A Sequence Similarity Search Algorithm Based on a Probabilistic Interpretation of an Alignment Scoring System

Philipp Bucher and **Kay Hofmann,**

Swiss Institute for Experimental Cancer Research (ISREC)
Ch. des boveresses 155, CH–1066 Epalinges, Switzerland
{pbucher, khofmann}@isrec-sun1.unil.ch

Abstract

We present a probabilistic interpretation of local sequence alignment methods where the alignment scoring system (ASS) plays the role of a stochastic process defining a probability distribution over all sequence pairs. An explicit algorithm is given to compute the probability of two sequences given an ASS. Based on this definition, a modified version of the Smith-Waterman local similarity search algorithm has been devised, which assesses sequence relationships by log likelihood ratios. When tested on classical examples such as globins or G-protein-coupled receptors, the new method proved to be up to an order of magnitude more sensitive than the native Smith-Waterman algorithm.

Introduction

The comparison of a new protein sequence against a database of known proteins is perhaps the most important computer application in molecular sequence analysis. It is generally accepted that the Smith-Waterman local similarity search algorithm (Smith & Waterman 1981) is the most sensitive technique to discover significant weak similarities between two sequences. The more frequently used heuristic algorithms implemented in the programs FASTA (Pearson 1990) and BLAST (Altschul et al. 1990) can be considered approximations or special cases of a full Smith-Waterman algorithm offering high speed in exchange for reduced sensitivity.

The Smith-Waterman algorithm maximizes an alignment scoring function over all possible local alignments between two sequences. The scoring function depends on a set of parameters referred to as an alignment scoring system, which consists of a residue substitution matrix and a linear gap penalty function. Optimal alignment scores computed by a Smith-Waterman or related sequence alignment algorithm have traditionally been viewed and interpreted as measures of similarity or distance (Smith, Waterman, & Fitch 1981). Concerning their mathematical properties, they were shown to define a metric over the sequence space (*e.g.* Sellers 1974).

Here, we advocate a different view of local sequence alignment methods, in which the scoring system plays the role of a stochastic process generating pairs of related sequences. Based on such an interpretation, we propose a modified version of a Smith-Waterman algorithm where the score computed for two sequences is a log likelihood ratio of two probabilities, one given the scoring system, and one given a null-model. The mathematical concepts underlying this approach is closely related to maximum likelihood estimation for global sequence alignments (*e.g.* Bishop & Thompson 1986). The performance of the new database search technique is assessed by test protocols previously used for the evaluation of alternative alignment scoring systems.

Review of the Native Smith-Waterman Algorithm

Let $\mathbf{a} = a_1 a_2 \cdots a_m$ and $\mathbf{b} = b_1 b_2 \cdots b_n$ be two sequences of residues from an alphabet $\mathbf{S}$ containing N elements. A local sequence alignment between two such sequences is defined by an alignment path represented as an ordered set of index pairs: $\mathbf{u} = (x_1,y_1), (x_2,y_2), \cdots (x_l,y_l)$. Each index pair identifies a pair of matched residues in the sequence alignment. A valid path $\mathbf{u}$ for sequences $\mathbf{a}$ and $\mathbf{b}$ satisfies the following conditions: $x_{k+1} > x_k$, $y_{k+1} > y_k$, $x_l \le m$, $y_l \le n$.

An alignment scoring system (ASS) consists of a substitution matrix and a gap weighting function. The substitution matrix defines substitution scores $s(a,b)$ for pairs of residues $(a,b) \in \mathbf{S}^2$. The gap weighting function assigns weights $w(k) = \alpha + \beta k$ to insertions and deletions of length $k \ge 1$. For gap length zero, we define $w(0) = 0$.

