2	-1	0	-4	-4	0	-3	-3	-1	0	0	-3	-2	-3	6	1	-1
Z	-1	1	0	2	-4	4	5	-2	0	-3	-3	2	-2	-3	-1	0	0	-2	-2	-2	1	5	-1
X	-1	-1	-1	-1	-3	0	-1	-2	-1	-1	-1	-1	-1	-1	-1	0	0	-1	-1	-1	-1	-1	0

Doublet Substitutions (Selected entries)

	L	1	2	3	4	5
AA	AA	2	0	2	1	0
AD	AD	2	2	1	1	1
AD	DA	4	3	3	3	2
DA	DA	1	1	2	3	2
DD	DD	0	3	3	3	2
CA	AD	3	0	1	2	0
CA	AC	7	3	5	2	3
CA	AQ	3	-1	0	1	-1
PI	LF	1	-1	0	-1	-1
PI	LP	5	4	3	2	0
PI	LS	2	3	1	1	0
RA	AA	0	1	-2	-2	-1
RA	AR	2	1	2	2	2
RA	AN	0	-1	0	1	1
PC	CG	10	6	6	16	2
PC	CL	8	4	4	8	3
PC	CK	14	3	6	14	-5
PC	CP	15	4	4	13	1

	L	1	2	3	4	5
CC	CA	-3	-1	-9	-1	0
CC	CR	0	2	-4	-1	2
CC	CN	-1	0	-11	-3	1
CC	CD	-1	-1	-10	-3	0
CC	CC	2	0	-3	-1	-2
CC	CQ	-2	0	-4	-3	1
CC	CE	0	0	-7	-3	0
CC	CG	-3	-2	-9	-3	-1
CC	CH	-4	-1	-5	-2	-1
CC	CI	-1	-2	-13	-2	-2
CC	CL	-3	-2	-10	1	-2
CC	CK	-1	3	-9	-1	3
CC	CM	-2	0	-13	2	-1
CC	CF	-4	-2	-16	7	-2
CC	CP	0	-4	-12	-3	-1
CC	CS	-2	-2	-10	-1	0
CC	CT	-1	-2	-10	1	1
CC	CW	-4	-2	-11	2	-3
CC	CY	-5	1	-2	6	0
CC	CV	-2	-4	-8	-2	-2

	L	1	2	3	4	5
ET	AA	0	0	-1	-1	0
ET	AR	0	-1	1	1	1
ET	AN	1	-2	0	1	0
ET	AD	1	0	1	1	1
ET	AC	1	1	2	0	2
ET	AQ	1	-1	0	1	0
ET	AE	2	0	1	2	1
ET	AG	0	0	-1	-2	-1
ET	AH	0	0	-1	0	0
ET	AI	0	-1	0	0	-1
ET	AL	0	1	-1	-1	0
ET	AK	-1	-2	0	2	0
ET	AM	0	-1	-2	-1	-2
ET	AF	0	0	0	-1	-1
ET	AP	1	0	0	0	0
ET	AS	-1	-1	0	0	1
ET	AT	0	1	-1	-1	-1
ET	AW	-1	0	-2	-1	-1
ET	AY	-1	0	0	1	1
ET	AV	0	-1	0	1	0

Fig. 2. BLOSUM65 singlet substitution matrix derived from the BLOCKS 13+ database (left) and selected elements of the corresponding doublet substitution matrices (right). Scores are in 1/4 bit units, rounded to the nearest integer. The average standard statistical error is ∼1/4 bits (i.e. ∼1 unit) for the doublet scores and essentially insignificant for the singlet scores, as judged by bootstrap resampling (See Section 2.3) The singlet scores are the log odds of observing the given substitution; positive scores are more likely, and negative score less likely to be observed than would be expected for uncorrelated sequences [Equation (2)]. Similarly, the doublet scores represent the log odds for observing pairs of substitutions, at various sequence separations, relative to the singlet substitutions likelihood [Equation (5)]. For example, the $L = 3$ column for ET AV (bottom right) indicates a score of zero for the alignment of ExxT in one sequence to AxxV in the other.

The widely used canonical BLOcks SUbstitution Matrices (BLOSUM) were generated from version 5 of the BLOCKS database (Henikoff and Henikoff, 1992). In order to generate a series of matrices representing different evolutionary divergences, the sequences in each block are clustered at a given level of sequence identity and the intercluster sequence correlations are collected. Thus BLOSUM100 (where only 100% identical sequences are clustered) represents a wide range, including low levels, of evolutionary divergence, whereas BLOSUM30 represents only correlations between very diverged sequences.

