**Match-Box_server: a multiple sequence alignment tool placing emphasis on reliability**

**Eric Depiereux¹, Guy Baudoux, Pascal Briffeuil, Isabelle Reginster, Xavier De Bolle, Carla Vinals and Ernest Feytmans**

---

**Abstract**

**Motivation:** The Match-Box software comprises protein sequence alignment tools based on strict statistical thresholds of similarity between protein segments. The method circumvents the gap penalty requirement: gaps being the result of the alignment and not a governing parameter of the procedure. The reliable conserved regions outlined by Match-Box are particularly relevant for homology modeling of protein structures, prediction of essential residues for site-directed mutagenesis and oligonucleotide design for cloning homologous genes by polymerase chain reaction (PCR).

**Results:** The method produces reliable results, as assessed by tests performed on protein families of known structures and of low sequence similarity. A reliability score is computed in relation to a threshold of similarity progressively raised to extend the aligned regions to their maximal length, up to the significance limit of matching segments. The score obtained at each position is printed below the sequences and allows a discriminant reading of each aligned region.

**Availability:** Sequences may be submitted to a Web server at [http://www.fundp.ac.be/sciences/biologie/bms/matchbox_submit.html](http://www.fundp.ac.be/sciences/biologie/bms/matchbox_submit.html) or sent by e-mail to matchbox@biq.fundp.ac.be (help available by just mailing help).

**Contact:** E-mail: eric.depiereux@fundp.ac.be

---

**Introduction**

In spite of the tremendous increase in the performance of the different approaches devoted to protein structure modeling, it remains evident that when the sequence alignment is in error, the comparative model is guaranteed to be wrong (Mosimann et al., 1995). When a homology modeling procedure is considered, the quality of the target and template sequence alignment is mostly determinant. The precise definition of the predicted structurally conserved regions and variable regions is requested in order correctly to assign the fragments of conserved structure to the target sequence. The slightest misalignment can dramatically shift the conformation of the model (e.g. change the orientation of a β strand). It is also important to consider a variable region as a distinct structural element, different from the template, instead of simply inserting one or a few residues in a rather speculative position along the existing template structure (Vinals et al., 1993, 1995).

Also, identification of the residues involved in a protein activity by site-directed mutagenesis requires an accurate delineation of the regions conserved in several related sequences (Delforge et al., 1993; Bertrand et al., 1997). A misalignment or a 'ghost' conserved pattern may bring about negative results after substantial experimental attempts.

Numerous powerful sequence alignment methods have been implemented (Argos, 1987; Barton and Sternberg, 1987; Feng and Doolittle, 1987; Taylor, 1988; Subbiah and Harrison, 1989; Waterman and Jones, 1990; Schuler et al., 1991; Vingron and Argos, 1991), and most of them are available through the Internet (Lipman et al., 1989; Smith et al., 1990; Lawrence et al., 1993; Bailey and Elkan, 1994; Huang, 1994; Thompson et al., 1994). However, the confidence of the results is not always compatible with the complete automation of the methods and accessible CPU requirements. Several methods lead to different results, most of them being very sensitive to gap weighting. For the user confronted with contradictory and unclear outputs, a critical analysis of the results obtained remains speculative.

Match-Box is an interactive package for the simultaneous alignment of several sequences, based on strict statistical thresholds, and is essentially characterized by the circumvention of the gap penalty parameter and by the reliability of its predictions (Depiereux and Feytmans, 1991, 1992). The method has been applied successfully in different applications requiring reliable predictions of structurally or functionally conserved regions in several protein families (Delforge et al., 1993; Vinals et al., 1993; Tibor et al., 1994; De Bolle et al., 1995). However, the interactive handling of the package requires a self-investment that users easily accessing different push-and-go programs accept harshly. Much effort has therefore been devoted to design a subset of Match-Box yielding both automation and high reliability of the final results. This objective is achieved by a new algorithm of Scanning designed to analyse the signal/noise ratio for each pair of sequences and a given score matrix, and to fix optimal values to the statistical thresholds in order to conserve a high rate of confidence (selectivity) and to optimize the rate of power (sensitivity). Accessible through...
Table I. Proteins of known structures used to evaluate the performance of the method. (a) Number of residues in the common core. (b) Smallest pairwise Blast score. (c) Percentage of identities in the structure alignment. (d) Power and (e) confidence of the sequence alignment computed according to equation (6). (f) Reference of the family: FSSP database (Holm et al., 1992), (1) (Bordo et al., 1994), (2) (Newman et al., 1993), (3) (McKenzie and White, 1991), (4) (Pastore and Lesk, 1990), (5) (Greer, 1990), (6) (Siezen et al., 1991)