The scoring system assigns an alignment score S_A to any local sequence alignment $(\mathbf{a},\mathbf{b},\mathbf{u})$:

$$S_A(\mathbf{a},\mathbf{b},\mathbf{u}) \;=\; \sum_{k=1}^{l} s(a_{x_k},b_{y_k}) \qquad (1)$$

$$- \sum_{k=1}^{l-1} w(x_{k+1}-x_k-1)+w(y_{k+1}-y_k-1)$$

Note that the alignment score can be described as the sum of two components, a sequence-dependent match score:

$$S_M(\mathbf{a},\mathbf{b},\mathbf{u}) \;=\; \sum_{k=1}^{l} s(a_{x_k},b_{y_k}) \qquad (2a)$$

and a sequence-independent gap score:

$$S_G(\mathbf{u}) \;=\; -\sum_{k=1}^{l-1} w(x_{k+1}-x_k-1)+w(y_{k+1}-y_k-1) \qquad (2b)$$

These two definitions will be useful in the formulation of the probabilistic version of the Smith-Waterman algorithm. The native Smith-Waterman algorithm computes the optimal local alignment score for two sequences:

$$SWscore(\mathbf{a},\mathbf{b}) \;=\; \max_{\text{paths } \mathbf{u}} S_A(\mathbf{a},\mathbf{b},\mathbf{u}), \qquad (3)$$

which serves as a measure of sequence similarity in database search applications. An efficient procedure to compute $SWscore(\mathbf{a},\mathbf{b})$ is described in (Gotoh 1982).

Probabilistic Smith-Waterman (PSW) Algorithm

The probabilistic version of the Smith-Waterman algorithm is based on an analogy between alignment scoring systems (ASSs) and hidden Markov models (HMMs), a class of statistical models that have recently been introduced to molecular sequence analysis (Baldi et al. 1994, Krogh et al. 1994). This analogy comprises the following ideas and assumptions:

(i) An HMM defines a probability distribution over the sequence space by means of a stochastic process involving a random walk through the model.
An ASS defines a probability distribution over the space of sequence pairs by means of a stochastic process involving a random walk through an alignment path matrix.

(ii) The probability of a sequence $\mathbf{a}$ being generated by an HMM is the sum of the probabilities of sequence $\mathbf{a}$ being generated via a particular path $\mathbf{u}$:

$$\mathrm{Prob}_{\mathrm{HMM}}(\mathbf{a}) \;=\; \sum_{\text{paths } \mathbf{u}} \mathrm{Prob}_{\mathrm{HMM}}(\mathbf{a},\mathbf{u})$$

over all possible paths through the model.
The probability of a sequence pair $\mathbf{a}$,$\mathbf{b}$ being generated by an ASS is the sum of the probabilities of sequence pair $\mathbf{a}$,$\mathbf{b}$ being generated via a particular path $\mathbf{u}$:

$$\mathrm{Prob}_{\mathrm{ASS}}(\mathbf{a},\mathbf{b}) \;=\; \sum_{\text{paths } \mathbf{u}} \mathrm{Prob}_{\mathrm{ASS}}(\mathbf{a},\mathbf{b},\mathbf{u})$$

over all valid local alignment paths.

(iii) In database applications, membership of a sequence $\mathbf{a}$ to an HMM-defined sequence class is estimated by an *HMMscore* which has the form of a log likelihood ratio:

$$HMMscore(\mathbf{a}) \;=\; \log\frac{\mathrm{Prob}_{\mathrm{HMM}}(\mathbf{a})}{\mathrm{Prob}_{\mathrm{null}}(\mathbf{a})}$$

where $\mathrm{Prob}_{null}(\mathbf{a})$ is the probability of sequence $\mathbf{a}$ given a specific null-model.
In a probabilistic Smith-Waterman search, an ASS-defined kind of similarity between two sequences $\mathbf{a}$,$\mathbf{b}$ will be estimated by a *PSWscore* of the following type:

$$PSWscore(\mathbf{a},\mathbf{b}) \;=\; \log\frac{\mathrm{Prob}_{\mathrm{ASS}}(\mathbf{a},\mathbf{b})}{\mathrm{Prob}_{\mathrm{null}}(\mathbf{a},\mathbf{b})}$$

What remains to be done in order to implement the method suggested by this analogy, is to chose an appropriate null-model for random sequence pairs, and to work out a reasonable definition for the probability of a local sequence alignment $\mathrm{Prob}_{\mathrm{ASS}}(\mathbf{a},\mathbf{b},\mathbf{u})$ based on the definition of the alignment score.

