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 except SABMARK. 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...
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 align-ments (10–13), it is not reasonable to expect sequence
methods to reproduce the arbitrary boundary between
structural similarity and dissimilarity produced by a par-
ticular 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 arrange-ments 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 some-
times 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 annota-
tions 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 con-
versely if it contains different folds it is likely to be incor-
crect. 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 fol-
dowing domain. Measures based solely on domain anno-
tations 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 struc-
ture 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 \texttt{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 \texttt{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 align-
ment, 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 mea-
surements on selected columns, e.g. core blocks, must also
be assumed to correlate only approximately with the
accuracy of the complete alignment. Core block agree-
ment presumably tends to over-estimate accuracy by mea-
suring only regions with relatively strong conservation
which are therefore more easily aligned by sequence sim-
ilarity. SABMARK v1.65 (11) was also constructed using
an automated protocol, but employed a more systematic
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 (d1b8_2, d181_1), and rows (3,4) are the reference alignment for (d1b8y_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.

Figure 1. Inconsistent SABMARK reference alignment. This 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 (d1b8_2, d181_1), and rows (3,4) are the reference alignment for (d1b8y_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.

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 as CS, 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.

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 CS, 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.

(1) d1a81_2 DWR!YFVT...FT--...CPYCLAVHKHAINTKARKILG
(2) d1a87_1 PK!VF!V...KED...QYCDQ7!LQGVLELYT--D.KLY
(3) d1a8y_2 E!K!L!YT!F...NQDSEYHAPF!AEDEPI...PYYIYPB
(4) d1a81_2 DWR!YFVT...FT--...CPYCLAVHKHAINTKARKILG

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 CS, 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.

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 as CS, 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.

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 CS, 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.

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 as CS, 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.
- \( S_\text{SF} \): Number of columns for which all letters are in the same superfamily.
- \( S_\text{F} \): Number of columns for which all letters are in the same fold.
- \( D_\text{F} \): Number of columns containing a pair of letters in two different superfamilies.
- \( D_\text{C} \): 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.

\[
CSF = S_\text{SF} / C_+
\]

Correct by Superfamily

\[
CFLD = S_\text{F} / C_+
\]

Correct by Fold

\[
ESF = D_\text{SF} / C_-
\]

Error rate by Superfamily

\[
EFLD = D_\text{F} / C_-
\]

Error rate by Fold

\[
ECLS = D_\text{C} / C_-
\]

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 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 BB20008 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 CIS_H_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.
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 superfamly. 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).
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 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) 15% and OXBENCH 25%. 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.
the full-length proteins used in BALIBASE do not have annotations and accuracy measures are applied. However, they could be useful in assessment providing that appropriate multi-domain proteins in its reference alignments, which buried regions that are generally understood to be reliably with conserved secondary structures in the sequence conservation criteria that do not correlate with BALIBASE core block identification is dominated explicitly consider secondary structure. This suggests generated by structural alignment methods that do not PREFAB (28%) and OXBENCH (23%), which were rate of secondary structure disagreement is higher than BALIBASE authors, 63% of core blocks contain loops (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 be accurately aligned, such as loop regions’ (2). 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.

### Table 2. BALIBASE alignment quality scores

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

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

