Adaptive algorithm of automated annotation

A. M. Leontovich, L. I. Brodsky, V. A. Drachev and V. K. Nikolaev

Belozersky Institute of Physico-Chemical Biology, Moscow State University, Moscow 119899, Russia and Quark Biotech Inc., 10265 Carnegie Avenue, Cleveland, OH 44106, USA

Received on July 7, 2001; revised on November 27, 2001; December 23, 2001; accepted on January 8, 2002

ABSTRACT

Motivation: It is common knowledge that the avalanche of data arriving from the sequencing projects cannot be annotated either experimentally or manually by experts. The need for a reliable and convenient tool for automated sequence annotation is broadly recognized.

Results: Here, we describe the Adaptive Algorithm of Automated Annotation (A4) based on a statistical approach to this problem. The mathematical model relates a set of homologous sequences and descriptions of their functional properties, and calculates the probabilities of transferring a sequence description onto its homologue.

The proposed model is adaptive, its parameters (distribution characteristics, transference probabilities, thresholds, etc.) are dynamic, i.e. are generated individually for the sequences and various functional properties (words of the description).

The proposed technique significantly outperforms the widely used test for frequency threshold, which is a special case of our model realized for the simplest set of parameters.

The prediction technique has been realized as a computer program and tested on a random sequence sampling from SWISS-PROT.

Availability: The automated annotation program based on the proposed algorithm is available through the Web browser at http://www.genebee.msu.su/services/annot/basic.html

Contact: nik@genebee.msu.su

1 INTRODUCTION

The genetic flow of data arriving from sequencing projects keeps growing. These data become valuable only after their annotation, or determining the function and structure of the sequences. The revealed properties (the sequence function, structure, etc.) are recorded in sequence databases such as SWISS-PROT (Bairoch and Apweiler, 2000), GenBank (Benson et al., 2000), PDB (Sussman et al., 1999) as their descriptions. These descriptions are divided into several fields. Specifically, SWISS-PROT descriptions include KW (Keywords), DE (Description), FT (Feature Table), etc. fields. The KW and DE fields describe the sequence as a whole, while words from the FT section apply to its individual sites.

The impossibility of annotating such data volumes either experimentally or manually using experts is generally accepted. The demand for developing a reliable and convenient tool for sequence automatic annotation was declared many times. At the same time, the results of such programs can not be absolutely reliable; the obtained functional and structural properties of the proteins can be used to select the pathway for much more lengthy, laborious, and expensive experiments with the real protein.

Now there are a considerable number of automated annotation systems (Gaasterland and Sensen, 1996; Frishman and Mewes, 1997; Harris, 1997; Bailey et al., 1998; Andrade et al., 1999). These systems include the subsystems of data and task control, multilevel object-oriented specifications (ontology), linguistic subsystems, and advanced graphical user interface—ample set of tools providing for multichannel data processing, etc. However, the key tools of such systems are the programs predicting the functional properties of a given sequence from analysis of homologous sequences with known properties.

In a broad sense, all annotation methods in use are based (directly or indirectly) on the correlation between similarity (homology) of the primary structures and similarity between functions (and, hence, the bank descriptions). In some articles (Devos and Valencia, 2000; Thornton et al., 2000; Wilson et al., 2000), such a correlation is examined, including how it is dependent on functional and structural properties of sequences (KW, Enzyme Commission classification, etc.).

The following approach is used for transference of functional properties to an annotated sequence (Andrade et al., 1999). The broad cellular function (Tamames et
al., 1998) is determined for the query sequence family by analyzing the set homologous to the query and using the keywords frequency as a test. In the case of pronounced homology (clear case) specific functional properties of the bank sequences are transferred to the annotated one.

Andrade (1999) introduces an empirical function estimating the likelihood of transferring a bank description word onto an annotated homologous sequence. Depending on the values of this function, the words of the description are assigned the reliability score of their presence in the annotation and the region of their application.

The pitfalls of the existing approaches are well known (Bork and Bairoch, 1996; Bork and Koonin, 1998). In our opinion, such drawbacks include the absence of a strict mathematical approach to the problem, and the forcedly static character of the parameters underlying the predictions.