The null-model we choose has the general form:

$$\mathrm{Prob}_{\mathrm{null}}(\mathbf{a},\mathbf{b}) \;=\; P_L(m,n)P_0(\mathbf{a},\mathbf{b}) \qquad (4)$$

where $P_L(m,n)$ is a length distribution over sequence length pair classes, and $P_0(\mathbf{a},\mathbf{b})$ is the null-model probability of sequence pair $\mathbf{a}$,$\mathbf{b}$ given the length pair pair class n,m, which is defined as follows:

$$P_0(\mathbf{a},\mathbf{b}) \;=\; \prod_{i=1}^{m} p(a_i) \prod_{j=1}^{n} p(b_j) \qquad (5)$$

where $p(a)$ denotes the null-model probability of residue a. The null-model thus essentially consists of a residue probability distribution over the alphabet $\mathbf{S}$.

The distribution of sequence length pair classes will be the same for the null-model and the ASS-defined probability distribution. For reasons that will become clear later, we require that there is a logarithmic base z such that:

$$\sum_{a \in \mathbf{S}, b \in \mathbf{S}} p(a) p(b) z^{s(a,b)} = 1. \qquad (6)$$

Note that log-odds substitution matrices, such as those from the PAM or BLOSUM series (Henikoff & Henikoff 1992), have known mutational equilibrium compositions and logarithmic bases satisfying the above condition. If a different residue composition is used as null-model, there will always be a unique solution for z solving the above equation, if the substitution matrix satisfies the conditions necessary for local sequence alignments (Altschul 1991).

The ASS-defined probability distribution has the general form:

$$\text{Prob}_{\text{ASS}}(\mathbf{a},\mathbf{b}) \;=\; P_L(m,n) P_A(\mathbf{a},\mathbf{b}) \qquad (7)$$

where $P_A(\mathbf{a},\mathbf{b})$ is the ASS-defined probability of sequence pair $\mathbf{a},\mathbf{b}$, given the length pair class n,m. The precise nature of the length pair class distribution is irrelevant, as long as it is the same for the null-model and the ASS, since the term $P_L(m,n)$ cancels itself in the definition of the $PSWscore$:

$$PSWscore \;=\; \frac{\text{Prob}_{\text{ASS}}(\mathbf{a},\mathbf{b})}{\text{Prob}_{\text{null}}(\mathbf{a},\mathbf{b})} \qquad (8)$$

$$= \frac{\sum_{\text{paths } \mathbf{u}} P_L(n,m)\; P_A(\mathbf{a},\mathbf{b},\mathbf{u})}{P_L(n,m) P_0(\mathbf{a},\mathbf{b})}$$

$$= \frac{\sum_{\text{paths } \mathbf{u}} P_A(\mathbf{a},\mathbf{b},\mathbf{u})}{P_0(\mathbf{a},\mathbf{b})}$$

For sequences of a given length pair class $\mathbf{a} \in S^m$, $\mathbf{b} \in S^n$, the probability of a sequence alignment $(\mathbf{a},\mathbf{b},\mathbf{u})$ will be defined as follows:

$$P_A(\mathbf{a},\mathbf{b},\mathbf{u}) \;=\; \frac{z^{S_A(\mathbf{a},\mathbf{b},\mathbf{u})}\; P_0(\mathbf{a},\mathbf{b})}{B(n,m)} \qquad (9)$$

where $S_A(\mathbf{a},\mathbf{b},\mathbf{u})$ is the local alignment score of $(\mathbf{a},\mathbf{b},\mathbf{u})$ as defined by equation 1, and $B(n,m)$ is a length normalization term whose functions is to ensure that the probabilities of all sequence pairs of length pair class m,n sum to one. The definition of the alignment probability is natural if one interprets Smith-Waterman scores as log likelihood ratios, as suggested by Altschul (Altschul 1991). The length normalization term obviously has to satisfy:

$$B(n,m) \;=\; \sum_{\mathbf{a} \in \mathbf{S}^m, \mathbf{b} \in \mathbf{S}^n, \text{paths } \mathbf{u}} z^{S_A(\mathbf{a},\mathbf{b},\mathbf{u})}\; P_0(\mathbf{a},\mathbf{b}) \qquad (10)$$