In principle, we should match the divergence inherent in the substitution matrix to the divergence of the pair of sequences we wish to align (Bishop and Thompson, 1986; Thorne *et al.*, 1991, 1992; Altschul, 1993). However, this is computationally expensive, and, in practice, a single matrix is chosen based on its ability to align remote homologs, on the grounds that matching close homologs is relatively easy (Brenner, 1996, 1998; Crooks and Brenner, 2005). In a recent evaluation of remote pairwise homology detection efficacy (Green and Brenner, 2002; Zachariah *et al.*, 2005), we discovered that the BLOSUM65 substitution matrix, reparameterized from the BLOCKS 13+ database, was more effective than any other reparameterized BLOSUM (BLOCKS 13+), classic BLOSUM (BLOCKS 5) or PAM (Dayhoff *et al.*, 1978) substitution matrix, and was comparable to the most effective VTML matrix (Müller *et al.*, 2002). Consequentially, we have built singlet and doublet substitution matrices from the BLOCKS 13+ database at 65% clustering, using an adaptation of the original BLOSUM clustering code (Henikoff and Henikoff, 1992). This provides $\sim 10^7$–10^8 independently aligned doublets, depending on the sequence separation l.

The estimated doublet target frequencies $q_l(i, i'; j, j')$, where smoothed and regularized by adding a pseudocount $\alpha(i, i'; j, j')$ to the raw count data, $n(i, j'; j, j')$. These pseudocounts are taken to be proportional to the marginal singlet target probabilities, $q_l(i; j)q_l(i', j')$.

$$q_l(i, i'; j, j') \approx \frac{\alpha(i, i'; j, j') + n(i, i'; j, j')}{A + N}, \tag{12}$$

$$\alpha(i, i'; j, j') = A \times q(i; j)q(i'; j'), \tag{13}$$

where, N is the total number of counts. Thus, if no data are available (the total number of counts is zero, $N = 0$), then all doublet scores would be zero, as can be seen from Equation (5). Here, A is a scale parameter that determines how much data are required to overcome the prior probability inherent in the pseudocount. Typically, such scale factors are picked empirically. However, in this case, we performed a full Bayesian analysis and determined that for doublet substitutions, reasonable values of A are $\sim 2 \times 10^6$, which can be compared with the 10^7–10^8 actual observations. The full details are given in the Supplementary materials. A representative subset of a doublet substitution matrix is shown in Figure 2.

Standard statistical errors were estimated by non-parametric Bayesian bootstrap resampling on sequence blocks (Efron, 1979; Rubin, 1981). Instead of assigning equal weight to every sequence block, each block is instead given a random weight drawn from a Dirichlet distribution. This random reweighting induces random changes in the estimated scores, thereby providing an estimate of the statistical errors caused by the finite size and inhomogeneity of the training data.

2.4 Evaluation of remote homology detection

We have previously developed and applied a sensitive strategy for evaluation of database search methods (Brenner *et al.*, 1998; Green and Brenner, 2002; Zachariah *et al.*, 2005; Price *et al.*, 2005). This strategy is made possible by the availability of a large collection of protein sequences whose evolutionary interrelations are known (primarily from structural information). In our approach, each sequence is aligned against every other sequence, and the alignment scores are used to determine putative homologs. We then consider the proportion of correctly identified homologs as a function of erroneous matches. Since the homology information derives from sequence-independent data, we avoid the circularity inherent in other evaluation approaches.

The collection of related sequences is derived from the structural classification of proteins (SCOP) database (Murzin *et al.*, 1995). We use the ASTRAL compendium (Chandonia *et al.*, 2004) of representative subsets of SCOP release 1.61 (Sept. 2002), filtered so that no two domains share more than 40% sequence identity. We partition every other SCOP fold into separate test and training subsets of approximately equal size, each containing ∼550 superfamilies, 2500 sequences, and 50 000 homologous sequence pairs. To avoid overfitting, adjustable parameters are optimized using the training set. Results of an all-versus-all comparison of the test set, using these optimized parameters, are reported as a plot of coverage (fraction of true relations found)

versus errors per query (EPQ), the total number of false relations divided by the number of sequences (Fig. 4). The raw, unnormalized coverage is the fraction of all true relations that are found.

Since the number of relations within a superfamily scales as the square of the size of the superfamily, and because SCOP superfamilies vary greatly in size, this reported coverage is dominated by the ability to detect relations within the largest superfamilies. To compensate for this unwarranted dependence, we also report the average fraction of true relations per sequence (linear normalization) and the average fraction of true relations per superfamily (quadratic normalization). In general, large superfamilies are more diverse, and the relationships within them are harder to discover (Green and Brenner, 2002). Thus, unnormalized coverage is typically less than the linearly normalized coverage, which in turn is less than quadratically normalized coverage. One important point of comparison for search results is 0.01 EPQ rate for linearly normalized results, the average fraction of true relations per database query at a false positive rate of 1 in 100. We report the observed difference in coverage of two methods at this selected EPQ, and determine standard statistical errors and confidence intervals using Bayesian bootstrap resampling (Rubin, 1981; Price *et al.*, 2005).

