Methodology article

OXBench: A benchmark for evaluation of protein multiple sequence alignment accuracy

GPS Raghava[2,3,4], Stephen MJ Searle[2,5], Patrick C Audley[1], Jonathan D Barber[1] and Geoffrey J Barton*[1,2,3]

Address: [1]School of Life Sciences, University of Dundee, Dow St., Dundee, DD1 5EH, Scotland, UK, [2]European Molecular Biology Laboratory: European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK, [3]University of Oxford, Laboratory of Molecular Biophysics, Rex Richards Building, South Parks Road, Oxford, OX1 3QU, UK, [4]Bioinformatics Centre, Institute of Microbial Technology, Sector 39A, Chandigarh, India and [5]Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

Email: GPS Raghava - raghava@imtech.res.in; Stephen MJ Searle - searle@sanger.ac.uk; Patrick C Audley - paudley@compbio.dundee.ac.uk; Jonathan D Barber - jon@compbio.dundee.ac.uk; Geoffrey J Barton* - geoff@compbio.dundee.ac.uk

* Corresponding author

Abstract

Background: The alignment of two or more protein sequences provides a powerful guide in the prediction of the protein structure and in identifying key functional residues, however, the utility of any prediction is completely dependent on the accuracy of the alignment. In this paper we describe a suite of reference alignments derived from the comparison of protein three-dimensional structures together with evaluation measures and software that allow automatically generated alignments to be benchmarked. We test the OXBench benchmark suite on alignments generated by the AMPS multiple alignment method, then apply the suite to compare eight different multiple alignment algorithms. The benchmark shows the current state-of-the art for alignment accuracy and provides a baseline against which new alignment algorithms may be judged.

Results: The simple hierarchical multiple alignment algorithm, AMPS, performed as well as or better than more modern methods such as CLUSTALW once the PAM250 pair-score matrix was replaced by a BLOSUM series matrix. AMPS gave an accuracy in Structurally Conserved Regions (SCRs) of 89.9% over a set of 672 alignments. The T-COFFEE method on a data set of families with <8 sequences gave 91.4% accuracy, significantly better than CLUSTALW (88.9%) and all other methods considered here. The complete suite is available from http://www.compbio.dundee.ac.uk.

Conclusions: The OXBench suite of reference alignments, evaluation software and results database provide a convenient method to assess progress in sequence alignment techniques. Evaluation measures that were *dependent* on comparison to a reference alignment were found to give good discrimination between methods. The STAMP S_c Score which is *independent* of a reference alignment also gave good discrimination. Application of OXBench in this paper shows that with the exception of T-COFFEE, the majority of the improvement in alignment accuracy seen since 1985 stems from improved pair-score matrices rather than algorithmic refinements. The maximum theoretical alignment accuracy obtained by pooling results over all methods was 94.5% with 52.5% accuracy for alignments in the 0–10 percentage identity range. This suggests that further improvements in accuracy will be possible in the future.

Background

Multiple sequence alignment is a central technique in molecular biology [1,2]. Alignments enhance the understanding of structure-function relationships by allowing common functional and structural regions in protein families to be identified [3]. Accurate alignment is also the essential first step in predicting a protein structure by homology modelling [4]. Many different techniques have been developed to align protein sequences [5–9]. For two sequences, dynamic programming guarantees a mathematically optimal alignment for a given set of parameters [10,11]. Dynamic programming can be extended to the alignment of more than two sequences (multiple alignment) [12], but this becomes computationally intractable for more than ≈ 3 sequences without adding complexity to the basic dynamic programming algorithm [7]. Most practical methods for multiple alignment work by following a guide tree to add sequences or clusters of sequences to an alignment [8,13], or by iteratively refining an initial alignment [14,15].

The quality of automatic alignments have been assessed on small sets of protein sequence families [5,14,16,17]. Barton and Sternberg [14] evaluated the quality of alignment on globin and immunoglobulin families by comparison to reference alignments from 3D (three-dimensional) structure comparison. McClure *et al.* [16] studied the performance of 12 different global and local methods of multiple protein sequence alignment on four protein families (hemoglobin, kinase, aspartic acid protease and ribonuclease H). Their criteria of assessment were based on the ability of the methods to identify correctly the ordered series of motifs that are conserved throughout each protein family. Gotoh [17] assessed the multiple sequence alignment method CLUSTALW [8]; and four of his own methods [18–20], on 54 families from the Joy 3.2 database [21] of alignments from 3D structure comparisons.

More recently, the BAliBASE database of sequence alignments [22] has been created and used to evaluate the accuracy of alignment methods. The set of 142 alignments in BAliBASE are divided into five types that aim to test different factors that affect alignment accuracy, which include large insertions, orphan sequences and N- or C-terminal extensions.

In this study, we describe a data set of reference alignments and software tools for benchmarking pairwise and multiple alignment methods. The benchmark data set is made up of domain families obtained from the 3Dee database of protein structural domains [23,24]. After filtering these families by different criteria, reference structural alignments were determined by the STAMP algorithm [25]. The initial reference data set of domain

family alignments was extended and subdivided in various ways to allow the study of different aspects of the protein sequence alignment problem. The reference alignments and tools were applied to the AMPS multiple alignment method [13,14] in order to identify the most informative test measures. The benchmark suite was then applied to six further methods for comparison and the detailed results stored in a database accessible *via* the WWW.

Results

The results of this study consider the development of a database of reference alignments; the definition of evaluation measures for multiple alignment accuracy; the identification of the most informative evaluation measures by application to the AMPS [13,14] multiple alignment method; the application of the training data set to find good parameters for the AMPS multiple alignment program and investigation of different features of this hierarchical alignment method; exploration of the accuracy of alignment for AMPS on the different OXBench test sets and application and comparison of the OXBench benchmark to eight different multiple alignment methods.

Development of reference alignments and evaluation measures

Structural alignments

Reference proteins for alignment were drawn from the 3Dee database of structural domains [23,24]. 3Dee contains domain definitions for proteins of experimentally determined three-dimensional structure in the Protein Data Bank (PDB) up to July 1998. The domains are organised into a hierarchy of structurally similar protein domain families classified by the "S_c score" [25] from the automatic multiple structure alignment program STAMP [25]. S_c scores greater than 3.0 indicate clear structural similarity. STAMP not only provides the multiple structure alignment, but also gives a measure of reliability to each structurally aligned position. Thus, STAMP alignments provide a convenient way to filter out positions that are not structurally equivalent or where structural alignment can be ambiguous.

We started with 729 domain structure families at the S_c 5.0 level which contained 9,015 domains. Families with only one member were removed, as were structures of resolution poorer than 3.2 Å and domains with less than 40 residues. Domains with more than 5% unknown residues and any domain for which the secondary structure could not be defined by DSSP [26] were also removed. The stereochemical quality of the structures was assessed by running PROCHECK v.3.4.4 on each chain [27]. PROCHECK examines a range of stereochemical features of protein structures and identifies torsion angles that deviate significantly from the distributions seen in protein structures

solved at a similar resolution. The PROCHECK G-factor encapsulates these quality measures in a single figure. Accordingly, we filtered the domains to exclude any protein with an overall PROCHECK G-factor ≤ -1. These refinements left 465 families containing 7,217 domains. All multiple segment domains were then excluded to leave 5,428 domains in 381 families.

