Published online 4 January 2010 *Nucleic Acids Research, 2010, Vol. 38, No. 7* ***2145–2153***
doi:10.1093/nar/gkp1196

Quality measures for protein alignment benchmarks

Robert C. Edgar*

Tiburon, CA 94920, USA

Received October 28, 2009; Revised November 19, 2009; Accepted December 8, 2009

ABSTRACT

Multiple protein sequence alignment methods are central to many applications in molecular biology. These methods are typically assessed on benchmark datasets including BALIBASE, OXBENCH, PREFAB and SABMARK, which are important to biologists in making informed choices between programs. In this article, annotations of domain homology and secondary structure are used to define new measures of alignment quality and are used to make the first systematic, independent evaluation of these benchmarks. These measures indicate sensitivity and specificity while avoiding the ambiguous residue correspondences and arbitrary distance cutoffs inherent to structural superpositions. Alignments by selected methods that indicate high-confidence columns (ALIGN-M, DIALIGN-T, FSA and MUSCLE) are also assessed. Fold space coverage and effective benchmark database sizes are estimated by reference to domain annotations, and significant redundancy is found in all benchmarks except SABMARK. Questionable alignments are found in all benchmarks, especially in BALIBASE where 87% of sequences have unknown structure, 20% of columns contain different folds according to SUPERFAMILY and 30% of 'core block' columns have conflicting secondary structure according to DSSP. A careful analysis of current protein multiple alignment benchmarks calls into question their ability to determine reliable algorithm rankings.

INTRODUCTION

Multiple protein sequence alignments (MPSAs) are ubiquitous in molecular biology. Automated alignment is an essential step in a wide-range of applications from phylogeny inference to function prediction. Development of MPSA algorithms is an active area of research, and many MPSA programs are available. Validation of these methods is required for assessment of new algorithms and for biologists to make informed decisions about which programs to use. Validation of MPSAs has a long history (1), and recent work has tended to focus on comparison with reference alignments in benchmarks such as BALIBASE (2–4). In the following, I describe new methods for alignment quality assessment and apply them to reference alignments in the benchmark databases and to selected programs. In this article, I focus exclusively on datasets that are (i) constructed explicitly for multiple alignment validation and (ii) based on biological data; this excludes simulated data such as IRMBASE (5) and alignments such as HOMSTRAD (6) that have been used for MPSA assessment but were not designed for this purpose. For a previous benchmark comparison, see (7).

ALIGNMENT CORRECTNESS

Generally, alignments are intended to indicate residues that are homologous or structurally equivalent and it should be noted that in more challenging cases these criteria will not always agree. Structures can be compared independently of sequence and can therefore be used for sequence alignment assessment. However, structural alignments become ambiguous as structures diverge and different structural alignment methods may disagree with each other at the scale of individual residues (8,9). Structure therefore has intrinsic limitations as a standard for residue alignment.

Over- and under-alignment

Presumably, local alignment methods have a tendency to under-align, i.e. to fail to align some residues or regions that are homologous or structurally similar, and global methods to over-align, i.e. to align regions that are not homologous or are structurally dissimilar. A measurable definition of alignment correctness is required in order to quantify these issues. It is natural to require that residues are aligned if and only if they are homologous, but residue homology cannot be experimentally determined without sequence comparison and therefore cannot be used for assessment. An alternative is to require residues to be aligned only if they are structurally equivalent, but there are no unique criteria for structural similarity of

*To whom correspondence should be addressed. Tel: +1 415 819 6005; Email: bob@drive5.com individual residues; rather, structural equivalence must be defined in the context of a particular alignment protocol and requires arbitrary parameters such as distance cutoffs. While attempts have been made to measure under- and over-alignment by direct comparison with structural alignments (10–13), it is not reasonable to expect sequence methods to reproduce the arbitrary boundary between structural similarity and dissimilarity produced by a particular structural alignment protocol.

DOMAINS

Proteins are assembled from independently folding units known as domains. Domains can be grouped into superfamilies, which have clear evidence of homology, and folds, which have similar secondary structure arrangements but lack convincing evidence for homology. Domains assigned to the same fold but different superfamilies may be distantly related, or alternatively exhibit a favorable conformation discovered multiple times by convergent evolution. In the context of MPSA benchmarking, it is reasonable to consider regions to be potentially alignable at the scale of individual residues only if they belong to the same superfamily. It is sometimes reasonable to attempt alignments of similar folds, e.g. when attempting structure prediction by threading. However, when homology is uncertain, structural superpositions are ambiguous and sequence similarity is low enough to be caused by chance. Fold-level alignments are therefore not suitable as trusted references.

Assessment by domain classification

SCOP (14) and CATH (15) annotate domains in solved protein structures and the evolutionary relationships between them. These annotations can be transferred to a sequence alignment, and the evolutionary relationships implied by the alignment can be compared with those reported by the domain databases. While domain annotations cannot determine whether individual residues are correctly aligned, they can be used to identify columns that are consistent or inconsistent with being correct. For example, if all residues in a column belong to the same superfamily this supports correctness, and conversely if it contains different folds it is likely to be incorrect. However, domain boundaries are not well-defined and different experts and different automated methods may disagree on how a protein should be divided into domains and whether evidence for homology between domains is convincing (16). In some cases, a secondary structure element may be assigned to the preceding or following domain. Measures based solely on domain annotations can therefore be informative, but not definitive.