3 RESULTS

3.1 Doublet substitution correlations

Various trends are evident within the doublet score matrix, as illustrated in Figure 2. Notably, exact conservations, such as AA↔AA, AD↔AD and DD↔DD, generally have positive scores. This is expected because the pairs of sequences used to build the BLOSUM have a variety of intersequence similarity, ranging from most conserved to very diverged. Thus the observation of a conserved residue suggests that the sequences are relatively undiverged, and therefore, that other aligned residues are also more likely than average to be conserved.

Also notable is that many (but far from all) exact swaps, such as DA↔AD, are significantly more likely that expected. Possibly, this is because the effect of a deleterious mutation X→Y can sometimes be ameliorated by the occurrence of the corresponding mutation Y→X, in the immediate sequence neighborhood. Partial swaps, where only one of the substitution pair is conserved, are also often positive. This might reflect alignment errors in the original dataset. The most highly positive scores (and therefore those events that are most overrepresented in the data relative to uncorrelated substitutions) are associated with the substitutions PC↔Cx, i.e. a translocation of a cystine, replacing a proline. The most relatively uncommon substitutions involve the mutation of one cystine in the cystine pair CxxC (second column), a widespread and important motif found, for example, in the thioredoxin family. However, these interesting particular cases are atypical. Most of the doublet substitution matrix is similar to the ET↔Ax substitutions displayed in the third column; the majority of the scores are not significantly different from zero, indicating that most possible substitution doublets are essentially uncorrelated.

We can place the above observations on a quantitative footing by considering the intersequence mutual information [Equation (6)], a measure of the correlation strength between aligned homologous sequences. The first order contribution is equal to the average singlet score, which is 0.31 bits per aligned residue for BLOSUM65 (BLOCKS13+). The corresponding average doublet score, the additional information encoded in intersite substitution covariation, is ∼0.04 bits at modest sequence separations (illustrated in Fig. 3). Thus, the intersite substitution correlations carry relatively little information. However, these correlations appear to persist

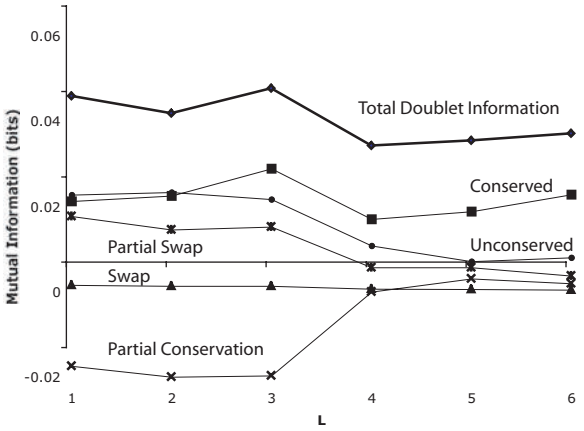

Fig. 3. The intersequence mutual information of homologs encoded in intersite correlations at increasing separation, L, i.e. the average doublet substitution scores [Equation (6)]. The top, dark line is the total information at various sequence separations. For comparison, the information encoded in the corresponding singlet substitutions (the average singlet matrix score) is 0.31 bits per residue. The remaining lines illustrate the relative contributions of different substitutions classes to this total information; these are exact conservation XY↔XY, partial conservation XY↔XZ, swaps XY↔YX, partial swaps XY↔ZX and unconserved, double substitutions XY↔ZU.

to non-local neighbors, which suggests that the total information from interactions at all sequence separations is substantial. However, Figure 3 also displays the contributions to this total information from various categories of substitution. The largest contribution, and the only contribution to persist above a sequence separation of four residues, represents exactly conserved pairs of residues. This is a rather trivial correlation which is persistent because all parts of two homologous sequences have the same chronological divergence. All other substitution classes, summing over all sequence separations, contribute no more than 0.1 bits per residue. This is not entirely insignificant, but it is still small compared with the singlet mutual information. Thus non-trivial correlations between substitutions are relatively weak.

3.2 Homology detection

The primary use for pairwise alignment methods is to search databases of previously characterized biological sequences for homologs of the sequence of interest. Therefore, the most powerful methods will perform this task most effectively by assigning true homolog significant statistical scores and assigning unrelated sequence low statistical scores. Our assessment methodology compares database search methods on this criteria.

We compared the `doublet` alignment algorithm against the standard Smith–Waterman algorithm. To perform a fair test, we converted raw scores to statistical scores for both algorithms using the same length normalized maximum-likelihood EVD parameter determination method (Bailey and Gribskov, 2002). Optimal parameters for gapping, matrix scaling and distance over which to consider dipeptide correlations were found using the training database described above. Then, the algorithms were evaluated by comparing the relative ability to detect remote homologs within the test dataset, using the parameters optimized on the training dataset. (Inset, Fig. 4).