We will show that the expression for $B(m,n)$ simplifies to

$$B(n,m) \;=\; \sum_{\text{paths } \mathbf{u}} z^{S_G(\mathbf{u})} \qquad (11)$$

where $S_G(\mathbf{u})$ is the gap score of alignment path $\mathbf{u}$ as defined by equation 2b, if the null-model satisfies the constraint imposed by equation 6. Let us first rewrite the expression for $B(m,n)$ as follows:

$$B(m,n) = \sum_{\text{paths } \mathbf{u}} z^{S_G(\mathbf{u})} \left[\sum_{\mathbf{a} \in \mathbf{S}^m, \mathbf{b} \in \mathbf{S}^n} z^{S_M(\mathbf{a},\mathbf{b},\mathbf{u})} P_0(\mathbf{a},\mathbf{b}) \right] \qquad (12)$$

where $S_M(\mathbf{a},\mathbf{b},\mathbf{u})$ is the match score of the sequence alignment $(\mathbf{a},\mathbf{b},\mathbf{u})$ as defined by equation 2a. The inner sum in the above expression can be rewritten as shown below, using the notation $v_1 v_2 \cdots v_{m-l}$, $w_1 w_2 \cdots w_{n-l}$ for the residues of sequences $\mathbf{a}$ and $\mathbf{b}$, which are not part of a matched residue pair defined by path $\mathbf{u}$:

$$\sum_{\mathbf{a} \in \mathbf{S}^m, \mathbf{b} \in \mathbf{S}^n} z^{S_M(\mathbf{a},\mathbf{b},\mathbf{u})}\; P_0(\mathbf{a},\mathbf{b})$$

$$= \underset{\mathbf{a} \in \mathbf{S}^m, \mathbf{b} \in \mathbf{S}^n}{sum} \prod_{k=1}^{m-l} p(a_{v_k}) \prod_{k=1}^{n-l} p(b_{w_k}) \prod_{k=1}^{l} p(a_{x_k}) p(b_{y_k}) z^{s(a_{x_k}, b_{y_k})}$$

$$= \left[\sum_{a \in \mathbf{S}} p(a) \right]^{m-l} \left[\sum_{b \in \mathbf{S}} p(b) \right]^{n-l} \left[\sum_{a \in \mathbf{S}, b \in \mathbf{S}} p(a) p(b) z^{s(a,b)} \right]^{l}$$

$$= 1\,.$$

Combining equations 8, 9, and 11, we obtain the following intuitively plausible expression for the $PSWscore$:

$$PSWscore(\mathbf{a},\mathbf{b}) \;=\; \frac{\sum_{\text{paths } \mathbf{u}} z^{S_A(\mathbf{a},\mathbf{b},\mathbf{u})}}{\sum_{\text{paths } \mathbf{u}} z^{S_G(\mathbf{u})}} \qquad (13)$$

Both sums in the log likelihood ratio can efficiently be computed by special cases of the forward algorithm used for computation of HMM scores (Krogh et al. 1994, Rabiner 1989). Numerical recipes are given in Figure 1.