Here, we propose such a strict mathematical (statistical) approach to this problem. The proposed mathematical model relates a set of mutually homologous sequences and their functional descriptions.

The central element of our model is a ‘transference probabilities’ from a sequence description onto its homologue. These probabilities are implemented into the model because of the idea that all sequences are results of an evolutionary process that goes through stochastic mutations and indels in the set of ancestor sequences. It follows that the transference of functional properties from sequence to sequence according to the phylogenetic tree of the process goes by probabilistic way. It seems naïve to base the algorithm on such simplified biological views. However, a mathematical model constructed in this manner does give good predictions.

In existing algorithms of sequence annotation, parameters of the procedure do not depend on the sequence that is annotated, and on functional properties that we want to predict. These algorithms are ‘static’. They are not adapted to the sequence under annotation or to its functional class. However, the ‘world’ of proteins is not homogeneous (Devos and Valencia, 2000). It means that the rate of evolutionary changes of primary sequences (mutations and indels) can differ from the rate of evolutionary changes of functional properties of the protein. The difference of rates (that refers to the relative importance of the sequence regions and active sites) can vary from sequence to sequence, and can vary much more from one functional group of proteins to another.

In contrast to the ‘static parameters’ approach, our algorithm is an ‘adaptive’ one. Its parameters—distribution characteristics, transference probabilities, and thresholds, etc. are dynamic, i.e. are generated individually for various sequences (and even their regions) and functional properties.

As already mentioned, words from KW and DE fields describe a SWISS-PROT sequence as a whole. However, there are grounds to believe that they apply to specific regions of these sequences, domains (Corpet et al., 2000). Our algorithm not only predicts every particular word in the description, but also points to a presumable region in the sequence described by the word.

Hence, the algorithm could be applied not only to the sequences with unknown description, but also to already annotated bank sequences—since the resulting predictions could refine the description by delimiting the regions described by the predicted words. Of course, such predictions should not be considered as absolutely reliable without biological validation.

The proposed algorithm is applicable to word prediction for all field types. However, prediction of FT words is a particular case (the prediction considers the positions of the sequence corresponding to a given word ω) and it is not considered below.

The algorithm was realized as a computer program and tested on a random sequence sampling from SWISS-PROT.

2 DESCRIPTION OF THE ALGORITHM

Let $KW(\pi)$, $DE(\pi)$ denote KW, DE description fields (set of words) of sequence $\pi$ correspondingly. Hereafter we will refer to the prediction of KW field descriptions.

Below we describe the simplest variant of the algorithm omitting some dispensable technical details.

Suppose we have query sequence $\pi_0$ with an unknown description. We want to restore (predict) its description using the available descriptions of relatively similar (homologous) sequences.

The algorithm consists of several stages. Figure 1 presents its framework.
Stage 1—forming the initial set of similar sequences and their primary motifs

At the first stage, we select a bank of sequences with available description fields. The sequences homologous to the query sequence are found: all motifs (and local alignments (Altschul et al., 1990) between the query sequence and each bank sequence by some procedure (BLAST, Altschul et al., 1990; FASTA, Pearson and Lipman, 1988; Smith and Waterman, 1981; DotHelix, Leontovich et al., 1993; etc.) are generated and motifs with the similarity above a certain threshold are selected (let us be reminded that under the motif between two sequences we mean a pair of homologous fragments of the same length that were taken from these sequences). This results in a set of primary motifs; the corresponding bank sequences will be referred to as similar ones.

Stage 2—cross motifs preparation

The cross motifs are found between the established similar sequences. For this, all pairs of primary motifs \( \mu_i \) and \( \mu_j \) are looked through. Since motifs \( \mu_i \) and \( \mu_j \) are projected on the same query sequence \( \pi_0 \), such a pair of motifs naturally creates a shift in the corresponding similar sequences. All motifs for sequences \( \pi_i \) and \( \pi_j \) for this shift and projecting on the (slightly extended) fragments of sequences \( \pi_i \) and \( \pi_j \) included in the corresponding primary motifs \( \mu_i \) and \( \mu_j \) are considered as the cross motifs.