Assessment by secondary structure

Alignment of different secondary structures, e.g. an alpha helix to a beta strand, is generally incorrect. This suggests annotating sequences according to a secondary structure classification and examining the agreement of each column. For this article, I chose the widely-used DSSP algorithm (17), which assigns residues to the following classes: ‘H’ (alpha helix), ‘B’ (isolated beta-bridge), ‘E’ (extended strand in beta ladder), ‘G’ (3/10 helix), ‘I’ (pi helix), ‘T’ (H-bonded turn), ‘S’ (bend) and ‘L’ (loop/irregular). As with domains, measures of secondary structure agreement should be considered informative rather than definitive. Different classifications are possible (18), and different methods may disagree (19). Also, depending on the precise definition of correctness, disagreement does not necessarily indicate an alignment error. However, in regions of alignments considered reliable enough to be used in assessment there should generally be consistent secondary structure, and this is the explicit intent for BALIBASE ‘core blocks’.

BENCHMARK DATASETS

BALIBASE is the most widely used benchmark. Version 3.0 contains 218 reference alignments that are constructed by a combination of structure and sequence methods with manual refinement. Most alignments (168) are provided in two versions: one where sequences are described as ‘truncated to homologous regions’ (which I shall call *trimmed*) and one described as having ‘full-length sequences’ (*untrimmed*). Some columns in the reference alignments (‘core blocks’) are annotated as ‘reliably aligned’ while the remaining regions are described as ‘ambiguous’. Two measures of accuracy are defined: the sum of pairs score (SPS), the fraction of aligned letter pairs in the reference alignment that are correctly reproduced in the tested alignment, and the column score (CS), the fraction of aligned columns that are correctly reproduced. The included *bali_score* program computes SPS and CS by comparison either with core blocks only or with all columns, which implies that non-core-block columns could be considered appropriate for assessment. PREFAB (20) was created using a fully automated protocol starting from an *ad hoc* set of pairs of related structures taken from several published sources. For each pair, up to 50 similar sequences were added by making a PSI-BLAST (21) search and selecting a subset with reduced redundancy. Reference alignments were created by aligning each structure pair using CE (22) and FSSP (23) and identifying the set of residue pairs on which the two structural alignments agreed. In version 4.0 there are 1681 reference alignments. Accuracy of a multiple alignment is measured on the structure pair alone by Q, the fraction of letter pairs that agree with both CE and FSSP. The accuracy of one pair of sequences is assumed to correlate with the accuracy of the whole alignment, and methods are ranked by averaging Q over all sets. While using a single pair introduces an unknown error into each measurement, it should be noted that measurements on selected columns, e.g. core blocks, must also be assumed to correlate only approximately with the accuracy of the complete alignment. Core block agreement presumably tends to over-estimate accuracy by measuring only regions with relatively strong conservation which are therefore more easily aligned by sequence similarity. SABMARK v1.65 (11) was also constructed using an automated protocol, but employed a more systematic

```
(1) d1a8l_2  DVRILVFVT...PT-...CPYCPLAVRMAHKFAIENTKAGKG.KILG
(2) d1a8l_1  PVKLIVFVR...KDH...CQYCDQLKQLVQELSELT-----D.KLSY
(3) d1a8y_2  EIKLIGYFK------NKDSEHYKAFKEAAEEFH----.....PYIPF
(4) d1a8l_2  DVRILVFVTPTCPYC---PLAVRMAHKFAIENTKAGK.....GKILG
```

Figure 1. Inconsistent SABMARK reference alignment. This typical example is from Superfamily set 198. Rows (1,2) are a segment from the reference alignment for the pair (d1a82_2, d181_1), rows (2,3) are the reference alignment for (d181_1, d18y_2), and rows (3,4) are the reference alignment for (d18y_2, d1a81_2). Dashes indicate gaps in the reference alignments, dots are gaps added as needed to align the reference alignments to each other via d1a8y_2. Note conflicts between the direct alignment (1,2) and transitive alignment (2,4) of (d1a82_2, d181_1) indicated by shaded letters in d1a81_2.

method for selecting sequences based on the ASTRAL database (24). All sequences have known structure, and reference alignments for all pairs were constructed as a consensus of CE and SOFI (25). These pair-wise alignments are not in general consistent with a multiple alignment (Figure 1). Structural alignments of distantly related proteins are inherently ambiguous, so inconsistency is expected and underscores the difficulty of defining alignment correctness by reference to structure. Accuracy scores are f_D (equivalent to SPS and Q), and f_M, the fraction of letter pairs in the test alignment that are correctly aligned in the reference, which penalizes over-alignment and rewards under-alignment with short correct regions (e.g. a test alignment consisting of a single correct column would achieve $f_M = 1$). Penalizing over-alignment implies that false positives can be meaningfully identified, but SABMARK fails to do this. The structural alignments are not consistent, i.e. contradict each other on putative true positives, and the distinction between alignable and non-alignable regions is based on arbitrary parameters in the structural alignment methods. SABMARK is divided into two subsets. Twilight Zone sets have an effective BLAST (26) *e*-value >1 for all pairs, which selects for relationships that are exceptionally difficult to detect by sequence similarity, in contrast to typical alignment problems which are biased in the opposite direction towards detectable similarity. The name 'Twilight Zone' is misleading as the sequence pairs have 12% average pair-wise identity, while the twilight zone for alignment is usually considered to be the range from 20% to 35% identity (27). OXBENCH (28) contains 672 multiple structure alignments generated by STAMP (29) from structures in the 3Dee database (30).

