BAliBASE 3.0: Latest Developments of the Multiple Sequence Alignment Benchmark

Julie D. Thompson,[1*] Patrice Koehl,[2] Raymond Ripp,[1] and Olivier Poch[1]

[1] *Département de Biologie et Génomique Structurales, Institut de Génétique et de Biologie Molculaire et Cellulaire, (CNRS/INSERM/ULP), Illkirch Cedex, France*

[2] *Genome Center and Department of Computer Science, University of California, Davis, Davis, California*

ABSTRACT **Multiple sequence alignment is one of the cornerstones of modern molecular biology. It is used to identify conserved motifs, to determine protein domains, in 2D/3D structure prediction by homology and in evolutionary studies. Recently, high-throughput technologies such as genome sequencing and structural proteomics have lead to an explosion in the amount of sequence and structure information available. In response, several new multiple alignment methods have been developed that improve both the efficiency and the quality of protein alignments. Consequently, the benchmarks used to evaluate and compare these methods must also evolve. We present here the latest release of the most widely used multiple alignment benchmark, BAliBASE, which provides high quality, manually refined, reference alignments based on 3D structural superpositions. Version 3.0 of BAliBASE includes new, more challenging test cases, representing the real problems encountered when aligning large sets of complex sequences. Using a novel, semiautomatic update protocol, the number of protein families in the benchmark has been increased and representative test cases are now available that cover most of the protein fold space. The total number of proteins in BAliBASE has also been significantly increased from 1444 to 6255 sequences. In addition, full-length sequences are now provided for all test cases, which represent difficult cases for both global and local alignment programs. Finally, the BAliBASE Web site (http://www-bio3d-igbmc.u-strasbg.fr/balibase) has been completely redesigned to provide a more user-friendly, interactive interface for the visualization of the BAliBASE reference alignments and the associated annotations. Proteins 2005;61:127–136. © 2005 Wiley-Liss, Inc.**

Key words: alignment accuracy; alignment reliability; reference alignment; program evaluation; program comparison; structure superposition

INTRODUCTION

Multiple alignment of protein sequences is one of the most widely used applications in molecular biology. Traditionally, it has been used in the identification of conserved motifs or key functional residues in a family of proteins[1] and in evolutionary studies to define the phylogenetic relationships between organisms.[2] Other important applications include the identification of functional domains[3] and 3D homology modelling.[4] More recently, new technologies such as complete genome sequencing, structural genomics, and proteomics have not only increased the amount of biological information publicly accessible, but have led to a paradigm shift in bioinformatics. The integration and analysis of large amounts of complex and heterogeneous information will be crucial to the detailed description of the function of a protein and the comprehension of its role, not only at the molecular level, but also at the higher levels of the macromolecular complexes, the cellular pathways, the cell, or the organ. In this context, multiple sequence alignment provides an ideal tool for the integration, crossvalidation, and analysis of biological information in the framework of the overall protein family.[5] Thus, multiple alignments now play a fundamental role in most of the computational methods used in genomic analyses or proteomics projects, from gene identification and validation to the characterization of the molecular and cellular functions of the protein.

However, the accuracy and reliability of these methods depend critically on the quality of the underlying multiple alignments. Numerous methods are now available for the multiple alignment of protein sequences. Until recently, the most popular method for the construction of multiple sequence alignments has been the progressive alignment procedure.[6] A multiple sequence alignment is built up gradually by aligning the two closest sequences first and successively adding in the more distant ones. A number of alignment programs based on this method exist, using either a global algorithm to align the complete sequences[7,8] or a local algorithm to align only the more conserved regions.[9] Many other algorithms have also been applied to the multiple alignment problem, including

Grant sponsor: the Institut National de la Santé et de la Recherche Médicale. Grant sponsor: the Centre National de la Recherche Scientifique. Grant sponsor: the Hôpital Universitaire de Strasbourg. Grant sponsor: the Fond National de la Science (GENOPOLE). Grant sponsor: the SPINE project; Grant number: QLG2-CT-2002-00988.

*Correspondence to: Julie D. Thompson, Département de Biologie et Génomique Structurales, Institut de Génétique et de Biologie Molculaire et Cellulaire (CNRS/INSERM/ULP), BP 10142, 67404 Illkirch Cedex, France. E-mail: poch@igbmc.u-strasbg.fr

Received 10 November 2004; Revised 27 January 2005; Accepted 21 February 2005

Published online 25 July 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20527

Hidden Markov Models (HMMs) in programs such as HMMT[10] or SAM,[11] Genetic Algorithms in SAGA,[12] segment-to-segment alignments in DIALIGN,[13] or iteration techniques, notably in the PRRP[14] program. In the postgenomics era, the evolution of the sequence and structure databases has lead to new challenges for sequence alignment programs. Given the ever-increasing amount of sequence and structure information available in the public databases, the size of the data sets that need to be routinely analyzed is increasing. Large multidomain proteins, in particular from eukaryotic organisms, are also becoming more prevalent. Furthermore, the incorporation of heterogeneous, error-prone data will require major changes to the fundamental alignment algorithms used to date. At the sequence level, it has been estimated recently that 44% of predicted proteins emanating from whole-genome shotgun sequencing projects and 31% of high-throughput cDNA (HTC) may contain errors in their intron/exon structure.[15] It is now clear that no single algorithm can cope with these highly complex relationships. There has been some renewed interest in the development of multiple alignment techniques, with current opinion moving away from a single all-encompassing algorithm to a more cooperative strategy. For example, 2D/3D structure information has been used to improve the quality of sequence alignment.[16,17] An alternative approach has been to combine local and global algorithms to produce a single multiple alignment.[18–22] The efficiency of multiple alignment methods has also been significantly improved by exploiting fast methods to identify regions of high similarity in the sequences and by restraining the alignment to include these regions.[23,24]

