PROTEINS: Structure, Function, and Genetics 24:165–177 (1996)

Defining a Similarity Threshold for a Functional Protein Sequence Pattern: The Signal Peptide Cleavage Site

Henrik Nielsen,[1] Jacob Engelbrecht,[1] Gunnar von Heijne,[2] and Søren Brunak[1]
[1]Center for Biological Sequence Analysis, Department of Physical Chemistry, The Technical University of Denmark, DK-2800 Lyngby, Denmark; and [2]Department of Biochemistry, Arrhenius Laboratory, Stockholm University, S-106 91 Stockholm, Sweden

ABSTRACT **When preparing data sets of amino acid or nucleotide sequences it is necessary to exclude redundant or homologous sequences in order to avoid overestimating the predictive performance of an algorithm. For some time methods for doing this have been available in the area of protein structure prediction. We have developed a similar procedure based on pair-wise alignments for sequences with *functional* sites. We show how a correlation coefficient between sequence similarity and functional homology can be used to compare the efficiency of different similarity measures and choose a nonarbitrary threshold value for excluding redundant sequences. The impact of the choice of scoring matrix used in the alignments is examined. We demonstrate that the parameter determining the quality of the correlation is the relative *entropy* of the matrix, rather than the assumed (PAM or identity) substitution model. Results are presented for the case of prediction of cleavage sites in signal peptides. By inspection of the false positives, several errors in the database were found. The procedure presented may be used as a general outline for finding a problem-specific similarity measure and threshold value for analysis of other functional amino acid or nucleotide sequence patterns.**

Key words: sequence data sets, similarity screening, redundancy reduction, signal peptides, database errors

INTRODUCTION

One of the recurring problems haunting the analysis of protein and DNA sequences is the redundancy of the data. Many entries in protein or gene databases represent members of protein and gene families or versions of homologous genes found in different organisms, and, therefore, many sequences are more or less closely related.

The use of a redundant data set implies at least three potential sources of error. First, if a data set of amino acid or nucleic acid sequences containing large families of closely related sequences is used for statistical analysis, the statistics will be biased for these families and will overrepresent features peculiar to them. Second, apparent correlations between different positions in the sequences may also be an effect of a biased sampling of the data. Finally, if the data set is being used for predicting a certain feature, and the sequences used for defining and calibrating the prediction method—"the training set"—are too closely related to the sequences used for testing the method, the apparent predictive performance may be an overestimate, reflecting the method's ability for reproducing its own particular input rather than the generalization power of the method.

For these reasons, it is necessary to avoid too closely related sequences in the data set. On the other hand, a too rigorous definition of "too closely related" may lead to valuable information being discarded from the data set. Thus, there is a tradeoff between size and nonredundancy, and the appropriate definition of "too closely related" may depend on the problem under consideration. In practice, this is rarely considered. Often the test data are described as being selected "randomly" from the complete data set, implying that great care was taken when preparing the data, even though redundancy reduction was not applied.[1] In many cases where redundancy reduction is applied, a more or less arbitrary homology threshold is used, or a "representative" data set is made by using a conventional list of protein or gene families and selecting one member from each family.

Sander and Schneider[2] pioneered the algorithmic investigation of the relationship between protein sequence similarity and structural similarity. In a plot of the alignment length versus the percentage of identical residues in the overlap, two domains could be discerned: one of almost exclusively structurally

Received November 22, 1994; revision accepted July 7, 1995.
Address reprint requests to Søren Brunak, Center for Biological Sequence Analysis, Department of Physical Chemistry, The Technical University of Denmark, DK-2800 Lyngby, Denmark.

homologous pairs (defined by more than 70% secondary structure assignment identity in the overlap), and one containing a mixture of homologous and nonhomologous pairs. The mixed region reflects the fact that the secondary structure identity may exceed 70% by chance, especially for very short overlaps, even in pairs of completely unrelated sequences.

The border between the two domains, and thereby the threshold for sequence similarity, measured in percentage identity, depends on the length of the aligned region (the overlap). Sander and Schneider defined a length-dependent threshold function: For overlap length $L<10$, no pairs are above the threshold, for $10<L<80$, the threshold is $290.15L^{-0.562}\%$, and for $L>80$, the threshold is 24.8%.

There is an urgent need for a generalization of the work of Sander and Schneider to other sequence prediction problems involving *functionality* as opposed to structure. Their article may be followed as an outline of finding a similarity measure and a threshold which can discriminate between one domain where the prediction problem in question can be solved by alignment alone and another where more advanced pattern recognition methods are needed.

In this paper, we are concerned with defining a nonarbitrary threshold for a specific sequence recognition problem: that of finding the cleavage site of signal peptides. Our approach may easily be generalized to other types of protein sequence analysis problems involving, e.g., glycosylation sites, transit peptides for chloroplasts and mitochondria, or cleavage sites of polyproteins; and to nucleotide sequence analysis problems such as intron splice sites in pre-mRNA, ribosome binding sites, and promoters.

The signal peptide (also known as the signal sequence) is the initial N-terminal part of any secretory protein. It serves as a signal for translocation of a protein across a membrane (in prokaryotes, the plasma membrane; in eukaryotes, the membrane of the endoplasmic reticulum) and is cleaved from the rest of the protein during translocation at a well-defined *cleavage site*.[3]

Presently, we are developing a method for predicting the cleavage site from the amino aid sequence.[4] For the reasons stated above, we want to avoid pairs of proteins in the data set that are too homologous. The appropriate threshold for homology, however, is not necessarily the same for the cleavage site location problem as for the secondary structure prediction problem.

In analogy with Sander and Schneider's "structural homology," we define a "functional homology," meaning that the position of a specific functional site in one sequence can be inferred from the other. In this case, it means that two signal peptide sequences are functionally homologous if it is possible to predict the cleavage site of one by aligning it to another with a known cleavage site.

DATA

The signal peptide data were taken from SWISS-PROT version 29 (1994).[5]

From a total of 38,303 entries, 5995 entries contained the keyword SIGNAL in the feature table. Entries suggesting questionable evidence for the cleavage site were discarded, i.e., where the signal peptide was incomplete, the cleavage site was unknown, question marks or comments such as "POTENTIAL" or "PROBABLE" were present, or an alternative cleavage site was suggested. Furthermore, all virus and phage genes were discarded. From the eukaryotic data set, proteins encoded by organellar (nonnuclear) genes were discarded (by excluding entries containing an "OG" line). From the prokaryotic data set, signal peptides cleaved by signal peptidase II (*Lsp*, a specific lipoprotein signal peptidase) were discarded, since the cleavage sites of these proteins differ considerably from those cleaved by the standard prokaryotic signal peptidase (*Lep*)[6]; this was done by excluding entries with a cross-reference to the PROSITE entry named "PROKAR_LIPOPROTEIN".[7]

This left a database of 2282 eukaryotic entries (619 human) and 579 prokaryotic entries (120 from *E. coli*).

From each entry, the sequence of the signal peptide only and the first 30 amino acids of the mature protein were included in the data set. (One entry having less than 30 amino acids after the cleavage site was deleted.) It would not be reasonable to give the entire protein sequence as background to the cleavage site, since the cleavage takes place while the protein is being translocated and the cleavage enzyme therefore hardly has the entire protein as a potential substrate. The value 30 is not arbitrary: several experimental results indicate that in *E. coli*, the first approximately 30 residues of the mature protein seem to have a function for protein export.[8,9]

METHODS

Alignment Algorithms

All pairwise optimal local alignments between the sequences in our data sets were computed using the fasta program (version 1.7).[10] fasta compares a query sequence to a sequence library in order to detect homologous sequences, employing several cuts to locate regions with high density of sequence identities before performing actual alignments. Thereby the use of the computationally expensive alignment algorithm is restricted to the most interesting region of the most interesting library sequences.

For the human data set, we also tried a rigorous Smith–Waterman alignment between all the sequences (using the program ssearch of the FASTA package[10]) for comparison. We found that this did not improve the results as measured by the correlation coefficient (see definition below).