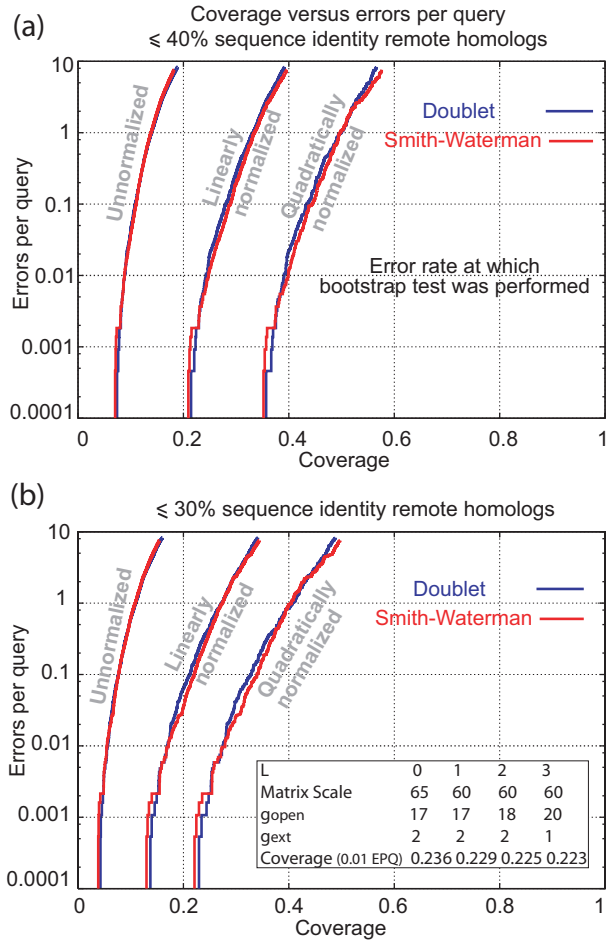

Fig. 4. These coverage versus EPQ plots show that including dipeptide covariation information in alignment determination (`doublet`) does not improve remote homolog detection. (**a**) Optimized matrix, gap and look-back parameters were used to search the test database with the `doublet` and Smith–Waterman algorithms. This database contains no sequence pairs that share >40% sequence identity. The number of correctly identified homologs is shown as a function of the number of errors made. Smith–Waterman outperforms `doublet` over all but extremely low error-rates. (**b**) Remote homolog test using only sequence pairs with <30% sequence identity. As above, Smith–Waterman correctly identifies more remote homologs than the `doublet` algorithm. Inset: Optimal matrix scale parameter, gap parameters and corresponding linearly normalized homology detection coverage at 0.01 EPQ, as a function of the covariation distance considered, L.

The results of a database search for Smith–Waterman and `doublet`, using only nearest neighboring dipetide covariations, are shown in Figrue 4a. Both the Smith–Waterman and `doublet` methods performed remarkably similarly over all error rates and normalization schemes. The linearly normalized coverage at 0.01 EPQ was slightly higher for Smith–Waterman than `doublet` (Inset, Fig. 4). From this, we conclude that including dipeptide covariation, information does not improve remote homology detection and, in fact, slightly degrades performance at this error rate. We also performed the same coverage versus EPQ analysis using only sequences with <30% sequence identity (Fig. 4b), as it was previously reported that dipeptide covariation information may be useful only for detecting these extremely remote evolutionary relationships (Jung and Lee, 2000). Our results, however, show that even at this evolutionary distance, dipeptide covariation scoring does not improve homology detection.

We used Bayesian bootstrap resampling to estimate statistical errors and to determine if the observed coverage difference was statistically significant. We found that a 95% confidence interval for the coverage difference at 0.01 EPQ comfortably contained zero difference. Therefore, we cannot distinguish between the remote homolog detection abilities of Smith–Waterman and `doublet`.

We also evaluated the effect of including covariation information over larger sequence separations. As can be seen in table of Figure 4, incorporating this additional information into alignment scores actually results in a slow degradation of homology detection efficacy.

4 DISCUSSION

We have developed, implemented and tested an alignment algorithm, `doublet`, that generates the optimal pairwise protein sequence alignment under a scoring scheme that includes dipeptide covariation information. Perhaps surprisingly, and in marked contrast to previous reports, we found that using this information provides no benefit to remote homolog detection. The performance of the `doublet` algorithm for detecting remote homologs is statistically indistinguishable from the standard Smith–Waterman algorithm.

The underlying explanation for this indifference of alignment to dipeptide covariation is that substitution correlations are weak on the average (Figs 2 and 3). Therefore, the average effect of these interactions is insignificant and including covariation in sequence alignment makes very little material difference to remote homology detection.