To objectively compare the quality of the numerous multiple sequence alignment methods available today, high-quality benchmarks are crucial. One of the first large scale benchmarks specifically designed for multiple sequence alignment was BAliBASE.[25,26] The alignment test cases in BAliBASE are based on 3D structural superpositions that are manually refined to ensure the correct alignment of conserved residues. The alignments are organized into reference sets that are designed to represent real multiple alignment problems. Reference 1 contains alignments of equidistant sequences and is divided into nine subsets, according to three different sequence lengths and three levels of sequence variability. Reference 2 contains families aligned with one or more highly divergent "orphan" sequences, Reference 3 contains divergent subfamilies, Reference 4 contains sequences with large N/C-terminal extensions, and Reference 5 contains sequences with large internal insertions. In addition, three separate Reference Sets, 6–8, are devoted to the particular problems posed by sequences with transmembrane regions, repeats, and inverted domains. A comparison of some of the alignment methods described above,[27] based on BAliBASE (version 1.0), revealed a number of specificities in the different algorithms. For example, although most of the programs successfully aligned sequences sharing >40% residue identity, an important loss of accuracy was observed for more divergent sequences with <20% identity. Another important discovery was the fact that global alignment methods in general performed better for sets of sequences that were of similar length, although local algorithms were more successful at identifying the most conserved motifs in sequences containing large extensions and insertions. Several other benchmarks for multiple sequence alignment algorithms have been proposed recently. Multiple alignments are provided in the OXBench benchmark suite,[28] based on automatic structure and sequence alignments. PREFAB[24] (version 3.0) contains 1932 multiple alignments, although alignment accuracy is only assessed on a single pair of PDB sequences in each alignment. SABmark[29] contains pairwise reference alignments derived from the SCOP protein structure classification, divided into two sets, twilight zone (Blast E-value ≥1) and superfamilies (residue identity ≤50%), but the optimal multiple alignments are not provided. All of these methods provide large sets of test alignments by using automatic structure superposition methods. However, the automation of the alignment construction process inevitably results in a lower alignment quality than expert-validated alignments. Furthermore, none of these benchmarks provide different test categories that allow a detailed exploration of the performance of multiple alignment algorithms under different test conditions. An alternative strategy for evaluating multiple alignment methods has been to use a semiautomatic structural alignment database such as HOMSTRAD,[30] although this database is not specifically designed as a benchmark.

To respond to the challenges of the postgenomic era and to evaluate the new approaches being developed to solve the multiple alignment problem, we have developed BAliBASE version 3.0. A crucial new development for this release is the introduction of a semiautomatic update protocol. The automation of many of the steps involved in the construction of the reference alignments will allow more frequent releases of the benchmark by incorporating new protein families and new sequences as they become available. Nevertheless, the high quality of the alignments, which was an important feature of the previous releases, has been maintained by the use of 3D structural superpositions combined with a final manual validation and refinement step. The new update protocol has been used to significantly increase both the number of alignments and the number of sequences in each alignment in Reference Sets 1–5, with the exception of Reference Set 1, subset V3 that contained sequences with >40% identity. This subset has been excluded from this release of the benchmark based on the previous observation[27] that comparisons between the programs at this level of sequence identity were indecisive. The update protocol also includes an automatic, objective definition of the "core block" regions in each alignment that can be reliably aligned, excluding the ambiguous regions that cannot be structurally superposed. Finally, to facilitate the automatic evaluation of multiple alignment programs, the core blocks and other annotations are now provided in standard data exchange XML format files, replacing the text files used previously.

Another important criterion in the development of a sequence alignment benchmark is the coverage of the protein fold space. One of the principal goals of many structural genomics initiatives is to identify the total repertoire of protein folds and to obtain a global view of the "protein structure universe." For this release of BAliBASE, we have therefore chosen representative protein families from as many different structural fold types as possible, with examples from the five main classes (excluding transmembrane, coiled coil, small proteins, and peptides) in the SCOP[31] protein classification. In addition to the rapid growth in the total number of structures available, large multidomain proteins are also becoming more and more prevalent, in particular from eukaryotic organisms. To test the performance of different alignment methods in the face of these highly complex proteins, BAliBASE version 3.0 now includes the full-length sequences for all the Reference Sets, as well as the alignment of the conserved domains. This new release of BAliBASE thus contains test cases covering most of the current multiple alignment problems, from alignment of single domains, for example, in the construction of protein domain databases to the alignment of full-length, complex sequences, such as those detected by the database searches routinely performed in automatic, high throughput genome analysis projects.

MATERIAL AND METHODS

The protocol used to construct the BAliBASE benchmark has been automated as much as possible. Automatic 3D structural superpositions provide the basis for the construction of the reference alignments, although a final validation and refinement step is required that includes manual expertise to ensure the high quality of the resulting multiple sequence alignments. BAliBASE version 2 contained a total of 141 multiple alignments, with 82 protein family alignments in Reference 1. The 532 PDB[32] sequences in these 82 Reference 1 alignments were used as the basis for the construction of version 3.0, with the addition of a number of families from the SCOP (Structural Classification of Proteins) multidomain class.

For each protein family included in the benchmark, a semiautomatic protocol has been developed to detect related sequences in the PDB and UniProt[33] databases and to construct a "primary" multiple alignment that is used to select suitable sequences for each Reference Set. A general overview of the new update protocol is shown in the flowchart in Figure 1.

Primary Structure Alignments

For each protein family, the first step involves the detection and superposition of those family members with a known 3D structure, using the following protocol: (1) The PDB database is searched using the PSI-BLAST program[34,35] with each of the known family members from the original Reference 1 alignment in turn. In the case of new families that were not included in the previous release, a single PSI-BLAST search is performed with a sequence selected from the SCOP classification database. The PSI-BLAST search is stopped after five iterations and all sequences detected with $E < 10^{-3}$ are initially selected. (2) In a previous study,[27] it was observed that all multiple alignment programs were generally successful when the sequences shared >40% residue identity. To create challenging test cases, sequences with >40% identity are therefore removed in a clustering step, based on a multiple sequence alignment constructed with the MAFFT[23] program. This sequence alignment is used uniquely for the removal of very similar, redundant PDB sequences. (3) The reference alignment of the selected, nonredundant sequences is constructed using the SAP 3D superposition program.[36] The SAP program was selected based on a number of reliability/functionality criteria. First, SAP is a reliable program, derived from SSAP,[37] which is the reference program used for the CATH[38] structure classification. Second, SAP is available for local installation and has a command-line interface, facilitating its integration in an automatic protocol. Third, the SAP program provides a sequence alignment based on the structural superposition, together with a reliability score for each pair of aligned residues.)4) The automatic structure alignment is