The steps of the fasta procedure are as follows (for each library sequence):

1. Identify initial regions with high density of identities using a lookup table method that limits the number of comparisons.
2. Rescan the 10 best regions from step 1 using the scoring matrix (giving the init1 score).
3. If possible, join some of the best regions from step 2 using a joining penalty (giving the initn score).
4. Only for the best-scoring library sequences: compute a Smith–Waterman alignment around the best region from step 3 using the scoring matrix (giving the opt score).

When using relatively short sequences such as our data, the joining step (step 3) is of little practical importance. We found that the initn score very rarely was (in less than 1% of the comparisons) larger than the init1 score. The optimization step is more important: the opt score was larger than the initn score in 91% of the comparisons when using the default PAM250 matrix and in 32% when using the (6,−3) identity matrix.

The lookup table in step 1 may be built from single amino acids (ktup = 1) or pairs of amino acids (ktup = 2, the default). Decreasing ktup increases sensitivity but reduces speed. We found that the ktup value gave very little difference for the results: with ktup = 1, the maximal correlation coefficients (see definition below) were approximately 0.01 lower than with ktup = 2.

The output from fasta and ssearch was parsed by a program which determined, for each alignment, whether the cleavage sites of both sequences were aligned, i.e., whether they both fell at the same position inside the region of overlap.

Substitution Matrices

The scoring matrix used in fasta or ssearch alignments specifies a set of scores s_{ij} for substituting amino acid i by amino acid j.

The default amino acid scoring matrix for all programs in the FASTA package (up to version 1.7) is the PAM250 matrix.[11] This is calculated from a simplified protein evolution model involving amino acid frequencies, p_i, and pairwise substitution frequencies, q_{ij}, observed in existing alignments of naturally occurring proteins. In a PAM matrix, a match involving a rarely occurring amino acid counts more than a match involving a common amino acid, while a mismatch between two easily exchangeable amino acids contributes a higher score than a mismatch between two functionally unrelated amino acids. A mismatch with a nonnegative score is known as a similarity or a conservative replacement.

The substitution or target frequencies q_{ij} depend on the amount of evolutionary divergence between the two sequences, giving rise to a range of different PAM matrices (1 PAM corresponds to the amount of evolutionary change resulting in substitutions in 1 percent of the sequence positions).

There is a wide variety of other types of substitution matrices, e.g., based on the relationships between the amino acids according to the genetic code or on physical and/or chemical properties of the amino acids. However, as shown by Altschul,[12] any amino acid substitution matrix is, either implicitly or explicitly, a matrix of logarithms of normalized target frequencies, since the substitution scores may be written as

$$s_{ij} = \frac{1}{\lambda}\left(\ln \frac{q_{ij}}{p_i p_j}\right) \tag{1}$$

where λ is a scaling factor. Changing λ will change the absolute value of the scores but not the relative scores of different local alignments, so it will not affect the alignments.

The simplest possible scoring matrices are *identity matrices,* where all the diagonal elements have the same positive value (the match score, s), and all the off-diagonal elements have the same negative value (the mismatch score, $\bar{s}$). An identity matrix may be derived from the simplest possible model for amino acid substitutions, where all 20 amino acids appear with equal probability, and the off-diagonal substitution frequencies are equal:

$$\begin{aligned} p_i &= 1/20 \quad \text{for all } i \\ q_{ij} &= \begin{cases} q & \text{for } i = j \\ \bar{q} & \text{for } i \neq j \end{cases} \end{aligned} \tag{2}$$

In other words, when an amino acid mutates, it has equal probabilities $\bar{q}$ for changing into any of the 19 other amino acids.

As is the case for the PAM matrices, there is a range of different identity matrices, depending on the ratio between the positive and negative scores, $s/\bar{s}$. If $s = -\bar{s}$, a local alignment must necessarily contain more matches than mismatches in order to yield a positive score, resulting in short and strong alignments, while if $s \gg -\bar{s}$, one match can compensate for many mismatches, resulting in long and weak alignments. The percentage identity p in gap-free local identity matrix alignment has a minimum value:

$$p > \frac{-\bar{s}}{s - \bar{s}} \tag{3}$$

We define $r = \bar{q}/q$, the mutability or the probability that a given position in the sequence has changed into a random amino acid (including the original one). $r = 0$ corresponds to no changes, while $r = 1$ corresponds to an infinite evolutionary distance.

Since the sum of all q_{ij} must be 1, we use the relation $20q + 380\bar{q} = 1$ to calculate the target frequencies:

$$q = \frac{1}{20 + 380r} \quad \text{and} \quad \bar{q} = \frac{r}{20 + 380r} \tag{4}$$

and the s_{ij} values may be calculated using Eq. (1). Since the score ratio, $s/\bar{s}$, is independent of λ and therefore a function of r, we can calculate r numerically from the score ratio.

The *relative entropy* of an amino acid substitution matrix was defined by Altschul[12]:

$$H = \sum_{i,j} q_{ij} s_{ij} \text{ bits} \tag{5}$$

where the s_{ij}s are normalized so that $\lambda = \ln 2$ [corresponding to using the base 2 logarithm in Eq. (1)]. The relative entropy of a matrix can be interpreted as the amount of information carried by each position in the alignment.

The shorter the evolutionary distance assumed in the calculation of the matrix, the larger H. At zero evolutionary distance (corresponding to PAM0 or $r = 0$), the mismatch penalty $\bar{s}$ is infinite, i.e., gaps are completely disallowed, and the relative entropy is equal to the entropy of the amino acid distribution: $H = \sum_i p_i \log_2 p_i$. In the identity model case, $H = \log_2 20 = 4.32$ bits, and the local alignment problem is reduced to the problem of finding the longest common substring between two sequences. Conversely, as the evolutionary distance approaches infinity (corresponding to high PAM numbers or $r \simeq 1$), all differences between the q_{ij} values disappear, and H approaches 0.

In this work, we have used three different PAM matrices, PAM250 and PAM120 taken from the FASTA package, and PAM20 from the Clustal W package.[13] In addition, we have used a range of different identity matrices: while the identity matrix in the FASTA distribution has $s = 6$ and $\bar{s} = -3$, we varied $\bar{s}$ from -1 to -1000 (i.e., practically infinity).

The gap penalties in all the matrices were -12 for gap initiation and -4 for gap elongation. Additionally, we have used versions with infinite gap penalties in order to estimate the effect of allowing gaps.

Correlation Coefficient

In order to be able to compare the discrimination abilities of different similarity measures and calculation methods, we have used a correlation coefficient as a measure of how well the threshold separates aligned and nonaligned examples. We use a correlation coefficient defined for predicted and observed discrete classes,[14] regarding the similarity score as a prediction of whether the cleavage sites of the two sequences will be correctly aligned. Thus, the equation for the correlation coefficient is

$$C = \frac{(P^t N^t) - (N^f P^f)}{\sqrt{(N^t + N^f)(N^t + P^f)(P^t + N^f)(P^t + P^f)}} \tag{6}$$

where P^t are true positives: the similarity is above the threshold and the cleavage sites are aligned; P^f are false positives: the similarity is above the threshold but the cleavage sites are not aligned; N^f are false negatives: the similarity is below the threshold but the cleavage sites are aligned; N^t are true negatives: the similarity is below the threshold and the cleavage sites are not aligned.

Data Set Extraction

For every threshold value tested, the extraction of the nonhomologous sequences from the data set was done using algorithm 2 of Hobohm et al.[15] This method counts the "neighbors"—i.e., sequences with a similarity larger than the threshold—to every sequence in the data set, discards the sequence with the largest number of neighbors, and repeats this procedure until no sequences with neighbors remain. In a final pass, discarded sequences which have no neighbors in the remaining set are reinstated. In accordance with the results of Hobohm et al., we found that this final pass did not have any practical importance—at most two sequences were reinstated and it happened only in those cases where more than half the data set was removed.

RESULTS

The relationship between overlap length and percentage of identical residues within the alignment is shown as two-dimensional scatter plots in Figure 1. Crosses denote sequence pairs where the cleavage sites of both sequences were aligned at the same position within the overlap, and squares denote sequence pairs where the cleavage sites were aligned to different positions or outside the overlap boundaries.

The scatter plots show, in the upper right part, a region of almost exclusively crosses, where the cleavage site can be located by pure alignment. Several of the squares in this region actually stem from database errors (see below). The rest of the plot is a mixture of squares and crosses, the crosses reflecting the fact that cleavage sites may randomly hit the same position even though the sequences are very dissimilar.