METHODS

Residue accuracy measures

Several measures of residue accuracy have been proposed. Most are based on comparison of an evaluated (*test*) alignment with a reference alignment (4,13,31), though alternatives have been suggested that compute measures on a structural superposition implied by the test alignment (28,32,33). In this section, a 'column' is understood to include all sequences in a multiple alignment. Measures based on a reference alignment vary depending on (i) whether all columns in the reference alignment are considered or a well-conserved subset of columns, (ii) whether aligned letter pairs or columns are counted as units and (iii) whether gapped positions are considered. Published measures that count gapped positions (12,28,31) can be misleading, primarily because the position of the gap is not considered, so a radically misplaced gap (e.g. terminal versus internal) can give a positive contribution to accuracy. If the reference alignment has long gaps, relatively high accuracy can be obtained even if many residues are misaligned. Also, by most definitions, gaps are uniquely fixed by specifying residue correspondences and introduce a biased double-counting. In well-conserved regions, different definitions of alignment correctness will tend to agree, while in more variable regions definitions may disagree and correctness is not uniquely definable by structure. Measures that consider test alignment letter pairs as units rather than reference columns, such SPS or Alignment Metric Accuracy (31), are less sensitive to alignment errors caused by a single misaligned sequence or correctly aligned subsets that are misaligned to each other (Figure 2). Consider, for example, a set S of 100 well-annotated sequences that are closely related. A distantly related sequence R is aligned to S in order to infer the location of critical residues that are known in S but unknown in R. If S is correctly aligned to itself but misaligned to R, then the critical residue assignments will be incorrect. In this case, a sum-of-pairs measure considers the alignment to be 98% correct and a column measure considers the alignment to be 0% correct. Which alignment accuracy measure is most informative depends on how the alignment will be used, but if a multiple alignment is required it is conservative to assume that correct columns are more important biologically than correct pairs of letters. These considerations suggest that a measure for general-purpose assessment should: (i) consider reference columns as units, (ii) should not count gaps and (iii) should be restricted to conservative columns. The simplest such measure is CS on structurally conserved regions, which will tend to correlate better with correctness by a wide range of definitions. It should be noted that CS is effectively a sensitivity measures, and when restricted to highly conserved regions it is expected to systematically over-estimate the true sensitivity. Also, a single residue error invalidates an entire column, so CSs for sets with different numbers of sequences are not directly comparable.

```
1aab_   MSAKEKGKFEDMAKADK--------------------------ARYEREMTKY
1j46_A  LTEAEKWPFFQEAQKLQAMHREKYPNYYTRP---RRKAKMLPK----------
1k99_A  LPEKKKMKYIQDFQREK--------------------------QEFERNLARF
2lef_A  LSREEQAKYYELARKERQLHMQLYPGWSARDNYGKKKKRKREK----------
        *****************!!!!!!                    !!!!!!
```

Figure 2. FSA alignment of BALIBASE set BB11001. The figure shows a segment of the FSA alignment of BB11001. Core block columns (yellow) that agree with the reference alignment are indicated by asterisks. Columns marked exclamatory are core block columns that have been split into two by FSA. Measures that consider columns in the test alignment as units, such as CSF, may assign this alignment a perfect score, and measures that consider pairs of sequences, such as SPS, may assign it a high score despite this problem, which is common in FSA and DIALIGN. Measures that are more sensitive to this type of problem include CS, which treats a reference column as a unit, and 'gappiness', which can be defined as the average fraction of gapped sequences per column.

Domain annotations

Domain annotations for structures in the benchmarks were obtained from SCOP and CATH. Most BALIBASE sequences (87%) do not have solved structures, and for these I used SUPERFAMILY (34), a library of hidden Markov models that generates annotations of predicted SCOP domains.

Correctness and error measures

Given an alignment and domain annotations of the sequences, agreement measures are defined using the following variables.

C_2 Number of columns containing at least two annotated letters.
C_+ Number of columns for which all letters are annotated.
S_S Number of columns for which all letters are in the same superfamily.
S_F Number of columns for which all letters are in the same fold.
D_S Number of columns containing a pair of letters in two different superfamilies.
D_F Number of columns containing a pair of letters in two different folds.
D_C Number of columns containing a pair of letters in two different classes.

Columns may be only partially annotated. One conflict is sufficient to establish that a column is not correct. Partially annotated columns are probably biased towards more variable regions and more distant or absent relationships, i.e. towards columns that are more likely to be incorrect, and it is therefore conservative to measure annotation agreement ('correctness') only on columns in which all letters are annotated. With these considerations in mind, correctness and error measures are defined as follows.