Highly similar domains (≥ 98% identity) provide limited information for assessing alignment quality and so were removed from the data set by the following procedure. Within each family, the domains were compared by pairwise sequence alignment and clustered by percentage sequence identity [14], then one domain whose structure was solved at high resolution was selected from the clusters formed at 98% identity. Thus, the data set reduced to 1,168 domains in 218 families; where no two sequences in a family share ≥ 98% identity. We chose this relatively high PID cut-off since obtaining accurate alignment of sequences that are very similar is of critical importance in protein modelling and function prediction studies.

Throughout this work the PID for two domains was calculated from the reference structural alignment as the number of identical amino acid pairs in the alignment divided by the length of the shortest sequence.

The STAMP multiple structure comparison algorithm [25] provides good reference alignments for testing sequence alignment methods since it can generate both pairwise and multiple alignments from structure and automatically identify SCRs (Structurally Conserved Regions). STAMP implements several alternative iterative hierarchical methods for finding the structural alignment of two or more proteins. All alternative methods were tried for all families, and the alignment with the highest structural similarity score (S_c) was selected [25]. Alignments produced by STAMP are usually at least as good as those by a human expert, but as structural similarity drops, alignments by any method become less easy to define [28,29]. For these reasons, the few alignments found with unusually high or low S_c values compared to their PID were carefully inspected and where structural alignments were thought to be in error, alternative STAMP parameters were tried to obtain more satisfactory results.

Structural alignments for every sequence pair in the families of the data set were also generated by STAMP as for the multiple structure alignments. This pairwise reference data set allows comparisons between pairwise and multiple alignment accuracies to be made.

Master data set
For some families in the unique data set of 218 families, the sequence identity between a subset of domains is <

10% and it is difficult for sequence alignment methods to align these families as a whole. An example is the immunoglobulin superfamily, where structure comparison puts C-type and V-type domains together, even though there is little sequence similarity. Although alignments of the complete families presents a useful test, alignments of sub-families within these families are also a challenge to methods. Accordingly, the families were sub-divided on the basis of sequence identity and structural similarity.

In order to generate sequence similar sub-families we first calculated the PID between every pair of sequences from its structural alignment. The family was then clustered on PID between domains by complete linkage with the program OC [30]. The domain clusters formed at PID cut-offs of 60, 40, 30, 20, 10 and 5 were used as sub-families as illustrated in Figure 1 for the dehydrogenase family (Family 10). The sub-families formed between the given PID cut-off were extracted as shown by the sub-divisions labelled A, B, C, D, E, F, G, H and I. For example, sub-family B comprises domain 1hya-AUTO and Ihyb-AUTO. A total of 391 sequence sub-families were created. The structural alignment of these sub-families was optimised by STAMP. In a similar manner, sub-families were generated on structural similarity at S_c cut-offs of 7, 6, 5, 4, 3 and 2.

The creation of sequence sub-families and structure sub-families were independent, so it was possible for there to be sub-families containing identical members. One of each pair of identical sub-families was removed to leave a total of 672 families and sub-families. This set included the 218 unique families and is referred to as the Master data set. Figure 2 summarises the further data sets and subsets that were derived from the Master data set and are described in the following sections.

The distribution of of the 218 families in percentage identity (PID) bins is shown in Table 1 and Figure 3. The families include a wide range of numbers of sequences (from 2 to 122) and a wide distribution of length and PID. The percentage of structurally conserved residues in the families ranges from 2.5% to 100%.

Extended data set
It has been observed in previous studies that a multiple alignment will often yield better alignments than a pair of sequences taken in isolation [14,17]. The Master data set only contains sequences of known three-dimensional structure, but for each family in the Master data set there may be many more known sequences. In order to understand the effect on alignment accuracy of increasing the number of sequences in an alignment, we extracted all clearly similar sequences to each family from the SWALL [31] sequence database.
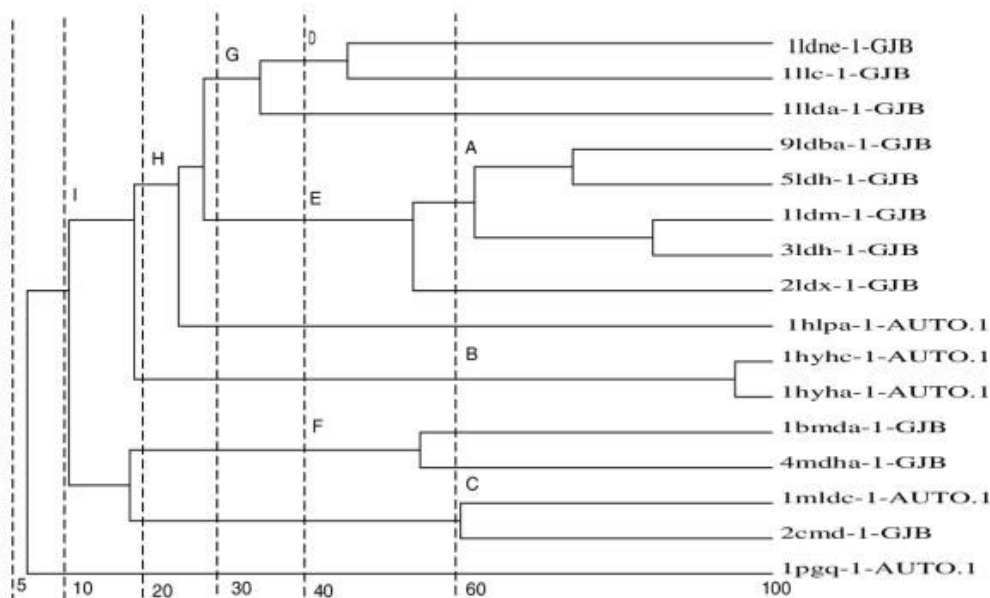
Figure 1
An example of the creation of sequence similar sub-families for Family 10 showing the families created at different cut-offs. For a full explanation see "Master data set" in the Results section.

The ratio of the number of sequences in the extended family to the number of sequences in the master family is shown in Figure 4. Approximately half the extended families are more than twenty times the size of the master families.

Full-length sequence data set
The above data sets are based on isolated domains. In practice, the domain boundaries may be unknown. Accordingly, we generated a data set of families which contain the full-length protein sequences rather than just the domain sequences. Full-length sequences were obtained by cross-reference to SWISS-PROT annotations

[32] and sequence comparison to the corresponding full-length sequences.