<table>
<thead>
<tr>
<th>PDB entry</th>
<th>Family name and protein</th>
<th>Sequence length</th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1crb</td>
<td>Fatty acid binding protein</td>
<td>136</td>
<td>115</td>
<td>78.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1opa</td>
<td>Cellular retinol-acid-binding protein</td>
<td>134</td>
<td>76</td>
<td>86.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2hmb</td>
<td>Fatty acid binding protein (holo form)</td>
<td>132</td>
<td>11%</td>
<td>FSSP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1lfb</td>
<td>Fatty acid binding protein (apo form 1)</td>
<td>131</td>
<td></td>
<td></td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>1hnd</td>
<td>Malate dehydrogenase with NADH</td>
<td>327</td>
<td>233</td>
<td>56.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1emd</td>
<td>Malate dehydrogenase</td>
<td>312</td>
<td>60</td>
<td>82.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1mdl</td>
<td>Malate dehydrogenase</td>
<td>314</td>
<td></td>
<td>15%</td>
<td>FSSP</td>
<td></td>
</tr>
<tr>
<td>4ndh</td>
<td>Cytoplasmic malate dehydrogenase</td>
<td>333</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1auj</td>
<td>Plastocyanin (apo form)</td>
<td>105</td>
<td>62</td>
<td>76.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1pcl</td>
<td>Plastocyanin (Cu2+, pH 6.0)</td>
<td>99</td>
<td>51</td>
<td>77.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1pza</td>
<td>Pseudoazurin (pH 8.0)</td>
<td>123</td>
<td>9%</td>
<td>FSSP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2plt</td>
<td>Plastocyanin</td>
<td>98</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1sdy</td>
<td>Superoxidismutase Cu, Zn superoxide dismutase</td>
<td>153</td>
<td>146</td>
<td>91.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1spd</td>
<td>Superoxidismutase Cu, Zn superoxide dismutase</td>
<td>153</td>
<td>379</td>
<td>98.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1sr5</td>
<td>Cu, Zn superoxide dismutase</td>
<td>154</td>
<td>44%</td>
<td>(1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2sod</td>
<td>Cu, Zn superoxide dismutase</td>
<td>151</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1mnp</td>
<td>Aspartic protease</td>
<td>361</td>
<td>291</td>
<td>42.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2apr</td>
<td>Acid protease</td>
<td>325</td>
<td>58</td>
<td>90.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3apr</td>
<td>Acid protease</td>
<td>323</td>
<td>9%</td>
<td>(2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4ape</td>
<td>Acid protease</td>
<td>330</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4cms</td>
<td>Chymosin B</td>
<td>323</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4lep</td>
<td>Pepsin</td>
<td>326</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1alc</td>
<td>Lysosome</td>
<td>122</td>
<td>113</td>
<td>89.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1t25</td>
<td>Lysosome</td>
<td>130</td>
<td>117</td>
<td>89.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2t26</td>
<td>Lysosome</td>
<td>129</td>
<td>25%</td>
<td>(3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2t27</td>
<td>Lysosome</td>
<td>129</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4pcya</td>
<td>Phycocyanin</td>
<td>162</td>
<td>160</td>
<td>67.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4pcyb</td>
<td>Phycocyanin</td>
<td>172</td>
<td>76</td>
<td>83.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4cmla</td>
<td>Phycocyanin</td>
<td>162</td>
<td>15%</td>
<td>(4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4cmlb</td>
<td>Phycocyanin</td>
<td>172</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1est</td>
<td>Serine protease</td>
<td>266</td>
<td>184</td>
<td>60.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1gtb</td>
<td>beta-Trypsin</td>
<td>223</td>
<td>62</td>
<td>82.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table I. Continued