As in the paper of Sander and Schneider,[2] the appropriate percentage identity threshold depends on the overlap length—neither of these two measures alone would make a suitable discrimination function between the two regions. For example, an alignment with close to 100% identities does not reliably imply homology, if the overlap is shorter than 10 amino acids.

A nonlinear curve separates the two regions better. An obvious candidate for such a curve is the product of overlap length and percentage identity, i.e., the *number* of identical residues within the overlap, henceforth denoted "*number of identities.*" The curves shown in the scatter plots (Fig. 1) represent constant numbers of identities, corresponding

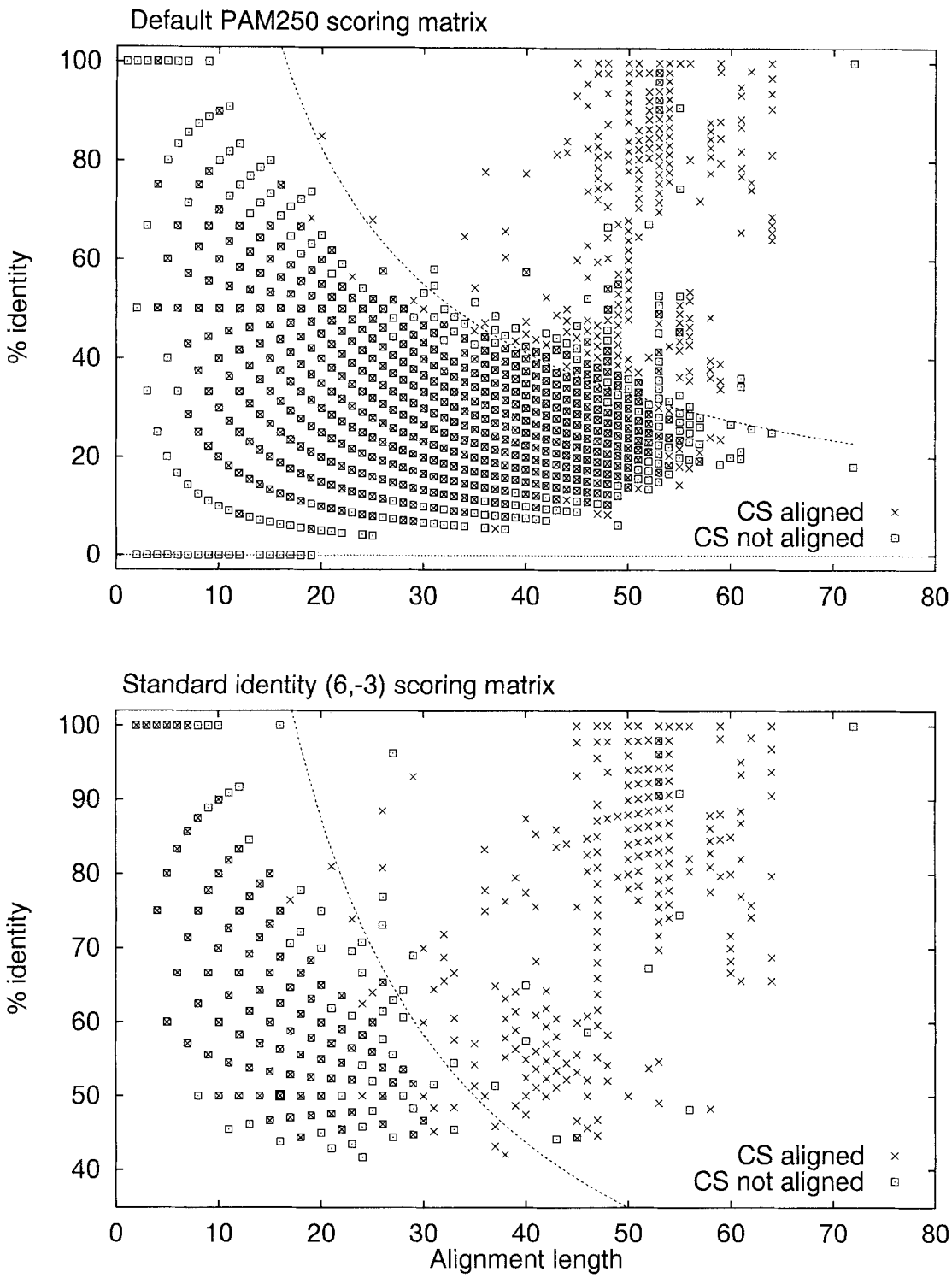

Fig. 1. Overlap length L versus percentage identity p for alignments of human signal peptides. Each data point represents one or more alignment(s) between two signal peptides with known cleavage sites, where the cleavage sites are aligned (crosses) or not aligned (squares) ("CS," Cleavage sites). The alignments were made by fasta using the default PAM250 matrix (**above**) and the standard (6,−3) identity matrix (**below**). Both plots show a region of high similarity where crosses predominate and a region of low similarity containing a mixture of crosses and squares. The dashed curves represent a possible separation of these regions using the number of identical amino acids in the overlap (i.e., the product of L and p) as a threshold. The expression for the dashed curves are $p = 16.5/L$ (**above**) and $p = 17.5/L$ (**below**) (see Fig. 2 for the derivation of these threshold values). Note: the pattern structure in the left-hand part of the plots is an effect of the fact that for every overlap length L, there is only $L + 1$ possible p values: 0, $1/L$, $2/L$, . . . 1.

to the threshold values found to be optimal by distribution and correlation coefficient analysis.

The upper scatter plot (Fig. 1) shows alignments made with the default PAM250 matrix. In this way, an overlap region may be defined mainly by similar but not identical residues, and thus have a low percentage of identities. In contrast, the alignments shown in the lower scatter plot are made with the (6,−3) identity scoring matrix. Therefore, a minimum percentage of identities—more than 33% according to Eq. (3)—is needed to define an overlap region. For the same reason, the overlaps tend to be shorter.

Judged from the scatter plots (Fig. 1), the (6,−3) identity matrix alignments seem to separate crosses from squares better than the default PAM250 matrix alignments. The scatter plots, however, do not show the density of the points, since each square or cross may represent many sequence pairs. Since there are 619 sequences in the human data set, there are $\frac{1}{2} \times 619 \times 618 = 191{,}271$ pairwise comparisons.

A clearer picture of the separation between aligned and nonaligned cleavage sites is seen from the distribution of all these 191,271 pairwise similarities measured by number of identities (Fig. 2). Note that the aligned cleavage sites show a two-peaked or three-peaked distribution—presumably, the first peak represents weak alignments where the cleavage sites are correctly aligned by chance, while the subsequent peaks represent cleavage sites aligned by functional homology. Here, it is apparent that the (6,−3) identity matrix gives a better separation—compared with the default PAM250 matrix, the fraction of aligned cleavage sites in the low-similarity part of the distribution is lower and the number of nonaligned cleavage sites above the crossing point is much lower.

Instead of the number of identities, the matching score from the scoring matrix may be used directly as a discrimination function. These two measures are highly correlated, and the matching score distribution curves (not shown) look very similar to the number of identities distribution curves (Fig. 2) and show the same differences between the two matrices.