Performance Evaluation of the PSW Algorithm

In order to compare the sensitivities of the PSW algorithm and the native Smith-Waterman algorithm, we performed parallel database searches on SWISS-PROT release 32 (Bairoch & Apweiler 1996) with prototype query sequences from known protein families and domains, including the globins, the hsp20 family, cytochrome C, the G protein-coupled receptor family, and SH2, SH3 as domain examples. The results of these tests are shown in Table 1.

$$PSWscore(\mathbf{a},\mathbf{b}) \;=\; \frac{\sum\limits_{\text{paths } \mathbf{u}} z^{S_A(\mathbf{a},\mathbf{b},\mathbf{u})}}{\sum\limits_{\text{paths } \mathbf{u}} z^{S_G(\mathbf{u})}}$$

Algorithm to compute: $\sum\limits_{\text{paths } \mathbf{u}} z^{S_A(\mathbf{a},\mathbf{b},\mathbf{u})}$

```
    $M_{1,1} \leftarrow z^{s(a_1,b_1)}$;
    $I_{1,1} \leftarrow 0$;
    $D_{1,1} \leftarrow 0$.
**for** $i \leftarrow 2$ **to** $m$:
    $M_{i,1} \leftarrow z^{s(a_i,b_1)}$;
    $I_{i,1} \leftarrow z^{-\alpha} z^{-\beta} M_{i-1,1} + z^{-\beta} I_{i-1,1}$;
    $D_{i,1} \leftarrow 0$.
**for** $j \leftarrow 2$ **to** $n$:
    $M_{1,j} \leftarrow z^{s(a_1,b_j)}$;
    $I_{1,j} \leftarrow 0$;
    $D_{1,j} \leftarrow z^{-\alpha} z^{-\beta} M_{1,j-1} + z^{-\beta} D_{1,j-1}$.
**for** $i \leftarrow 2$ **to** $m$;  $j \leftarrow 2$ **to** $n$:
    $M_{i,j} \leftarrow z^{s(a_i,b_j)} \left[ 1 + M_{i-1,j-1} + I_{i-1,j-1} + D_{i-1,j-1} \right]$;
    $I_{i,j} \leftarrow z^{-\beta} \left[ z^{-\alpha} M_{i-1,j} + I_{i-1,j} + z^{-\alpha} D_{i-1,j} \right]$;
    $D_{i,j} \leftarrow z^{-\beta} \left[ z^{-\alpha} M_{i,j-1} + D_{i,j-1} \right]$.
```

$$\sum_{\text{paths } \mathbf{u}} z^{S_A(\mathbf{a},\mathbf{b},\mathbf{u})} \;=\; \sum_{1 \le i \le m, 1 \le j \le n} M_{i,j}$$

Algorithm to compute: $\sum\limits_{\text{paths } \mathbf{u}} z^{S_G(\mathbf{u})}$

```
    $M_{1,1} \leftarrow 1$;
    $I_{1,1} \leftarrow 0$;
    $D_{1,1} \leftarrow 0$.
**for** $i \leftarrow 2$ **to** $m$:
    $M_{i,1} \leftarrow 1$;
    $I_{i,1} \leftarrow z^{-\alpha} z^{-\beta} + z^{-\beta} I_{i-1,1}$;
    $D_{i,1} \leftarrow 0$.
**for** $j \leftarrow 2$ **to** $n$:
    $M_{1,j} \leftarrow 1$;
    $I_{1,j} \leftarrow 0$;
    $D_{1,j} \leftarrow z^{-\alpha} z^{-\beta} + z^{-\beta} D_{1,j-1}$.
**for** $i \leftarrow 2$ **to** $m$,  $j \leftarrow 2$ **to** $n$:
    $M_{i,j} \leftarrow 1 + M_{i-1,j-1} + I_{i-1,j-1} + D_{i-1,j-1}$;
    $I_{i,j} \leftarrow z^{-\beta} \left[ z^{-\alpha} M_{i-1,j} + I_{i-1,j} + z^{-\alpha} D_{i-1,j} \right]$;
    $D_{i,j} \leftarrow z^{-\beta} \left[ z^{-\alpha} M_{i,j-1} + D_{i,j-1} \right]$.
```

$$\sum_{\text{paths } \mathbf{u}} z^{S_G(\mathbf{u})} \;=\; \sum_{1 \le i \le m, 1 \le j \le n} M_{i,j}$$

Figure 1: Algorithm to compute the *PSWscore* for two sequences $\mathbf{a} \in \mathbf{S}^m$, $\mathbf{b} \in \mathbf{S}^n$.

