LEON: multiple aLignment Evaluation Of Neighbours

Julie Thompson, Véronique Prigent, Olivier Poch

Published online February 24, 2004

1298–1307 *Nucleic Acids Research, 2004, Vol. 32, No. 4*
DOI: 10.1093/nar/gkh294

LEON: multiple aLignment Evaluation Of Neighbours

Julie D. Thompson, Véronique Prigent and Olivier Poch*

Laboratoire de Biologie et Genomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS/INSERM/ULP, BP 163, 67404 Illkirch Cedex, France

Received December 11, 2003; Revised January 16, 2004; Accepted January 29, 2004

ABSTRACT

Sequence alignments are fundamental to a wide range of applications, including database searching, functional residue identification and structure prediction techniques. These applications predict or propagate structural/functional/evolutionary information based on a presumed homology between the aligned sequences. If the initial hypothesis of homology is wrong, no subsequent application, however sophisticated, can be expected to yield accurate results. Here we present a novel method, LEON, to predict homology between proteins based on a multiple alignment of complete sequences (MACS). In MACS, weak signals from distantly related proteins can be considered in the overall context of the family. Intermediate sequences and the combination of individual weak matches are used to increase the significance of low-scoring regions. Residue composition is also taken into account by incorporation of several existing methods for the detection of compositionally biased sequence segments. The accuracy and reliability of the predictions is demonstrated in large-scale comparisons with structural and sequence family databases, where the specificity was shown to be >99% and the sensitivity was estimated to be ~76%. LEON can thus be used to reliably identify the complex relationships between large multidomain proteins and should be useful for automatic high-throughput genome annotations, 2D/3D structure predictions, protein–protein interaction predictions etc.

INTRODUCTION

Multiple sequence comparisons or alignments are one of the cornerstones of modern molecular biology. Since their introduction in the 1970s they have been used in a wide range of molecular biology applications, including the identification of key functional residues in a family of proteins, the prediction of 2D/3D structural features and evolutionary studies to define the phylogenetic relationships between organisms. All these applications predict or propagate biological information between the sequences in the multiple alignment based on a presumed homology. The hypothesis is that homologous sequences, i.e. sequences that have descended from the same ancestor, often share the same structure and function. In this article, the term 'homologous sequences' will therefore refer to proteins that have similar functions or that share regions of similar structure. A fundamental step in these so-called 'homology-based' methods is the determination of the extent of homology (i.e. the extent of shared structure/function) between the aligned sequences. Without this initial crucial step, the subsequent applications that rely on an accurate multiple alignment cannot be expected to yield high-quality results.

Homology-based methods generally begin with a search for similar proteins in the public sequence databases using tools such as BLAST (1), PSI-BLAST (2) or domain/motif databases such as Interpro (3). However, simply finding a medium or weak sequence similarity to an experimentally determined protein is not sufficient for an accurate transfer of information from the known protein to the unknown one (4–6). While a direct relationship between sequence similarity and conservation of protein structure has been clearly established (7–10), the relation between fold and function is more controversial. A number of authors have investigated the relation between sequence similarity and protein function via the Enzyme Classification (EC) (11–13). Precise function appears to be conserved down to ~40% sequence identity, whereas broad functional class is conserved to ~25%. The simple transfer of information from the most closely related hits found by a database search has led to a number of errors, notably in automatic high-throughput genome annotation projects. The level of error in genome functional annotations has been estimated to be ~5–8% for more general enzymic functions (5,14) to >30% for specific functions, such as substrate specificity (14). More recently, Rost (15) has suggested that the conservation of enzyme function has been overestimated and consequently, that the percentage of errors in annotations may have been underestimated. Local sequence similarities in relatively short regions and/or transferring annotations for different domains in multidomain proteins were identified as the cause of most misclassifications. Another study (16) also identified some common causes of questionable predictions such as: (i) non-critical use of annotations from existing database entries; (ii) taking into account only the annotation of the best database hit; (iii) insufficient masking of low-complexity regions (e.g. non-globular domains) in protein sequences, resulting in spurious database hits obscuring relevant ones; (iv) ignoring multidomain organization of the query proteins and/or the database hits. Clearly, new cooperative and integrative algorithms are now required in

*To whom correspondence should be addressed. Tel: +33 3 88 65 32 00; Fax: +33 3 88 65 32 01; Email: poch@igbmc.u-strasbg.fr

order to develop a comprehensive picture of the complex relationships that exist between large multidomain proteins.

Global multiple alignments of complete sequences (MACS) provide an ideal basis for more in-depth analyses of protein family relationships (17). By placing the sequence in the context of the overall family, MACS permit not only a horizontal analysis of the sequence over its entire length, but also a vertical view of the evolution of the protein. Nevertheless, MACS have often been considered unsuitable for automatic high-throughput projects because of their unreliability in the face of complex, often non-collinear, proteins and the relatively long calculation times required. Fortunately, recent advances in multiple alignment algorithms (18–20) now allow fast reliable global multiple alignment of large families of complete sequences. However, the determination of the extent of homology between the sequences in the multiple alignment remains a fundamental problem. The degree to which the sequences in a multiple alignment are related can be estimated by an analysis of positional conservation (21–23) or by measuring the statistical significance of the alignment (24). Cline *et al*. (25) tested four different predictors of alignment position reliability and concluded that near-optimal alignment information was the best predictor, removing 70% of the substantially misaligned positions. Thompson *et al*. (26) used the NorMD objective function to remove unrelated or badly aligned sequences from multiple alignments. Errami *et al*. (27) analysed the agreement between predicted secondary structures of the aligned sequences to detect and discard unrelated sequences. Tress *et al*. (28) used sequence profiles generated from PSI-BLAST alignments to predict reliable regions between remotely related pairs of proteins. These methods work well when the sequences to be compared are homologous over their full lengths, but large multidomain proteins are becoming more and more prevalent in the sequence databases with the arrival of numerous new genome sequences, in particular from eukaryotic organisms. In the face of these highly complex proteins, new more sensitive methods are needed to detect sequences with either full-length or partial homology to the query sequence.