The plot of the correlation coefficient as a function of threshold (Fig. 3) confirms that the (6,−3) identity matrix gives a much better separation between aligned and nonaligned pairs than does the PAM250 matrix. However, results for the additional four matrices included in this plot show that the parameter determining the quality is the matrix entropy, rather than the assumed substitution model. Low entropy matrices, such as PAM250 or the (6,−1) identity, have a lower performance compared to any of the higher entropy matrices PAM20, PAM120, the (6,−3) identity, and the (6,−6) identity. Again, it does not make any significant difference whether

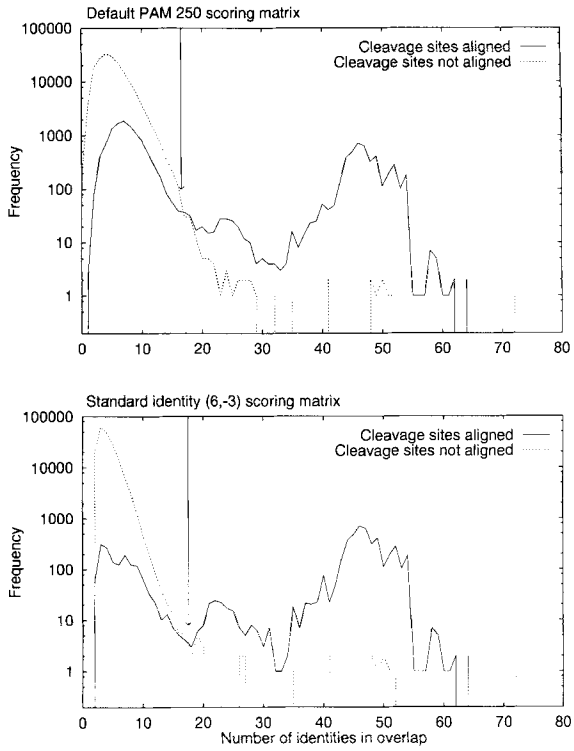

Fig. 2. The distribution (logarithmic scale) of pairwise similarities between human signal peptide sequences measured by the number of identical amino acids in the overlap. The solid and dotted curves show the distributions of alignments with and without correctly aligned cleavage sites, corresponding to the crosses and squares in Figure 1. The distributions may be divided into a region of high similarity where almost all alignments include correctly aligned cleavage sites, and a mixed region where the similarity does not indicate whether the cleavage sites are aligned. The arrows show a possible definition of the border between these regions: the points where the two curves cross, i.e., the threshold above which there are more aligned than nonaligned cleavage sites for a given similarity value. The dashed curves in Figure 1 correspond to these arrows. The alignments were made by fasta using the default PAM250 matrix (**above**) and the standard (6,−3) identity matrix (**below**). Apparently, the identity matrix gives a better separation: the amount of nonaligned cleavage sites to the right of the arrow is smaller, as is the proportion of aligned cleavage sites in the mixed region. Note: several of the isolated dotted lines in the high-similarity region represent errors in the database annotations.

the number of identities or the matching score is used as discrimination function (data not shown).

In order to test the upper limit of the relative entropy range, we also tried the identity matrices (6,−12), (6,−24), and (6,−1000) with relative entropies of 4.29, 4.32, and 4.32, respectively. In practice, the (6,−1000) matrix completely disallows mismatches as described in the Methods section. In these matrices, we avoided gaps by setting set the gap initiation penalty to −1000; if the gap initiation penalty is smaller than the mismatch penalty, the mismatches can simply be replaced by pairs of gaps.

While the (6,−12) identity matrix did not show any significant difference to the (6,−6) matrix (max-

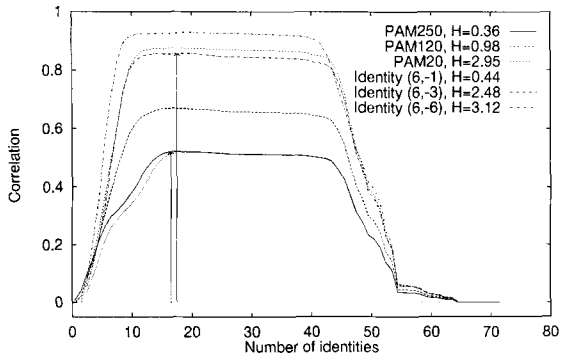

Fig. 3. The correlation coefficient for separation of alignments with and without aligned cleavage sites, shown as a function of the chosen threshold value. A perfect separation would yield a correlation coefficient of 1.0. Six different amino acid substitution matrices have been used, three of the PAM model type and three identity matrices. The relative entropy H of each matrix is indicated. The high entropy matrices give a better separation than the low entropy matrices, regardless of substitution model. The arrows show the threshold values for the PAM250 and (6,−3) matrices, corresponding to the crossing points of the distribution curves (Fig. 2). Note that they coincide with the maximal correlation coefficient values.

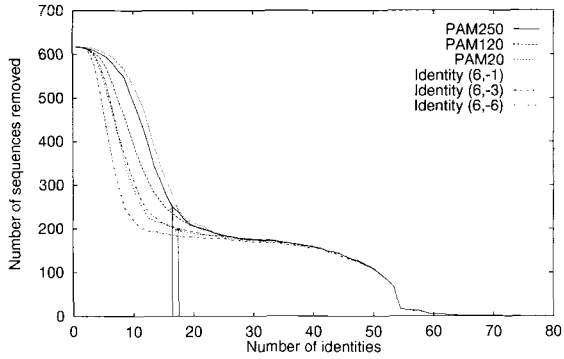

Fig. 4. The number of sequences removed shown as a function of the threshold value used in the Hobohm dilution procedure.[15] Six different amino acid substitution matrices have been used. The arrows show the threshold values for the PAM250 and (6,−3) matrices, corresponding to the crossing points of the distribution curves (Fig. 2). Using a high entropy matrix removes fewer sequences and thus leaves the largest data set.

imal correlation value = 0.93), the (6,−24) matrix was found to perform slightly worse (0.91) and the the (6,−1000) matrix considerably worse (0.62). For the (6,−1000) matrix, the correlation curves did not follow the general pattern (data not shown).

In order to test whether this drop in performance was simply an effect of disallowing gaps, we tried versions of all the PAM and identity matrices with practically infinite gap penalties. This generally *raised* the maximal correlation values by a small amount, but the effect was smaller for matrices with higher entropy (data not shown). This was to be expected since the short alignments produced by the high entropy matrices contain very few gaps—in fact, only 1.6% of the alignments made by the (6,−3) identity matrix contained gaps.

The curves of the correlation coefficient have a broad plateau, i.e., there is a range where the precise choice of threshold value does not make much of a difference for the correlation coefficient. This range—approximately 15–40 identities—corresponds to the "valley" in the distribution of the aligned cleavage sites (Fig. 2). It also corresponds to a relatively flat area in the curves showing the number of entries removed by the Hobohm dilution procedure (see Fig. 4).

It is not important in which part of this range the threshold should be placed. One approach is to use the maximum point of the correlation coefficient, in order to maximize to predictive power of the threshold. An alternative is to determine the point where the distribution curves (Fig. 2) cross, i.e., the similarity value above which the chance of the cleavage sites being aligned, for a given value of the alignment quality, is larger than 0.5. For the human data set, these were found to give identical results. The optimal threshold defined in this way lies between 16 and 17 identities when using the PAM250 matrix, and between 17 and 18 identities when using the (6,−3) identity matrix. As can be seen from the correlation curves (Fig. 3), these threshold values are in the lower part of the correlation coefficient plateau.

However, the "best" threshold value may depend on the goal of the homology analysis. For our purposes, we want to guarantee that no pair of signal peptides below the threshold is functionally homologous, but for different purposes one might want signal peptide sequences guaranteed to be functionally homologous when the similarity is above the threshold—just as the goal of Sander and Schneider[2] was to build a database containing structural homologous proteins. In the latter case, the threshold should be placed in the upper end of the correlation plateau or distribution valley, so that virtually all alignments above the threshold imply correctly aligned cleavage sites [at, say, 29 identities when using the PAM250 matrix or 21 identities when using the (6,−3) identity matrix, see Fig. 2].

The maximum correlation coefficients of the two original matrices are shown with arrows in Figure 3: When using the (6,−3) identity matrix it is 0.857 for a threshold of 17.5 identities, while PAM250 yields 0.521 for a threshold of 16.5 identities. In spite of the poorer separation, more sequences would be removed by the Hobohm procedure using the PAM250 matrices, as seen in Figure 4. With the (6,−3) identity matrix, 201 sequences are removed, in contrast to 250 sequences with PAM250.

We have selected the standard (6,−3) identity matrix with the 17.5 identities threshold for data set reduction. This matrix has near-optimal performance and is included in the standard FASTA dis-

tribution. After scanning the human database for erroneous or questionable cleavage site annotations (see below), five sequences were discarded from the data and the Hobohm procedure was repeated, removing 198 out of 614 sequences. The resulting reduced human data set contains 416 sequences; see Table I for a list of SWISS-PROT identifiers.