The ratio of the number of residues in each full-length sequence data set family to the number of residues in the equivalent master sequence family is shown in Figure 5. For most sequences, the ratio is between 1:1 and 5:1. The full-length sequence data set contains fewer families (605) than the master data set because it was not possible to identify full-length sequences for all the domain sequences in the master data set.
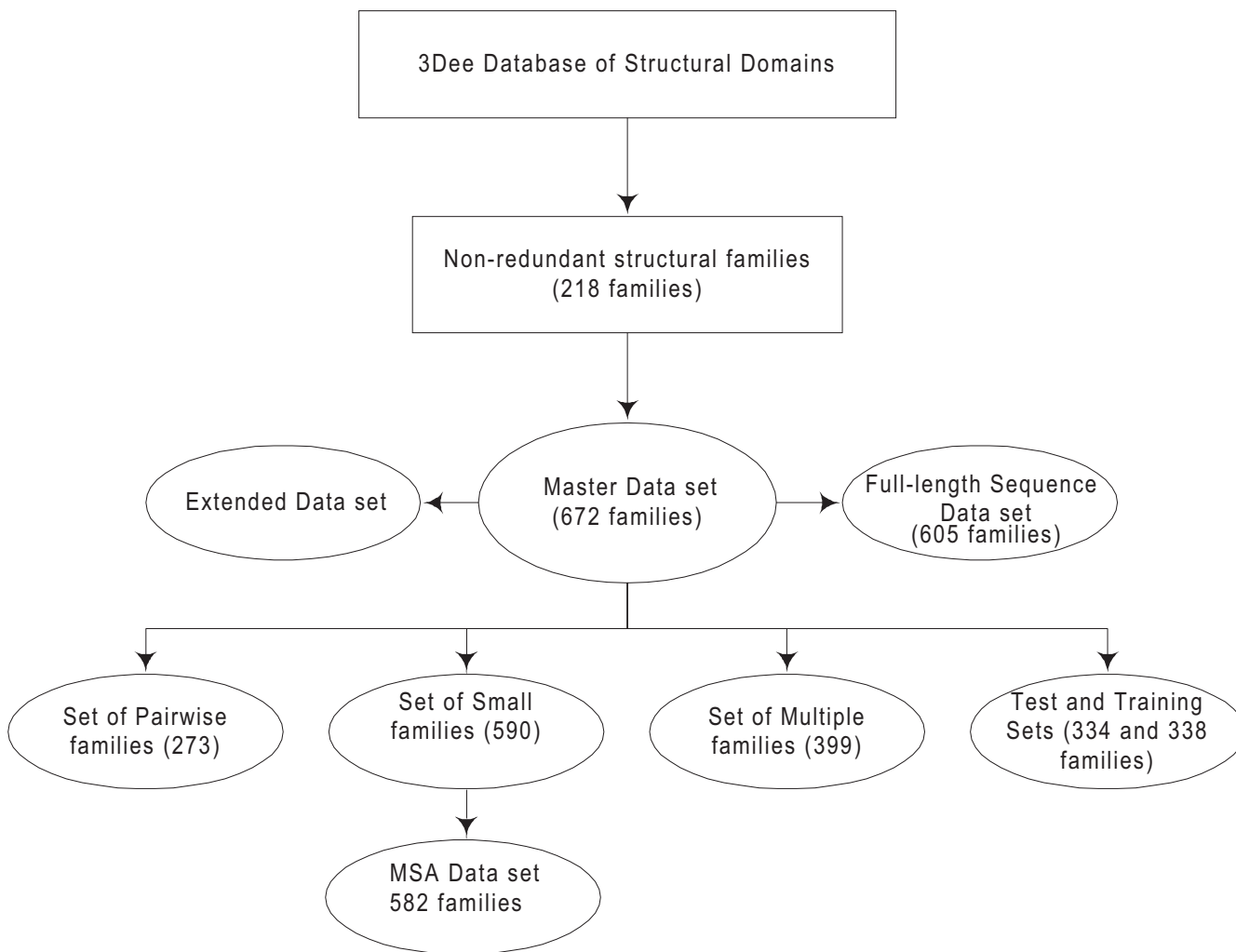
Figure 2
Flowchart outlining the relationship between the OXBench data sets and subsets. A non-redundant set of 218 structural domain families leads to the Master data set of 672 families by following the method outlined in Figure 1. The Master Data Set has additional sequences added to it to make the "Extended Data Set" and the sequences in the Master Data Set are made full-length in order to create the "Full-length Sequence Data Set". The Master Data Set is subdivided to create the test and training data sets as well as a set of two-sequence families (Pairwise Families), families with 8 or less sequences (MSA Data Set) and a set of families with more than two sequences in each family (Multiple Families). These families provide a range of different test data for multiple and pairwise alignment methods.

Set of pairwise families
The set of 273 families which contain only two sequences was extracted from the Master data set. This set may be used to evaluate alignment methods that work for only two sequences (pairwise methods) and the performance of multiple sequence alignment techniques when aligning pairs.

Set of multiple families
The set of 399 families with more than two members was extracted from the Master data set. This set allows the study of alignment algorithms and parameters on families having more than 2 sequences.

Set of small families
The set of 590 families, containing eight or fewer domains was extracted from the Master data set. This set can be

Table 1: Summary statistics for the Master data set. NDom: Number of domains. LenAln: Length of alignment. PID$_a$: Average pairwise percentage identity. PID$_w$: Percentage identity across all members of a family. S$_c$: The structural similarity score. PSCR: Percentage of positions in a structurally conserved region.

	Min	Max	Mean	Median
NDom	2	122	5.7	3
LenAln	42	598	157.6	129
PID$_a$	5.1	98	53.6	52.2
PID$_w$	0.0	98.9	39.3	32.4
S$_c$	2.6	10.0	8.1	8.5
PSCR	2.5	100.0	74.5	81.0
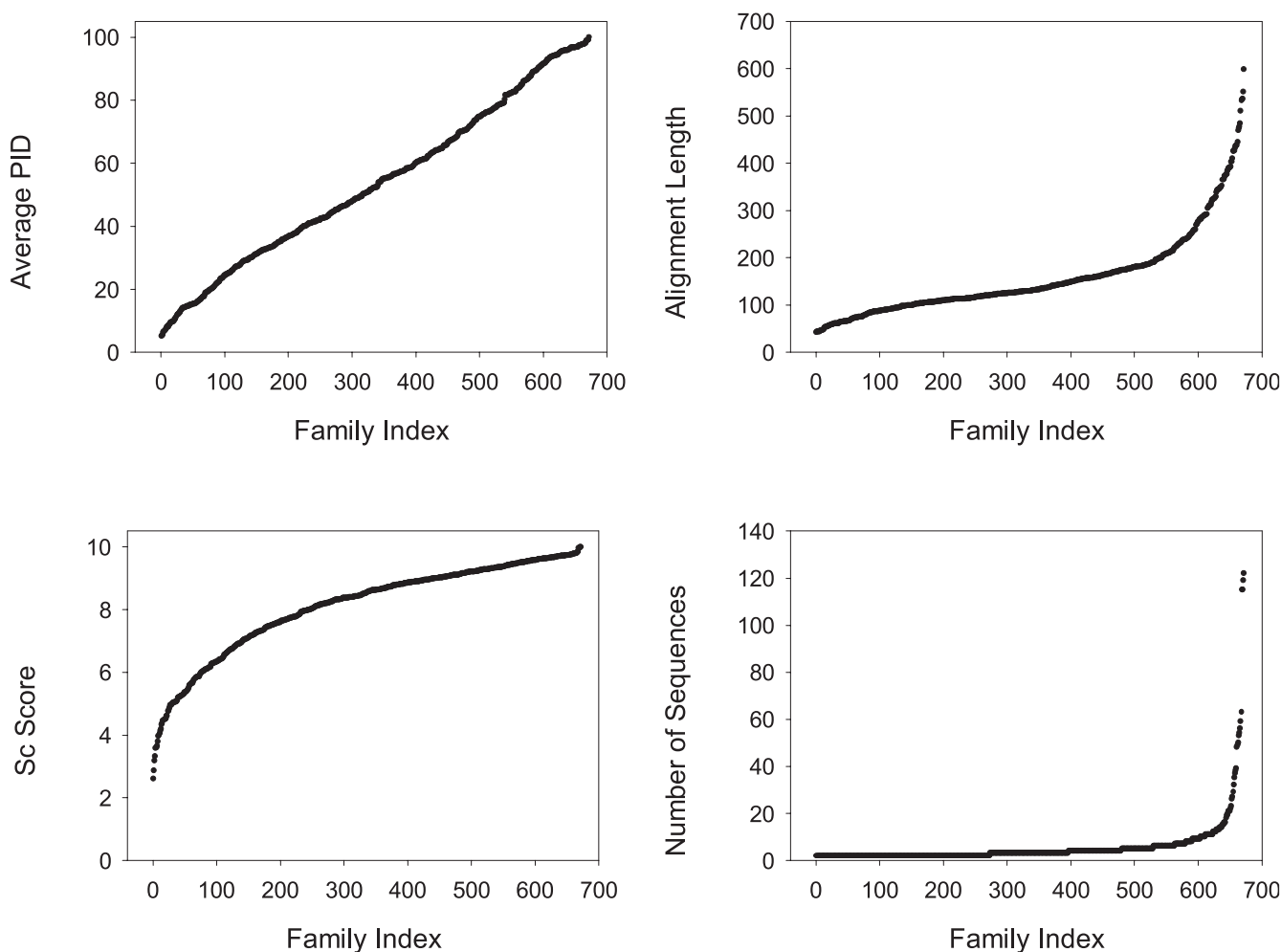
Figure 3
The families in the Master Data Set ordered by a) Percentage Identity (PID); b) STAMP [25] S$_c$ structural similarity score; c) length of alignment; and d) number of domains/sequences in the family.
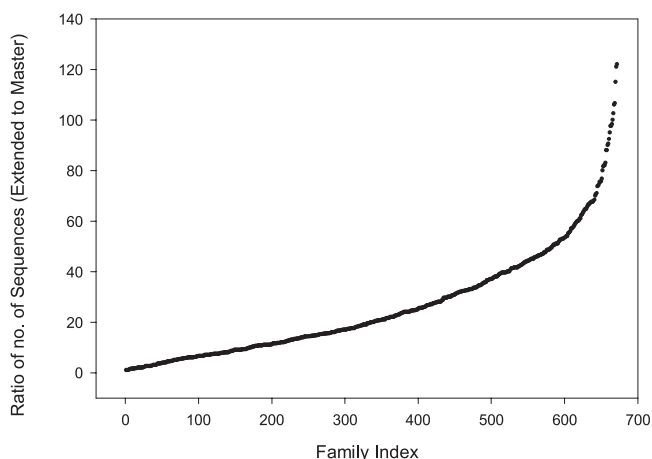
Figure 4
Graph showing the ratio of number of sequences in each family extended by adding additional sequences to the sequences in the family in the Master Data Set.