Here we propose a new method, LEON (aLignment Evaluation Of Neighbours), to predict homologous regions in MACS with respect to a user-defined reference or 'query' sequence. LEON incorporates some of the latest developments in multiple alignment analysis, including sequence clustering (29) and the identification of locally conserved motifs or 'core blocks' (30). Taking advantage of the transitive nature of homologous relationships, information from intermediate sequences (31,32) is used to help define the conserved core blocks for more divergent sequences. The conserved blocks for each subfamily are then chained together to form contiguous regions. Groups of motifs are now often used instead of single motifs, for example in sequence searches (2) or motif searches (33–35), because they inherently offer improved diagnostic reliability by virtue of the mutual context provided by motif neighbours. Finally, the amino acid composition of the sequences is taken into account by the incorporation of a number of algorithms for the detection of compositionally biased segments (36–38).

The specificity and sensitivity of the LEON method are demonstrated in two separate large-scale tests. The first involves a large set of 106 multiple alignments which is divided into separate training and test sets, each consisting of 53 multiple alignments. This set was based on a previous benchmark set specifically designed to assess the validity of inheriting protein function by homology (6). As all the sequences in this test had known 3D structures, an objective definition of homology could be used, i.e. that the two proteins share at least one domain with the same 3D fold. The results obtained by LEON for the 53 test alignments were compared with two structural family databases, CATH (39) and MMDB (40). The specificity in these tests was shown to be >99%, and the sensitivity was estimated to be ~76%. The second test set consists of a set of 174 enzymes, constructed by selecting representative sequences with different EC numbers. This test set contained examples of some of the known pitfalls encountered by homology-based methods, namely multidomain sequences, sequences with transmembrane regions and sequences with low-complexity regions. The results are compared with the manually curated protein sequence family database Pfam (41).

In the final LEON alignment, the sequences that are predicted to be homologous to the user's query sequence are ranked according to their similarity to the query sequence. Sequences containing no regions with significant scores are filtered from the alignment. The homologous regions within each sequence are delimited and thus can be used for reliable function annotation, fold classification, 2D/3D structure predictions, domain determination, etc.

MATERIALS AND METHODS

LEON consists of a suite of programs, all written in ANSI C. The programs were installed and tested on a DEC Alpha 6100 computer running OSF Unix. A UNIX shell script is provided that calls the C programs. The Secator program (http: //www-bio3d-igbmc.u-strasbg.fr/~wicker/Secator/secator.html) is required for sequence clustering. The NCOILS program (http://russell.embl-heidelberg.de/coils/coils.tar.gz) is also required for the prediction of coiled-coil regions. LEON takes multiple alignments in either MSF or FASTA format as input and outputs a refined alignment in either MSF or FASTA format, as requested by the user. The refined alignment contains only those sequences predicted to contain homologous regions. The positions of the conserved core blocks and conserved regions are output to a formatted text file. The source code for LEON is freely available from ftp://ftp-igbmc.u-strasbg.fr/pub/Leon.

Construction of the training and test sets

A first test set of multiple alignments was constructed based on a benchmark specifically defined for a previous study (6) to assess the validity of inheriting protein function by homology. The benchmark consisted of 106 chains from the Protein Data Bank (PDB) database (42) sharing <25% mutual sequence identity and included 95 enzymes and 11 non-enzymes. For each chain in the benchmark set, a PSI-BLAST (2) search of the PDB database was performed. In order to include a maximum number of related sequences, but also a certain number of unrelated sequences, 15 iterations of PSI-BLAST were performed and the *E* value threshold for including matches in the PSI-BLAST model was set to 10.0. As this process often resulted in a large number of closely related sequences, an automatic method was used to select a smaller