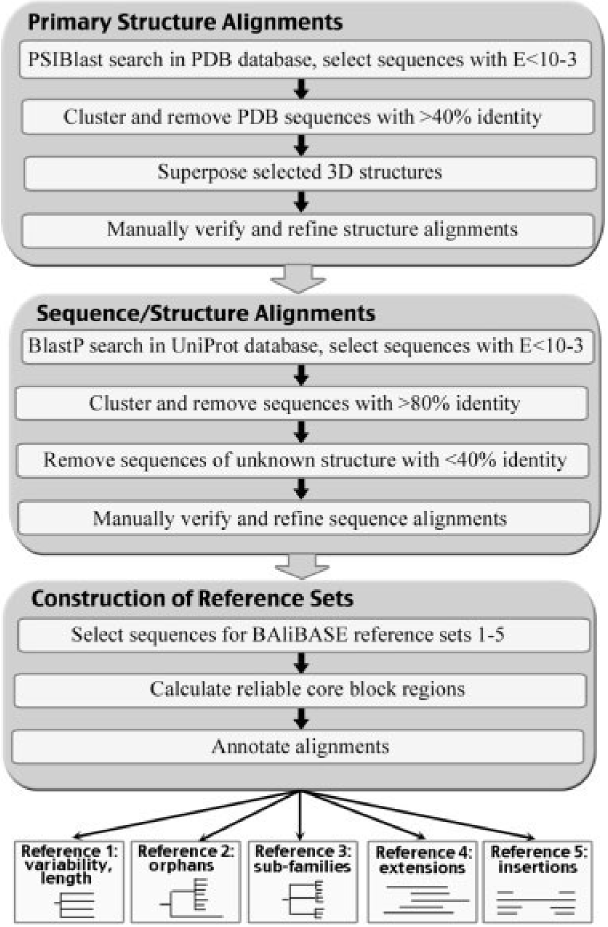
Primary Structure Alignments
PSIBlast search in PDB database, select sequences with E<10-3
Cluster and remove PDB sequences with >40% identity
Superpose selected 3D structures
Manually verify and refine structure alignments
Sequence/Structure Alignments
BlastP search in UniProt database, select sequences with E<10-3
Cluster and remove sequences with >80% identity
Remove sequences of unknown structure with <40% identity
Manually verify and refine sequence alignments
Construction of Reference Sets
Select sequences for BAliBASE reference sets 1-5
Calculate reliable core block regions
Annotate alignments
Reference 1: variability, length
Reference 2: orphans
Reference 3: sub-families
Reference 4: extensions
Reference 5: insertions

Fig. 1. Flow-chart showing the three major steps of the semiautomatic protocol used to construct the BAliBASE reference sets.

then manually verified and refined to assure the alignment of conserved residues.

Sequence–Structure Alignments

The Primary structure alignment for each protein family consists of a relatively small set of divergent sequences from the PDB database, which can be accurately aligned using 3D structure superpositions. To create a larger test set, we incorporate sequences of unknown structure from the Uniprot database. Sequences are included that are more closely related to at least one of the PDB sequences and can therefore be accurately aligned based on the primary sequence alone. The following protocol is used to identify suitable sequences: (1) BlastP[34] searches are performed using each of the PDB sequences in the Primary structure alignment and all sequences detected with $E < 10^{-3}$ are initially selected. (2) Redundant sequences sharing >80% residue identity are removed in a clustering step, similar to that described for the structure alignments above. (3) Because divergent sequences cannot be aligned accurately without using structural information as a guide, Uniprot sequences of unknown structure are only included in the alignment if they share >40% identity with at least one PDB sequence. (4) The automatic sequence alignment is then manually verified and refined to correct any badly aligned sequences or locally misaligned regions.

The complete set of PDB sequences and their related Uniprot sequences constitute the Primary multiple alignment, used in the construction of Reference Sets 1–5.

Construction of Reference Sets

The reference alignments in Reference Sets 1–5 are constructed by automatically selecting the sequences from the Primary multiple alignment that correspond to the criteria defined for each Reference Set.

1. For Reference 1, a set of equidistant PDB sequences is selected, in which any two sequences share <20% identity and sequences having large internal insertions (>35 residues) are excluded. Those protein families for which at least four structures are available are included in the V1 subset of Reference 1. Among the remaining families, those that contain a set of at least four equidistant sequences, in which any two sequences share 20–40% identity, are included in the new V2 subset. Again, sequences having large internal insertions are excluded. For Reference 2, a family is selected in which the sequences all share >40% identity and for which at least one 3D structure is known. “Orphan” sequences are then chosen that share <20% identity with all members in the family. Only PDB sequences are selected as orphans to guarantee an accurate alignment of these highly divergent sequences. For Reference 3, subfamilies are selected such that the sequences within a given subfamily share >40% identity, but any two sequences from different subfamilies share <20% identity. Each subfamily must also contain at least one PDB sequence. For References 1–3, the percent identity is calculated over the homologous region only, and no sequences contain large internal insertions. For References 4 and 5, sequences are selected that share >20% identity with at least one other sequence, but sequences are included that contain large N/C-terminal extensions or internal insertions respectively.
2. For each alignment in BAliBASE, core blocks are defined that correspond to the regions that are reliably aligned. An automatic method has been developed to objectively identify the reliable regions, based on a combination of secondary structure superposition and sequence conservation. The sequence conservation is measured using the NorMD[39] program in a sliding window analysis (window length = 5) along the length of the alignment (Fig. 2). Briefly, a 20-dimensional continuous sequence space is defined and, for a given alignment column, the residues observed in the column are assigned a point S in this space. The weighted mean distance (MD) between each pair of residues is then calculated and the conservation score of the column is defined as the MD score, normalized in the range of 0 to 100. A core block is then defined as a region in the alignment consisting of at least three columns with no gaps, in which either (a) the sequences with known 3D structure all share the same secondary structure (i.e., either all helix or all beta strand) and the NorMD score is >0.1 or (b) the NorMD is >0.2. A stricter NorMD cutoff is used in the second criterion to include regions that are conserved even though they do not correspond to either helix or beta-strand secondary structure elements. The core block definition was validated with reference to the core blocks in BAliBASE version 2.1, which were manually defined. For each of the alignments in BAliBASE version 2.1, References 1–5, the core blocks were recalculated using the new method and compared to the manually defined blocks for the same alignment. The correlation between the two sets of core blocks for a given alignment was calculated as follows:

$$\text{correlation} = \frac{\sum_{i=1}^{n} f(s_i^a, s_i^m)}{n}$$

where n is the length of the alignment, $s_i^a = 1$ if the ith column is defined as being in a core block by the automatic method and 0 otherwise, $s_i^m = 1$ if the ith column is defined as being in a core block by the manual method and 0 otherwise. $f(i,j)$ is a function taking the value of 1 if $I = j$ and 0 otherwise.
3. The final step in the construction process is the incorporation of structural/functional information from external sources for display on the BAliBASE web site alignment pages. First, secondary structure elements are calculated for each PDB sequence using the DSSP program.[40] Second, for the Uniprot sequences, the FT (feature table) lines are parsed for specific entries corresponding to domains, signal sequences, potential transmembrane regions, binding sites, active sites, or posttranslational modifications of a residue. The Pfam[39]