Stage 3—the compiling of a list of forecast words

A set of all words included in descriptions of the similar sequences \( \pi_1, \ldots, \pi_k \)—a list of forecast words—is generated. These words are used for the prediction in our algorithm.

Let the total number of words included in the similar sequences \( \pi_1, \ldots, \pi_k \) descriptions equal \( N \). Let \( N_{t+}, N_{t-}, N_{f+}, \) and \( N_{f-} \) denote the number of true positive (correctly predicted), true negative (correctly unpredicted), false positive (erroneously predicted), and false negative (erroneously unpredicted) predictions, respectively. Then \( N = N_{t+} + N_{t-} + N_{f+} + N_{f-} \) and \( N_{t+} + N_{t-} \) is equal to the number of words should be included in the KW(\( \pi_0 \)).

The following prediction errors can be determined from these values: first and second kind errors \( P^1, P^2 \), and proportion of false predictions \( P^f \):

\[
P^1 = \frac{N_{f-}}{N_{t+} + N_{f-}}, \quad P^2 = \frac{N_{f+}}{N_{t+} + N_{f+}}, \quad P^f = \frac{N_{f+}}{N_{t+} + N_{f+}}.
\]

The values \( 1 - P^1 = \frac{N_{t+}}{N_{t+} + N_{f-}}, \) \( 1 - P^2 = \frac{N_{t-}}{N_{t+} + N_{f+}}, \) and \( 1 - P^f = \frac{N_{t+}}{N_{t+} + N_{f+}} \) are called sensitivity, selectivity, and specificity, respectively (Baldi et al., 2000).

Note that the second kind error \( P^2 \) is not very sensible (hence, it is not used to evaluate the prediction quality, see Section 3). Indeed, if the number of similar sequences involved in the prediction is increased, the total number of predictions \( N \) increases. Most likely \( N_{t+}, N_{t-}, \) and \( N_{f+} \) values do not significantly change while \( N_{f-} \) sharply increases. Hence, the second kind error formally decreases but the prediction quality is clearly not improved. Nevertheless, the second kind error \( P^2 \) is used to determine the threshold for statistics Formula (4).

Let \( \omega \) be one of predicted words.

Stage 4—the detection of regions

Each such region can be defined as a connected component (with certain extension) of a union of fragments of the initial sequence included in the primary motifs corresponding to the similar sequences with this word \( \omega \) in their description. Several regions can correspond to one word. Thus, the total number of predictions can exceed the number of forecast words.

For simplicity, we assume that a single region corresponds to the word \( \omega \); this region matches the whole initial sequence.

Stage 5—the transference probabilities calculation

The transference probabilities of the words of descriptions are calculated on the basis of cross motifs. These probabilities show how stable functional properties of the given word are. These probabilities were calculated starting from the group of similar sequences that includes the given one.

We assume that each pair of sequences \( \pi_1 \) and \( \pi_2 \) and each word \( \omega \) corresponds to four transference probabilities — conditional probabilities

\[
\begin{align*}
    p_{1|\omega} & = p_{1|\omega} (\omega; \pi_2/\pi_1), \\
    p_{2|\omega} & = p_{2|\omega} (\omega; \pi_2/\pi_1), \\
    p_{1|\omega} & = p_{1|\omega} (\omega; \pi_2/\pi_1), \\
    p_{2|\omega} & = p_{2|\omega} (\omega; \pi_2/\pi_1),
\end{align*}
\]

where: \( p_{1|\omega} (\omega; \pi_2/\pi_1) \) is the probability of word \( \omega \) being found in \( \pi_2 \) on condition that it is present in \( \pi_1 \) description; \( p_{2|\omega} (\omega; \pi_2/\pi_1) \) is the probability of word \( \omega \) being found in \( \pi_2 \) on condition that it is not present in \( \pi_1 \) description;

\[
\begin{align*}
    p_{1|\omega} & = 1 - p_{1|\omega} (\omega), \\
    p_{2|\omega} & = 1 - p_{2|\omega} (\omega).
\end{align*}
\]

Since we observe a single realization of the probabilistic distribution of words in the descriptions true transference probabilities cannot be calculated. However, certain natural assumptions on these probabilities can provide for their plausible estimates.