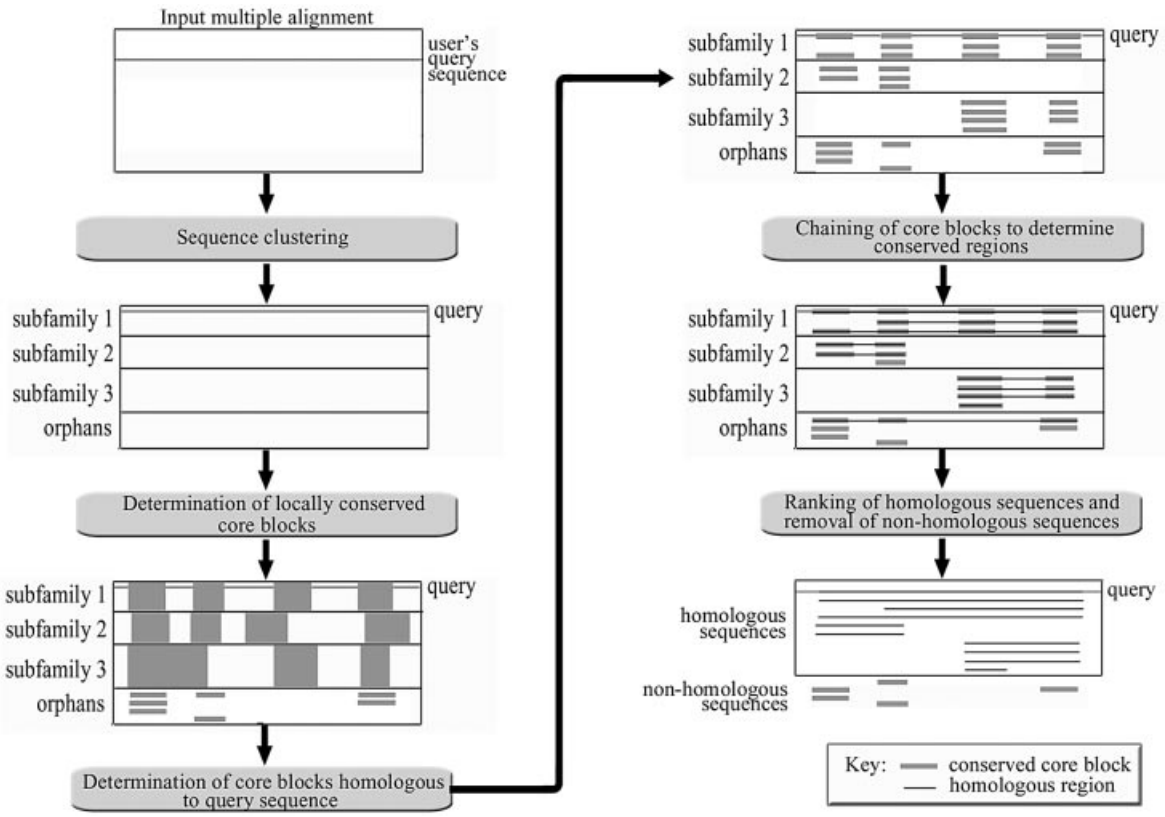

Figure 1. Flowchart showing the four major steps of the LEON algorithm. The input to the algorithm consists of a multiple sequence alignment, in which the user has identified a reference or 'query' sequence. The final result is a multiple alignment in which the sequences predicted to be homologous are ranked according to their similarity to the query. Non-homologous sequences are excluded from the alignment.

subset of representative sequences (43). Briefly, the range of E values found in the BlastP output file is analysed and a significance threshold is determined for the ratio of E values between two sequence entries. Then, if the E value ratio between two consecutive sequences is greater than the threshold, only the first sequence is selected. The result is a non-redundant subset of sequences selected from the full set of all sequences detected by PSI-BLAST with $E < 10.0$. For each query sequence, a multiple alignment of the non-redundant subset was constructed using DbClustal (18) and then refined using RASCAL (30). The resulting 106 multiple alignments contained a total of 11 736 sequences of known 3D structure. The benchmark set was then divided into separate training and test sets, each containing 53 multiple alignments.

A second test set was then constructed from the BRENDA Enzyme Information System (44) by selecting one example from each of the EC classifications at the third digit level. The test set consisted of 174 enzymes and included multidomain proteins as well as proteins containing transmembrane or coiled-coil regions. For each enzyme, a BlastP search of the SWISS-PROT/TrEMBL (45) and PDB databases was performed. The automatic method described above was again used to select a non-redundant subset of all sequences found with $E < 10.0$. A multiple alignment of the non-redundant subset was then constructed using DbClustal and refined using RASCAL. The resulting 174 multiple alignments contained a total of 11 550 full-length sequences.

Overview of the algorithm

Given a multiple sequence alignment, LEON determines, for each sequence in the alignment, the regions that are homologous to a specified query sequence. The method consists of four major steps, outlined in Figure 1. First, the sequences in the multiple alignment are clustered into subfamilies using the Secator program. Any highly divergent or 'orphan' sequences in the alignment are identified and are excluded from the subfamily clustering. Secondly, for each subfamily, locally conserved regions or 'core blocks' are determined using the RASCAL method. Briefly, RASCAL uses the NorMD objective function in a sliding window analysis to determine locally conserved segments or 'core blocks' in an alignment. Each core block is then represented by a Gribskov profile (46), based on the observed residue frequencies in the block and the Gonnet 250 matrix (47). Core blocks in the orphan sequences that are unclustered by Secator are then identified using sequence profile scores for each core block in each subfamily. Thirdly, the core blocks specific to each subfamily or orphan sequence are compared in a pairwise fashion in order to identify all core blocks that match the query sequence. In this way, intermediate sequences can be used to match core blocks between the more divergent sequences and the query. Finally, for each sequence, the conserved blocks that match the query sequence are chained together to form 'regions' and a score is calculated for each homologous region based on the length and score of the associated core blocks and

the composition of the sequences. Any sequences with no homologous regions scoring higher than a threshold are removed from the alignment. As the first two steps have been described in detail elsewhere, the following sections describe only the final two stages.

Matching core blocks between subfamilies

Pairwise profile–profile scores are calculated for all core blocks within a subfamily with overlapping core blocks in all other subfamilies. The raw score for two profiles is defined as the sum of the scores for each pair of aligned columns and the calculation of the score for two columns is the same as that used in ClustalW (48). The raw profile–profile scores are then normalized for the length of the overlap of the two core blocks and the similarity of the sequences within each core block. The normalized score S for matching two core blocks b_i, b_j in the ith and jth subfamilies, respectively, is calculated by

$$S(bi, bj) = \frac{100 * s * f(bi, bj)}{pcid * len}$$

where s is the raw profile–profile score, *len* is the length of overlap of the two core blocks, *pcid* is the mean percentage identity of the sequences in the core blocks and $f(b_i,b_j)$ equals 0.75 if either b_i or b_j overlap a segment with a biased residue composition (see below) or equals 1 otherwise. This normalization allows the definition of a threshold score t above which a match between two core blocks is considered to be significant. The next step is to identify those core blocks that match the subfamily containing the query sequence. A core block b_i matches a core block b_q in the query subfamily if either $S(b_i,b_q) > t$ or a subfamily j exists for which $S(b_i,b_j) > t$ and $S(b_j,b_q) > t$.

Thus intermediate subfamilies are used to match core blocks between the more divergent sequences and the query. In this way even though a core block in a remote sequence may not directly have a significant score when compared with the query, it may be linked through an intermediate whose similarities to both the query and the remote sequence are above the threshold.

Detection of sequence segments with biased composition

LEON incorporates several existing algorithms for the detection of compositionally biased segments and a heuristic procedure is employed to reduce the score for core blocks that contain these segments (see above). We use the SEG algorithm (38) to detect low-complexity sequence segments and the NCOILS program to detect coiled-coil regions (37). Potential transmembrane segments are identified using a sliding-window analysis of residue hydrophobicity scores (36). These algorithms were chosen for their efficiency and simplicity of implementation. Although they may not be the most accurate methods currently available, in the large-scale tests performed here they have proved to be sufficient for this application (data not shown).