<table>
<thead>
<tr>
<th>PDB entry</th>
<th>Family name and protein</th>
<th>Sequence length</th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1gdi</td>
<td>Trypsin</td>
<td>223</td>
<td>12%</td>
<td>(5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1ton</td>
<td>Tonn</td>
<td>235</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2chb</td>
<td>alpha Chymotrypsin a (toxylated)</td>
<td>245</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rp2</td>
<td>Rat mast cell protease II</td>
<td>224</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1cse</td>
<td>Subtilisin carlsberg complex with eglin-c</td>
<td>274</td>
<td>208</td>
<td>79.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2prk</td>
<td>Protenase k</td>
<td>279</td>
<td>158</td>
<td>79.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1tec</td>
<td>Thermilase complex with eglin-c</td>
<td>279</td>
<td>21%</td>
<td>(6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2sni</td>
<td>Subtilisin novo</td>
<td>275</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The sequences must also be retrieved by a search in a sequence database by the procedure BLAST (Altschul et al., 1990) according to the following criteria: a test family must include at least four proteins of known structure, sharing a reliability score.

Systems and methods

Programs were written in Fortran 77 and compiled on Indigo II RS4000. Silicone graphics, under Irix 5.3. The interface for the Web server is written in HTML 2.0 (with frames) interpretable by Netscape 2.0. Routines for the incoming and outgoing mailing management are written in Fortran 77, C and C-shell.

Structures used for testing the performances have been extracted from the Brookhaven database and are identified by their PDB entry code (Table I). Nine families have been selected partly in the FSSP database (Holm et al., 1992) and partly in specific publications (Greer, 1990; McKenzie and White, 1991; Siezen et al., 1991; Newman et al., 1993; Bordo et al., 1994) according to the following criteria: a test family must include at least four proteins of known structure, sharing a representative common core and low sequence similarity. The sequences must also be retrieved by a search in a sequence database by the procedure BLAST (Altschul et al., 1990) when one of them is defined as target. So, each test case is a representative example of a set of sequences obtained by users fetching related sequences in databases. The less related pair of sequences in each set is defined by the Smallest Pairwise Blast Score (SPBS).

The common core of each family is defined as a set of structurally conserved regions (SCRs), initiated by superimposing the amino acid of the major elements of the secondary structure of the less similar pair of sequences. Afterwards, each SCR is extended as far as the root mean square (RMS) computed between α-carbons on the whole SCR remains <1.8 Å. Then the other sequences of the family are progressively aligned and the common SCRs are limited to the regions for which all the pairwise comparisons produce a RMS < 1.8 Å. SCRs usually include the major elements of...
secondary structure, some turns and stretches of amino acids forming the active site. Structural superimposition and RMS computation have been performed using the INSIGHT© and HOMOLOGY© programs of BIOSYM Technologies (San Diego). The percentage of identities has been computed on the structural alignment as the number of residues conserved in each sequence at the same position, divided by the longest sequence length.

Alignment tests have been performed by Match-Box_server 1.1, using the default score matrix Blosum62 (Henikoff and Henikoff, 1992).

**Algorithms**

The basic algorithm of Match-Box is the scanning. It initiates three main applications: a significance test, a pairwise similarity analysis and the alignment itself. This algorithm has been described elsewhere (Depiereux and Feytmans, 1991, 1992). It is controlled by two parameters: the shift allowed between the segments to be compared and the threshold of similarity required to match them. The specificity developed in the Match-Box_server is the ability of the algorithm in self-determining optimal values for these parameters.