Here are these assumptions. First, the higher the true similarity, the higher should be the transference probability \( p_{1|\omega} \) and the lower should be the transference
probability $p_{t/f}(ω)$. Second, these transference probabilities (with constant true similarity between sequences $π_1$ and $π_2$ and constant word $ω$) do not depend much on sequences $π_1$ and $π_2$, as long as we remain in about the same region of the phylogenetic tree. Hence, we assume that the estimated transference probabilities depend only on the value $M$ of true similarity (with fixed word $ω$): $p_{t/f}(ω; M)$. Owing to the first assumption, this relationship between the transference probability and $M$ is monotone. Owing to the second assumption (locality), estimation of these values relies on the frequency of transferring word $ω$ from one sequence to another providing that these sequences are similar to sequence $π_2$.

From these propositions such a procedure follows (to be precise, estimation) transference probabilities $p_{t/f}(ω; M) = p_{t/f}(M)$ and $p_{t/f}(ω; M) = p_{t/f}(M)$.

The total set of similarity values is divided into a number of intervals so that the probabilities $p_{t/f}(M)$ and $p_{t/f}(M)$ are piecewise constant within these intervals, $p_{t/f}(M)$ and $p_{t/f}(M)$ functions monotonically increase and decrease, respectively, and the condition $p_{t/f}(M) ≥ p_{t/f}(M)$ is satisfied.

Suppose we will have interval $[M_1, M_2)$ (assuming $M_1$ belongs to the interval while $M_2$ does not). Let: $N^i_f([M_1, M_2])$ is the number of cross motifs with the similarity within the $[M_1, M_2)$ interval and word $ω$ true for both sequences; $N^i_f([M_1, M_2))$ is the number of cross motifs with the similarity within the $[M_1, M_2)$ interval and word $ω$ true for one sequence and false for another. By the same way $N^i_f([M_1, M_2))$, $N^i_f([M_1, M_2))$ are defined.

We assume:

$$p^i_f([M_1, M_2)) = \frac{N^i_f([M_1, M_2))}{N^i_f([M_1, M_2]) + N^i_f([M_1, M_2))})$$

$$p^i_f([M_1, M_2)) = \frac{N^i_f([M_1, M_2))}{N^i_f([M_1, M_2]) + N^i_f([M_1, M_2))}).$$

Hence, $p^i_f([M_1, M_2))$ and $p^i_f([M_1, M_2))$ values can be calculated from formula (1) for each $[M_1, M_2)$ interval; these values can be presented as sample values of $p_{t/f}(M)$ and $p_{t/f}(M)$ averaged by $M ∈ [M_1, M_2)$.

Furthermore, since we suppose that transference probabilities are ordered monotonically, the procedure of monotonous ordering of $p^i_f([M_1, M_2))$ and $p^i_f([M_1, M_2))$ values is applied.

In the case of $p_{t/f}(M)$, we will be building a system of intervals $\{[M^i, M^{i+1})\}$ satisfying the following conditions:

$$p^i_f([M^i, M^{i+1}) < p^i_f([M^{i+1}, M^{i+2})],$$

where

$$M^1 < M^2 < \ldots M^n < M^{n+1} = \infty.$$
where coefficients \( a_j = a(M_j) \), \( M_j \) is maximum similarity between primary motifs where \( j \)th sequence participates.

Values of these coefficients \( a_j \) are selected to minimize prediction errors. Namely, they are determined to minimize, by the gradient method, the probability of error \( \bar{P} = P^1 + P^2 \) assuming that \( \eta \) has normal distribution and \( \xi_j \) values from Formula (3) are independent.

The simple variant of the algorithm is based on the previous assumptions. The statistic that is used is the estimation of the ratio of errors of the first and second kinds:

\[
\gamma = \frac{\Phi\left(\frac{\eta-M_j \eta}{\sqrt{D_j \eta}}\right)}{1 - \Phi\left(\frac{\eta-M_j \eta}{\sqrt{D_j \eta}}\right)}, \quad (4)
\]

where \( \Phi \) is the standard normal distribution function:

\[
? (v) = \int_{-\infty}^{v} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx,
\]

\( M_j \eta = \sum_{j} a_j p_{ij}(M_j), \)

\( D_j \eta = \sum_{j} a_j^2 p_{ij}(M_j)|[1 - p_{ij}(M_j)], \ldots \quad (5)\)

Since for transference probabilities the following inequality \( p_{ij}(M) > p_{ij}(M) \) is true, then \( M_j \eta < M_j \eta \) is also true.