Determination of conserved regions

Once the core blocks for each sequence that match the query have been determined, the next step is to chain the core blocks together to form conserved 'regions', provided certain constraints are satisfied. This step, outlined in Figure 2, is performed individually for each sequence. First, the core blocks for the sequence are ordered according to their position in the sequence. Then, for each sequence, let $B = b_1,b_2,\ldots,b_n$ be the set of core blocks found in the sequence that match the query sequence. Then, let r_i be the position of the first residue in the ith block, let q_i be the corresponding residue in the query sequence and let l_i be the length of the block. Two core blocks b_i,b_j are chained together if they satisfy all of the following constraints:

$$S_i >; S_j > x; r_j - (r_i + l_i) < d; q_j - (q_i + l_i) < d$$

where S_i is the normalized profile score for the ith core block that matches the query, x is a minimum score for the core blocks and d is a maximum length of insertion between the two blocks. The chaining continues until all possible blocks are chained. The score for a conserved region is defined as the sum of the scores for the core blocks within the region. Finally, regions are predicted to be homologous if they score higher than a threshold score T and have a total core block length $>L$.

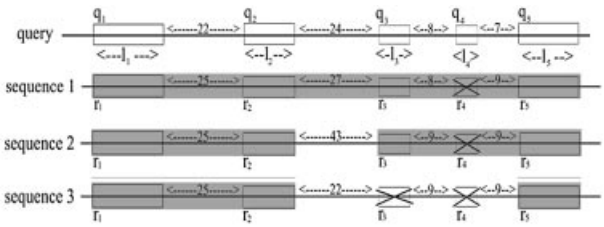

Figure 2. Determination of conserved regions, illustrated by the chaining of the core blocks in a subfamily of three sequences. The core blocks for each sequence are indicated by boxes outlined in black. Core blocks scoring less than the minimum score ($x = 5$) are marked with a black cross. The position of the first residue in the ith core block is denoted r_i and the corresponding position in the query is denoted q_i. The length of the ith core block is indicated by l_i. The numbers between the dotted lines indicate the number of residues between the two adjacent core blocks. For each sequence, two core blocks are chained if the insertion length between them is less than the maximum length ($d = 40$). The regions formed by the chaining are indicated by grey shading.

Optimization of threshold parameters

The method described above includes a number of threshold parameters, namely x, d, T and L. In order to optimize the values used for these parameters, the LEON method was used to detect unrelated sequences in the 53 multiple alignments included in the training set (see Materials and Methods). LEON was trained by comparison with the results obtained with the CATH Protein Structure Classification database (49), which is a manually maintained high-quality database of protein domain structure families. Proteins in the same homologous superfamily in CATH are thought to share a common ancestor and were therefore assumed to be true homologues. In order to determine the optimal parameter settings, an iterative search of a constrained parameter space was performed. By a manual examination of the alignments in the training set, the four parameters were limited as follows: $3 \leqslant x \leqslant 6$, $30 \leqslant d \leqslant 50$, $260 \leqslant T \leqslant 320$, $18 \leqslant L \leqslant 24$. The four parameters were systematically modified within these limits, incrementing x,L by 1 and d,T by 10. At each iteration, the homologous regions in the 53 training alignments were

recalculated and compared with the CATH classifications. Of all the parameter combinations tested, the settings that resulted in a specificity of less than 100% were immediately rejected. The parameters corresponding to the maximum sensitivity were then selected for use in the subsequent tests. The final threshold parameters were set to $x = 5$, $d = 40$, $T = 280$ and $L = 21$. These settings provide a specificity of 100%, i.e. no false positives were detected, and an estimated sensitivity of 79% (415 of the 2035 homologous sequences in CATH were not correctly identified).

RESULTS

The algorithm developed in LEON shows some interesting parallels with the latest gapped BLAST program. The central idea of the BLAST algorithm is that a statistically significant alignment (HSP) is likely to contain a high-scoring pair of aligned words. BLAST first scans the database for 'words' (typically of length 3 for proteins) that score higher than a certain threshold when aligned with some word within the query sequence. Any aligned word pair satisfying this condition is called a 'hit'. A gapped alignment is then generated only if two non-overlapping hits are found within a specified distance of one another. This 'two-hit' heuristic was shown to be more sensitive than the original 'one-hit' method. By default, the minimum score for the two hits is 11 and the distance between them should not exceed 40. The threshold scores for LEON and BLAST are not directly comparable because they are based on different residue comparison matrices. Nevertheless, it is interesting to note that the optimal value of $d = 40$, corresponding to the maximum distance allowed between two consecutive core blocks, is the same as the BLAST default value determined by manual inspection of 100 000 model HSPs. The second step of the BLAST algorithm checks whether each hit lies within an HSP with score sufficient to be reported, which could be compared to the combination of the core blocks in LEON to form longer regions. The final scoring schemes of the two methods are different, however. Gapped BLAST scores all the aligned residues in the HSP using a residue comparison matrix, with penalties for gap opening and extension. The score for a region in LEON is based only on the conserved core blocks within the region and gap costs are not needed in this approach.