We might reasonably question if the training data are at fault. Indeed, the slight degradation of homology detection, as more distant correlations are included (Inset table, Fig. 4), does indicate that the doublet substitution matrices contain anomalies, perhaps owing to the training or alignment of the BLOCKS sequences, or perhaps because of the different sampling of sequences included in BLOCKS compared with those included in SCOP. The BLOCKS database that we use to train the doublet substitution matrices contains ungapped alignments, many of shorter length than the average SCOP protein domain. Fukami-Kobayashi *et al.* (2002) showed that the covariation signal is strongest within single secondary structure elements. The poor performance of `doublet`, then, may be the result of its applying the covariation model too bluntly across the entire protein sequences when it is only applicable within secondary structure elements. However, we note that the BLOCKS database has been used to derive very effective singlet substitution matrices (Green and Brenner, 2002), and therefore, it is implausible that the substitution signals within the BLOCKS database are substantially erroneous. On the contrary, the observed degradation simply reinforces the idea that neighboring substitutions are weakly correlated, particularly when compared with single substitution correlations, and therefore, the doublet signal is readily degraded by minor anomalies in the data.

Another line of evidence comes from examining the intersite amino acid correlation of single protein sequences (Yčas, 1958; Weiss *et al.*, 2000; Crooks and Brenner, 2004; Crooks *et al.*, 2004). Neighboring amino acids are almost entirely uncorrelated; the nearest neighbor mutual information has been estimated as only 0.006 bits (Crooks and Brenner, 2004). This lack of sequence correlation is consistent with (but does not require) small intersite substitution correlations.

It should be emphasized, however, that the observation of weak average dipeptide covariation does not negate the possibility of strong, interesting covariation in particular instances, such as CP↔Cx, or within particular families. Moreover, it is conceivable that covariation information could be used more judiciously, thereby improving alignment results. For example, as previously discussed, one might include doublet-type scoring information only for residue pairs that are likely to be within the same secondary structural element. Similarly, one might examine the covariation of residues that are proximate in the tertiary structure, rather than along the sequence (Rodionov and Johnson, 1994; Lin *et al*., 2003). However, residues that are proximate in space are also only weakly correlated (Cline *et al*., 2002; Crooks *et al*., 2004), and the inter-residue mutual information is not improved by foreknowledge of the local structure environment (Crooks and Brenner, 2004; Crooks *et al*., 2004). Therefore, we suspect that such approaches will also not have dramatic effects on protein sequence alignment.

In conclusion, the ubiquitous assumption that neighboring sites along a protein sequence evolve independently appears to be generally appropriate. This leads to fast, elegant and effective algorithms for protein sequence alignment and homology detection.

ACKNOWLEDGEMENTS

R.E.G. and G.E.C. jointly conceived and designed the `doublet` alignment algorithm and co-wrote this paper, with guidance from S.E.B.; G.E.C. was responsible for creating the doublet BLOSUM substitution matrices and R.E.G. for the statistical comparison of `doublet` to Smith–Waterman. The authors would like to thank Emma Hill, Sandrine Dudoit and Jeff Thorne for their helpful discussions and suggestions. This work was supported by the National Institutes of Health (1-K22-HG00056) and an IBM Shared University Research grant. G.E.C. received funding from the Sloan/DOE postdoctoral fellowship in computational molecular biology. S.E.B. is a Searle Scholar (1-L-110). Funding to pay the Open Access publication charges for this article was provided by the N.I.H.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, **219**, 555–565.

Altschul,S.F. (1993) A protein alignment scoring system sensitive at all evolutionary distances. *J. Mol. Evol.*, **36**, 290–300.

Altschul,S.F. and Erickson,B.W. (1986) Optimal sequence alignment using affine gap costs. *Bull. Math. Biol.*, **48**, 603–616.

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bailey,T.L. and Gribskov,M. (2002) Estimating and evaluating the statistics of gapped local-alignment scores. *J. Comput. Biol.*, **9**, 575–593.

Bishop,M.J. and Thompson,E.A. (1986) Maximum likelihood alignment of DNA sequences. *J. Mol. Biol*, **190**, 159–165.

Brenner,S.E. (1996). *Molecular propinquity: Evolutionary and structural relationships of proteins*. PhD thesis, Cambridge University.

Brenner,S.E. *et al*. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.

Chandonia,J.-M. *et al*. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res.*, **32**, 189–192.

Cline,M.S. *et al*. (2002) Information-theoretic dissection of pairwise contact potentials. *Proteins*, **49**, 7–14.

Cover,T.M. and Thomas,J.A. (1991) *Elements of Information Theory*. Wiley, New York.

Crooks,G.E. and Brenner,S.E. (2004) Protein secondary structure: entropy, correlations and prediction. *Bioinformatics*, **20**, 1603–1611.

Crooks,G.E. and Brenner,S.E. (2005) An alternative model of amino acid replacement. *Bioinformatics*, **21**, 975–980.

Crooks,G.E. *et al*. (2004) Measurements of protein sequence-structure correlations. *Proteins*, **57**, 804–810.

Dayhoff,M.O. *et al*. (1978) A model of evolutionary change in proteins. *Atlas of Protein Sequences and Structure*, **5**(Suppl 3), 345–352.