Let us select a certain threshold \( \gamma^{th} \) value. The prediction is positive (providing that \( \omega \in KW(\pi_0) \)) for \( \gamma \geq \gamma^{th} \) and negative for \( \gamma < \gamma^{th} \).

The value of the threshold has to be chosen in such a way that the sum of errors will be minimized. The ‘theoretical’ value \( \gamma^{th} \) is calculated from the condition that the sum of errors of the first and second kinds (\( P^1 + P^2 \)) is minimal. From the normality of \( \eta \) and mutual independence of \( \xi_j \) variables, it follows that \( \gamma^{th} = 1.0 \). However, it is better to calculate this threshold by direct minimization of \( P^1 + P^2 \) according to the set of test sequences.

**Stage 7—a calculation of the reliability score**

The higher the statistic \( \gamma \) value, the more reliable the prediction. Our algorithm gives an estimation of the reliability score. This score is evaluated in advance using a certain set of test samples. We take a set and consider only the predictions with the statistic exceeding \( \gamma \) for each \( \gamma \) value; the proportion of correct predictions is calculated for them. This proportion is the mean of the reliability scores for the statistic exceeding \( \gamma \). Monotonization of this proportion, considered as a function of \( \gamma \), yields a function which can be considered as evaluation of the reliability score for a given value of the statistic. This function is calculated and included in the prediction.

**Other variants of the prediction.** There are different variants of the algorithm. The described simplest variant relies on \( \gamma \) statistic Formula (4)—estimation of the ratio between first and second kind errors. Other statistics can evaluate the proportion between either first kind error and the fraction of false predictions, or the number of true unpredicted and false predicted words in addition to this estimate. Consequently, these evaluations can rely on different approaches to the evaluation of the distribution function of the major statistics \( \eta \). One may assume normal distribution of \( \eta \). We can be using the experimental distribution functions of \( \tilde{P}_{ij}(i) \) statistics, determined from certain sample series with known descriptions of the query sequences. Another option would be the model distribution functions generated by computer simulation with \( \tilde{P}_{ij}(i) \) and \( \tilde{P}_{ij}(i) \) probabilities (in these experiments \( \xi_j \) values are played out independently). Hence, three tests and three distribution functions yield nine different statistic variants (see Figure 2).

**The prediction for the degenerative case.** Thus, the prediction was made for non-degenerate cases. For degenerate cases when transference probabilities cannot be determined, a rough prediction relies on the proportion of similar sequences with word \( \omega \) in their descriptions among all similar sequences. This corresponds to the prediction case when all \( a_j = \text{const.} \)

**3 RESULTS**

Five hundred probe sequences were selected at random from the SWISS-PROT bank. The prediction was carried out for these sequences using our algorithm, and the predictions obtained were compared with the true (SWISS-PROT) ones.

---

**Table 1. Total prediction quality for amino acids sequences**

<table>
<thead>
<tr>
<th></th>
<th>KW</th>
<th>DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>p1 + p2</td>
<td>p1</td>
</tr>
<tr>
<td>Freq</td>
<td>29.8</td>
<td>21.9</td>
</tr>
<tr>
<td>4A</td>
<td>11.5</td>
<td>9.6</td>
</tr>
</tbody>
</table>

The A4 row represents the prediction quality obtained by the A4 algorithm for optimal \( \gamma^{th} \) threshold. The Freq row represents the prediction quality obtained from the frequency of a word occurrence. Just this prediction will be made in the degenerate case. It approximately corresponds to the method used by Andrade et al. (1999).