Comparison with the CATH and MMDB structural databases

The sensitivity and specificity of the LEON method were estimated using the test set of 53 multiple sequence alignments from the PDB database (see Materials and Methods), which contains only sequences with known 3D structures and therefore provides an objective definition of homology, i.e. that the two proteins share at least one domain with the same 3D fold. Table 1 shows the details of the comparison between the LEON predictions of homology and the structural classifications in the CATH manually curated database. The specificity of LEON in this test is 99.6% (seven of the 1845 non-homologous sequences in the tests were predicted to be homologous to the query sequence). The sensitivity of LEON is estimated to be 78.2%. However, 2306 or 36% of the sequences in the test multiple alignments had not yet been classified in the CATH database, so the sensitivity of the method could not be measured accurately. A more complete structural classification is provided in the MMDB database (40), which is updated monthly to reflect the complete PDB database. Structure neighbours of each entry in the MMDB database are identified automatically using the VAST algorithm (50). A comparison of the predictions made by LEON and by VAST is shown in Table 2. A surprising result from this comparison was the 273 'false-positive' predictions where LEON predicted homologous sequences which were not defined as structural neighbours in the MMDB database. A more detailed investigation showed that, in 246 of the 273

Table 1. Comparison of the results of LEON for a test set of 53 multiple alignments of the sequences detected in a search of the PDB by PSI-Blast with $E < 10$

	CATH homologues[a]	CATH non-homologues[b]	Sequences not yet classified in CATH[c]	Total sequences[d]
Predicted homologues	1788	7	665	2460
Predicted non-homologues	498	1845	1641	3984

Multiple alignments were constructed using DbClustal and RASCAL.
[a]Sequences with at least one domain in the same CATH superfamily as the query.
[b]Sequences in CATH database with no domain in the same CATH superfamily as the query.
[c]Sequences with known 3D structure but not yet included in the CATH database.
[d]Total number of sequences in multiple alignments.

Table 2. Comparison of the results of LEON for the test set of 53 multiple alignments used in Table 1

	VAST homologues[a]	VAST non-homologues[b]	Total sequences[c]
LEON homologues	2187	273	2460
LEON non-homologues	672	3312	3984

[a]Sequences defined as a 'structure neighbour' of the query in VAST database.
[b]Sequences are either not defined as a 'structure neighbour' of the query or are not in the VAST database (see explanation in text).
[c]Total number of sequences in multiple alignments.

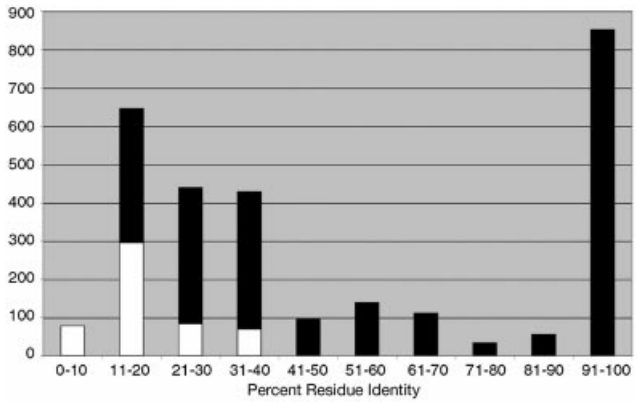

Figure 3. A histogram of the number of sequences predicted by LEON to be homologous (in black) and non-homologous (in white). The *x* axis represents the percentage residue identity calculated by the VAST structure comparison method.

cases, the sequence shared >35% residue identity with another chain that was identified as a structural neighbour of the query. This is the result of a known inconsistency in the VAST neighbour processing protocol (T. Madej, personal communication). The true number of false positives detected by LEON in these tests is therefore estimated to be 27 sequences, leading to a specificity of >99% and a sensitivity of 76.5%. We also investigated the 672 'false-negative' predictions made by LEON. The MMDB database includes various statistics that describe the homologous regions found by the VAST algorithm and so allows a more in-depth comparison. Figure 3 shows a breakdown of the LEON predictions, classified by the percentage residue identity of the homologous region, as calculated by the VAST algorithm. It can be seen that all structures predicted by VAST to contain a homologous region with >40% identity and approximately half of the regions with 11–20% identity, were also predicted to be homologous by LEON. Of the 429 VAST structure neighbours with 31–40% identity, LEON failed to identify homologous regions in 70 sequences. These 70 VAST structure neighbours corresponded mainly to short regions, with a mean length of 71 residues.

Comparison with the Pfam manually curated sequence database

In order to test the accuracy of the LEON predictions for a wide variety of different full-length proteins, a set of representative sequences with different EC numbers at the third-digit level was used (see Materials and Methods). The test set consisted of 174 enzymes and included multidomain proteins, proteins containing transmembrane or coiled-coil regions. For each test case, LEON was used to predict homologous regions and to detect unrelated sequences in the multiple alignment of the complete sequences detected by a BLAST database search. Table 3 shows details of the comparison between the LEON predictions and the Pfam protein domain database, which is a manually maintained high-quality database. Of the 7646 sequences that had matches with the query sequence in the Pfam database, LEON failed to predict any homologous regions for 173 of them. In contrast, 104 sequences were identified as having homologous regions

Table 3. Comparison of the results of LEON for a set of 174 multiple sequence alignments containing sequences detected in a BlastP search with $E < 10$

	Pfam homologues[a]	Pfam non-homologues[b]	Sequences with no Pfam matches[c]	Total sequences[d]
LEON homologues	7573	104	2921	10569
LEON non-homologues	173	408	371	981

Multiple alignments were constructed using DbClustal and RASCAL.
[a]Sequences with at least one Pfam match in common with query.
[b]Sequences with matches in Pfam database, but with no match in common with query.
[c]Sequences with no matches in release 10.0 of the Pfam database.
[d]Total number of sequences in multiple alignments.

that had no matches to the query in Pfam. An example of one of these 'false-positive' predictions is shown in Figure 4, which shows part of the multiple alignment constructed using the L-lactate dehydrogenase from *Escherichia coli* (P33232) as the query. Two subfamilies are included in the alignment, one containing the query sequence which matches the Pfam family of FMN-dependent dehydrogenases (PF01070) and the other containing sequences (e.g. P50098) that match the Pfam family of inosine monophosphate dehydrogenases (PF00478). These two families were predicted by LEON to be homologous and, in fact, they are both known to contain domains with the same TIM barrel fold (CATH homologous superfamily 3.20.20.90). The orphan PDB sequence, 1P0N_B, which was also predicted to be a homologue of the query, has not yet been classified by CATH, but is defined in MMDB as a structural neighbour of the first subfamily (mean RMSD, 2.3; mean % identity, 23.7) and of the second subfamily (mean RMSD, 2.6; mean % identity, 17.7). Nevertheless, the two families do have different functions and different enzyme classifications (EC 1.1.2.3 and EC 1.1.2.05 respectively), and an analysis of the active sites (shown in Fig. 4) described in the SWISS-PROT entries for the proteins revealed that the catalytic residues of the two subfamilies are not found within the conserved core block regions defined by LEON.