Figure 6
Distribution of families in the two test/training subsets of the Master data set sorted by percentage identity (PID).

Figure 5
Graph showing the ratio of the number of residues in the family in the full-length sequence family to the number of residues in the family in the Master Data Set.
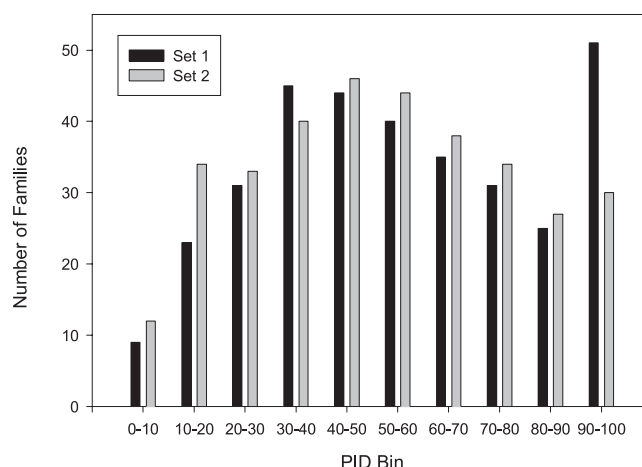
PID<10 or length of alignment >150 which suggests that MSA is not suitable for aligning sequences with low PID or long sequences. Accordingly, we generated a further data set that excludes these 8 families and call this the MSA data set (582 families.)
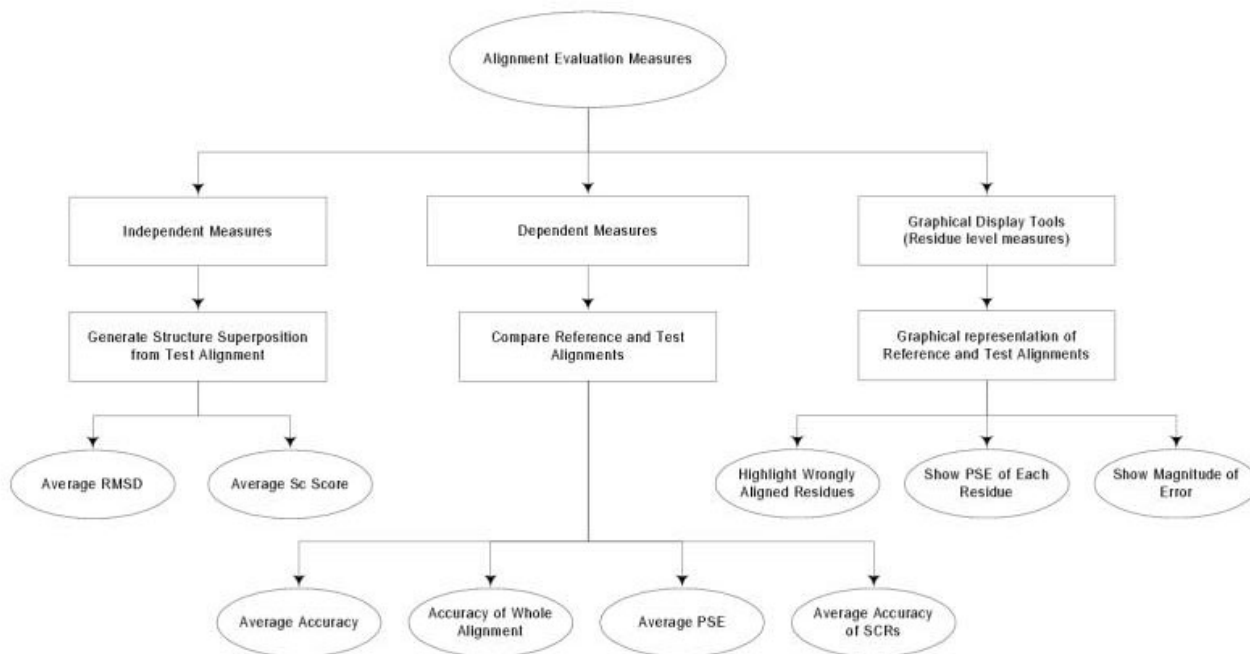
Test and training sets
So that a fair assessment of performance may be performed, it is necessary that separate, independent training and testing data sets should exist. Ideally a full, leave-one-out jack-knife test would be performed but for multiple sequence alignments this would normally be too time consuming. Accordingly, a simple two-fold cross-validation method was adopted.

The Master data set was split into two sets in such a way that there was no domain in one set that shared sequence similarity with domains in the other set and the PID distribution in both sets were equal. The two sets were created as follows: the PIDs across the whole alignment of each family of the master data set were computed. These families were sorted into ascending order by PID, then the families with odd indices were placed in one set and the complement in the other.

The number of families in each of the test and training sets when the families are binned on PID is shown in Figure 6.

Alignment accuracy evaluation measures
The three different approaches to the evaluation of alignment accuracy developed in this study are summarised in Figure 7. The three approaches are: (i) *dependent measures* that compare an alignment to a reference alignment; (ii)

used to assess multiple alignment methods that are not suitable for large families or are too time consuming to use on large families [7,33].