**Scanning**

A segment of consecutive residues is delineated in a sequence by a window of constant length \( w \) and compared with all the segments of same length of each other sequence. The cumulated distance between two segments defined by an initial and a running window at position \( p \) is easily defined as:

\[
D_p = \sum_{k=1}^{w} y_{ijk}
\]

where \( y_{ijk} \) is the distance between the pair of amino acids \( i \) and \( j \) found at position \( k \) of the window \( p \). A cut-off can be fixed on \( D_p \) to determine when segments are considered as significantly similar. Then, \( D_p < \) cut-off defines the matches between two segments and the global minimum of \( D_p \), if \( D_p < \) cut-off, defines the best match. The 210 scores corresponding to the pair of amino acids \( i \) and \( j \) are provided by the chosen score matrix.

Considering a correct match defined from a structural alignment, the scanning generates three typical results: (i) the global minimum of \( D_p \) corresponds to the correct match which may be detected by overall scanning on the whole sequence length; (ii) a local minimum of \( D_p \) corresponds to the correct match, which may be detected by a restrained scanning between anchor points; (iii) the global or a local minimum of \( D_p \) corresponds to an incorrect structural match and represents a random noise to be filtered and discarded.

**Test of significance**

This analysis allows one to determine whether the submitted set of sequences contains at least some sequences more similar than expected by chance or to detect non-related pairs of sequences among this set. The scanning is performed on the whole sequences and the frequency distribution of matches is computed for equidistant cut-offs covering the whole range of observed distances. Then the residues of each sequence are shuffled and the frequency distribution is computed again. Figure 1 illustrates graphically how frequency distributions of related sequences depart from randomness (1a) and unrelated sequences do not (1b).

For each class of distance, the statistic

\[
\chi = \frac{f_{obs} - f_{rnd}}{\sqrt{f_{rnd}}}
\]

is computed to test whether the number of matches observed in the sequences \( f_{obs} \) is significantly higher than the one expected by chance \( f_{rnd} \). The statistic is not squared to conserve the sign of the difference in order to discard non-relevant differences \( f_{rnd} > f_{obs} \). The maximum value of \( \chi \) observed for each pair of sequences is recorded. The lowest of these \( \chi \) values determines the less related pair of sequences. An overall frequency distribution is computed for all the
Analysis of similarity between the sequences

This analysis is designed to help the user in delineating relevant subsets of related sequences. For a given cut-off and a pair of sequences $i$ and $j$, this procedure computes the proportion $r_{ij}$ of initial windows that match at least once with a running window. Let $R$ be the similarity matrix obtained after scanning each pair of the $n$ sequences of the set, with $0 \leq r_{ij} \leq 1$. This matrix is not symmetrical: if a sequence $i$ is shorter than a sequence $j$, then the whole sequence $i$ can be very similar to a part of the sequence $j$, but only a part of the sequence $j$ can be similar to the sequence $i$, and $r_{ij} > r_{ji}$.

The matrix $R$ is then analysed by a principal coordinates analysis (Gower in Sneath and Sokal, 1973) in order to obtain a graphical representation of the sequence similarities in a plane. Before computation, $R$ is transformed into a symmetrical matrix by:

$$r_{ij} = \max(r_{ij}, r_{ji}) \forall ij$$  \hspace{1cm} (3)

When a short sequence $i$ is very similar to part of a sequence $j$, this convention enhances the similarity between $i$ and $j$ and masks the difference due to the part of $j$ not present in $i$. Figure 2 shows the representation of a set of sequences in the plane of the factors 2 and 3 of the principal coordinates analysis of the matrix $R$ obtained by scanning. The factor 1 is generally trivial and omitted. This example shows on the $x$-axis a group of four closely related sequences of superoxide dismutases ($1SDY, 1SRD, 1SOD, 1SPD$) well distinct from a group of aspartic proteases. On the $y$-axis, this last group appears more heterogeneous, with a group of acid proteinases ($4APE, 3APP$ and a less related $2APR$), a group with pepsin ($4PEP$) and chymosin B ($4CMS$), less related to another pepsin ($1MPP$).