The sensitivity of LEON in the face of complex multidomain proteins is illustrated by the example shown in Figure 5. The NADH-dependent nitrate reductase (NR) sequence from *Arabidopsis thaliana* (P11035) was used as the initial query sequence for the BlastP search. NR catalyses the first step in nitrate assimilation, a pathway that is of key importance for plant nutrition. Plant NR has been shown to have a homodimeric structure, containing three functional domains, heme, fad/nadh and molybdenum cofactor (51), represented by five different Pfam families. The BlastP search detected a number of different proteins that contained one or more of the three domains found in NR. One notable subfamily comprised several sulfite oxidases (SUOX), the enzyme that catalyzes the terminal reaction in the sulfur amino acid degradation pathway. In humans, defects in SUOX (mutations are shown in Fig. 5) are known to cause isolated sulfite oxidase deficiency (ISOD) (52), a very rare hereditary metabolic disorder, which often leads to death at an early age.

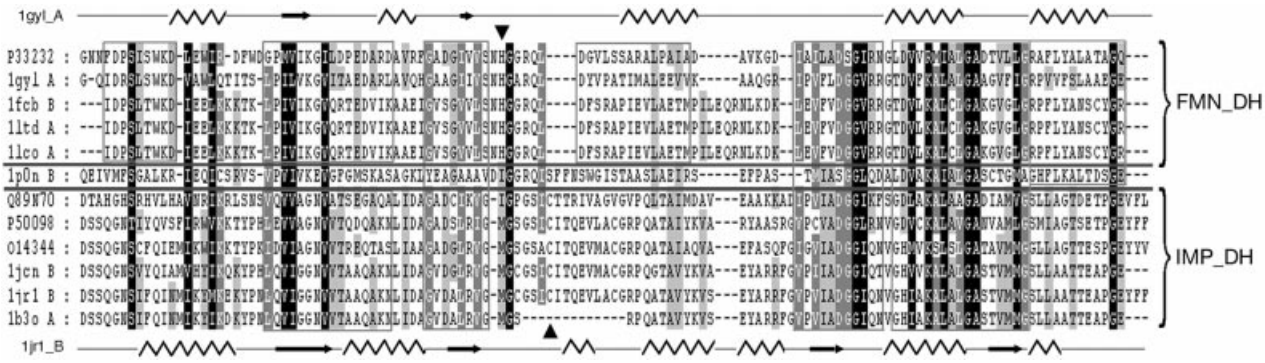

Figure 4. Part of a multiple alignment of the sequences detected by BlastP with $E < 10$ using the L-lactate dehydrogenase from *E.coli* (P33232) as the query. Conserved columns are shaded (black, 100%; dark grey, 80%; light grey, 60%). Two subfamilies containing FMN-dependent dehydrogenases (FMN_DH) and inosine monophosphate dehydrogenases (IMP_DH) and one orphan sequence (1P0N_B) are shown. The secondary structure elements of two representative PDB sequences are shown above and below the alignment (coil, α helix; arrow, β sheet). Grey outlined boxes correspond to conserved core blocks. Black triangles indicate active sites in the two subfamilies.

DISCUSSION

The determination of homology is a crucial problem for a wide range of homology-based applications and poses particular problems in automatic high-throughput genome analysis and annotation projects. A number of methods exist that estimate homology based on a multiple sequence alignment (21–28). These methods work well when the sequences to be compared share global homology. But they generally look for features shared amongst all or most of the sequences. For example, when searching with a fused protein containing two domains, A and B, many of the sequences detected will contain only domain A and others will contain only domain B. Thus, no globally conserved features will exist and the homology detection methods described above will fail.

Here, we have presented LEON, a new method for the automatic reliable estimation of the homology shared between protein sequences based on MACS. The rationale of the LEON method is the integration of a number of new algorithmic ideas in a cooperative knowledge-based system. In the context of multiple sequence alignment programs, we showed previously that no single algorithm was capable of providing accurate multiple alignments for all the cases studied (53). Subsequent research (18–20) showed that the combination of information from both local and global alignment algorithms significantly improved the quality of multiple sequence alignments.

Using the same philosophy, LEON first identifies local 'core blocks', which may be conserved in either the complete protein family or in clustered subfamilies. Core blocks are compared using a sensitive profile scoring scheme (48) in order to determine their similarity with the query sequence. Information from intermediate sequences is used to help define the conserved blocks for more divergent sequences. This technique is similar to the Intermediate Sequence Search (ISS) method (31,32), which takes advantage of the transitive nature of homologous relationships to successfully improve the sensitivity of database searches. These local core blocks define single conserved motifs that may not always be significant when considered in isolation. A more global view is obtained by chaining the core blocks together into longer contiguous regions if they satisfy certain constraints. Although these contiguous regions often correspond to structural domains (e.g. Fig. 4), this is not guaranteed to be always true. For example, a structural domain may consist of several non-contiguous sequence segments. Another important feature is the calculation of core block scores that take into account the composition of the sequences. It is known that sequence segments with low residue complexity or composition bias, such as transmembrane regions or coiled-coil regions, can give rise to spurious scores in sequence comparisons. For example, the accuracy of the PSI-BLAST program (2) was improved by tuning the position-specific scoring system for each database sequence to the sequence's amino acid composition (54).