The data for words from KW and DE are given individually. Here \( P^1 \) is the first kind error and \( P^2 \) is the fraction of predicted false words in all predicted ones.
Adaptive algorithm of automated annotation

Fig. 2. The relationship between the prediction quality and selected $\gamma^\text{th}$ threshold. All predicted words are arranged by the corresponding statistic values so that each word and the corresponding statistic value have the rated number in the above arranged data array. Abscissa: rating; ordinate: prediction quality for the selected $\gamma^\text{th}$ threshold of the statistic corresponding to this rating. The dotted line represents the statistics using normal distribution and ratio between fractions of I and II kind errors (corresponds to Formula (4)). This statistics gives the best results. The thin lines represent all other variants of statistics.

Fig. 3. The probability of a word prediction as a function of its rating. Abscissa: rating of the predicted words; ordinate: the reliability score. Fine and heavy lines indicate reliability score for the words from KW and DE fields, respectively.

The main testing results are presented in Table 1, which is the total quality of the prediction by our algorithm for statistics $\gamma$. The threshold $\gamma^\text{th}$ was calculated through the minimisation of $P^1 + P^f$ statistics.

The query sequence was excluded from the list of similar sequences. Leaving it in the list, naturally, improves the evaluation of the prediction (10 and 17% instead of 21 and 29%, respectively).

To check the reliability of test results, the set of 500 sequences was randomly sliced into 5 subsets of equal sizes. The quality of the prediction was calculated for every subset separately. These results show that Table 1 data is reliable. KW results vary in the interval $21 \pm 4$, and for DE—in the interval $29 \pm 3$.

The quality of the prediction is very dependent on the number of sequences with good similarity: the quality is better when we have more sequences with good similarity. In the case of small number of good similarities ($<50$), the quality of prediction becomes two times worse than the average quality.

Figure 2 shows the prediction quality (total errors $P^1 + P^f$) for KW words plotted against $\eta^\text{th}$ threshold value corresponding to different statistics. The prediction quality is plotted on the ordinate. In order to compare the prediction quality for different statistic types, the prediction rating with the statistic threshold are laid off as abscissa. The lower the plot, the higher the quality of prediction for the selected threshold. The minimum of the graph corresponds to the optimal threshold.

The figure demonstrates that the results have a weak dependence on statistic type.

Figure 3 plots the reliability score for words from KW and DE fields against its rating for statistic $\gamma$ Formula (4).
The same prediction groups can be seen here as in Figure 2. The most interesting middle group corresponds to the steepest part of the plot. Figure 4 represents results of the prediction for one of the test series.

One can see that errors tend to group together, rather than being evenly distributed.

The prediction quality is much worse for the nucleotide sequences (GenBank): \( P^1 = 18, P^f = 63, \) and \( P^f + P^1 = 81 \) for optimal selection of the threshold. We believe that such poor quality of the prediction is caused by the inadequate quality of descriptions in the GenBank.

4 DISCUSSION

Table 1, demonstrates quite good results of the algorithm. Nevertheless, there are many errors. Below, we discuss the reasons for these errors, and possible ways to reduce them.

Most errors are due to poor quality of the available bank descriptions. These include direct spelling errors, and descriptions of insufficient formalization: the same notion can often be described by different words and different abbreviations in different sequences. This is particularly true for GenBank. Apparently, they should be solved by composing lists of bad (unpredictable) words as well as dictionaries of synonyms and abbreviations; i.e. syntactic and semantic analysis should be carried out for the available bank descriptions. In addition, we, as well as some biology experts, believe that a predicted word can sometimes correctly describe a sequence property although it is absent from the actual description. (Hence, our algorithm can be carefully used to correct the available descriptions.)

Below, we present examples of such formally incorrect but basically true predictions for the SWISS-PROT bank.

**TMPC_TREPA**, words MEMBRANE and LIPOPROTEIN were not predicted in the field DE but were predicted in the KW field.

**C551_CHRVI**, true words C-551 and C551 were not predicted due to different spelling.

**IL10_PIG**, true word IL-10 was not predicted; however, another word, INTERLEUKIN-10 specifying the same notion has been predicted.

**HEM4_SYNP7**, true word UROPORPHYRINOGEND不同单词和不同缩略词在不同序列中的使用。
III as well as ‘false’ word UROPORPHYRIN-III have been predicted.