With a single query sequence, one typically finds between 50% and 90% of all true positives in these examples. The number of missed sequences depends on the divergence of the sequence family, as well as on stringency of the significance criterion applied. For cross-standardization, we define equivalent stringency levels by a fixed number of false positives accepted. Such an approach may not be totally representative of a real application where one does not know the status of the sequences in advance, but it constitutes the only way of ensuring a fair comparison between methods that express similarity scores on genuinely different scales.

For each family and domain, we compared the results obtained with three different database search programs: BLAST (Altschul et al. 1990), SSEARCH (Pearson 1991) implementing the native Smith-Waterman algorithm, and an experimental program implementing PSW. BLAST was used with the default substitution matrix BLOSUM 62. The scoring system used for SSEARCH and PSW consisted of a $10Log_{10}$-scaled BLOSUM 45 matrix and a gap weighting function $w(k)=8+4k$. The BLOSUM 45 was chosen for the latter two methods because it performed better than the default BLOSUM 62 matrix used by SSEARCH. The gap weighting function chosen corresponds to the default settings in SSEARCH.

The results shown in Table 1 document a robust trend of increased sensitivity of the probabilistic Smith-Waterman algorithm over the native version. The gain in performance is particularly impressive for the globin and HSP20 families where the sequence similarities often extend over the entire length of the proteins compared. Note that in these examples, one would have to accept a 10 times higher false positive rate with the native Smith-Waterman algorithm in order to find as many true members as with the PSW algorithm. The increased sensitivity is less evident, or even debatable, in the domain examples SH2 and SH3, where a superior performance of PSW over SSEARCH is only observed at the highest selectivity level. These results, if conformed by further experiments, indicate that PSW is the currently most sensitive method to detect weak similarities between two protein sequences.

Experiment		# of true positives missed at *N* false positives accepted			
Family	**Method**	**N=0**	**N=10**	**N=100**	**N=1000**
Globin	PSW	36	17	10	4
P02023	Smith-Waterman	59	39	19	8
674 members	BLAST	95	77	62	50
HSP20	PSW	18	8	2	1
P02515	Smith-Waterman	33	16	11	3
129 members	BLAST	43	30	19	14
Cytochrome C	PSW	132	113	87	71
P00001	Smith-Waterman	131	131	112	82
243 members	BLAST	129	118	103	80
GPC receptors	PSW	76	73	65	52
P30542	Smith-Waterman	86	74	67	48
497 members	BLAST	108	95	77	60
SH2-domain	PSW	14	9	0	0
P12931 (150-247)	Smith-Waterman	17	5	0	0
119 members	BLAST	21	20	13	1
SH3-domain	PSW	20	14	1	1
P12931 (83-144)	Smith-Waterman	25	10	6	1
131 members	BLAST	30	22	9	2

Table 1. Comparative performance evaluation of PSW, Smith-Waterman, and BLAST algorithms.

Data and methods: Database searches were performed on SWISS-PROT release 32 (Bairoch & Apweiler, 1996) containing 49340 entries. The following entries were used as query sequences: P02023: human β-globin, P02515: *Drosophila m.* heat shock protein 22, P00001: human Cytochrome C, P30542: human adenosine A1 receptor, P12931: human proto-oncoprotein Src. The classification of the sequence families and domains is based on PROSITE release 32 (Bairoch, Bucher, & Hofmann 1996). The list of G-protein-coupled (GPC) receptors was compiled from the four PROSITE entries PS00237, PS00649, PS00979, PS00238. Five additional putative proteins from *C. elegans* corresponding to SWISS-PROT entries P41590, P34488, P46564, P46568, P46567, were also considered true positives. Blast searches were performed with the default parameter settings (BLOSUM 62 matrix). PSW and Smith-Waterman searches were performed with a BLOSUM 45 matrix and the default gap weights of the program SSEARCH (*see* text).