If the matching score was used as similarity measure instead of the number of identities, the maximum correlation coefficients were insignificantly lower. The maximum points of the correlation coefficient still coincided with the crossing points of the similarity distributions, and the differences between the various matrices persisted. Furthermore, using the init1 or initn scores instead of the opt score yielded virtually identical results.

Differences Between Data Sets

To see whether the threshold found for the human signal peptide data set is appropriate for signal peptide data sets in general or specific for the human data set, we have repeated the analysis [using fasta with the (6,−3) identity matrix only] for three other signal peptide data sets: the eukaryotic set (including the human set), the *E. coli* set (smaller than the human set), and the prokaryotic set (including the *E. coli* set, approximately the same size as the human set).

For the data set of all eukaryotic signal peptides, the maximal correlation coefficient was considerably lower than for the human data set (see Fig. 5): 0.539. However, the maxima occurred at the same number of identities, and the distribution curves (not shown) crossed at the same point. In other words, the threshold values found for the human subset were consistent with those found when using the entire eukaryotic set.

The prokaryotic data set showed even lower maximal correlation coefficients (see Fig. 5), and the maximum occurred at a much lower similarity value: 0.224 for a threshold of 8 identities. The *E. coli* correlation coefficient curves were more noisy, but similar to the curves of all prokaryotes.

The lower separation ability of the method with the prokaryotic data is also seen from the distribution curves (Fig. 6, here shown for the entire prokaryotic set only). Instead of the two-peaked distribution observed for the human aligned cleavage sites, there is one peak of very short alignments and a more or less flat tail of relatively few strong alignments. There is no well-defined crossing-point, but instead an extended region where the numbers of aligned and nonaligned cleavage site pairs are approximately equal. For the entire prokaryotic set, this range is approximately 13–21 identities, and for *E. coli* it is 10–11 identities.

These regions do not correspond exactly with the maxima of the correlation coefficient, but they are within the range of relatively high correlation. Since the goal of the data set reduction is to avoid pairs of sequences where there is a high chance of finding the cleavage sites by alignment alone without discarding too many sequences, we choose to place the threshold in the upper part of the distribution crossing range for each data set, i.e., 21 identities for the entire prokaryotic set.

The results for the prokaryotic data are shown for the (6,−3) identity matrix only. Although it performed better than the PAM250 matrix, the difference between the various matrices was not nearly as large as observed with the human data set.

Locating Errors in the Database

When using the (6,−3) identity matrix alignments and a threshold of 17.5 identities, there were 25 false positives, i.e., sequence pairs above the threshold without aligned cleavage sites, in the human data set. In order to see whether some of these examples are caused by erroneously assigned cleavage sites, the references in SWISS-PROT to the "worst" examples were checked, and in those cases where errors were found in the SWISS-PROT annotations, the corresponding entries were changed in our data set.

There are also false positives in the prokaryotic data set and the nonhuman part of the eukaryotic data set, but we have not yet checked the references for these. Here, we only report errors found in annotations concerning human signal peptides.

The sequences of FCGB_HUMAN and FCGC_HUMAN (immunoglobulin γ Fc receptor II B and C) were completely identical in the entire cleavage site region (72 identities), but the cleavage site was given at position 42 in FCGB_HUMAN and at position 45 in FCGC_HUMAN. Of the two references to FCGB_HUMAN, one[16] predicted a signal peptide of 44 aa without specifying the prediction method, while the other[17] predicted a signal peptide of 42 aa by homology to the corresponding mouse gene. The only reference to FCGC_HUMAN[18] likewise predicted a signal peptide of 42 aa by homology to the mouse gene—i.e., the signal peptide length of 45 aa could not be supported by the reference and must be an error. In this reference, it was also mentioned that a methionine in position 28 may be the correct initiation codon, making the signal peptide only 15 aa long. Since none of the references mentioned experimental evidence for the cleavage site, both FCGB_HUMAN and FCGC_HUMAN were removed from the database.

The sequence of HA22_HUMAN (HLA class II histocompatibility antigen, DQ(2) α-chain) with the cleavage site given at position 26 had very high similarity values (48–52 identities) to five other sequences (HA21-, -23-, -25-, -26-, and -27_HUMAN) with the cleavage site given at position 23. This turned out to be an error, since the only reference to HA22_HUMAN[19] indicated a signal peptide of 23 aa.

TABLE I. SWISS-PROT ID's of the 416 Entries Comprising the Homology-Reduced and Error-Corrected Data Set of Human Signal Peptides