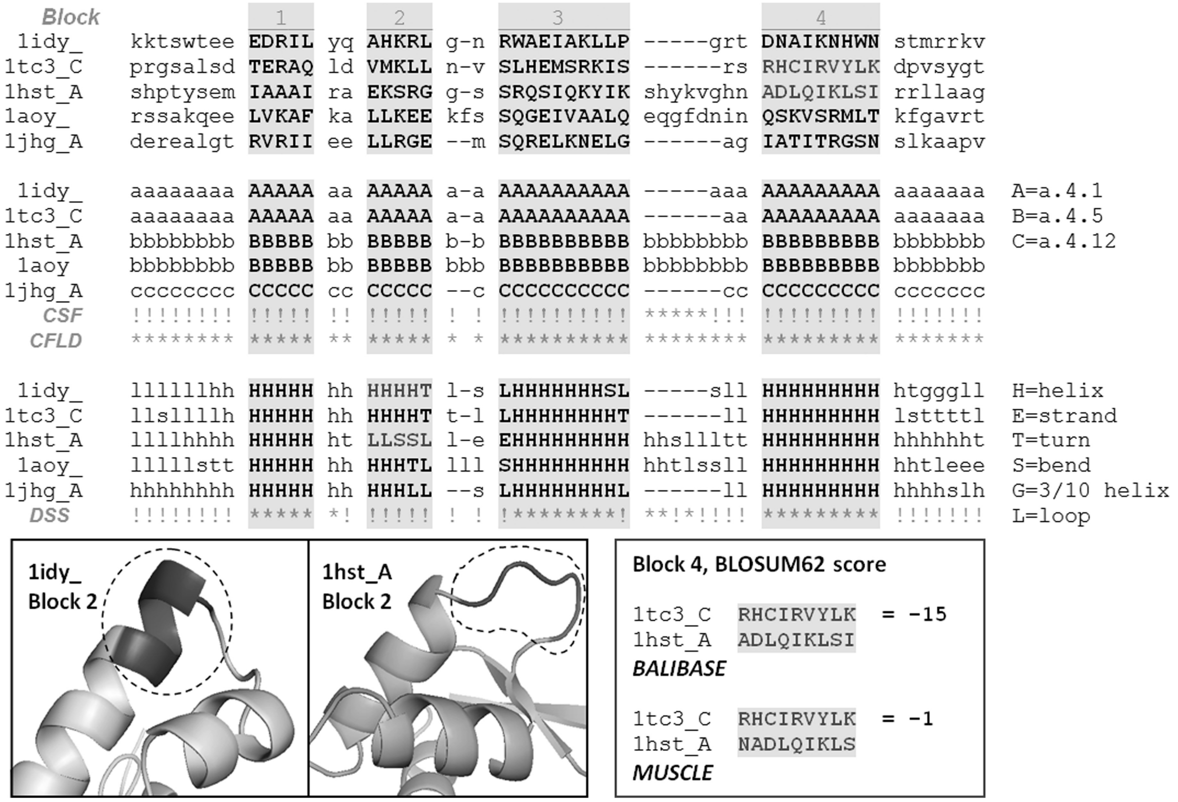
$CSF = S_S/C_+$ Correct by Superfamily
$CFLD = S_F/C_+$ Correct by Fold
$ESF = D_S/C_2$ Error rate by Superfamily
$EFLD = D_F/C_2$ Error rate by Fold
$ECLS = D_C/C_2$ Error rate by Class

Coverage (sensitivity) measures are defined as follows. Intuitively, the goal is to measure the fraction of residues belonging to a given domain that are aligned to other residues belonging to the same domain. The *depth* (d) of a label (superfamily, fold or class) is the number of sequences containing at least one letter with that label, and a column is *full* with respect to a label if it has d letters. *Sensitivity by Area* (SA) is then defined as the total number of letters in columns with $CSF = 1$ divided by the number T of labeled letters, and *Sensitivity by Area Depth* (SAD) is the number of letters in full columns divided by T. Note that if sequences have varying numbers of letters in a given domain then there will necessarily be columns that are not full and all possible alignments will have $SAD < 1$. The DSSP disagreement score, DSS, is defined as the fraction of columns having two or more annotated letters that contain at least two classes (Figure 3).

RESULTS

Reference alignments

The domain and secondary structure agreement measures were applied to reference alignments in the benchmarks, with results shown in Table 1. In BALIBASE, 29% of core block columns contain conflicting DSSP secondary structure assignments, and more than 10% of core blocks columns contain residues assigned to different superfamilies by SUPERFAMILY. Core block columns do not reliably correspond to conserved secondary structures in buried protein cores: 63% of core blocks contain at least one loop residue, many of which have high solvent accessibility, and 36% of core blocks contain at least one loop residue aligned to regular secondary structure (H, B, E, G or I). Outside core blocks, 20% of BALIBASE columns contain residues belonging to different SUPERFAMILY folds, and 70% of columns contain conflicting DSSP secondary structure assignments. This suggests that many non-core block columns and a significant fraction of core blocks are incorrectly aligned. This is confirmed by detailed examination of sets with low scores such as BB20008 and BB40011 (Figure 4). BB40011 has 39 sequences, the majority of which (28) are also found in BB20008, and has ECLS = 46% by SUPERFAMILY. Twelve sequences in BB40011 and ten in BB20008 are SH2 domain structures. The 'full-length' sequences of the structures are isolated SH2 domains rather than complete proteins. The remaining sequences in these two sets are full-length proteins of unknown structure from Uniprot. Core blocks for these sets are found in the SH2 domains, and the trimmed set BBS20008 is truncated to SH2 domains. Some domains flanking SH2 are not homologous, but are aligned to each other by the untrimmed reference alignments. For example, CSK_CHICK contains an SH3 domain at the N terminal, which is not homologous to the tandem SH2 in CSW_DROME (35). C terminals contain the following domains: SOCS box (in CISH_CHICK), protein kinase (in CSK_CHICK), and tyrosine phosphatase (in CSW_DROME). The SOCS box domain is natively disordered but folds into a helical conformation when bound (36,37). It is thus radically different from tyrosine phosphatase, which is a large, stable domain (38). The BB4 sets are described as having 'long terminal extensions', a description that is equally applicable to BB20008 and to many others outside BB4. Other BALIBASE examples are described in the Supplementary Data. OXBENCH alignments have relatively high scores by all measures in regions annotated as structurally conserved (SCRs). However, outside SCRs there is a high level of secondary structure conflict (59%), showing that, as for BALIBASE, a comparison with all columns in the reference alignments should not be expected to reliably assess prediction of either structural similarity or per-residue homology by an MPSA method.



```
Block                  1          2          3                          4
 1idy_     kktswtee EDRIL yq AHKRL g-n RWAEIAKLLP -----grt DNAIKNHWN stmrrkv
1tc3_C     prgsalsd TERAQ ld VMKLL n-v SLHEMSRKIS ------rs RHCIRVYLK dpvsygt
1hst_A     shptysem IAAAI ra EKSRG g-s SRQSIQKYIK shykvghn ADLQIKLSI rrllaag
1aoy_      rssakqee LVKAF ka LLKEE kfs SQGEIVAALQ eqgfdnin QSKVSRMLT kfgavrt
1jhg_A     derealgt RVRII ee LLRGE --m SQRELKNELG ------ag IATITRGSN slkaapv

 1idy_     aaaaaaaa AAAAA aa AAAAA a-a AAAAAAAAAA -----aaa AAAAAAAAA aaaaaaa  A=a.4.1
1tc3_C     aaaaaaaa AAAAA aa AAAAA a-a AAAAAAAAAA ------aa AAAAAAAAA aaaaaaa  B=a.4.5
1hst_A     bbbbbbbb BBBBB bb BBBBB b-b BBBBBBBBBB bbbbbbbb BBBBBBBBB bbbbbbb  C=a.4.12
 1aoy      bbbbbbbb BBBBB bb BBBBB bbb BBBBBBBBBB bbbbbbbb BBBBBBBBB bbbbbbb
1jhg_A     cccccccc CCCCC cc CCCCC --c CCCCCCCCCC ------cc CCCCCCCCC ccccccc
   CSF     !!!!!!!! !!!!! !! !!!!! ! ! !!!!!!!!!! *****!!! !!!!!!!!! !!!!!!!
  CFLD     ******** ***** ** ***** * * ********** ******** ********* *******

 1idy_     llllllhh HHHHH hh HHHHT l-s LHHHHHHHSL -----sll HHHHHHHHH htgggll  H=helix
1tc3_C     llslllll HHHHH hh HHHHT t-l LHHHHHHHHT ------ll HHHHHHHHH lsttttl  E=strand
1hst_A     llllhhhh HHHHH ht LLSSL l-e EHHHHHHHHH hhsllltt HHHHHHHHH hhhhhht  T=turn
 1aoy_     lllllstt HHHHH hh HHHTL lll SHHHHHHHHH hhtlssll HHHHHHHHH hhtleee  S=bend
1jhg_A     hhhhhhhh HHHHH hh HHHLL --s LHHHHHHHHL ------ll HHHHHHHHH hhhhslh  G=3/10 helix
   DSS     !!!!!!!! ***** *! !!!!! ! ! !********! **!*!!!! ********* !!!!!!!  L=loop
```