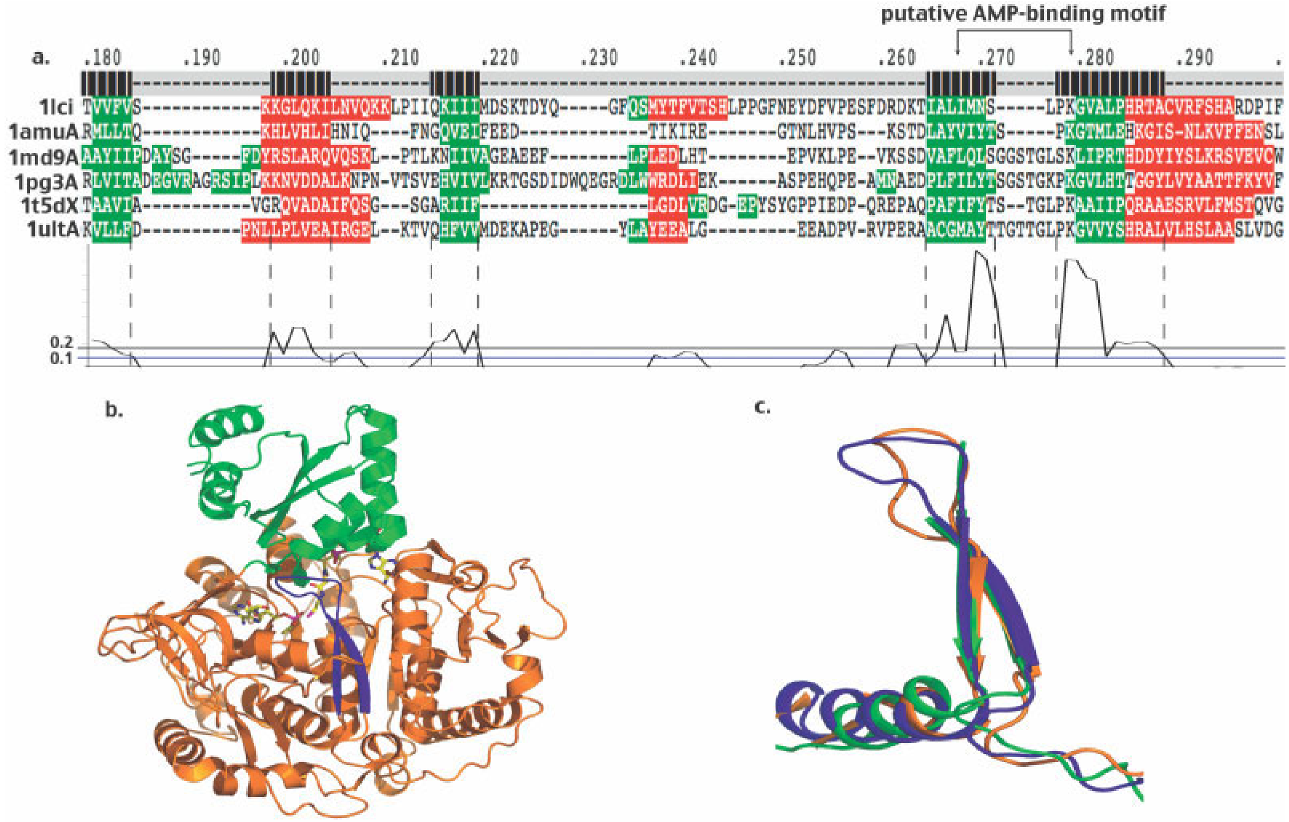
putative AMP-binding motif
a.
.180 .190 .200 .210 .220 .230 .240 .250 .260 .270 .280 .290
1lci TVVFVS-------------KKGLQKILNVQKKLPIIQKIIIMDSKTDYQ-----GFQSMYTFVTSHLPPGFNEYDFVPESFDRDKTIALIMNS-----LPKGVALPHRTACVRFSHARDPIF
1amuA RMLLTQ-------------KHLVHLIHNIQ----FNGQVEIFEED--------------TIKIRE--------GTNLHVPS---KSTDLAYVIYTS-----PKGTMLEHKGIS-NLKVFFENSL
1md9A AAYIIPDAYSG-----FDYRSLARQVQSKL--PTLKNIIVAGEAEEF---------LPLEDLHT---------EPVKLPE--VKSSDVAFLQLSGGSTGLSKLIPRTHDDYIYSLKRSVEVCW
1pg3A RLVITADEGVRAGRSIPLKKNVDDALKNPN-VTSVEHVIVLKRTGSDIDWQEGRDLWWRDLIEK-------ASPEHQPE-AMNAEDPLFILYTSGSTGKPKGVLHTTGGYLVYAATTFKYVF
1t5dX TAAVIA------------VGRQVADAIFQSG---SGARIIF------------------LGDLVRDG-EPYSYGPPIEDP-QREPAQPAFIFYTS--TGLPKAAIIPQRAAESRVLFMSTQVG
1ultA KVLLFD-----------PNLLPLVEAIRGEL--KTVQHFVVMDEKAPEG------YLAYEEALG----------EEADPV-RVPERAACGMAYTTGTTGLPKGVVYSHRALVLHSLAASLVDG
0.2
0.1
b.
c.

Fig. 2. (**a**) Part of a multiple alignment of 6 AMP-binding enzymes, including the putative AMP-binding signature motif. Secondary structure elements are highlighted in red (helix) or green (beta-strand). The norMD column scores used in the determination of the core blocks are shown below the alignment. The corresponding core blocks are indicated by black boxes above the alignment. (**b**) Structure of Acetyl coa synthetase bound to CoA and adenosine-5′-propylphosphate (PDB:1pg3). The larger N-terminal domain is shown in brown and the C-terminal domain is shown in green. The putative AMP-binding loop is highlighted in blue. (**c**) 3D structural superposition of the region containing the AMP-binding signature motif in 1pg3A (blue), 1md9A (brown), 1lci (green). All structure cartoons were produced with Pymol (http://www.pymol.org).