MSA data set
In testing on the set of small families, DIALIGN and T-COFFEE were able to align all 590 families but MSA failed to align 8 families. Most of these families had either

Figure 7
Summary of the measures used to evaluate alignment quality that are discussed in this paper. "Independent Measures" are those that do not compare an alignment to a reference alignment, but compare the superimposed structures implied by an alignment. "Dependent Measures" compare an alignment to a reference alignment. The Graphical Display Tools highlight differences in alignment between the reference alignment and a test alignment. An example output is shown in Figure: 8. For definitions of terms used in this Figure, see Results.

independent measures that compare the three dimensional structures implied by an alignment; and (iii) *visualisation* tools that highlight differences between alignments on a colour display.

Dependent measures: evaluation of the complete alignment
Dependent measures which compare an alignment to a reference alignment have long been used in the evaluation of alignment quality [14,17].

Multiple alignments can be assessed either by considering the alignment as a whole, or by examining the quality of each pairwise alignment within the multiple alignment. Thus, the accuracy of multiple alignment as a whole (AC_w) and average accuracy AC_a of all pair alignments were computed by Equations 1, 2 & 3.

$$AC_w = 100 \frac{Number\ of\ correctly\ aligned\ positions}{Length\ of\ reference\ alignent} \tag{1}$$

$$ACP = 100 \frac{Number\ of\ correctly\ aligned\ positions}{Length\ of\ aligned\ region} \tag{2}$$

$$AC_a = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{ACP_{ij}}{n(n-1)/2} \tag{3}$$

where n is the number of sequences in the alignment. The parameters AC_w or AC_a only consider correctly aligned residues, not the magnitude of the error. The Position Shift Error (PSE) was introduced at the Second Meeting on the Critical Assessment of Techniques for Protein Structure Prediction (CASP2) [34] to measure the magnitude of error in alignments. However, the PSE described in CASP2 does not consider the gaps in an alignment, so here we calculate PSE as follows. An index value is assigned to each residue in the alignment (Table 2). For example, **1** for the first residue and **n** for the **n**th residue of the sequence. A gap is assigned to the mean of the index

Table 2: Example calculation of Position Shift Error (PSE). See Equation: 4 and associated text for explanation.

a) Reference Alignment							
I	2	3	3.5	4	5	6	7
G	K	L	-	M	R	A	Y
G	K	P	H	K	L	R	L
I	2	3	4	5	6	7	8

b) Test Alignment							
I	2	3	4	4.5	5	6	7
G	K	L	M	-	R	A	Y
G	K	P	H	K	L	R	L
I	2	3	4	5t	6	7	8

values for first left and right residues. The PSE was calculated by Equation 4

$$PSE_p = \frac{\sum_{i=1}^{LA_p} | IR_i - IT_i |}{LA_p} \qquad (4)$$

where *IR* and *IT* are the index values for the reference and test alignments corresponding to the index position **i**. PSE_p and LA_p are the mean position shift error and alignment length of the pair respectively. The PSE_p for the alignment shown in Table 2 would be $\frac{0+0+0+1+0+0+0+0}{8} = 0.125$ if we take the first sequence as the reference and would be $\frac{0+0+0+0.5+0.5+0+0+0}{8} = 0.125$ if we take the second sequence as reference. The mean PSE is calculated over all $\frac{n(n-1)}{2}$ pairs of sequences.

Dependent measures: evaluation of structurally conserved regions
Assessing accuracy of the complete alignment assumes that every position in the reference alignment is equally valid. However, not every position in a protein sequence has an equivalence in a homologue. In particular, a loop region may be structurally very different and so no alignment is valid at that position. For this reason we also calculate the accuracy of alignment only within SCRs. SCRs were obtained directly from STAMP [25]. The regions with STAMP P_{ij} [25] > 6.0 for 3 or more residues were considered as SCRs. The average accuracy of each SCR was calculated within each alignment, then the average Acc_{SCR} was calculated over all SCRs in an alignment.

Independent measures
One limitation of the dependent accuracy measures is that the quality of evaluation is dependent upon the quality of the reference alignment. Errors in the reference alignment may reflect badly on a good sequence alignment method. For this reason evaluation measures that are independent of any reference alignment were also developed. To do this, the structure superposition implied by the test alignment was computed. The quality of the test alignment was then calculated from this structure superposition by computing the RMSD [35,36] and S$_c$ [25] values.

For multiple alignment the RMSD between each pair of domains was calculated as well as the average RMSD over all the pairs in the alignment. Since the RMSD is dependent on the number of atoms fitted [28,29], the percentage of equivalenced atoms (Pfit) was also calculated.

The Rossmann and Argos [37] probability is an alternative criterion for computing similarity between structures [25,37] that combines both distance and conformational terms to give the probability, P_{ij} that residue *i* in one structure and residue *j* in the other structure are equivalent. This value was computed for all the equivalent residues of the superimposed structures that were obtained from the sequence alignment by least squares fitting of the main chain C$_\alpha$ atoms. The average sum of pairs (S_p) was calculated from Equation 5:

$$S_p = \frac{\sum_{i=1}^{n} P_{ij}}{n}. \qquad (5)$$

Since these scores are a function of alignment length, it is necessary to normalise them so that domains of different size may be compared [25]. Equation 6 provides a more useful measure of alignment quality:

$$S_c = \left(\frac{S_p}{n} \right) \left(\frac{n - I_a}{L_a} \right) \left(\frac{n - I_b}{L_b} \right). \qquad (6)$$

where S_c, n, L_a, L_b, I_a, I_b are structural similarity score, number of equivalent residues, length of alignment of sequence **a**, length of alignment of sequence **b**, length of gap introduced in **a** and length of gap introduced in **b** respectively.

Substitution matrices and statistics

The substitution matrices investigated by Vogt *et al.* [38] were tested. To be consistent with early work on AMPS and the work of Vogt *et al.*, matrices containing negative scores were made positive by subtracting the most negative number in the matrix from all elements. The scores were made integer by multiplying all the elements by an appropriate power of 10.

Significance in the difference of performance between two sets of alignment parameters or two weight matrices was determined by the Wilcoxon Signed Rank Pair test as coded in the statistics package R [39].

Web server and database

An important goal of this work was to make the reference alignment sets, evaluation test alignments and results, and analysis software readily available to developers and users. Accordingly, we have developed WWW tools to permit the reference alignment data sets to be interrogated and to allow new alignments to be compared to the reference alignments. The facilities available from the OXBench web server (accessible *via* http://www.comp bio.dundee.ac.uk) are described in detail in the on-line documentation, but summarised here.

The server includes a database of more than 400 sets of alignments generated by different sequence alignment methods where each set consists of alignments of the Master data set of 672 families. The database also contains full statistics on these sets of alignments. All the evaluation results and family statistics are stored in a relational database managed by the PostgreSQL system, while the sequence alignments are stored as flat files.

The reference alignment of an individual family or the whole data set can be downloaded. In addition, the sequences of any family, or of all families are also available for download. The details of each family in the OXBench reference data set are available on the server. This information can be searched by PDB code, PDB header and PDB compound. The OXBench evaluation software can be downloaded from the server, and used locally to study alignment quality. The code is written C and R and was developed on the Linux platform.

Identification of the most informative evaluation measures

In this section we sought to find which of the different evaluation measures we examined were most able to discriminate between different alignments. To do this, we applied the AMPS multiple alignment method with a range of parameters chosen deliberately to generate very varied alignments.