Block 4, BLOSUM62 score

```
1tc3_C  RHCIRVYLK  = -15
1hst_A  ADLQIKLSI
```

BALIBASE

```
1tc3_C  RHCIRVYLK  = -1
1hst_A  NADLQIKLS
```

MUSCLE

Figure 3. BALIBASE set BBS11013. Analysis of BBS11013, with three different one-letter annotations of the sequences: top, letters are amino-acid codes; middle, letters are domain assignments by SCOP; bottom, letters are secondary structure assignments by DSSP. Keys for letters in the middle and bottom alignments are given to the right. In all three cases, sequences are aligned according to the database reference. Core blocks are indicated by upper-case letters. In each column, annotation agreement by superfamily (CSF), fold (CFLD) or secondary structure (DSS) is indicated by asterisks and disagreement by exclamation marks. The second core block has secondary structure disagreements in every column, confirmed by the Pymol ribbon diagrams (below left), in which core block 2 is highlighted for 1idy_ and 1hst_A. Block 2 contains a surface loop in 1hst_A which should be excluded by the stated criteria for core blocks. Lower-right the MUSCLE and BALIBASE alignments of 1tc3_C and 1hst_A in core block 4 are compared. The BLOSUM62 substitution scores are negative for all pairs in the BALIBASE alignment and sum to −15, while scores in the MUSCLE alignment sum to −1, reflecting a higher degree of biochemical similarity and suggesting that the MUSCLE alignment may be more accurate by some criteria.

Table 1. Reference alignment domain and secondary structure agreement scores

Benchmark	Cols	DSS	Ann	CSF	CFLD	ESF	EFLD	ECLS
BALIBASE	Core	28.8	SF	89.4	93.0	11.5	8.1	5.9
	Non-core(U)	69.4	SF	83.9	86.0	22.0	20.4	15.0
	Non-core(T)		SF	90.9	94.1	10.4	7.1	4.2
PREFAB	Ref	28.4	SCOP	96.2	98.5	3.8	1.5	0.5
			CATH	95.4	97.9	4.6	2.1	0.6
OXBENCH	SCR	22.9	SCOP	96.0	99.3	3.9	0.6	0.2
			CATH	96.3	99.2	3.8	0.8	0.0
	Non-SCR	58.7	SCOP	96.5	99.4	3.5	0.6	0.2
			CATH	96.6	99.3	3.4	0.7	0.0

Annotation agreement scores for the benchmark reference alignments. Cols: subset of columns measured. Core = BALIBASE core blocks, non-core(U) = other columns in untrimmed sets, non-core(T) = other columns in trimmed sets, Ref = reference columns, SCR = structurally conserved columns. Ann = annotation database. SF = SUPERFAMILY. SABMARK is excluded as the measures cannot be computed on its inconsistent pair-wise alignments. The non-core block columns in untrimmed BALIBASE alignments have exceptionally high error rates, with 69% disagreement on secondary structure and 20% disagreement on fold.

SUPERFAMILY accuracy

An estimate of the accuracy of SUPERFAMILY predictions was made by comparing with SCOP on the subset of BALIBASE with known structure, where 96% of residues annotated by both methods were assigned to the same superfamily. This high rate of agreement supports the use of SUPERFAMILY for sequences of unknown structure.

Transitive consistency of SABMARK

The pair-wise reference alignments in SABMARK are not in general consistent with a multiple alignment (Figure 1).