10KS_HUMAN	CASB_HUMAN	ENPL_HUMAN	IHA_HUMAN	LV6E_HUMAN	PTPG_HUMAN
1B05_HUMAN	CASK_HUMAN	EPOR_HUMAN	IHBA_HUMAN	LYC_HUMAN	PZP_HUMAN
5NTD_HUMAN	CATD_HUMAN	EPO_HUMAN	IL11_HUMAN	LYSH_HUMAN	REL2_HUMAN
7B2_HUMAN	CATE_HUMAN	F13B_HUMAN	IL1R_HUMAN	MABC_HUMAN	RENI_HUMAN
A1AH_HUMAN	CATH_HUMAN	FA12_HUMAN	IL1X_HUMAN	MAG_HUMAN	RETB_HUMAN
A1AT_HUMAN	CATL_HUMAN	FA5_HUMAN	IL2A_HUMAN	MCPI_HUMAN	RIB1_HUMAN
A2AP_HUMAN	CBG_HUMAN	FA8_HUMAN	IL2B_HUMAN	MCP_HUMAN	RIB2_HUMAN
A2HS_HUMAN	CBP1_HUMAN	FBLB_HUMAN	IL2_HUMAN	MDP1_HUMAN	RNKD_HUMAN
A4_HUMAN	CBPB_HUMAN	FCEA_HUMAN	IL3_HUMAN	MG24_HUMAN	SAA_HUMAN
AACT_HUMAN	CBPC_HUMAN	FCG1_HUMAN	IL4_HUMAN	MGP_HUMAN	SABP_HUMAN
ABP_HUMAN	CBPN_HUMAN	FETA_HUMAN	IL5R_HUMAN	MI1B_HUMAN	SAMP_HUMAN
ACET_HUMAN	CCKN_HUMAN	FGF7_HUMAN	IL5_HUMAN	MI2B_HUMAN	SAP3_HUMAN
ACE_HUMAN	CD14_HUMAN	FGR3_HUMAN	IL6R_HUMAN	MK_HUMAN	SAP_HUMAN
ACHA_HUMAN	CD1A_HUMAN	FIBA_HUMAN	IL6_HUMAN	MLCH_HUMAN	SCF_HUMAN
ACHB_HUMAN	CD1D_HUMAN	FIBB_HUMAN	IL7R_HUMAN	MOTI_HUMAN	SEM2_HUMAN
ACHE_HUMAN	CD1E_HUMAN	FIBH_HUMAN	IL7_HUMAN	MPRD_HUMAN	SG1_HUMAN
ACHG_HUMAN	CD28_HUMAN	FINC_HUMAN	IL8_HUMAN	MPRI_HUMAN	SIAL_HUMAN
ACHN_HUMAN	CD2_HUMAN	FKB3_HUMAN	IL9_HUMAN	MYP0_HUMAN	SLIB_HUMAN
ACRO_HUMAN	CD30_HUMAN	FOL2_HUMAN	INA7_HUMAN	NAGA_HUMAN	SMS1_HUMAN
ALBU_HUMAN	CD3D_HUMAN	FSA_HUMAN	INB_HUMAN	NCA2_HUMAN	SODE_HUMAN
ALK1_HUMAN	CD3E_HUMAN	FSHB_HUMAN	INGR_HUMAN	NDDB_HUMAN	SOMW_HUMAN
ALS_HUMAN	CD3G_HUMAN	GA6S_HUMAN	ING_HUMAN	NEC2_HUMAN	SPRC_HUMAN
AMYP_HUMAN	CD3Z_HUMAN	GELS_HUMAN	INIG_HUMAN	NEU2_HUMAN	SRCH_HUMAN
ANF_HUMAN	CD45_HUMAN	GL6S_HUMAN	INIP_HUMAN	NEUB_HUMAN	SSBP_HUMAN
ANGI_HUMAN	CD4X_HUMAN	GLCM_HUMAN	INSR_HUMAN	NGFR_HUMAN	STAT_HUMAN
ANGT_HUMAN	CD4_HUMAN	GLHA_HUMAN	INS_HUMAN	NIDO_HUMAN	STS_HUMAN
ANPA_HUMAN	CD52_HUMAN	GLPE_HUMAN	IPSP_HUMAN	NMZ2_HUMAN	TCO1_HUMAN
ANPC_HUMAN	CD59_HUMAN	GLUC_HUMAN	IPST_HUMAN	OMGP_HUMAN	TCO2_HUMAN
ANT3_HUMAN	CD5_HUMAN	GLYP_HUMAN	IRBP_HUMAN	ONCM_HUMAN	TENA_HUMAN
APA1_HUMAN	CD7_HUMAN	GMCR_HUMAN	ITA2_HUMAN	P4HA_HUMAN	TETN_HUMAN
APA2_HUMAN	CD82_HUMAN	GONL_HUMAN	ITA4_HUMAN	PA21_HUMAN	TFPI_HUMAN
APA4_HUMAN	CD8A_HUMAN	GP1A_HUMAN	ITA6_HUMAN	PA2M_HUMAN	TF_HUMAN
APC1_HUMAN	CERU_HUMAN	GP1B_HUMAN	ITAB_HUMAN	PAHO_HUMAN	TGR3_HUMAN
APC2_HUMAN	CETP_HUMAN	GP39_HUMAN	ITAL_HUMAN	PAI1_HUMAN	THBG_HUMAN
APC3_HUMAN	CFAI_HUMAN	GPIX_HUMAN	ITAV_HUMAN	PBGD_HUMAN	THY1_HUMAN
APD_HUMAN	CHLE_HUMAN	GR78_HUMAN	ITAX_HUMAN	PDGB_HUMAN	THYG_HUMAN
APE_HUMAN	CLUS_HUMAN	GRA1_HUMAN	ITB1_HUMAN	PEC1_HUMAN	TIM2_HUMAN
APOA_HUMAN	CMGA_HUMAN	GRAA_HUMAN	ITB2_HUMAN	PENK_HUMAN	TNFB_HUMAN
APOH_HUMAN	CO2_HUMAN	GRP2_HUMAN	ITB4_HUMAN	PEPA_HUMAN	TNR1_HUMAN
ARSA_HUMAN	CO3_HUMAN	GUAN_HUMAN	ITB7_HUMAN	PEPC_HUMAN	TNR2_HUMAN
ASM_HUMAN	CO4_HUMAN	HA25_HUMAN	KAL_HUMAN	PERF_HUMAN	TRFE_HUMAN
ASPG_HUMAN	CO6_HUMAN	HA2R_HUMAN	KFMS_HUMAN	PF4L_HUMAN	TRFL_HUMAN
AXO1_HUMAN	CO7_HUMAN	HA2Z_HUMAN	KHEK_HUMAN	PGDR_HUMAN	TRFM_HUMAN
B2MG_HUMAN	CO8G_HUMAN	HB23_HUMAN	KKIT_HUMAN	PGDS_HUMAN	TRKA_HUMAN
B61_HUMAN	COG1_HUMAN	HB2A_HUMAN	KMET_HUMAN	PGH1_HUMAN	TRY2_HUMAN
B71_HUMAN	COG7_HUMAN	HB2Q_HUMAN	KNL_HUMAN	PGSG_HUMAN	TRYA_HUMAN
BAL_HUMAN	COG8_HUMAN	HC_HUMAN	KV4B_HUMAN	PLFV_HUMAN	TSHB_HUMAN
BFR2_HUMAN	COG9_HUMAN	HEP2_HUMAN	KV5A_HUMAN	PLMN_HUMAN	TSHR_HUMAN
BGAM_HUMAN	COL_HUMAN	HEXA_HUMAN	LAG3_HUMAN	PLR2_HUMAN	TSP1_HUMAN
BGLR_HUMAN	CR1_HUMAN	HGFA_HUMAN	LBP_HUMAN	PP11_HUMAN	TTHY_HUMAN
BLSA_HUMAN	CR2_HUMAN	HGF_HUMAN	LCAT_HUMAN	PP14_HUMAN	TVA2_HUMAN
C1QA_HUMAN	CRFB_HUMAN	HIS3_HUMAN	LCA_HUMAN	PPA5_HUMAN	TVA3_HUMAN
C1QC_HUMAN	CRTC_HUMAN	HPT2_HUMAN	LDLR_HUMAN	PPAL_HUMAN	TVB2_HUMAN
C1R_HUMAN	CSF1_HUMAN	HRG_HUMAN	LEM1_HUMAN	PPAP_HUMAN	TVC_HUMAN
C1S_HUMAN	CSF2_HUMAN	HV1B_HUMAN	LEM3_HUMAN	PPB3_HUMAN	TYRR_HUMAN
C4BB_HUMAN	CSF3_HUMAN	HV2H_HUMAN	LEUK_HUMAN	PPBT_HUMAN	UPAR_HUMAN
C4BP_HUMAN	CTRB_HUMAN	HV2I_HUMAN	LFA3_HUMAN	PRIO_HUMAN	UROK_HUMAN
CA11_HUMAN	CYPB_HUMAN	HV3C_HUMAN	LIF_HUMAN	PRL_HUMAN	UROM_HUMAN
CA13_HUMAN	CYRG_HUMAN	I12A_HUMAN	LIPG_HUMAN	PRN3_HUMAN	VEGF_HUMAN
CA14_HUMAN	CYTC_HUMAN	I12B_HUMAN	LIPH_HUMAN	PROP_HUMAN	VIP_HUMAN
CA18_HUMAN	CYTS_HUMAN	I309_HUMAN	LIPL_HUMAN	PRPC_HUMAN	VTDB_HUMAN
CA19_HUMAN	DAF2_HUMAN	IAC2_HUMAN	LIPP_HUMAN	PRTC_HUMAN	VTNC_HUMAN
CA21_HUMAN	DEFN_HUMAN	IBP1_HUMAN	LITH_HUMAN	PRTP_HUMAN	VWF_HUMAN
CA24_HUMAN	DOPO_HUMAN	IBP2_HUMAN	LMB1_HUMAN	PRTS_HUMAN	WNT1_HUMAN
CA25_HUMAN	DRN1_HUMAN	IBP3_HUMAN	LMB2_HUMAN	PRTZ_HUMAN	ZA2G_HUMAN
CAH4_HUMAN	E2_HUMAN	IBP4_HUMAN	LMP1_HUMAN	PS2_HUMAN	ZP2_HUMAN
CAH6_HUMAN	EGFR_HUMAN	IC1_HUMAN	LMP2_HUMAN	PSPA_HUMAN	
CAMA_HUMAN	EL2B_HUMAN	ICA1_HUMAN	LPH_HUMAN	PSSP_HUMAN	
CAML_HUMAN	ELS_HUMAN	ICA2_HUMAN	LSHB_HUMAN	PTHY_HUMAN	
CAP7_HUMAN	EMBP_HUMAN	IGF2_HUMAN	LV0A_HUMAN	PTN_HUMAN	

The cleavage site of HA22_HUMAN was changed accordingly.

The sequences of SOMV_HUMAN and SOMW_HUMAN (growth hormone variant I and II) which were completely identical in the cleavage site region and had the cleavage site at position 25, aligned well (50 and 41 identities respectively) with SOMA_HUMAN (somatotropin) and PLC_HUMAN (lactogen or chorio-

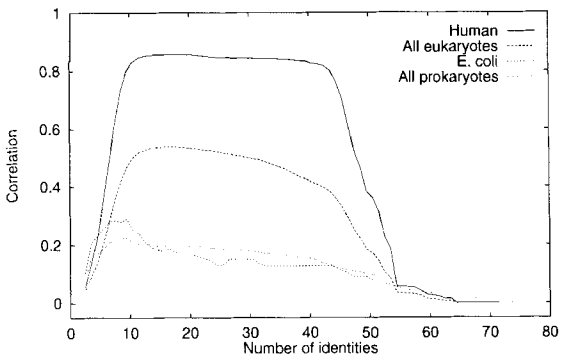

Fig. 5. Correlation coefficients for all four data sets shown as a function of the chosen threshold value (number of identities). All alignments were made with fasta using the standard (6,−3) identity matrix.