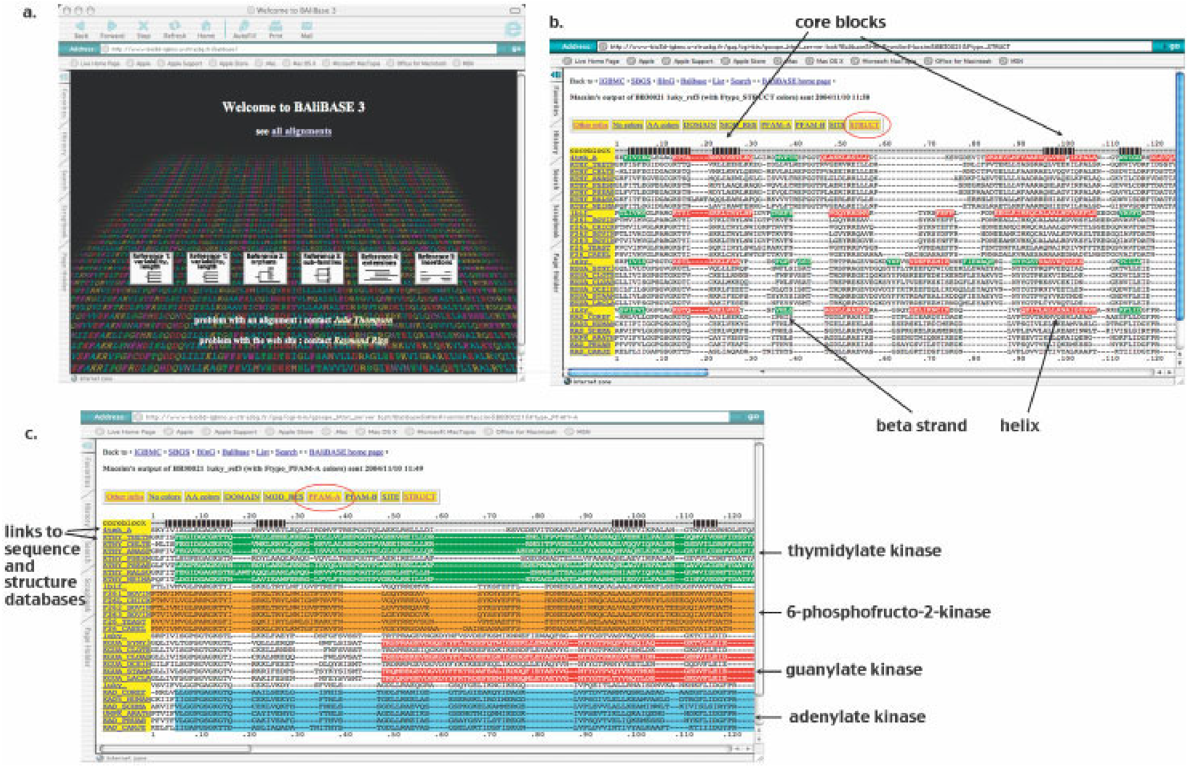

Fig. 3. BAliBASE Web site. (**a**) Entry page with links to each reference set. (**b**) Part of a multiple alignment of four subfamilies of P-loop containing kinases. Secondary structure elements are highlighted in red or green. Core blocks are indicated by black boxes above the alignment. (**c**) The same alignment coloured according to Pfam database entries. Links to the Uniprot and PDB databases are provided by clicking on the sequence names on the left.

database is also searched for protein family domains in the Uniprot sequences in each alignment.

Availability

All the reference alignments in BAliBASE are available for viewing on the WWW at http://www-bio3d-igbmc.u-strasbg.fr/balibase and are also provided in MSF format. The annotated alignments, including the core block definitions, are also provided in XML format. For Reference Sets 1–3, which contain alignments of sequences that are colinear, both the sequences corresponding to the homologous regions only and the full-length sequences are provided. A C program is also provided to estimate the quality of the multiple alignments constructed by different alignment programs compared to the BAliBASE references. Two different alignment scores are calculated. The sum-of-pairs score is the percentage of correctly aligned pairs of residues in the alignment produced by the program. It is used to determine the extent to which the programs succeed in aligning some, if not all, of the sequences in an alignment. The column score is the percentage of correctly aligned columns in the alignment, which tests the ability of the programs to align all of the sequences correctly. The evaluation program requires the Expat XML parser, which is freely available from http://expat.sourceforge.net/. The complete database and the evaluation program are available for downloading by ftp from ftp://www.igbmc.u-strasbg.fr/BioInfo/.

RESULTS AND DISCUSSION

Previous releases of BAliBASE have proved to be a useful benchmark for the evaluation of multiple alignment methods and have been widely adopted by the community.[23,24,42–50] This is most probably due to the hierarchical organization of the reference alignments in a series of Reference Sets, representing many of the problems encountered when performing multiple alignments, such as highly divergent sequences, overrepresentation of certain members of a family, and proteins with large N/C-terminal extensions or internal insertions. Reference Sets are also available for the particular problems of proteins with transmembrane proteins, repeats, or circular permutations. With the development of new techniques for multiple sequence alignment, new benchmarks are now needed. The purpose of this work is to provide larger, more challenging test cases that allow detailed evaluation and statistical comparison of these new alignment methods.