**Alignment**

The goal of this third application of the scanning procedure is to collect reliable matches in order to propose an optimal simultaneous alignment of the whole set of sequences. Results of type (i) are selected in a first run allowing long shifts, with a low cut-off on $D_p$ [equation (1)] to exclude random noise. Results of type (ii) are added by three subsequent runs, with a progressively relaxed cut-off, to enlarge the aligned regions by a restrained scanning. Both represent success. Remaining results of type (iii) represent ‘false-positive’ matches (type I error) which decrease the confidence of the method. Correct matches remain undetected (type II error) when they do not correspond to a minimum of $D_p$ or when the minimum of $D_p$ remains higher than the most permissive cut-off. This error decreases the power of the method.

For each initial window, the global minimum of $D_p$ is retained as a potential match, and the random noise is discarded by the four filters described below.

(i) An initial window is retained only if it matches with a running window in every other sequence, according to the cut-off. A new algorithm determines the optimal cut-offs for each pair of sequences in the following way. For a given cut-off, let $BM$ be the total number of best matches and $M$ the total number of matches recorded in a pairwise scan. The signal quality is defined as

$$SQ = \frac{BM}{M}$$  \hspace{1cm} (4)

The procedure of scanning applied to each pair of sequences allows the relationship between the signal quality and the cut-off to be drawn. Figure 3a shows a typical sigmoidal decrease in $SQ$ in relation to the progressive increase in the cut-off. The probability of getting a correct match remains high as long as $SQ = 1$, then decreases dramatically. A first threshold is fixed at the appearance of the signal, a second at the first distance at which $SQ < 1$.

The number of best matches increases with the cut-off to reach a maximum value corresponding to the number of initial windows $(L_s - w + 1)$ where $L_s$ represents the length of the sequence $i$ (circles in Figure 3b). The estimated number of
**Match-Box server: multiple sequence alignment tool**

---

**C8**

**OJD**

**35**

**550**

**600**

**650**

**700**

**150**

**T**

**ooooooooooooooo**

**BM**

**Cutoff level 4**

**Distance between segments**

![Graph](https://academic.oup.com/bioinformatics/article-abstract/13/3/249/423156)

**Fig. 3.** Determination of the cut-offs for ISPY and 4APE. **a**. Variation of the signal quality computed according to expression (4) as a function of the distance between segments \( D_p \). (1) Dotted lines point out the first occurrence of best matches (cut-off level 1) and the last point before the increase in random noise (cut-off level 2). (2) Variation in the number of best matches (circles) and in the estimated correct matches computed according to expression (5) (diamonds) as a function of the distance between segments. Dotted lines point out the maximum number of estimated correct matches (cut-off level 3) and the maximum number of best matches (cut-off level 4). **b**. Cumulated number of matches.  

---

The number of estimated best matches as a function of the cut-off increases to reach a maximum and then decreases according to the increase in random noise (Figure 3b, diamonds). The third threshold is the distance corresponding to the maximum value observed for ECM, and the fourth is the shortest distance corresponding to the maximum value observed for BM.

The procedure is applied to each pair of sequences, each of them being characterized by four levels of cut-off for a given score matrix. The alignment is run in four loops and the cut-off level incremented at each loop. The comparisons between an initial window and running windows are performed with respect to the respective cut-off of the pair of sequences concerned.

**Implementation**

Algorithms are organized in two main programs: EXPLORE for pairwise similarity analysis and ALIGN for the optimal multiple alignment computation.

The EXPLORE procedure analyses the pairwise similarities between the whole set of sequences, and presents outputs to illustrate the test of significance performed for the whole set of sequences and for the less related pair of sequences (Figure 1) and the classification of the set of sequences (Figure 2). If at least two sequences in the set are not significantly similar, the reliability of the procedure ALIGN may be affected and a warning invites the user to submit more relevant subsets of sequences.