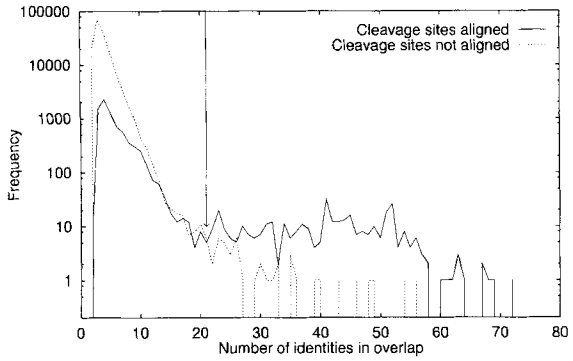

Fig. 6. The distribution (logarithmic scale) of pairwise similarities between prokaryotic signal peptide sequences measured by the number of identical amino acids in the overlap. The alignments were made with fasta using the standard (6,−3) identity matrix. Compared to Figure 2, it is evident that the procedure yields a poorer separation for these data. The crossing point is not well defined; the arrow shows our preferred placement of the threshold.

mammotropin) which both had the cleavage site at position 26. This also seems to be be an error, since one of the references to both SOMV_HUMAN and SOMW_HUMAN[20] indicated the cleavage site at position 26. The cleavage sites of these two entries were changed accordingly.

These three changes alone removed the 10 "worst" false positives. A few additional examples of errors or questionable evidence were found among the less significant false positives:

The cleavage site of ELNE_HUMAN (leukocyte elastase or neutrophil elastase or medullasin) is given at position 29, but three of the references[21–23] speculate that this may be a 27 aa signal peptide plus a 2 aa propeptide which is cleaved off in a later step. This is inferred from cleavage site consensus and homology to a protein with an experimentally confirmed 2 aa propeptide, CATG_HUMAN (cathepsin G). This would correct the conflicting alignments with CATG_HUMAN (26 identities) and PRN3_HUMAN (proteinase 3 or myeloblastin, 18 identities), but since the evidence is not experimental, ELNE_HUMAN was removed from the database.

Finally, FCGA_HUMAN and FCG3_HUMAN (immunoglobulin γ Fc receptors II and III-1) which produced a false positive (27 identities) had no experimental evidence for their cleavage sites in any of the eight references. FCGA_HUMAN had the cleavage site given at position 36; this was in agreement with one of the references[24] which used analogy with the mouse gene and a consensus method[25]; but another[26] placed the cleavage site at position 34, curiously also by analogy with the mouse gene; while the two remaining references[16,27] placed the cleavage site at position 35 without indicating the prediction method. In addition, there was disagreement about the start of the signal peptide—one of the references[26] mentioned three possible initiation codons. FCG3_HUMAN had the cleavage site given at position 18 in agreement with two of the references[28,29] which cited a weight matrix cleavage site prediction method,[30] while the two remaining references[31,32] did not contain any signal peptide information. For the lack of experimental evidence, FCGA_HUMAN and FCG3_HUMAN were removed from the database.

All these changes abolished 15 of the 25 false positives, 6 due to removal of sequences and 9 due to changes in cleavage site position. The strongest remaining false positive is PZP_HUMAN (pregnancy zone protein) versus A2MG_HUMAN (α_2-macroglobulin) with 35 identities, but this is not necessarily in error—a close inspection of the alignment shows that the most significant difference between the two sequences occurs precisely in the cleavage site region. The other remaining false positives with 18–27 identities have their region of high identity either upstream or downstream of the cleavage site.

A few additional false positives which were found in a pilot study using SWISS-PROT version 27 did not appear when using version 29. The entry KGFR_HUMAN (keratinocyte growth factor receptor), which had a cleavage site in disagreement with BFR2_HUMAN and FGR2_HUMAN (fibroblast growth factor receptor), has been merged into FRG2_HUMAN. The entries EPIA_HUMAN and EPIB_HUMAN (episialin variant A and B), which were identical except for a 9 aa gap in the cleavage site region of EPIB_HUMAN, have been merged into MUC1_HUMAN (mucin 1). In both cases, the cleavage sites have been marked "POTENTIAL."

DISCUSSION

The results show how a *functional* sequence pattern can be related to the sequence similarity. By using distribution curves and correlation coefficient calculations of the similarity function, it is possible to compare different alignment algorithms and sim-

ilarity measures, and to determine the appropriate threshold value. The similarity measure found to be optimal—the number of identities in an alignment—is very simple.

Significance of the Substitution Matrix Entropy

When the objective is to divide a set of pair-wise alignments in two subsets by a sequence similarity threshold, the overall determining parameter is the substitution matrix entropy. We found that alignments made by a simple protein identity scoring matrix gave a similar separation performance as those made by the more sophisticated PAM matrices, provided that the matrix entropies were comparable. This is in contrast to Sander and Schneider,[2] who claim that it is important to use a more sophisticated measure of sequence similarity in producing the alignments, and that a more refined measure in calibrating the homology threshold presumably would result in fewer false positive assignments.

The significance of the matrix entropy may be explained by the fact the we are dealing with alignments of short (local) sequence patterns, as being the case not only for signal peptide cleavage sites, but also for many other functional sites. Low entropy matrices produce long weak alignments, as opposed to high entropy matrices which are well suited for the detection of strong localized similarities. The extended overlaps detected by low entropy matrices do not relate in particular to the functional site—here the cleavage site—but contribute false positives lowering the correlation coefficient. It is the relative weight between the score of matches and mismatches which is important, not the difference between conservative and nonconservative replacements, as implicitly taken for granted in the Sander and Schneider argument.

Further, identity matrices are theoretically a natural choice from our point of view, because we use sparse encoding of the amino acids (i.e., regard them as equidistant like the identity matrices do) in the signal peptide prediction method under development.[4]

Possible Fast Version of the Similarity Analysis

As mentioned in the Methods section, we found the efficient fasta algorithm performed as well as a full Smith–Waterman alignment. In addition, we found that the initial matching scores from fasta could yield the same performance as the optimized score. The matching score, when used as a similarity measure, was in most cases found to yield a slightly lower performance than the number of identities in the alignment; but it has the advantage that it is possible to calculate an estimate of it (the initial scores) using fasta without actually performing the computationally expensive alignment.

As mentioned in Methods, it is necessary to perform all pairwise alignments in order to be able to determine whether the cleavage sites are aligned and to calculate the number of identities in the alignment. However, once an appropriate threshold score is found, the initial scores may be used for data set reduction without performing all the alignments. The fact that the analysis of the human data set resulted in the same threshold score value as its eukaryotic superset suggests the following scheme: First, the computationally expensive analysis where all alignments are needed is carried out on a subset of the data in order to find the appropriate threshold score, and then fasta can be used on the entire data set in its fast mode, i.e., without requesting optimization of all alignments. This makes the calculation of a complete pairwise similarity matrix feasible within reasonable computing time even for very large data sets.

However, these results may be dependent on the fact that our sequences are relatively short and show a comparatively small variation in length. With longer sequences, it may not be the case that initial score correlates well with optimized score.

Generalization of the Procedure

Our approach may be generalized to data sets of other functional sequence patterns. The definition of functional homology may, however, be more complicated in other cases. In the case of signal peptides, we can assume that there is precisely one cleavage site in each sequence, which makes the functional similarity concept very simple: The cleavage sites of a sequence pair are either aligned or not.

For other problems, such as splice sites in pre-mRNA or glycosylation sites of proteins, there are several sites per sequence. One way of addressing this problem is to split each sequence into a number of subsequences, one for each potential site, and then use our approach on the collection of subsequences. Alternatively, the fraction of aligned sites per alignment may be used as a functional homology measure, in analogy with the structural homology used by Sander and Schneider[2] (the percentage of identical secondary structure assignments in the alignment). In this case, a threshold value for functional homology—analogous to the 70% structural homology threshold used by Sander and Schneider—must be defined before the similarity threshold can be calculated.