Automatic Update Protocol

One of the drawbacks of the previous releases of BAliBASE was the time and effort required to manually create and annotate the reference alignments. With the current explosion of the sequence and structure databases, the automatic update of the benchmark has become a crucial issue, although a final manual validation step is still required to maintain the high quality of the multiple alignments. For this reason, a number of bioinformatics algorithms and tools have been incorporated in a new semiautomatic update protocol, designed to handle the vast amounts of data now available in the public databases. For each protein family in the benchmark, PSI-Blast is used to perform in-depth searches for family members in the PDB database whose 3D structure is known. Accurate multiple alignments of these proteins can be constructed, even for divergent sequences, based on their 3D structural superposition. These structural alignments, known as Primary structure alignments, are used as the basis for the construction of the different Reference Sets. First, to provide larger test sets, protein sequences from the Uniprot database, whose 3D structure is not yet known, are automatically incorporated in the Primary multiple alignment. In this case, only proteins that are more closely related to at least one of the PDB sequences are included, because more divergent sequences cannot be reliably aligned based only on primary sequence information. Then, Reference Sets 1–5 are constructed by selecting the sequences that satisfy the predefined criteria from the Primary multiple alignments. Finally, core blocks are automatically defined using a newly developed method to identify the reliable regions in the alignments. This method relies on the presence of conserved secondary structures, combined with a sequence conservation score for each position in the alignment. The NorMD objective function is used to estimate the conservation as it provides normalized column scores, with completely conserved columns scoring 1.0. The method was validated by comparison to the core blocks in BAliBASE 2.1, which were manually defined, based on visual inspection of the 3D structural superpositions. The correlation between the manual and automatic methods (see Methods) was calculated to be 77%, with only 5% of the manually defined core blocks excluded by the automatic protocol. The reliability of the alignments in the core blocks has also been estimated in terms of the coordinate root-mean-square (cRMS) of the structural alignments. The mean cRMS in the homologous domains is 4.6, compared to a mean cRMS of 2.7 when only the residues in the core blocks are taken into account. Thus, the core blocks calculated by the automatic method successfully identify the regions in the alignment that can be structurally superposed. The identification of the reliably aligned core blocks is a crucial part of the BAliBASE benchmarking system. The core blocks are designed to exclude the sequence stretches that cannot be accurately aligned, such as loop regions. These nonsuperposable regions often represent a significant proportion of the multiple alignment, particularly in the case of very divergent proteins, and may significantly effect the scores obtained when evaluating a multiple alignment program.

New Challenges

In general, the role of a benchmark is to provide a set of tests to compare the performance of alternative tools or technologies. The benchmark should provide realistic test cases that reproduce the types of problems likely to be experienced in practice. With the evolution of the sequence and structure databases resulting from high throughput technologies, the multiple alignment of large numbers of complex, multidomain sequences has become a standard

TABLE I. BAliBASE Statistics

		Reference 1 Equidistant Sequences			Reference 2 Family with "Orphans"	Reference 3 Divergent Subfamilies	Reference 4 Large Extensions	Reference 5 Large Insertions	Total
		V1 <20%	V2 20–40%	V3 40–60%					
No. of alignments	version2	23	31	28	23	12	12	12	141
	version3	38	45	N/A	41	30	48	16	217
No. of sequences	version2	125	119	123	487	292	150	148	1444
	version3	265	411	N/A	1896	1882	1317	483	6255

The number of alignments and the total number of sequences in each reference set in this version of BAliBASE, compared to the previous release. Reference 1, subset V3 has been excluded from the latest version (N/A = not applicable).

requirement. Sequence alignment benchmarks must now evolve to accurately represent the new requirements, but also to avoid overfitting of the methods to a particular set of test cases. BAliBASE release 3.0 is designed to respond to these challenges. The size of the alignments in the BAliBASE benchmark has been increased in release 3.0 to reflect the ever-growing sequence and 3D structure databases. A comparison between versions 2.1 and 3.0 is provided in Table I. The number of alignments has been increased from 141 to 217 and perhaps more significantly, the number of sequences in the benchmark has been increased from 1444 to 6255. Furthermore, the variability of the sequence test sets has been increased by excluding the most conserved alignments. In a previous comparison of 10 different multiple alignment algorithms, significant differences in the methods were observed in all the Reference Sets, with the exception of Reference 1, subset V3. The sequences in this subset, which are colinear and share 40–60% residue identity, were successfully aligned by most of the algorithms included in the comparison and no difference could be discerned between the performances of the various methods. The remaining subsets, V1 and V2, contain alignments with low and medium variability respectively, that represent a real challenge for today's multiple alignment methods.

Alignment Complexity

Reference Sets 1–3 provide protein domain alignments, designed specifically to test the effect of sequence length and variability, the presence of "orphan" sequences and the overrepresentation of some members of the protein family on alignment quality. The sequences in these alignments are truncated to include only the homologous domains and therefore they do not always correspond to the full-length sequences found in the PDB or Uniprot databases. In BAliBASE 3.0, the alignments of the full-length sequences are now provided, in addition to the truncated version, to test the performance of multiple alignment methods in the presence of "noise" in the form of nonconserved regions in the sequences. In BAliBASE version 3.0, the core blocks represent 39% of the columns in the domain alignments and only 19% of the alignments of the full-length sequences. These full-length sequence sets therefore provide ideal test cases for the evaluation of many different types of alignment techniques, including both local and global alignment algorithms.

Coverage of the Protein Fold Space

Apart from providing test cases for different types of alignment problems, the protein families included in a multiple alignment benchmark should also cover, as far as possible, the protein fold space. One possibility would be to incorporate the complete PDB database in the benchmark; however, the structure set available in the PDB has been shown to be biased.[51] It is therefore desirable to select a subset of protein families that constitute a more balanced population, while still covering as many different fold types as possible. This approach has been successfully adopted for benchmarks used for fold recognition and 3D superposition.[52–54] To demonstrate the coverage of BAliBASE, we assessed the protein families in each Reference Set with reference to the SCOP fold classification (Table II), although a direct comparison with SCOP is difficult, because BAliBASE contains full-length protein sequences that often correspond to more than one structural domain. According to the SCOP classification, there are ~700 different folds currently known, although many of these are unique proteins. In order to identify potential protein families that verify the selection criteria for BAliBASE, one PDB sequence was selected from each of the 700 SCOP fold classifications and, for each sequence, the first part of the automatic BAliBASE update protocol was performed, that is, the PDB database was searched for similar sequences, a multiple alignment was constructed and sequences sharing >40% identity were removed. Of the 700 protein families, only 102 contained at least 4 sequences sharing <40% identity, which is the minimum number of sequences required to create a BAliBASE test case. Of these 102 potential families with >4 known structures, 38 (37%) have been included in subset V1, and a further 45 (44%) are included in subset V2. As can be seen in Table II, the five main classes in the SCOP classification are represented in each of the Reference Sets 1–5. One important class of proteins, namely the SCOP multidomain fold class is new to this release. As an example, Figure 2(a) shows part of an alignment of six AMP-forming enzymes (SCOP classification e.23.1) from Reference 1, subset V1. The AMP-forming family includes acyl-CoA synthetases (PDB:1pg3A, 1ultA), firefly luciferase (PDB: 1lci), 2,3-dihydroxybenzoate-amp ligase (PDB:1md9), 4-chlorobenzoyl-coa ligase (PDB:1t5dX) and the phenylalanine adenylation domain of gramicidin synthetase (PDB:1amuA).