The LEON algorithm was tuned using the CATH structural classification (39) as a standard of truth. The threshold parameters used in the subsequent tests were set to provide maximum specificity. This is obviously important for automatic genome analysis projects, where inclusion of non-homologous proteins in the multiple alignment could lead to erroneous functional or structural annotations. However, it is possible for an experienced user to lower the threshold parameters in order to achieve a higher sensitivity, at the expense of some loss in specificity. In a separate large-scale test, LEON was compared with both the CATH database and the automatic structural neighbour method VAST (50), and the specificity of LEON was shown to be >99%. The predicted homologous regions therefore provide a reliable basis for the many applications that rely on accurate multiple alignments, e.g. functional residue identification, 2D/3D structure prediction or homology modelling.

Work is now in progress to incorporate LEON in the PipeAlign protein family analysis www server (55). Future developments will also include the integration of other information in LEON, such as functional information in the form of textual annotations, 2D/3D structures or known domains from the manually curated databases. This knowledge will be exploited for automatic refinement of the multiple alignments and correction of local alignment errors. The LEON homology predictions in combination with this high-quality knowledge-based multiple alignment will pro-

A.

Accession	Description	Organism	Domain	Domain organisation
P11035	Nitrate reductase [NADH] 2	Arabidopsis thaliana	Eukaryota	
1sox_B	Sulfite oxidase	Gallus gallus	Eukaryota	
Q9vwp4	Probable sulfite oxidase, mitochondrial precursor	Drosophila melanogaster	Eukaryota	
P51687	Sulfite oxidase, mitochondrial precursor	Homo sapiens	Eukaryota	
Q87H22	Putative molybdenum containing oxidoreductase.	Vibrio parahaemolyticus.	Bacteria	
1ndh	nadh-cytochrome b5 reductase	Bos taurus	Eukaryota	
Q94E53	Putative cytochrome b5 reductase.	Oryza sativa	Eukaryota	
Q876w2	Reductase.	Gibberella zeae	Eukaryota	
Q8BTI5	Flavohemoprotein B5/B5RR.	Mus musculus	Eukaryota	
Q82WR8	Putative reductase oxidoreductase protein.	Nitrosomonas europaea.	Bacteria	
Q9K0M8	Na(+)-translocating NADH-quinone reductase	Neisseria meningitidis	Bacteria	
Q92TG4	Putative ferredoxin reductase electron transfer	Rhizobium meliloti	Bacteria	
1j0q_A	Cytochrome b5.	Bos taurus	Eukaryota	
P40312	Cytochrome b5	Saccharomyces cerevisiae	Eukaryota	

Key to Pfam domains:
- Oxidoreductase molybdopterin binding domain
- Oxidoreductase FAD binding domain
- CS domain (unknown function)
- Mo-co oxidoreductase dimerisation domain
- Oxidoreductase NAD binding domain
- Iron-sulfur cluster binding domain
- Cytochrome b5-like Heme/steroid binding domain

B.

Figure 5. Multiple alignment of the sequences detected by BlastP with $E < 10$, using the NADH-dependent nitrate reductase from *A.thaliana* (P11035) as the query. (**A**) Domain organization of some of the top-scoring sequences. Domains were identified in the Pfam protein family database (%id, residue percentage identity). (**B**) A global multiple alignment of the full-length sequences. Residues are coloured according to the Pfam domain colouring scheme in (A). The homologous regions predicted by LEON are outlined in black. Black triangles indicate mutations in the human sulfite oxidase enzyme (P51687) that cause ISOD.

vide a powerful tool for the characterization of new or unknown proteins.

In conclusion, LEON is a fully automatic method which reliably detects homologous regions in multiple sequence alignments and which can be applied in a wide variety of situations, including difficult cases such as distantly related sequences, multidomain sequences or transmembrane sequences. It can therefore be incorporated in high-throughput protein analysis protocols and provides a basis for reliable cross-validation, propagation and prediction of structural/functional information.

ACKNOWLEDGEMENTS

We would like to thank Luc Moulinier, Frédéric Plewniak and Nicolas Wicker for stimulating discussions and Jean Claude Thierry and Dino Moras for their continued support. This work was supported by institute funds from the Institut National de la Santé et de la Recherche Médicale, the Centre National de la Recherche Scientifique, the Hôpital Universitaire de Strasbourg, the Fond National de la Science (GENOPOLE) and the SPINE project (E.C. contract number QLG2-CT-2002–00988).