Doolittle,R.F. (1992) Reconstructing history with amino acid sequences. *Protein Sci.*, **1**, 191–200.

Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological sequence analysis*. Cambridge University Press.

Eddy,S.R. (2001) *HMMER: Profile hidden Markov models for biological sequence analysis*.

Efron,B. (1979) Bootstrap methods: another look at the jacknife. *Ann. Stat.*, **7**, 1–26.

Fukami-Kobayashi,K. *et al*. (2002) Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. *J. Mol. Biol.*, **319**, 729–743.

Goldman,N. *et al*. (1998) Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, **149**, 445–458.

Gonnet,G.H. *et al*. (1994) Analysis of amino-acid substitution during divergent evolution—the 400 by 400 dipeptide substitution matrix. *Biochem. Biophys. Res. Comm.*, **199**, 489–496.

Green,R.E. and Brenner,S.E. (2002) Bootstrapping and normalization for enhanced evaluations of pairwise sequence comparison. *Proc. IEEE*, **90**, 1834–1847.

Henikoff,J.G. *et al*. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.*, **28**, 228–230.

Henikoff,S. and Henikoff,J.G. (1992) Amino-acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Jung,J.S. and Lee,B. (2000) Use of residue pairs in protein sequence-sequence and sequence-structure alignments. *Protein Sci.,* **9**, 1576–1588.

Karplus,K. *et al*. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.

Lin,K. *et al*. (2003) Testing homology with Contact Accepted mutatiOn (CAO). *Comput. Biol. Chem.*, **27**, 93–102.

Müller,T. *et al*. (2002) Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Mol. Biol. Evol.*, **19**, 8–13.

Murzin,A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Park,J. *et al*. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.,* **284**, 1201–1210.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Price,G.A. *et al*. (2005) Statistical evaluation of pairwise protein sequence comparison with the Bayesian bootstrap. *Bioinformatics*, doi:10.1093/bioinformatics/bti627

Rodionov,M.A. and Johnson,M.S. (1994) Residue-residue contact substitution probabilities derived from aligned three-dimensional structures and the identification of common folds. *Protein Sci.*, **3**, 2366–2377.

Rubin,D.B. (1981) The Bayesian bootstrap. *Ann. Stat.*, **9**, 130–134.

Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Thorne,J.L. *et al*. (1996) Combining protein evolution and secondary structure. *Mol. Biol. Evol.*, **13**, 666–673.

Thorne,J.L. *et al*. (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.,* **33**, 114–124.

Thorne,J.L. *et al*. (1992) Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol.,* **34**, 3–16.

Topham,C.M. *et al*. (1997) Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. *Protein Eng.*, **10**, 7–21.

Weiss,O. *et al*. (2000) Information content of protein sequences. *J. Theor. Biol.*, **206**, 379–386.

Yu,Y.K. *et al*. (2003) The compositional adjustment of amino acid substitution matrices. *Proc. Natl Acad. Sci. USA*, **100**, 15688–15693.

Yčas,M. (1958) The protein text. In *Symposium on Information Theory in Biology*, Pergamon Press, NY, pp. 70–101.

Zachariah,M.A. *et al*. (2005) A generalized affine gap model significantly improves protein sequence alignment accuracy. *Proteins*, **58**, 329–338.

Pairwise alignment incorporating dipeptide covariation
Supplemental Appendix:
Estimating probabilities from counts with a prior of uncertain reliability

Gavin E. Crooks*, Richard E. Green and Steven E. Brenner
Dept. of Plant and Microbial Biology,
111 Koshland Hall #3102,
University of California, Berkeley,
CA 94720-3102, USA

August 29, 2005

A common problem is that of estimating a discrete probability distribution, $\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$, given a limited number of samples drawn from that distribution, summarized by the count vector $n = \{n_1, n_2, \ldots, n_k\}$, and a reasonable *a priori* best guess for the distribution $\theta \approx \pi = \{\pi_1, \pi_2, \ldots, \pi_k\}$. (For a general introduction, see Durbin *et al.* 1998.) This guess may simple be the uniform probability, $\pi_i = 1/k$, which amounts to asserting that, as far as we know, all possible observations are equally likely. At other times, we may know some some more detailed approximation to the distribution θ.

For example, in the present case we wish to estimate the probabilities of substituting a pair of amino acid residues by another residue pair, given the number of times that this substitution has been observed in the training dataset. This probability is hard to estimate reliably since the distribution is very large with $20^4 = 160,000$ dimensions. Moreover, many of the possible observations occur very rarely. However, substitutions at different sites are not strongly correlated, and therefore we may approximate the doublet substitution probabilities by a product of single substitution probabilities. Since the dimensions of these marginals are relatively small we can accurately estimate them from the available data, and thereby construct a reliable and reasonable initial guess for the full doublet substitution distribution.