Differences Between Data Sets

The poorer separation performance of the alignments on prokaryotic data compared with eukaryotic data may reflect differences both in data distribution and in the sequence patterns. On one hand, the missing second peak in the distribution of the aligned prokaryotic cleavage sites possibly represents a smaller number of truly related proteins,

since protein or gene families are not as abundant in prokaryotes as in eukaryotes. On the other hand, the local cleavage site consensus is known to be stronger in prokaryotes—i.e., the prokaryotic cleavage sites carry more information—while the global signal peptide variability is larger.[4] This may increase the chance of finding both cleavage sites in a very short local alignment of two completely unrelated signal peptides, which will reduce the correlation between sequence similarity and apparent functional homology. In this case, a global alignment of the signal peptide sequences might yield a better separation.

This serves as a demonstration of the complexity of the functional homology problem. Great caution should always be taken when applying a threshold value found for one data set to another. Even for the same type of sequence pattern, the appropriate threshold values may differ depending on the amount of information in the pattern. For splice sites in pre-mRNA, e.g., the situation may be different for donor (5′) and acceptor (3′) sites, although there must be a one-to-one correspondence between them in the sequences, since they have different extension and information content.[33]

Eliminating Database Errors

An important result is that close inspection of the false positive alignments can be useful in eliminating some database errors which otherwise might lead to inconsistencies in the data set. By checking the references we were able to explain more than half of the false positives by errors in SWISS-PROT. In addition to erroneously placed cleavage sites, we found several examples of cleavage sites that were predicted or inferred by homology without this being mentioned in the database entry. Many entries simply lack information about the quality of the evidence, as we have previously seen from comparison of SWISS-PROT to the manually compiled signal peptide database SIGPEP[34] (results not published). In other words, our data selection procedure described in the data section reduces the amount of cleavage sites which are not experimentally determined, but it does not eliminate them. This serves as an illustration of another of the recurring problems haunting the analysis of protein and DNA sequences: the quality of database annotations.

Signal Peptide Data Available by Anonymous FTP

The redundancy reduced and error corrected data set is available by anonymous FTP from the Center for Biological Sequence Analysis. It is deposited in the pub/signalp directory on the machine virus.cbs.dtu.dk.

ACKNOWLEDGMENTS

This work was funded by the Danish National Research Foundation and the Swedish Medical Research Council. We thank Ole Lund and other coworkers at Center for Biological Sequence Analysis in Lyngby for inspiring comments.

REFERENCES

1. Prechelt, L. A study of experimental evaluations of neural network learning algorithms: Current research practice. Technical Report 19/94 Fakultät für Informatik, Universität Karlsruhe 1994.
2. Sander, C., Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9:56–68, 1991.
3. von Heijne, G. The signal peptide. J. Membrane Biol. 115: 195–201, 1990.
4. Nielsen, H., Brunak, S., Engelbrecht, J., von Heijne, G. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Submitted.
5. Bairoch, A., Boeckmann, B. The SWISS-PROT protein sequence data bank: Current status. Nucleic Acids Res. 22: 3578–3580, 1994.
6. von Heijne, G. The structure of signal peptides from bacterial lipoproteins. Protein Eng. 2:531–534, 1989.
7. Bairoch, A. PROSITE: A dictionary of sites and patterns in proteins. Nucleic Acids Res. 20:2013–2018, 1992.
8. Rasmussen, B. A., Silhavy, T. J. The first 28 amino acids of mature LamB are required for rapid and efficient export from the cytoplasm. Genes Dev. 1:185–196, 1987.
9. Andersson, H., von Heijne, G. A 30-residue-long "export initiation domain" adjacent to the signal sequence is critical for protein translocation across the inner membrane of *Escherichia coli*. Proc. Natl. Acad. Sci. U.S.A. 88:9751–9754, 1991.
10. Pearson, W. R. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. 183:63–98, 1990.
11. Dayhoff, M., ed. "Atlas of Protein Sequence and Structure," Vol. 5, Suppl. 3. Silver Spring, MD: National Biomedical Research Foundation, 1978.
12. Altschul, S. F. Amino acid substitution matrices from an information theoretic perspective. J. Mol. Biol. 219:555–565, 1991.
13. Thompson, J., Higgins, D., Gibson, T. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680, 1994.
14. Mathews, B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim. Biophys. Acta 405:442–451, 1975.
15. Hobohm, U., Scharf, M., Schneider, R., Sander, C. Selection of representative protein data sets. Protein Sci. 1:409–417, 1992.
16. Brooks, D., Qiu, W., Luster, A., Ravetch, J. Structure and expression of human IgG FcRII (CD32). J. Exp. Med. 170: 1369–1385, 1989.
17. Engelhardt, W., Geerds, C., Frey, J. Distribution, inducibility and biological function of the cloned and expressed human βFc receptor II. Eur. J. Immunol. 20:1367–1377, 1990.
18. Stuart, S., Simister, N., Clarkson, S., Kacinski, B., Shapiro, M., Mellman, I. Human IgG receptor (hFcRII; CD32) exists as multiple isoforms in macrophages, lymphocytes and IgG-transporting placental epithelium. EMBO J. 8:3657–3666, 1989.
19. Chang, H.-C., Moriuchi, T., Silver, J. The heavy chain of human B-cell alloantigen HLA-DS has a variable N-terminal region and a constant immunoglobulin-like region. Nature (London) 305:813–815, 1983.
20. Cooke, N., Ray, J., Emery, J., Liebhaber, S. Two distinct species of human growth hormone-variant mRNA in the human placenta predict the expression of novel growth hormone proteins. J. Biol. Chem. 263:9001–9006, 1988.
21. Farley, D., Travis, J., Salvesen, G. The human neutrophil

elastase gene. Biol. Chem. Hoppe-Seyler 370:737–744, 1989.
22. Takahashi, H., Nukiwa, T., Yoshimura, K., Quick, C., States, D., Holmes, M., Whang-Peng, J., Knutsen, T., Crystal, R. Structure of the human neutrophil elastase gene. J. Biol. Chem. 263:14739–14747, 1988.
23. Okano, K., Aoki, Y., Shimizu, H., Naruto, M. Functional expression of human leukocyte elastase (HLE)/medullasin in eukaryotic cells. Biochem. Biophys. Res. Commun. 167: 1326–1332, 1990.
24. Hibbs, M., Bonadonna, L., Scott, B., Mckenzie, I., Hogarth, P. Molecular cloning of a human immunoglobulin G Fc receptor. Proc. Natl. Acad. Sci. U.S.A. 85:2240–2244, 1988.
25. von Heijne, G. Patterns of amino acids near signal sequence cleavage sites. Eur. J. Biochem. 133:17–21, 1983.
26. Stuart, S., Trounstine, M., Vaux, D., Koch, T., Martens, C., Moore, K. Isolation and expression of cDNA clones encoding a human receptor for IgG (FcγRII). J. Exp. Med. 166: 1668–1684, 1987.
27. Stengelin, S., Stamenkovic, I., Seed, B. Isolation of cDNAs for two distinct human Fc receptors by ligand affinity cloning. EMBO J. 7:1053–1059, 1988.
28. Simmons, D., Seed, B. The Fcγ receptor of natural killer cells is a phospholipid-linked membrane protein. Nature (London) 333:568–570, 1988.
29. Peltz, G., Grundy, H., Lebo, R., Yssel, H., Barsh, G., Moore, K. Human FcγRIII: Cloning, expression, and identification of the chromosomal locus of two Fc receptors for IgG. *Proc.* Natl. Acad. Sci. U.S.A. 86:1013–1017, 1989.
30. von Heijne, G. A new method for predicting signal sequence cleavage sites. Nucleic Acids Res. 14:4683–4690, 1986.
31. Ravetch, J., Perussia, B. Alternative membrane forms of FcγRIII(CD16) on human natural killer cells and neutrophils. J. Exp. Med. 170:481–497, 1989.
32. Simmons, D., Seed, B. Erratum. Nature (London) 340:662, 1989.
33. Engelbrecht, J., Knudsen, S., Brunak, S. G+C-rich tract in 5′ end of human introns. J. Mol. Biol. 227:108–113, 1992.
34. von Heijne, G., Abrahmsén, L. Species-specific variation in signal peptide design. FEBS Lett. 244:439–446, 1989.