AMPS implements multiple sequence alignment methods based on a progressive approach [13,14]. Multiple alignment is achieved in three steps. First, all pairs of sequences are compared. Next, the order in which they should be aligned (i.e. most similar pair through to least similar pair) is determined by cluster analysis. Finally, the sequences are aligned in that order by performing two-sequence, sequence-to-profile or profile-profile comparisons as required. AMPS has a number of options. Any substitution matrix may be employed, but the default before the work described in this paper was PAM250 [40]. The order of alignment may either follow the simple addition of one sequence at a time (single-order) or follow a tree (tree-order). Ordering may be determined by a range of alternative measures such as PID, SD score (Z-score) or Normalised Alignment Score (NAS) [14]. Optionally, for single-order, the method can iterate to refine the alignment [14].

In identifying the most informative evaluation measures we were keen to work with the complete Master data set of alignments since this provided the largest sample. Clearly, working with all the data presented a possible problem with optimisation on the test data when subsequently evaluating AMPS alongside other methods. In order to minimise the risk of bias, we tested the evaluation measures by applying AMPS with its original standard protocol of PAM250 [40] matrix and following a guide tree calculated from Z scores obtained from 100 randomisations. This protocol had been found to be effective in 10 years experience of using the program in practical applications to many different sequence families and from tests on a small number of alignments [13,14].

Since big changes in alignment can happen when the gap-penalty is varied, we varied this single parameter over a large range and examined the effect on alignment as judged by the different evaluation measures. This was not an exhaustive attempt to find the best gap-penalty for AMPS with PAM250, but rather to generate a set of scores to highlight the differences in the evaluation measures.

Evaluation of dependent measures of alignment quality

Each family in the Master data set was aligned by AMPS [14] with the PAM250 weight matrix and a range of gap-penalties. Table 3 summarises the comparison of performance for AMPS at different gap-penalties. A gap-penalty of

Table 3: Dependent measure accuracies (i.e. comparison of test to reference alignment) for AMPS run on the Master data set at various gap-penalties sorted by Acc$_{SCR}$ – Clustering was performed on SD score (from 100 randomisations) with the PAM250 matrix. Acc$_{SCR}$: Accuracy of Structurally Conserved Regions. AC$_a$: Accuracy average (pairwise). AC$_w$: Accuracy of whole alignment. PSE : Position Shift Error. NA in columns marked p indicates the highest accuracy in the preceding column, p gives the Wilcoxon Signed Rank test probability that the difference to the highest scoring row occured by chance.

Gap-penalty	Acc$_{SCR}$	p	AC$_a$	p	AC$_w$	p	PSE	p
9	88.60	NA	82.30	0.1	74.29	0.09082	0.99	0.8841
12	88.48	0.46	82.49	NA	74.52	NA	0.96	NA
6	88.31	0.032	81.77	6.3e-05	73.64	7.076e-07	1.02	0.9658
15	88.08	0.031	82.31	0.13	74.34	0.2743	0.97	0.0956
3	87.47	5.7e-09	80.39	<2.2e-16	71.82	<2.2e-16	0.98	0.1704
18	87.44	9.0e-05	81.89	0.00020	73.82	0.0002315	1.01	0.000374
21	87.06	5.6e-07	81.68	4.2e-06	73.49	2.507e-07	1.02	4.4e-06
24	86.53	1.8e-10	81.30	5.5e-10	73.00	1.972e-10	1.04	7.565e-08
27	86.02	4.9e-13	80.97	7.6e-12	72.58	1.132e-13	1.12	5.164e-12
30	85.40	<2.2e-16	80.50	6.4e-16	71.85	<2.2e-16	1.17	1.94e-13
0	84.71	<2.2e-16	77.16	<2.2e-16	67.40	<2.2e-16	1.04	1.103e-13

Table 4: Independent accuracy measures for AMPS runs on the Master data set at various gap-penalties. Clustering was performed on significance score (from 100 randomisations) with the PAM250 matrix. S$_c$: Scoring method based on STAMP S$_c$ [25]. RMSD: Score based on Root Means Squared Deviation. Pfit: Percentage of aligned positions without gap used in fit.

Gap-penalty	S$_c$	p	RMSD	p	Pfit
0	7.082	<2.2e-16	3.342	<2.2e-16	92.52
3	7.208	2.213e-05	3.176	0.0001652	93.28
6	7.238	0.1558	3.145	NA	93.50
9	7.245	NA	3.156	0.807	93.62
12	7.239	0.01863	3.152	0.7806	93.69
15	7.221	0.0006053	3.151	0.2370	93.77
18	7.192	1.481e-07	3.189	0.02599	93.82
21	7.169	7.09e-11	3.211	0.002619	93.85
24	7.144	4.038e-15	3.230	0.0001391	93.88
27	7.111	<2.2e-16	3.304	5.281e-07	93.92
30	7.081	<2.2e-16	3.333	1.044e-08	93.95

9 produced the most accurate alignments as measured in the structurally conserved regions (Acc$_{SCR}$) with an average accuracy of 88.60%. The results for this penalty were not significantly better than those for a penalty of 12 (p = 0.46), but were significantly better than for all other penalties. The worst alignments were produced with gap-penalties of 0 and 30 at the two extremes, with average accuracies of 84.71% and 85.40% respectively.

The AC$_a$ and AC$_w$ show similar trends in this test, with the best AC$_a$ of 82.49% and best AC$_w$ of 74.52% both for a penalty of 12, but not significantly different to accuracies for a penalty of 9. The PSE did not discriminate between the different alignments quite as well as Acc$_{SCR}$, with only the alignments generated with penalties of 18, 21, 24, 27,

30 and 0 showing significantly worse PSE values than the best alignments with a penalty of 12.

These results suggest that all the measures provide a useful ranking of quality for the alignments. However, since some regions of protein sequences usually share no common structural features and are unalignable, in the following studies, we chose the Acc$_{SCR}$ as the primary measure of quality.

Evaluation of independent measures of alignment quality
The performance of AMPS with default parameters judged by the independent measures RMSD, S$_c$ and Pfit on the Master data set are summarised in Table 4. As might be expected, the RMSD does not discriminate as well between the different alignment sets as S$_c$, with penalties
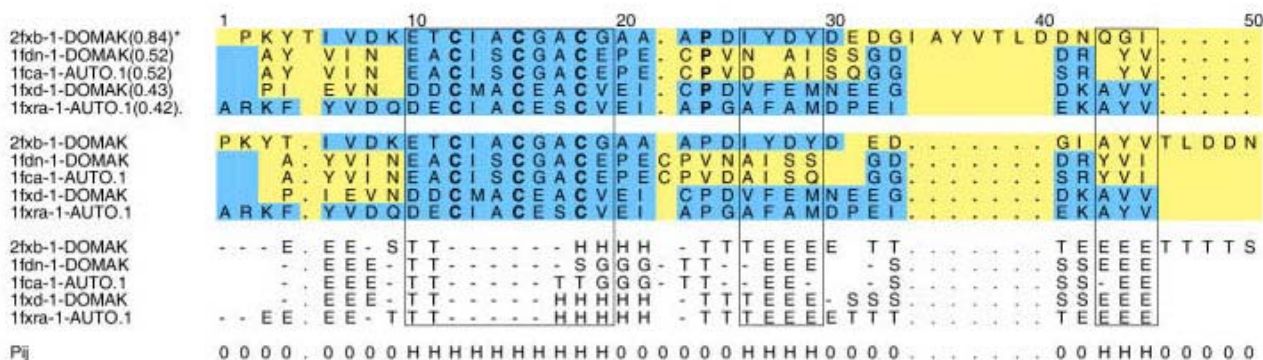
Figure 8
Comparison of an alignment generated by AMPS (PAM250 matrix, gap-penalty 6, tree order) and the reference structural alignment of Family 75 (Ferredoxin-like). The top block shows the AMPS alignment which contains the name of each domain, average PSE (in brackets) and the multiple alignment. The second block shows the reference multiple alignment obtained by 3D-structure comparison with STAMP [25]. The third block shows the secondary structure as determined by DSSP [26] and aligned in the same way as for the reference alignment. The Structurally Conserved Regions (SCRs) as determined by the STAMP multiple structure alignment program are boxed. Symbol 'H' in the "Pij" row indicates a STAMP [25] P_{ij} value of 10 or higher when aligning the least similar pair of structural sub-familes in the alignment. Thus, the boxed regions show regions where the reference alignment is most reliable. Outside these regions, the proteins either do not share the same conformation, or STAMP will not label them as confidently aligned. Residues where the alignment agrees with the reference are shown with a blue background, while residues that disagree are shown with a yellow background. The Figure is produced by ALSCRIPT [41] from commands generated by OXBench software. See text for further discussion of this alignment.