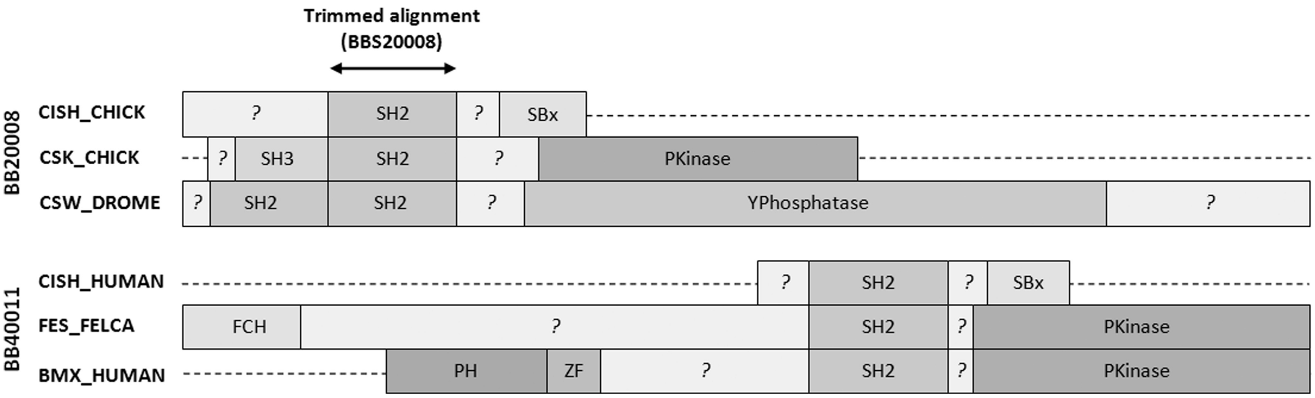


Figure 4. Non-homologous domains aligned by BALIBASE. BALIBASE reference alignments BB20008 and BB40011 are shown for selected sequences with domain annotations according to Uniprot. Unannotated regions are indicated by question marks; long terminal gaps are indicated by dashed lines. Structures are not known for any of these proteins. In both sets, core blocks are in an SH2 domain, and in the case of BB20008 the trimmed set is limited to SH2 (trimmed versions are not provided for BB4). Many columns in these reference alignments contain residues that are definitively not homologous or structurally similar. Some terminal gaps are very long, e.g. 872 gapped columns follow the C terminal of CSK_CHICK in BB20008. This may explain the improvements achieved by reduced penalties for long gaps in methods such as PROBCONS (45), which uses a 'double-affine' penalty.

The degree of inconsistency was quantified as follows. For each pair A–B in a given set, every possible transitive alignment A–X–B via a third sequence X was constructed from the reference alignments A–X and X–B, and the fraction *Q* of A,B letter pairs that agreed between A–X–B and A–B was computed. The *consistency* (*c*) of a set is then defined to be the average of *Q* over all transitive alignments. The average *c* for all sets was found to be 80%. Seventeen sets were found to be maximally inconsistent ($c = 0$). In at least one case (Superfamily set 169), this is because there are no aligned letter pairs [*sic*] in any of the pair-wise reference alignments.

Coverage of fold space and redundancy

Fold space coverage was quantified using the number of different SCOP folds found in each benchmark, which was found to be: PREFAB 321, SABMARK 256, OXBENCH 129 and BALIBASE (trimmed) 94. The untrimmed sets in BALIBASE have a total of 233 folds; most of the additional 137 appear in terminal regions in which unrelated folds are aligned to each other. The fraction of alignments in which the most-used fold appears was found to be: SABMARK 3%, PREFAB 8%, BALIBASE (trimmed) 8%, BALIBASE (untrimmed) 15% and OXBENCH 25%. Thus OXBENCH has the lowest coverage of fold space and also the strongest bias to a single fold. To measure an effective size for each database I constructed a graph in which nodes are sets and an edge is a SCOP fold found in both sets. The number of connected components in this graph is the effective size, defined as number of subsets with no folds in common, which was found to be: OXBENCH 93, SABMARK 92, PREFAB 82, BALIBASE (trimmed) 28 and BALIBASE (untrimmed) 9.

Sequence divergences and set sizes

I believe that the twilight zone of 20–35% sequence identity is generally the most appropriate for MPSA benchmarking. Sequences with higher identity are relatively easy to align and therefore less able to discriminate between methods, while proteins with lower identities have increasingly ambiguous structural alignments that are inappropriate for trusted references. The fractions of alignments in each benchmark having identities below/in/above the twilight zone are as follows: BALIBASE 18/60/22%, OXBENCH 4/13/82%, PREFAB 49/36/15% and SABMARK 43/44/13%. Thus BALIBASE emphasizes the twilight zone while OXBENCH emphasizes higher identities and PREFAB and SABMARK are comparable in this regard, both having a substantial fraction of sets with low identity. The minimum/mean/maximum number of sequences per set is: SABMARK 3/8/25, BALIBASE 4/28/142, PREFAB 2/45/50 and OXBENCH 4/122/395, showing that SABMARK has significantly fewer sequences per set.

Sensitivity and specificity

A few MPSA methods attempt to distinguish reliably aligned columns, including ALIGN-M v2.3 (10), DIALIGN-TX v1.0.2 (39), FSA 1.14.5 (12) and MUSCLE v4.0.128, which implements a new algorithm based on conditional random fields (R.C. Edgar, submitted for publication). While sequences in other benchmarks are generally trimmed to a single structural domain, untrimmed BALIBASE alignments contain non-homologous regions and can therefore be used to assess the ability of such methods to identify them, as shown in Table 2. The MUSCLE alignments have higher coverage (SA and SAD), better domain agreement (CSF, CFLD) and fewer domain disagreements (ECLS, EFLD, ESF) than the BALIBASE core block alignments. ALIGN-M also performs well by these measures, although lower per-residue accuracy is indicated by its SPS and CS scores. The apparently high sensitivities and low error rates of FSA and DIALIGN-TX are artifacts of a tendency to correctly align closely related sequences to each other while failing to assemble them into a complete multiple alignment (Figure 2), which is indicated by a high percentage of gaps and lower CS scores.