REFERENCES

1. 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.
2. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
3. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
4. Doerks,T., Bairoch,A. and Bork,P. (1998) Protein annotation: detective work for function prediction. *Trends Genet.*, **14**, 248–250.
5. Brenner,S.E. (1999) Errors in genome annotation. *Trends Genet.*, **15**, 132–133.
6. Aloy,P., Querol,E., Aviles,F.X. and Sternberg,M.J. (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.*, **311**, 395–408.
7. Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.
8. Wood,T.C. and Pearson,W.R. (1999) Evolution of protein sequences and structures. *J. Mol. Biol.*, **291**, 977–995.
9. Yang,A.S. and Honig,B. (2000) An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. *J. Mol. Biol.*, **301**, 679–689.
10. Koehl,P. and Levitt,M. (2002) Sequence variations within protein families are linearly related to structural variations. *J. Mol. Biol.*, **323**, 551–562.
11. Wilson,C.A., Kreychman,J. and Gerstein,M. (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.*, **297**, 233–249.
12. Hegyi,H. and Gerstein,M. (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.*, **11**, 1632–1640.
13. Todd,A.E., Orengo,C.A. and Thornton,J.M. (2001) Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.*, **307**, 1113–1143.
14. Devos,D. and Valencia,A. (2001) Intrinsic errors in genome annotation. *Trends Genet.*, **17**, 429–431.
15. Rost,B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.*, **318**, 595–608.
16. Galperin,M.Y. and Koonin,E.V. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol.*, **1**, 55–67.
17. Lecompte,O., Thompson,J.D., Plewniak,F., Thierry,J. and Poch,O. (2001) Multiple alignment of complete sequences (MACS) in the post-genomic era. *Gene*, **270**, 17–30.
18. Thompson,J.D., Plewniak,F., Thierry,J. and Poch,O. (2000) DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res.*, **28**, 2919–2926.
19. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
20. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
21. Livingstone,C.D. and Barton,G.J. (1993) Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.*, **9**, 745–756.
22. Pei,J. and Grishin,N.V. (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, **17**, 700–712.
23. Wang,L. and Xu,Y. (2003) SEGID: identifying interesting segments in (multiple) sequence alignments. *Bioinformatics*, **19**, 297–298.
24. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
25. Cline,M., Hughey,R. and Karplus,K. (2002) Predicting reliable regions in protein sequence alignments. *Bioinformatics*, **18**, 306–314.
26. Thompson,J.D., Plewniak,F., Ripp,R., Thierry,J.C. and Poch,O. (2001) Towards a reliable objective function for multiple sequence alignments. *J. Mol. Biol.*, **314**, 937–951.
27. Errami,M., Geourjon,C. and Deleage,G. (2003) Detection of unrelated proteins in sequences multiple alignments by using predicted secondary structures. *Bioinformatics*, **19**, 506–512.
28. Tress,M.L., Jones,D. and Valencia,A. (2003) Predicting reliable regions in protein alignments from sequence profiles. *J. Mol. Biol.*, **330**, 705–718.
29. Wicker,N., Perrin,G.R., Thierry,J.C. and Poch,O. (2001) Secator: a program for inferring protein subfamilies from phylogenetic trees. *Mol. Biol. Evol.*, **18**, 1435–1441.
30. Thompson,J.D., Thierry,J.C. and Poch,O. (2003) RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics*, **19**, 1155–1161.
31. Park,J., Teichmann,S.A., Hubbard,T. and Chothia,C. (1997) Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.*, **273**, 349–354.
32. Salamov,A.A., Suwa,M., Orengo,C.A. and Swindells,M.B. (1999) Combining sensitive database searches with multiple intermediates to detect distant homologues. *Protein Eng.*, **12**, 95–100.
33. Attwood,T.K., Avison,H., Beck,M.E., Bewley,M., Bleasby,A.J., Brewster,F., Cooper,P., Degtyarenko,K., Geddes,A.J., Flower,D.R. *et al.* (1997) The PRINTS database of protein fingerprints: a novel information resource for computational molecular biology. *J. Chem. Inf. Comput. Sci.*, **37**, 417–424.
34. Grundy,W.N., Bailey,T.L., Elkan,C.P. and Baker,M.E. (1997) Meta-MEME: motif-based hidden Markov models of protein families. *Comput. Appl. Biosci.*, **13**, 397–406.
35. Henikoff,J.G., Greene,E.A., Pietrokovski,S. and Henikoff,S. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.*, **28**, 228–230.
36. Engelman,D.M., Steitz,T.A. and Goldman,A. (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.*, **15**, 321–353.
37. Lupas,A., Van,D.M. and Stock,J. (1991) Predicting coiled coils from protein sequences. *Science*, **252**, 1162–1164.
38. Wootton,J. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.*, **17**, 149–163.
39. Pearl,F.M., Bennett,C.F., Bray,J.E., Harrison,A.P., Martin,N., Shepherd,A., Sillitoe,I., Thornton,J. and Orengo,C.A. (2003) The CATH

database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.*, **31**, 452–455.
40. Chen,J., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al*. (2003) MMDB: Entrez's 3D-structure database. *Nucleic Acids Res.*, **31**, 474–477.
41. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
42. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Feng,Z., Gilliland,G.L., Iype,L., Jain,S. *et al*. (2002) The Protein Data Bank. *Acta Crystallogr. D*, **58**, 899–907.
43. Errami,M. PhD Thesis: Analyse statistique des structures tridimensionelles de proteines et validation de familles structurales a bas taux d'identite. Universite Claude Bernard Lyon 1.
44. Schomburg,I., Chang,A. and Schomburg,D. (2002) BRENDA, enzyme data and metabolic information. *Nucleic Acids Res.*, **30**, 47–49.
45. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
46. Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.
47. Gonnet,G.H., Cohen,M.A. and Benner,S.A. (1992) Exhaustive matching of the entire protein sequence database. *Science*, **256**, 1443–1445.
48. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) 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.
49. Orengo,C.A., Pearl,F.M., Bray,J.E., Todd,A.E., Martin,A.C., Lo,C.L. and Thornton,J.M. (1999) The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res.*, **27**, 275–279.
50. Madej,T., Gibrat,J.F. and Bryant,S.H. (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.
51. Crawford,N.M., Smith,M., Bellissimo,D. and Davis,R.W. (1988) Sequence and nitrate regulation of the *Arabidopsis thaliana* mRNA encoding nitrate reductase, a metalloflavoprotein with three functional domains. *Proc. Natl Acad. Sci. USA*, **85**, 5006–5010.
52. Johnson,J.L., Coyne,K.E., Garrett,R.M., Zabot,M.T., Dorche,C., Kisker,C. and Rajagopalan,K.V. (2002) Isolated sulfite oxidase deficiency: identification of 12 novel SUOX mutations in 10 patients. *Hum. Mutat.*, **20**, 74.
53. Thompson,J.D., Plewniak,F. and Poch,O. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.*, **27**, 2682–2690.
54. Schaffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
55. Plewniak,F., Bianchetti,L., Brelivet,Y., Carles,A., Chalmel,F., Lecompte,O., Mochel,T., Moulinier,L., Muller,A., Muller,J. *et al*. (2003) PipeAlign: a new toolkit for protein family analysis. *Nucleic Acids Res.*, **31**, 3829–3832.