TABLE II. Comparison with the SCOP Protein Structure Classification

SCOP class	SCOP Total	SCOP Subset	Reference 1 Equidistant Sequences V1:<20%	Reference 1 Equidistant Sequences V2:20–40%	Reference 2 Family with "Orphans"	Reference 3 Divergent Subfamilies	Reference 4 Large Extensions	Reference 5 Large Insertions
All alpha (a)	179	30	6	6	7	7	7	1
All beta (b)	126	21	7	11	7	3	11	3
a/b	121	28	10	11	19	13	13	9
a+b	234	21	10	9	12	8	12	4
Multidomain	38	2	2	4	3	3	5	1
Total	698	102	35	41	48	34	48	18

The number of BAliBASE alignments containing domains with a given SCOP fold class. The "SCOP total" column represents the total number of superfamilies in the SCOP database in each SCOP class. The "SCOP subset" column gives the number of superfamilies in the SCOP database satisfying the BAliBASE reference 1 criteria, that is, having at least four members with <40% identity.

The region shown in the multiple alignment includes the putative AMP-binding domain signature; a region rich in glycine, serine and threonine, followed by a conserved lysine, as defined in the PROSITE database[55] (PDOC00427). The proteins share a large N-terminal domain and an ~110-residue C-terminal domain, shown in Figure 2(b). Figure 2(c) shows the structural superposition of the loop supposedly involved in AMP-binding. The equivalence list for the pairwise structural alignments was constructed from the BAliBASE reference alignment, based on the method of Kabsch[56,57] using in-house software (P. Koehl, unpublished results).

Web Interface and Output File Formats

For this new release of the BAliBASE benchmark, the Web site has been completely redesigned. The top page allows entry directly to the different Reference Sets [Fig. 3(a)]. The alignments are shown in a scrolling window and can be colored according to a number of different criteria, for example, residue conservation, secondary structure elements, Pfam domains, or functional sites. The example display in Figure 3 shows part of an alignment of 4 subfamilies of P-loop-containing kinases from Reference 3. In this case, although the subfamilies correspond to different Pfam entries, they all share the same 3D structural fold (SCOP classification c.37.1). The conserved core blocks representing the reliable regions are displayed above the multiple alignment. The first core block contains the Walker motif A GxxxxGK[ST] located at the end of the first beta strand and including the first half-turn of the following alpha helix in the sequence, which corresponds to the phosphate binding loop (P-loop), a nucleotide binding site present in many ATPase or GTPase activity-exhibiting proteins.[58] The Walker B motif hhhhD is located in core block 4 at columns 116–120. Links to the Uniprot entries for each sequence are available by clicking on the sequence names on the left-hand side of the display.

In addition to the interactive Web-based display, all the multiple alignments in BAliBASE are available in a new standard data exchange XML format. The XML format allows parsing of the alignments and the associated annotations by processing applications that can then be incorporated in automatic evaluation and comparison systems.

CONCLUSIONS

The field of multiple sequence alignment is currently evolving, with the development of new, more sophisticated algorithms designed to cope with the large amounts of complex information now available in the protein sequence and 3D structure databases. New multiple alignment benchmarks are now required to keep up with these developments and to avoid optimization of the tools on a particular set of tasks. Here, we have presented the latest release of BAliBASE, the most widely used benchmark specifically designed for the evaluation of multiple sequence alignment programs.

A number of important factors have influenced the development of this latest release. First, the tests in the benchmark are designed to represent the tasks that multiple alignment tools or techniques are now expected to solve in the postgenomic era. To reflect the recent explosion of the sequence and structure databases, the Reference Sets in BAliBASE 3.0 have been increased in terms of both the number of alignments and the number of sequences in each alignment and the benchmark now covers most of the known protein fold space. Second, the complexity of the alignments has also been significantly increased with the addition of alignments containing full-length sequences for all the Reference Sets. These full-length alignments provide a large number of difficult tests for both global and local alignment algorithms. Third, a semiautomatic update protocol has been developed that facilitates the incorporation of new protein families and new Reference Sets and will allow more frequent releases of the benchmark in the future. Finally, the benchmark needs to be easy to obtain and to use, otherwise few people will be likely to use it. Therefore, all BAliBASE alignments and associated annotations are freely available on the WWW or by ftp. Furthermore, the alignments and their associated annotations are now available in a standard data exchange XML format that should facilitate the development of automatic procedures for the evaluation and comparison of new multiple alignment methods.

The widespread adoption of a benchmark can produce technically interesting results, but perhaps more significantly, it can lead to a better understanding of the research problem and the tools and techniques being developed. A previous study of multiple alignment algorithms based on BAliBASE 1.0 identified some of the strong and weak points of the 10 different algorithms compared. For example, although the iterative techniques were shown to be generally more accurate than progressive alignment algorithms, in certain specific cases, the iterative algorithms lead to errors. Another important result was the observation that local and global alignment methods were complementary techniques that could be used cooperatively to improve the accuracy and reliability of multiple alignments. We hope this new release of BAliBASE will lead to equally exciting developments in the future. One such development that is envisaged is a cooperative, knowledge-based, diagnostic process that detects the presence of specific alignment problems and selects the most appropriate algorithm for the alignment task.

ACKNOWLEDGMENTS

We would like to thank Luc Moulinier and Yann Brelivet for stimulating discussions and Jean Claude Thierry and Dino Moras for their continued support.