A significant number of errors are caused by the intrinsic shortcomings of the algorithm. These can be divided into principal ‘ideological’ limitations (related either to imperfect assumptions underlying the algorithm, i.e. limitations of purely statistic approach, or excessive simplification of these assumptions and, hence, their inaccuracy) and technical defects related to particular realisation of the ideology underlying the algorithm.

The limitation of the model is expressed in the fact that the model in no way accounts functionally important sequence regions responsible for its function. Hence, similarity with these regions are much more important as compared to similarities with the regions unrelated to the sequence functional properties. Use of sequence bank descriptions (specifying position of domains etc.) allows us to evaluate the functional importance of the revealed similarities. Clearly, the banks of patterns, profiles, fingerprints, and 3D structures can be useful here. Thus, the algorithm should be made multichannel to use many databases and search types.

Let us now consider limitations of the algorithm related to excessive simplification and inaccuracy of the underlying assumptions.

This is primarily applicable to the assumption on independence of $\xi_j$ values or, stated differently, to realization of the prediction by simple superposition of individual predictions corresponding to different similar sequences.

Clearly, this assumption cannot be true. For instance, values $\xi_i$ and $\xi_j$ can be mutually related if a powerful cross motif exists for similar sequences $i$ and $j$, since these most likely homologous sequences have similar descriptions.

Actually, the algorithm should account for the relationship between $\xi_j$ values, i.e. make the prediction from the whole $\xi_j$ column or even a set of columns (matrix) rather than from individual $\xi_j$.

Although the independence assumption is false, apparently this has no dramatic consequences, but solely decreases the statistic since a set of very similar sequences can be considered as a single one (which decreases the ‘effective’ number of similar sequences).

Separate calculation of the prediction for each word (i.e. independence of predictions for different words) is another ideological limitation. Actually, appearance of certain words is coordinated, and hence, joint prediction of certain word sets should be done. Iteration could the useful here (which is particularly applicable in the case of FT when predictions for neighbouring positions should be strongly mutually dependent).

It could be helpful to construct a phylogenetic tree of sequences that are similar to query sequence $\pi_0$. In this case, measuring similarity between sequences on the basis of such a phylogenetic tree should be more correct and the degree of similarity between very similar sequences should be measured by percent identity (Wilson et al., 2000).

At present the algorithm thresholds $\gamma^{th}$ are taken equally for all sequences, while they should be adapted for each sequence.

Let us now consider technical defects.

The assumption of $\eta$ normality clearly cannot be true. If for no other reason we know that the discreteness of $\eta$ is often limited to a small number of values. That is why we considered a statistic with a simulated distribution function of $\eta$ (which, however, has not yet improved the results). However, even in the cases with the model (or experimental) distribution functions, the $a_j$ coefficients were calculated on the basis of the normal distribution. This is clearly incorrect. Model distribution is better than normal distribution. However, it requires application of a more complex gradient algorithm.

A high number of technical errors is related to the calculation of transference probability.

Primarily, there is a biased error related to calculation of these probabilities only by cross motifs. While similar motifs should be searched for every similar sequence available and used to calculate the transference probability. However, this has not been implemented in the algorithm, in view of the significantly longer calculation required. For instance, if word $\omega$ is present in descriptions of all but one similar sequences, formally this is a degenerate case and transference probabilities are not determined. However, finding individual similarities for each similar sequence will most likely allow calculation of transference probability and positive prediction. (In the current version this results from rough prediction.)

Another inaccuracy in calculating transference probabilities results from the values overestimated ($p_{1/1}$ probabilities in the case) by monotontization.

The algorithm implements correction of transference probabilities. On the one hand, this is absolutely essential and significantly improves the prediction quality. This is chiefly because the $p_{1/1}$ is often overestimated at unity (a result of low statistics), which leads to numerous overpredictions. The correction algorithm has to be improved. The applied correction is too rough, which leads to miscalculation of expectations and variances Formula (5). We always assume $\alpha = 0.1$ in Formula (2), while making it adaptive, dependent on the motif similarity, could be particularly useful.

ACKNOWLEDGEMENTS

This work was supported by the Quark Biotech Inc., Russian Foundation for Basic Research, and the Ludwig Institute of Cancer Research. Our thanks are due to E.Feinstein and M.Shtutman for helpful discussion and valuable advice.
REFERENCES