Table 2. BALIBASE alignment quality scores

Method	SPS	CS	CSF	CFLD	ESF	EFLD	ECLS	SA	SAD	Gaps
BALIBASE (all)			85.2	87.7	20.2	18.4	13.4	71.1	55.3	36.0
BALIBASE (core)			89.4	93.0	11.5	8.1	5.9	32.8	32.8	0.0
MUSCLE (*P* = 0.5)	88.9	64.2	92.7	93.6	7.8	7.0	5.0	39.7	34.4	2.8
ALIGN-M	80.3	46.9	90.8	93.1	10.5	8.1	6.0	41.4	32.6	5.3
FSA	80.2	46.8	95.6	96.1	5.4	5.0	3.6	80.4	33.4	54.5
FSA-maxsn	86.3	57.9	90.3	91.7	12.1	10.6	7.4	75.3	46.2	44.4
DIALIGN-T	78.8	45.6	86.8	88.8	15.6	13.7	10.0	69.0	40.5	43.6

Quality scores for methods that distinguish reliable from unreliable columns. Untrimmed BALIBASE alignments were used. SPS and CS are by comparison with BALIBASE core blocks considering all columns in the tested alignments; for other measures only columns annotated as reliable were included. The reference alignments themselves are measured on core blocks (core) and all columns (all). For MUSCLE, a posterior threshold of 0.5 was used; FSA was used with default parameters and with the–maxsn (maximum sensitivity) option. Measures are shown as percentages. 'Gaps' is the fraction of sequences that are gapped in columns annotated as reliable.

DISCUSSION

BALIBASE is widely used and is widely believed to be of high quality. However, these results show that its alignments and core block annotations should not be considered reliable or to be independent of sequence methods. Many BALIBASE sets contain structures with uncertain homology, and in such cases reliable residue correspondences cannot be determined by any method. Only 13% of sequences in BALIBASE reference alignments have known structure. One set (BB11037) has no known structures, and 17 sets have only one structure. Most sequences (87%) were aligned by primary sequence alone with the help of methods including BLASTP (21) and NORMD (40), both of which use gap penalties and substitution matrices. Where structures were available, they were aligned by SAP (41), which uses the Needleman–Wunsch algorithm (42) to identify residue correspondences given a structural superposition. Bias is therefore a concern as some MPSA methods could tend to have higher agreement with BALIBASE alignments due to the use of substitution matrices or gap penalty functions similar to those in BLASTP, NORMD or SAP, and I have recently presented evidence that BLOSUM62 may be favored (43). Core blocks are defined by 'the presence of conserved secondary structures combined with a sequence conservation score for each position in the alignment . . . [this is] designed to exclude the sequence stretches that cannot be accurately aligned, such as loop regions' (2). According to DSSP, which was also used by the BALIBASE authors, 63% of core blocks contain loops and almost one in three (29%) core block columns contain conflicting secondary structure assignments. This rate of secondary structure disagreement is higher than PREFAB (28%) and OXBENCH (23%), which were generated by structural alignment methods that do not explicitly consider secondary structure. This suggests that BALIBASE core block identification is dominated by sequence conservation criteria that do not correlate reliably with conserved secondary structures in the buried regions that are generally understood to be the core of a protein. BALIBASE is unique in having multi-domain proteins in its reference alignments, which could be useful in assessment providing that appropriate annotations and accuracy measures are applied. However, the full-length proteins used in BALIBASE do not have solved structures and do not have globally alignable domain organizations, and in practice this aspect of the benchmark has resulted in misleading assessments. Many non-core-block columns in BALIBASE reference alignments contain non-homologous domains (e.g. Figure 4). Such columns are incorrect by any reasonable definition, but have been used in practice, e.g. to validate FSA (12). When columns not in core blocks are excluded from accuracy measurement, the use of untrimmed sets is more defensible, but it is important to note that in general they are only locally alignable to a short segment of each sequence. Validations using untrimmed sets to rank global algorithms, e.g. (44), should therefore note the use of input data for which a program was not designed so that results can be interpreted accordingly. While all four benchmarks contain structures with uncertain homology and have comparable rates of secondary structure conflicts in their reference alignments, all except BALIBASE are derived from structures of single domains having similar folds. These folds are approximately globally alignable by structure, and residue correspondences therefore cannot be shown to be definitively correct or incorrect. The benchmark alignment quality issues identified here motivate attempts at improvements, including: constructing multiple alignments from a consistent subset of SABMARK columns, discarding sets with identities that are too high to be informative, discarding diverged structures having unambiguous residue correspondences or uncertain homology, aligning structures with methods that do not use amino-acid similarity or gap penalties, assessment on regions with more highly conserved secondary structure, reducing redundancy and increasing coverage of fold space by reference to SCOP and CATH, eliminating sequences with unknown structure from reference alignments, and identifying multi-domain proteins with globally alignable domain organizations and assessing their alignments by reference to solved structures. These approaches can be used to develop new benchmarks and to investigate whether more robust rankings of MPSA methods can be achieved (manuscript in preparation). The present results show that protein alignment assessment is more challenging than generally realized, and skepticism is appropriate for claims that method rankings or advances can be reliably measured by current benchmarks.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

I am grateful to Steven Brenner and Arend Sidow for helpful discussions.

Conflict of interest statement. None declared.