of 9, 12 and 15 showing no significance differences ($p >$ 0.05). In contrast, the S_c independent measure provided a better level of discrimination. The best alignments judged by S_c are for a penalty of 9 (S_c = 7.245) and show significant differences in S_c to alignments from all gap-penalties considered, except those with a gap-penalty of 6.

Table 3 and Table 4 show that the S_c independent measure is comparable to Acc_{SCR} at discriminating alignment quality. However, since it is simpler to understand, we focus on the dependent measure Acc_{SCR} in the remainder of this paper.

Visualisation of alignment differences
The various measures of alignment quality provide an overall picture of different performance between methods, or parameter combinations. However, even small changes to an alignment can be critical to its utility, so a straightforward way of visualising differences between alignments is important when identifying possible improvements. We have developed scripts to to generate either HTML or input for the PostScript alignment annotating program ALSCRIPT [41] that highlights differences between alignments. Figure 8 shows one example of ALSCRIPT output, which illustrates a comparison of the

sequence alignment, obtained from AMPS with the PAM250 matrix at gap-penalty 6, and the reference alignment of family **75** (Ferredoxin-like). The residues aligned differently in the two alignments are shown in yellow. The two alignments agree in the large SCR at positions 10 to 19, but disagree in the small SCRs at positions 26 to 29 and 43 to 45 where gaps have been inserted by the sequence alignment algorithm in the middle of a β-strand. Insertions in the middle of secondary structures are unusual unless they form a β-bulge and alignment quality can in general be improved by reducing the likelihood of gaps in secondary structures [5].

The SCRs reported by STAMP and exploited in this benchmark are deliberately conservative to avoid the need to inspect every structural alignment for errors. As a consequence, some SCRs in Figure 8 could be extended by one or two residues. For example, positions 21 and 22 in Figure 8 are structurally equivalent, as is position 25 (column: DNDDG). However, positions 22–24 are less straightforward to align structurally due to the insertion in 1fdn and 1fca. Inspection of the structure superposition in this region shows that despite its position in the sequence, the proline at position 24 in 2fxb is not structurally equivalent to the prolines in 1fdn and 1fca, while the valine at

Table 5: The performance of AMPS on the Master data set with 1987 defaults (PAM250, open gap-penalty 8, single order) and optimised (BLOSUM75, gap-penalty 10, tree order) parameters. Percentage SCR is the percentage of residues in the structurally conserved regions. Acc_{SCR}: accuracy of alignment in SCRs. *p*: Wilcoxon Signed Rank test probability

PID Average 1	Number of Families 2	Percentage SCR 3	Acc_{SCR} (PAM250/8) 4	Acc_{SCR} (BLOSUM75/10) 5	Difference (5–4)	*P*
0–10	21	40.5	18.3	22.2	3.9	0.0571
10–20	57	45.4	55.9	60.2	4.3	0.0427
20–30	64	54.1	76.1	81.3	5.2	0.00214
30–50	175	68.6	91.3	92.3	1.0	0.00404
50–100	355	87.8	98.8	98.9	0.1	0.000196
Total	672	74.5	88.5	89.9	1.4	7.4e-09

position 24 in 1fdn and 1fca should be shown as an insertion.

Application of the training data set to find good parameters for the AMPS multiple alignment program

As a test of OXBench, we applied the suite to the AMPS program to see if pair-score matrices developed since 1986 might give better alignments with AMPS than the original defaults of PAM250 [40] and 8 [14].

The alignment of families in the training set was generated by AMPS with various combinations of pair-score matrix and gap-penalty by following a tree generated by clustering on Z-scores. In summary, the BLOSUM75 [42] matrix with gap-penalty 10 gave the maximum Acc_{SCR} of 89.90% on the training set while the next best combination was the BENNER74 matrix [43] with a gap-penalty of 100 (89.7%). Application of AMPS with the best parameters from the training set (BLOSUM75 matrix, gap-penalty 10 and tree mode) gave an average Acc_{SCR} for the test set of 90.3%. Exchanging the test and training sets did not alter the parameters that gave the best performance. While one would normally keep test and training data completely separate, since the performance of the AMPS alignments appeared to be unaffected by which training set was used, we felt it was safe for results on this method to be reported for the complete data set in all subsequent discussions and comparisons. The accuracy on the complete Master set for these parameters was lower (89.9%) than on either test set (90.3%) which gave us further confidence that comparison on the complete Master set was unlikely to enhance the apparent accuracy of AMPS over the result seen for the test data alone. The various problems of training and testing sets when benchmarking are returned to in the Discussion.

Table 5 summarises the comparison of Acc_{SCR} obtained with the original 1987 published AMPS default protocol (single-order alignment based on Z-score, PAM250 and penalty of 8) and with the matrix and gap-penalty from optimisation of gap-penalty and choice of pair-score matrix on the training set. The average accuracy improvement over the complete set of alignments was 1.4% (*p* = 7.4×10^{-9}) with significant improvements (*p* < 0.05) in all but the 0–10 average PID range. The largest average improvement was seen for the range 20–30 (5.2%, *p* = 0.002.)

As well as pair-score matrix and gap-penalty, the other adjustable parameters in AMPS are the choice of following a tree or using a single order for multiple alignment, the number of iterations, and the method used to calculate pairwise scores from which the tree is constructed. The minimum number of randomisations necessary has been considered elsewhere [44] and so was not investigated here. However, early studies on a small set of alignments suggested that the Normalised Alignment Score could be a faster to calculate and good approximation to the Z-score for clustering [14]. Accordingly, we investigated alternative strategies for clustering sequences prior to multiple alignment with AMPS.

Effect of alternative clustering methods on alignment accuracy

For the AMPS alignment results shown in Table 5, clustering was performed on SD score calculated from 100 randomisations. The accuracies of AMPS alignment based on the faster to compute PID and NAS scores for clustering were also computed and the differences summarised in Table 6 for families that contain more than two sequences (the set of Multiple Families). The order of performance for the three clustering methods as judged by average Acc_{SCR} was SD > NAS > PID with 90.5%, 90.3% and 89.8% accuracy respectively, but the difference in accuracy between NAS and SD ordering was insignificant in all PID ranges, *p* > 0.1 (data not shown). In contrast, the difference in accuracy between clustering based on PID and SD was significant overall, with an improvement of 0.7% on average and 7% in the 10–20% identity range.