Discussion

We have presented preliminary evidence that current methods for pairwise sequence alignments can be improved by interpreting an alignment scoring system as a probabilistic model, applying concepts and methods that have been introduced to molecular sequence analysis in the context of hidden Markov models. The fact that we observe an increase of sensitivity of the PSW algorithm over the native Smith-Waterman algorithm using a scoring system that has only been optimized for the latter one, lends further credibility to the this conclusion. It appears likely that an additional improvement of the performance of PSW can be achieved by fine-tuning the alignment parameters specifically for this method.

Increased sensitivity is not the only benefit resulting from a probabilistic interpretation of a sequence alignment method. Another advantage is the availability of simple statistical tests, *e.g.* Milosavljević's algorithmic significance test (Milosavljević & Jurka 1993), to assess the significance of database search scores. Moreover, the fact that PSW scores are scaled as absolute log likelihood ratios of two probabilities, and thus are not influenced by sequence length effects or the relative log-scale of the scoring system, facilitates the optimization of the gap weighting parameters. The probabilistic framework also suggests methods to correct for residue composition effects, both regional and global, via comparison to null-models that retain certain statistical properties of the analyzed sequences.

It needs to be mentioned that we are not the first to formulate a probabilistic sequence alignment method. Maximum likelihood approaches have been applied to various problems in the field of molecular evolution, including the estimation of the divergence time for two DNA sequences (Bishop & Thompson 1986), the optimization of the parameters of a likelihood model of sequence evolution (Thorne, Kishino, & Felsenstein 1991, 1992), and the assessment of the reliability of molecular sequence alignments (Thorne & Churchill 1995). The computations performed in these applications rely on the same principles as the PSW algorithm in that they achieve the summation of probabilities over an astronomically large number of sequence alignments by means of an efficient recursive procedure proven to yield the identical result. McCaskill's method (McCaskill 1990) to compute global and restricted partition functions for RNA secondary structures falls also into the same class of algorithms and is conceptually related to probability through the notion of entropy.

Despite obvious parallels to earlier work, our presentation of a probabilistic sequence alignment algorithm includes several important innovations. One is the nontrivial extension from global to local alignment mode. Another one is its application to a new problem, sequence similarity search. The question addressed by the PSW algorithm is fundamentally different from those addressed by maximum likehood methods applied in molecular evolution. The goal of our method is to reach a qualitative decision between two alternatives, presence or absence of significant local sequence similarity between two sequences, whereas previous maximum likelihood applications were aimed at quantitative estimation of evolutionary parameters such as divergence time or the ratio of indel versus substitution mutation frequencies. This difference explains the presence of a null-model as integral part of our approach, as well as the consistent absence of such a null-model in previous work.

Another important difference in our formulation of the probabilistic alignment problem compared to previous ones is the absence of an underlying model of sequence evolution. By dissociating the alignment concepts from a mandatory phylogenetic interpretation we remove an important conceptual obstacle to its applicability in contexts where sequence similarity does not necessarily mean homology. Therefore, the theoretical framework of the PSW algorithm can accommodate amino acid substitution matrices derived from an evolutionary model (*e.g.* Dayhoff, Schwartz, & Orcutt 1978) as well as indel frequency models based on structural superpositions (Pascarella & Argos 1992).

Sequence similarity search is of course not the only possible application of a probabilistic alignment method generalized to local alignment mode. Most questions addressed by previous applications of maximum likelihood, or of the mathematically related partition function calculations, can also be formulated with regard to the local alignment problem. Logical combinations of this kind include the design of an expectation-maximization (EM) algorithm to optimize the parameters of a local alignment scoring system for distantly related protein sequences, as well as the development of a dot matrix method reflecting residue association probabilities rather than local segment pair similarities.

Implementation Notes

The results shown in this paper were generated with an experimental program implementing a space-efficient version of the algorithms shown in Fig. 1, making use of the recursion introduced by Gotoh (Gotoh 1982).