In the common and conventional pseudocount approach, we assume that the distribution π was estimated from A previous observations. These pseudocounts, $\alpha_i = \pi_i A$, are then proportionally averaged with the real observations ($N = \sum_i n_i$) to provide an estimate of θ;

$$\theta_i = \frac{\alpha_i + n_i}{A + N}. \tag{1}$$

This prescription is intuitively appealing. When the total number of real counts is much less than the number of pseudocounts ($N \ll A$) the prior dominates, and the estimated distribution is determined by our initial guess, $\theta \approx \pi$. In the alternative limit that the real observations greatly outnumber the pseudocounts ($N \gg A$) the estimated distribution is given by the frequencies $\theta_i = n_i/N$. However, it is not immediately obvious how to select A, although many heuristics have been proposed, including $A = 1$, $A = k$ (Laplace), and $A = \sqrt{N}$ (e.g. Lawrence *et al.*, 1993; Durbin *et al.*, 1998; Nemenman *et al.*, 2001). Essentially, this total pseudocount parameter represents our confidence that the initial guess $\theta \approx \pi$ is accurate, since the larger the total pseudocount the more data is required to overcome this assumption.

Within a Bayesian approach we can avoid this indeterminacy by admitting that, *a priori*, we do not know how confidant we are that π approximates θ. The probability $P(n|\theta)$ of independently sampling a particular set of observations, n, given the underlying sampling probability, θ, follows the multinomial distribution, the multivariate generalization of the binomial distribution;

$$\mathcal{M}(n|\theta) = \frac{1}{M(n)} \prod_{i=1}^{k} \theta_i^{n_i}, \quad M(n) = \frac{\prod_i n_i!}{(\sum_i n_i)!}. \tag{2}$$

The prior probability of the sampling distribution $P(\theta)$ is typically modeled with a Dirichlet distribution,

$$\mathcal{D}(\theta|\alpha) = \frac{1}{Z(\alpha)} \prod_{i=1}^{k} \theta_i^{(\alpha_i - 1)}, \quad Z(\alpha) = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma(A)}. \tag{3}$$

where $\sum_i \theta = 1$, $\alpha_i > 0$ and $A = \sum_i \alpha_i$. Note that the mean of a Dirichlet is

$$\mathrm{E}[\theta_i] = \frac{\alpha_i}{A}. \tag{4}$$

Therefore, we may fix the parameters of the Dirichlet prior by equating our initial guess, π, with the mean prior distribution: $\pi = \alpha/A$. If we can fix the scale factor A, then we can combine the prior and observations using Bayes' theorem.

$$P(\theta|n) = \frac{P(n|\theta)P(\theta)}{P(n)}. \tag{5}$$

Because the multinomial and Dirichlet distributions are naturally conjugate, the posterior distribution $P(\theta|n)$ is also Dirichlet.

$$P(\theta|n) \quad \propto \quad \mathcal{M}(n|\theta)\mathcal{D}(\theta|A\pi)$$

*gec@compbio.berkeley.edu

$$\begin{aligned} &\propto \prod_{i=1}^{k} \theta_i^{(A\pi_i+n_i-1)}, \\ &= \mathcal{D}(\theta|A\pi+n) \end{aligned} \tag{6}$$

The last line follows because the product in the previous line is an unnormalized Dirichlet with parameters $(A\pi + n)$, yet the probability $P(\theta|n)$ must be correctly normalized.

Given multinomial sampling and a Dirichlet prior, the probability of the data is given by the under-appreciated multivariant negative hypergeometric distribution (Johnson & Kotz, 1969; Durbin *et al.*, 1998, Eq. 11.23);

$$\begin{aligned} P(n) &= \int d\theta\ P(n|\theta)P(\theta), \\ &= \int d\theta\ \mathcal{M}(n|\theta)\mathcal{D}(\theta|A\pi), \\ &= \frac{1}{Z(A\pi)}\frac{1}{M(n)}\int d\theta \prod_{i=1}^{20}\theta_i^{(A\pi_i+n_i-1)}, \\ &= \frac{Z(A\pi+n)}{Z(A\pi)M(n)} \equiv \mathcal{H}'(n|A\pi+n). \end{aligned} \tag{7}$$

Again, the last line follows because the product in the previous line is an unnormalized Dirichlet with parameters $(A\pi + n)$. Therefore, the integral over θ must be equal to the corresponding Dirichlet normalization constant, $Z(A\pi + n)$. Note that, confusingly, the negative hypergeometric distribution is sometimes called the inverse hypergeometric, an entirely different distribution, and vice versa.