The ALIGN procedure matches the most similar 9-residue segments in a scan over the whole sequences. Residues included in selected boxes are printed in lower case (Figure 4). The reliability score (recorded up to nine) is written below each position of the boxes, the lowest scores corresponding to the highest reliability of the alignment. Residues in upper case are not aligned, and thus gaps are placed arbitrarily before the next box.

Several runs are performed on the whole set of sequences in order to settle similarity significance, specific relevant cut-offs, to detect reliable anchor points separated by shifts as long as theoretically possible and to enlarge them to the extreme limit of the similarity significance. Programs are thus rather time consuming. Several by-passes allow one to save time depending on sequence similarity; thus, the relationship...
Fig. 4. Alignment output displayed by the server for the aspartic protease family. Aligned residues (included in boxes) are printed in lower case. Other residues (uppercase) are not aligned. Only the multiple alignment of the whole set of sequences is performed. A score of −9 is written below each position in the boxes. It is related to the statistical significance of the alignment at this position. The lower the score, the higher the reliability of the alignment.

between problem size and the time required is not straightforward. Roughly, a set of 5–10 proteins is analysed within some minutes of CPU time on a RS-4000 and sets of 40–50 proteins require 1 or 2 h. Information on several not too redundant sequences is required to filter the random noise...
submitted to the server and the results obtained compared with the structural alignment in the following way. For each test family, let $S$ be the total length of the SCRs, $s$ be the total number of residues included in boxes (predicted SCRs) and $l$ the number of positions correctly predicted. The power and the confidence of the method are evaluated by the following relationships:

$$\text{power} = \frac{l}{S} \times 100$$

$$\text{confidence} = \frac{l}{S} \times 100$$

Table I indicates the performances in terms of power and confidence of the global alignment obtained by Match-Box_server on the nine test families. The average confidence is 85%, with a minimum of 77% in the plastocyanin set. The average power is 71%, with a minimum of 42% in the aspartic protease family. Both families leading to the lowest performances share only 9% identity. On the other hand, confidence remains high (~80% and upper) for nearly all the sets. Moreover, when this evaluation is performed separately by level of the reliability score (Table II) the average confidence for the positions aligned is always 100% at level 1 and slows down progressively to 80% as the score increases. Results obtained on 1628 aligned residues show a linear relationship ($P < 0.001$) between the reliability score and the actual confidence computed from sequences of known structures. The lowest confidence (77%) is observed with a reliability score of six.

Match-Box_server offers automatic programs to evaluate a priori the reliability of the alignment and tools for delineating relevant subsets of sequences. Unrelated sequences submitted to the server produce an alignment of sparse boxes associated with a high reliability score, and a warning clearly informs the user about the presence of a non-significant relationships between the submitted sequences. Scores allow the reliability of each position aligned to be evaluated. The risk of aligning unrelated regions thus remains very low.

However, it remains evident that some structures may show similar structural regions without sharing any significant similarity of sequences. The rate of success, expressed as the capability to delineate structural conserved regions, decreases as a function of the sequence similarity and the level of confidence required. The average rate of success observed in the tests is ~70%, with a minimum of 40%. This method thus places emphasis on reliability and is mainly devoted to applications for which the accuracy of the delineation of conserved regions is the principal criterion of efficiency, such as homology modelling and site-directed mutagenesis. This method is also particularly suited for the design of oligonucleotide sequences able to amplify by polymerase chain reaction (PCR) parts of the gene coding for a given protein, aligned with a set of homologous sequences. Indeed, the high reliability of the boxes of level 1 allows one to delineate precisely the highly conserved regions in the whole protein family, determining the most appropriate oligonucleotides needed to initiate a PCR.

Acknowledgements

This project is supported by the Walloon Government, Federal State of Belgium. The authors want to thank all the 'sequence aligners' who have agreed to test the program and to provide useful comments. and those who will do so in the future.

References

E. Depiereux et al


Received on September 9, 1996; revised on December 4, 1996; accepted on December 10, 1996.