REFERENCES

1. McClure,M.A., Vasi,T.K. and Fitch,W.M. (1994) Comparative analysis of multiple protein-sequence alignment methods. *Mol. Biol. Evol.*, **11**, 571–592.
2. Thompson,J.D., Koehl,P., Ripp,R. and Poch,O. (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, **61**, 127–136.
3. Bahr,A., Thompson,J.D., Thierry,J.C. and Poch,O. (2001) BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic Acids Res.*, **29**, 323–326.
4. Thompson,J.D., Plewniak,F. and Poch,O. (1999) BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics*, **15**, 87–88.
5. Subramanian,A.R., Weyer-Menkhoff,J., Kaufmann,M. and Morgenstern,B. (2005) DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, **6**, 66.
6. Mizuguchi,K., Deane,C.M., Blundell,T.L. and Overington,J.P. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, **7**, 2469–2471.
7. Blackshields,G., Wallace,I.M., Larkin,M. and Higgins,D.G. (2006) Analysis and comparison of benchmarks for multiple sequence alignment. *In Silico Biol.*, **6**, 321–339.
8. Hasegawa,H. and Holm,L. (2009) Advances and pitfalls of protein structural alignment. *Curr. Opin. Struct. Biol.*, **19**, 341–348.
9. Godzik,A. (1996) The structural alignment between two proteins: is there a unique answer? *Protein Sci.*, **5**, 1325–1338.
10. Van Walle,I., Lasters,I. and Wyns,L. (2004) Align-m–a new algorithm for multiple alignment of highly divergent sequences. *Bioinformatics*, **20**, 1428–1435.
11. Van Walle,I., Lasters,I. and Wyns,L. (2005) SABmark – a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, **21**, 1267–1268.
12. Bradley,R.K., Roberts,A., Smoot,M., Juvekar,S., Do,J., Dewey,C., Holmes,I. and Pachter,L. (2009) Fast statistical alignment. *PLoS Comput. Biol.*, **5**, e1000392.
13. Sauder,J.M., Arthur,J.W. and Dunbrack,R.L. Jr (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins*, **40**, 6–22.
14. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
15. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH–a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
16. Jones,S., Stewart,M., Michie,A., Swindells,M.B., Orengo,C. and Thornton,J.M. (1998) Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci.*, **7**, 233–242.
17. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
18. Etchebest,C., Benros,C., Hazout,S. and de Brevern,A.G. (2005) A structural alphabet for local protein structures: improved prediction methods. *Proteins*, **59**, 810–827.
19. Colloc'h,N., Etchebest,C., Thoreau,E., Henrissat,B. and Mornon,J.P. (1993) Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Eng*, **6**, 377–382.
20. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
21. 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.
22. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
23. Holm,L. and Sander,C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.*, **26**, 316–319.
24. Brenner,S.E., Koehl,P. and Levitt,M. (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.*, **28**, 254–256.
25. Boutonnet,N.S., Rooman,M.J., Ochagavia,M.E., Richelle,J. and Wodak,S.J. (1995) Optimal protein structure alignments by multiple linkage clustering: application to distantly related proteins. *Protein Eng.*, **8**, 647–662.
26. 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.
27. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
28. Raghava,G.P., Searle,S.M., Audley,P.C., Barber,J.D. and Barton,G.J. (2003) OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, **4**, 47.
29. Russell,R.B. and Barton,G.J. (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**, 309–323.
30. Siddiqui,A.S., Dengler,U. and Barton,G.J. (2001) 3Dee: a database of protein structural domains. *Bioinformatics*, **17**, 200–201.
31. Schwartz,A.S., Myers,E.W. and Pachter,L. Alignment metric accuracy,arXiv:q-bio/0510052v1.
32. O'Sullivan,O., Zehnder,M., Higgins,D., Bucher,P., Grosdidier,A. and Notredame,C. (2003) APDB: a novel measure for benchmarking sequence alignment methods without reference alignments. *Bioinformatics*, **19(Suppl. 1)**, i215–i221.
33. Armougom,F., Moretti,S., Keduas,V. and Notredame,C. (2006) The iRMSD: a local measure of sequence alignment accuracy using structural information. *Bioinformatics*, **22**, e35–e39.
34. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.
35. Yu,H., Rosen,M.K., Shin,T.B., Seidel-Dugan,C., Brugge,J.S. and Schreiber,S.L. (1992) Solution structure of the SH3 domain of Src and identification of its ligand-binding site. *Science*, **258**, 1665–1668.
36. Bullock,A.N., Debreczeni,J.E., Edwards,A.M., Sundstrom,M. and Knapp,S. (2006) Crystal structure of the SOCS2-elongin C-elongin B complex defines a prototypical SOCS box ubiquitin ligase. *Proc. Natl Acad. Sci. USA*, **103**, 7637–7642.
37. Babon,J.J., Sabo,J.K., Soetopo,A., Yao,S., Bailey,M.F., Zhang,J.G., Nicola,N.A. and Norton,R.S. (2008) The SOCS box domain of SOCS3: structure and interaction with the elonginBC-cullin5 ubiquitin ligase. *J. Mol. Biol.*, **381**, 928–940.
38. Barford,D., Flint,A.J. and Tonks,N.K. (1994) Crystal structure of human protein tyrosine phosphatase 1B. *Science*, **263**, 1397–1404.
39. Subramanian,A.R., Kaufmann,M. and Morgenstern,B. (2008) DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms Mol. Biol.*, **3**, 6.
40. 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.
41. Taylor,W.R. (2000) Protein structure comparison using SAP. *Methods Mol. Biol.*, **143**, 19–32.

42. Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.
43. Edgar,R.C. (2009) Optimizing substitution matrix choice and gap parameters for sequence alignment. *BMC Bioinformatics*, **10**, 396.
44. Roshan,U. and Livesay,D.R. (2006) Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics*, **22**, 2715–2721.
45. Do,C.B., Mahabhashyam,M.S., Brudno,M. and Batzoglou,S. (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res.*, **15**, 330–340.