Since we do know a reasonable value for the scale factor A we cannot use a simple Dirichlet prior. As an alternative, we explicitly acknowledge our uncertainly about A by building this indeterminacy into the prior itself. Rather than a single Dirichlet, we use the Dirichlet mixture;

$$P(\theta|\pi) = \int_0^\infty dA\ \mathcal{D}(\theta|A\pi)P(A). \tag{8}$$

The distribution $P(A)$ is a hyperprior, a prior distribution placed upon a parameter of the Dirichlet prior. Following the same mathematics as Eqs. 5-7, we find that the posterior distribution is the Dirichlet mixture

$$P(\theta|n) = \int_0^\infty dA\ \mathcal{D}(\theta|A\pi+n)P(A|n)\,, \tag{9}$$

where

$$P(A|n) = \frac{P(A)\mathcal{H}'(n|A\pi+n)}{\int_0^\infty dA\ P(A)\mathcal{H}'(n|A\pi+n)}\,. \tag{10}$$

In principle, we have to select and parameterize a functional form for the hyperprior, $P(A)$. For example, an exponential distribution, $P(A) = \lambda\exp(-\lambda A)$, with mean $1/\lambda$, might be appropriate. Fortunately, we can often avoid selecting an explicit hyperprior. In practice, given sufficient data, the probability of that data $P(n|A)$ is a smooth, sharply peaked function of A. This is illustrated in figure 1 using 10^7 observations of the 160,000 dimensional doublet substitution probability, where the mean prior distribution is taken to be the product of singlet substitutions probabilities. If the prior distribution of A is reasonable, and neither very large nor very small over the range of interest, then the posterior distribution $P(A|n)$ will also be very strongly peaked. Moreover, the location of that peak will be almost totally independent of the prior placed on A. In this limit the posterior Dirichlet mixture (Eq. 9) reduces to the single component that maximizes the probability of the data;

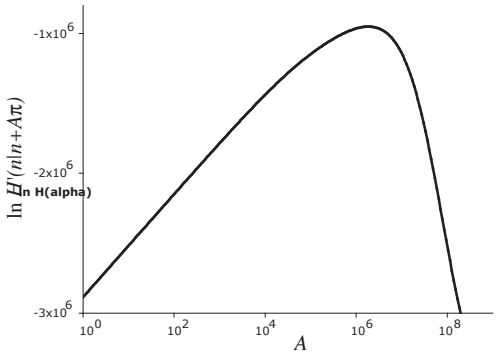

Figure 1: The likelihood of observations as a function of the scale parameter A. With multinomial sampling and a Dirichlet prior the likelihood of the data follows the negative hypergeometric distribution, $H'(n|A\pi + n)$, where n is the count vector of observations, π is the mean prior estimate of the sampling distribution, and A is a scale parameter (Eq. 7). Given a large number of observations (here, $N = \sum n_i$ is about 10^7) the probability of the data is a smooth and very sharply peaked function of the scale parameter A.

$$\begin{aligned} P(\theta|n) &\approx \mathcal{D}(\theta|A\pi+n), \\ A &= \mathrm{argmax}_A P(A|n) \approx \mathrm{argmax}_A P(n|A), \\ P(n|A) &= \mathcal{H}'(n|A\pi+n). \end{aligned} \tag{11}$$

Here, $\mathrm{argmax}_x f(x)$ is the value of x that maximizes that function $f(x)$.

Given any function of θ, the average of the function across the posterior distribution (the posterior mean estimate (PME) or Bayes' Estimate) minimizes the mean squared error of that estimate. In particular, the posterior mean estimate of θ (Eq. 4) is

$$\theta_i^{\mathrm{PME}} = \frac{A\pi_i + n_i}{A+N}. \tag{12}$$

Taken altogether, our practice is to take the raw doublet substitution counts and construct a mean prior distribution π based upon the approximation that substitutions on neighboring sites are uncorrelated. We then find the scaling factor A that maximizes the negative hypergeometric probability $\mathcal{H}'(n|A\pi + n)$. For our data the total number of observations N is around 10^7, for which the optimal scale factor A was found to be about 10^6. The posterior mean estimate of the doublet substitution distribution is then used to construct the doublet substitution matrix. Code for constructing doublet substitution matrices using this procedure and for finding the optimal prior and posterior, given any set of observations and π, a best guess for the true distribution θ, is

available from our web site (`http://compbio.berkeley.edu`), along with other code and data for this work. Our programs make extensive use of the Open Sourced GNU Scientific Library (GSL) (Gough, 2003; Matsumoto & Nishimura, 1998).

References

Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. (1998) *Biological sequence analysis*. Cambridge University Press.

Gough, B., ed. (2003) *GNU Scientific Library Reference Manual*. 2nd edition,, Network Theory Ltd.

Johnson, N. L. & Kotz, S. (1969) *Discrete Distributions*. John Wiley, New York.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. & Wootton, J. C. (1993) Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science,* **262** (5131), 208–214.

Matsumoto, M. & Nishimura, T. (1998) Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Trans. Model. Comput. Simul.,* **8** (1), 3–30.

Nemenman, I., Shafee, F. & Bialek, W. (2001). Entropy and inference, revisited. arXiv:physics/0108025.