REFERENCES

1. del Sol Mesa A, Pazos F, Valencia A. Automatic methods for predicting functionally important residues. J Mol Biol 2003;326: 1289–1302.
2. Phillips A, Janies D, Wheeler W. Multiple sequence alignment in phylogenetic analysis. Mol Phylogenet Evol 2000;16:317–330.
3. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJ, Vaughan R, Zdobnov EM. The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 2003;31:315–318.
4. Al-Lazikani B, Jung J, Xiang Z, Honig B. Protein structure prediction. Curr Opin Chem Biol 2001;5:51–56.
5. Lecompte O, Thompson JD, Plewniak F, Thierry J, Poch O. Multiple alignment of complete sequences (MACS) in the post-genomic era. Gene 2001;270:17–30.
6. Feng DF, Doolittle RF. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol 1987;25:351–360.
7. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and matrix choice. Nucleic Acids Res 1994;22;4673–4680.
8. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997;22:4673–4680.
9. Smith RF, Smith TF. Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. Protein Eng 1992;5:35–41.
10. Eddy SR. Multiple alignment using hidden Markov models. ISMB 1995;3:114–120.
11. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;10:846–856.
12. Notredame C, Higgins DG. SAGA: sequence alignment by genetic algorithm. Nucleic Acids Res 1996;24:1515–1524.
13. Morgenstern B, Dress A, Werner T. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proc Natl Acad Sci USA 1996;3:12098–12103.
14. Gotoh O. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. J Mol Biol 1996;264:823–838.
15. Bianchetti L, Thompson JD, Lecompte O, Plewniak F, Poch O. vALId: validation of protein sequence quality based on multiple alignment data. JBCB Bioinformatics 2005. Forthcoming.
16. Heringa J. Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. Comput Chem 1999;23:341–364.
17. Jennings AJ, Edge CM, Sternberg MJ. An approach to improving multiple alignments of protein sequences using predicted secondary structure. Protein Eng 2001;14:227–231.
18. Thompson JD, Plewniak F, Thierry JC, Poch O. DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. Nucleic Acids Res 2000;28:2919–2926.
19. Notredame C, Higgins DG, Heringa J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000;302:205–217.
20. Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. Bioinformatics 2002;18:452–464.
21. Morgenstern B. DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. Nucleic Acids Res 2004;32:W33–W36.
22. Do CB, Brudno M., Batzoglou S. PROBCONS: probabilistic consistency-based multiple alignment of amino acid sequences. Genome Res 2005;2:330–340.
23. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 2002;30:3059–3066.
24. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 2004;5: 113.
25. Thompson JD, Plewniak F, Poch O. BaliBASE: A benchmark alignment database for the evaluation of multiple sequence alignment programs. Bioinformatics 1999;1:87–88.
26. Bahr A, Thompson JD, Thierry JC, Poch O. BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. Nucleic Acids Res 2001;1:323–326.
27. Thompson JD, Plewniak F, Poch O. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res 1999;27:2683–2690.
28. Raghava GP, Searle SM, Audley PC, Barber JD, Barton GJ. OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. BMC Bioinformatics 2003;4:47.
29. Walle IV, Lasters I, Wyns L. SABmark—a benchmark for sequence alignment that covers the entire known fold space. Bioinformatics. 2004.
30. Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci 1998;**7**:2469–2471.
31. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32:D226–D229.
32. Bourne PE, Addess KJ, Bluhm WF, Chen L, Deshpande N, Feng Z, Fleri W, Green R, Merino-Ott JC, Townsend-Merino W, Weissig H, Westbrook J, Berman HM. The distribution and query systems of the RCSB Protein Data Bank. Nucleic Acids Res 2004;32:D223–D225.
33. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS. UniProt: the universal protein knowledgebase. Nucleic Acids Res 2004;32:D115–D119.
34. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;17: 3389–3402.
35. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF.Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 2001;14:2994–3005.
36. Taylor WR. Protein structure comparison using SAP. Methods Mol Biol 2000;143:19–32.
37. Orengo CA, Taylor WR. SSAP: sequential structure alignment program for protein structure comparison. Methods Enzymol 1996;266:617–635.

38. Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grant A, Lee D, Akpor A, Maibaum M, Harrison A, Dallman T, Reeves G, Diboun I, Addou S, Lise S, Johnston C, Sillero A, Thornton J, Orengo C. The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. Nucleic Acids Res 2005;33:D247–D251.
39. Thompson JD, Plewniak F, Ripp R, Thierry JC, Poch O. Towards a reliable objective function for multiple sequence alignments. J Mol Biol. 2001:937–951.
40. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;12:2577–2637.
41. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C, Eddy SR The Pfam protein families database. Nucleic Acids Res 2004;32:D138–D141.
42. Wang Y, Li KB. An adaptive and iterative algorithm for refining multiple sequence alignment. Comput Biol Chem 2004;2:141–148.
43. Marsden B, Abagyan R. SAD—a normalized structural alignment database: improving sequence-structure alignments. Bioinformatics 2004;15:2333–2344.
44. Nguyen HD, Yoshihara I, Yamamori K, Yasunaga M. Aligning multiple protein sequences by parallel hybrid genetic algorithm. Genome Inform Ser Workshop Genome Inform. 2002;13:123–132.
45. Althaus E, Caprara A, Lenhof HP, Reinert K. Multiple sequence alignment with arbitrary gap costs: computing an optimal solution using polyhedral combinatorics. Bioinformatics 2002;Suppl 2:S4–S16.
46. Lassmann T, Sonnhammer EL. Quality assessment of multiple alignment programs. FEBS Lett 2002;1:126–130.
47. Heringa J. Local weighting schemes for protein multiple sequence alignment. Comput Chem 2002;5:459–477.
48. Holmes I, Rubin GM. An expectation maximization algorithm for training hidden substitution models. J Mol Biol 2002;5:753–764.
49. Holmes I, Bruno WJ. Evolutionary HMMs: a Bayesian approach to multiple alignment. Bioinformatics 2001;9:803–820.
50. Karplus K, Hu B. Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. Bioinformatics 2001;8:713–720.
51. Peng K, Obradovic Z, Vucetic S. Exploring bias in the Protein Data Bank using contrast classifiers. Pac Symp Biocomput 2004:435–446.
52. Domingues FS, Lackner P, Andreeva A, Sippl MJ. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. J Mol Biol 2000;4:1003–1013.
53. Panchenko AR, Marchler-Bauer A, Bryant SH.Combination of threading potentials and sequence profiles improves fold recognition. J Mol Biol 2000;5:1319–1331.
54. Sommer I, Zien A, von Ohsen N, Zimmer R, Lengauer T. Confidence measures for protein fold recognition. Bioinformatics 2002; 6:802–812.
55. Hulo N, Sigrist CJA, Le Saux V, Langendijk-Genevaux PS, Bordoli L, Gattiker A, De Castro E, Bucher P, Bairoch A. Recent improvements to the PROSITE database. Nucleic Acids Res 2004;32:D134–D137.
56. Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. Acta Crystallogr 1978;A34:827–828.
57. McLachlan AD. Gene duplications in the structural evolution of chymotrypsin. J Mol Biol 1979;1:49–79.
58. Saraste M, Sibbald PR, Wittinghofer A. The P-loop—a common motif in ATP- and GTP-binding proteins. Trends Biochem Sci 1990;11:430–434.