The program, which is written in FORTRAN 77 and runs on various UNIX platforms, accepts as command line parameters a query sequence, a substitution matrix, two gap probability parameters, a maximum length of database sequences to be processed, and a cut-off value for PSW scores to be reported in the output. The sequence database is read from the standard input. The program starts by computing the sequence-independent quantity $\sum_{\text{paths}\,\mathbf{u}} z^{S_G(\mathbf{a},\mathbf{b},\mathbf{u})}$ for all relevant length pair combinations, making a single path through the algorithm shown on the right side of Figure 1, and then processes the sequence database sequentially by applying the algorithm shown on the left side of Figure 1 to each individual sequence.

Since no effort has been invested to optimize the speed of the program, no realistic time-efficiency comparisons between the PSW and the native Smith-Waterman algorithm can be made at this point. The current experimental program is at least ten times slower than Pearson's program SSEARCH (Pearson, 1991). Comparison with an HMM search program (Hughey & Krogh 1995) performing essentially the same type of computations, suggests that a time-efficiency trimmed PSW implementation sacrificing some arithmetic precision will only take between two and three times more time than publicly available Smith-Waterman database search programs.

The experimental PSW implementation used in this work is available upon request from the authors. We are in the process of preparing a faster public version of the program which will be made available from our ftp site (URL ftp://ulrec3.unil.ch).

Acknowledgements

This work was supported in part by grant 31-37687.93 from the Swiss National Science Foundation.

References

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

Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

Bairoch, A., and Apweiler, R. 1996. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucl. Acids Res.* 24:21-25.

Bairoch, A.; Bucher, P.; and Hofmann, K. 1996. The PROSITE database, its status in 1995. *Nucl. Acids Res.* 24:189-196.

Baldi, P.; Chauvin, Y.; Hunkapiller, T; and McClure, M.A. 1994. Hidden Markov models of biological primary sequence information. *Proc. Natl. Acad. Sci. USA* 91:1059-1063.

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

Dayhoff, M.O.; Schwartz, R.M.; and Orcutt, B.C. 1978. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure* (Dayhoff, M.O., ed.) vol 5, suppl. 3, pp. 345-352. Washington DC: National Biomedical Research Foundation.

Gotoh, O. 1982. An improved algorithm for matching biological sequences. *J. Mol. Biol.* 162:705-708.

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

Henikoff, S., and Henikoff, J.G. 1993. Performance evaluation of amino acid substitution matrices. *Proteins* 17:49-61.

Hughey, R., and Krogh, A. 1995. SAM: Sequence alignment and modeling software system. Technical Report, USCS-CRL-95-7, University of California, Santa Cruz.

Karlin, S., and Altschul, S.F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* 87:2264-2268.

Krogh, A.; Brown, M.; Mian, I.S.; Sjölander, K.; and Haussler, D. 1994. Hidden Markov models in computational biology. *J. Mol. Biol.* 235:1501-1531.

McCaskill, J.S. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 29:1105-1119.

Milosavljević, A., and Jurka, J. 1993. Discovering simple DNA sequences by the algorithmic significance method. *CABIOS* 9:407-411.

Pascarella, S., and Argos, P. (1992). Analysis of insertions/deletions in protein structures. *J. Mol. Biol.* 224:461-471.

Pearson, W.R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* 183:63-98.

Pearson, W.R. 1991. Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* 11:635-650.

Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77:257-286.

Sellers, P.H. 1974. On the theory and computation of evolutionary distances. *SIAM J. Appl. Math* 26:787-793.

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

Smith, T.F.; Waterman, M.S.; and Fitch, W.M. 1981. Comparative biosequence metrics. *J. Mol. Evol.* 18:38-46.

Thorne, J.L., and Churchill, G.A. 1995. Estimation and reliability of molecular sequence alignments. *Biometrics* 51:100-113.

Thorne, J.L.; Kishino, H.; and Felsenstein, J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* 33:114-124.

Thorne, J.L.; Kishino, H.; and Felsenstein, J. 1992. Inching toward reality: An improved likelihood model of sequence evolution. *J. Mol. Evol.* 34:3-16.