Table 6: The effect of parameters used for clustering on accuracy of SCRs. Evaluation was performed on families of the Master data set which have more than two domains (the set of multiple familes). Acc_{SCR} (SD): Accuracy when clustered by SD score. Acc_{SCR} (NAS): Accuracy when clustered by Normalised Alignment Score [14]. Acc_{SCR} (PID): Accuracy when clustered by percentage identity. Difference(SD-PID): Difference in SCR accuracy between clustering on SD and PID. p : Wilcoxon Signed Rank test probability (SD-PID).

PID Average	Number of Families	Acc_{SCR} (SD)	Acc_{SCR} (NAS)	Acc_{SCR} (PID)	Difference (SD – PID)	p
0–10	6	25.9	24.2	24.0	1.9	0.584
10–20	22	57.0	54.8	50.0	7.0	0.00604
20–30	42	76.8	76.7	75.9	0.9	0.379
30–50	130	91.3	91.1	90.9	0.4	0.0163
50–100	199	98.5	98.5	98.4	0.1	0.969
Total	399	90.5	90.3	89.8	0.7	0.000559
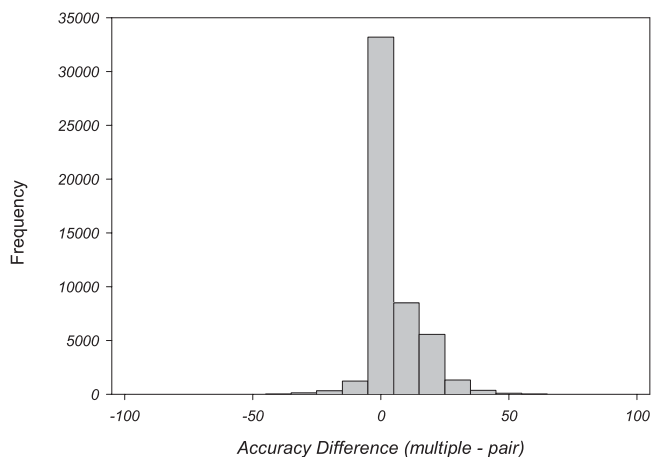
Figure 9
Graph showing the difference between multiple alignment accuracy and pairwise alignment accuracy for AMPS for all pairs from families with more than two members in the Master data set. A positive difference indicates better accuracy on multiple alignment.

This result confirmed our own experiences of using the AMPS package as a practical alignment tool.

Accuracy of alignment on different OXBench data sets
Comparison of multiple to pairwise alignment accuracy
Early tests of multiple sequence alignment methods on small numbers of families, showed an improvement in accuracy of alignment for multiple when compared to pairwise alignments [14]. In order to test if this trend held for the much larger data set developed here, pairwise alignments were performed on all pairs in each family in the Master data set and compared to the multiple align-

ment results. Figure 9 shows the difference in Acc_{SCR} for each pair aligned either individually or as part of the multiple alignment. The average improvement in alignment accuracy on multiple alignment was 4.7% ($p < 10^{-16}$) which supports the view that multiple alignment is generally beneficial.

Effect of adding additional sequences
The Extended data set provides a means of understanding the effect of extra sequences in a family on the performance of the sequence alignment method. An alignment of each family in the Extended and Master data sets was obtained by AMPS run with optimised parameters. In order to save time, clustering was performed on NAS instead of SD score. The performance of AMPS on the Master and Extended data sets is summarised in Table 7. The addition of similar sequences to families improved the accuracy of alignment by 13.6% on average in the 0–10% bin but no significant improvement was seen for higher PID. For the 50–100% bin, accuracy reduced from 98.9% to 98.8% ($p = 0.0026$). This drop in accuracy was presumably due to the extra diversity of sequences introduced in the Extended data set.

The effect of additional similar sequences was also examined on families with only two sequences (pairwise families). As shown in Table 8, the trends for pairwise families were similar to those shown in Table 7 with the only significant changes seen in the 0–10 and 50–100% ranges.

As in the comparison of multiple with pairwise alignment accuracy shown in Figure 9, these results confirm the early work on globin and immunoglobulin families that showed an improvement in alignment accuracy upon multiple alignment [14].

Table 7: The effect of additional similar sequences in a family, on the performance of AMPS applied to the Master Data Set. Clustering was performed on NAS instead of SD for efficiency with large alignments. Acc_{SCR} (Master Data Set): Accuracy for AMPS clustered on Normalised Alignment Score (NAS) for the Master data set. Acc_{SCR} (Extended data set): Accuracy for alignments on the data set with additional sequences. p : Wilcoxon Signed Rank Pair test significance

PID Average 1	Number of Families 2	Acc_{SCR} (Master Data set) 3	Acc_{SCR} (Extended data set) 4	Difference in Accuracy (4–3)	p
0–10	21	21.7	35.3	13.6	0.00947
10–20	57	59.4	63.3	3.9	0.0719
20–30	64	81.2	82.6	1.4	0.283
30–50	175	92.2	92.2	0.0	0.899
50–100	355	98.9	98.8	-0.1	0.00255
Total	672	89.7	90.5	0.8	0.238

Table 8: The effect of additional similar sequences in a family, for the set of pairwise families only. Headings as for Table 7.

PID Average 1	Number of Families 2	Acc_{SCR} (Pairwise Families) 3	Acc_{SCR} (Extended data set) 4	Difference in Accuracy (4–3)	p
0–10	15	20.7	34.7	14.0	0.0238
10–20	35	62.3	67.8	5.5	0.104
20–30	22	89.8	91.4	1.6	0.276
30–50	45	95.4	96.3	0.9	0.155
50–100	156	99.5	99.4	-0.1	0.0150
Total	273	89.0	90.6	1.6	0.0128

Effect of aligning full-length sequences

The accuracy of alignment of domain families and of full-length sequence families in different PID bins is shown in Table 9 and the accuracy difference between the Master data set alignments and full-length sequence data sets is plotted against average PID in Figure 10. As expected, domains within the full-length sequence families are less accurately aligned, a drop of 6.8%, than the domain sequence families, with the difference in accuracy significant in all but the 50–100 PID bin.

Application of the benchmark to compare 8 multiple alignment methods

The OXBench suite was applied to compare and contrast AMPS (with BLOSUM75 matrix and penalty of 10, tree order from Z-scores) and 7 further methods for multiple alignment with their default parameters. The methods considered were: PILEUP, based on the progressive method of Feng and Doolittle [6] as implemented in the Wisconsin Package Version 8.1-UNIX, (August 1995). CLUSTALW 1.7 [8] which also bases alignment on a tree, but includes a number of sophistications such as variable gap-penalties and variable pair-score matrices. PRRP [17] a DNR (doubly nested randomised iterative) method for

aligning multiple sequences. MSA [7,45] which computes an optimal multiple alignment with respect to a multiple alignment scoring system that considers all sequences simultaneously. HMMER [46] hidden Markov model method, trained first on unaligned sequences then applied to align these sequences. PIMA [47,48] which is based on a pattern construction algorithm. DIALIGN [33] is based on segment-to-segment comparison instead of residue-to-residue comparison. It constructs multiple alignments from local pairwise alignments. T-COFFEE [9] is the newest of the methods considered. T-COFFEE locates the most consistent alignments within a set of alignments and has been shown in previous studies [9] to out-perform CLUSTALW [8].

Comparison of alignment methods on the Master data set

AMPS [13,14], CLUSTALW [8], HMMER [46], PILEUP [6], PIMA [48], and PRRP [17] are able to align large numbers of sequences and so were tested on all families of the Master data set. The performance of these methods in different percentage identity ranges is shown in Table 10. With the exception of PIMA [48], all the methods were able to align all families. PIMA was unable to align 12 large families (>90 sequences). With the exception of
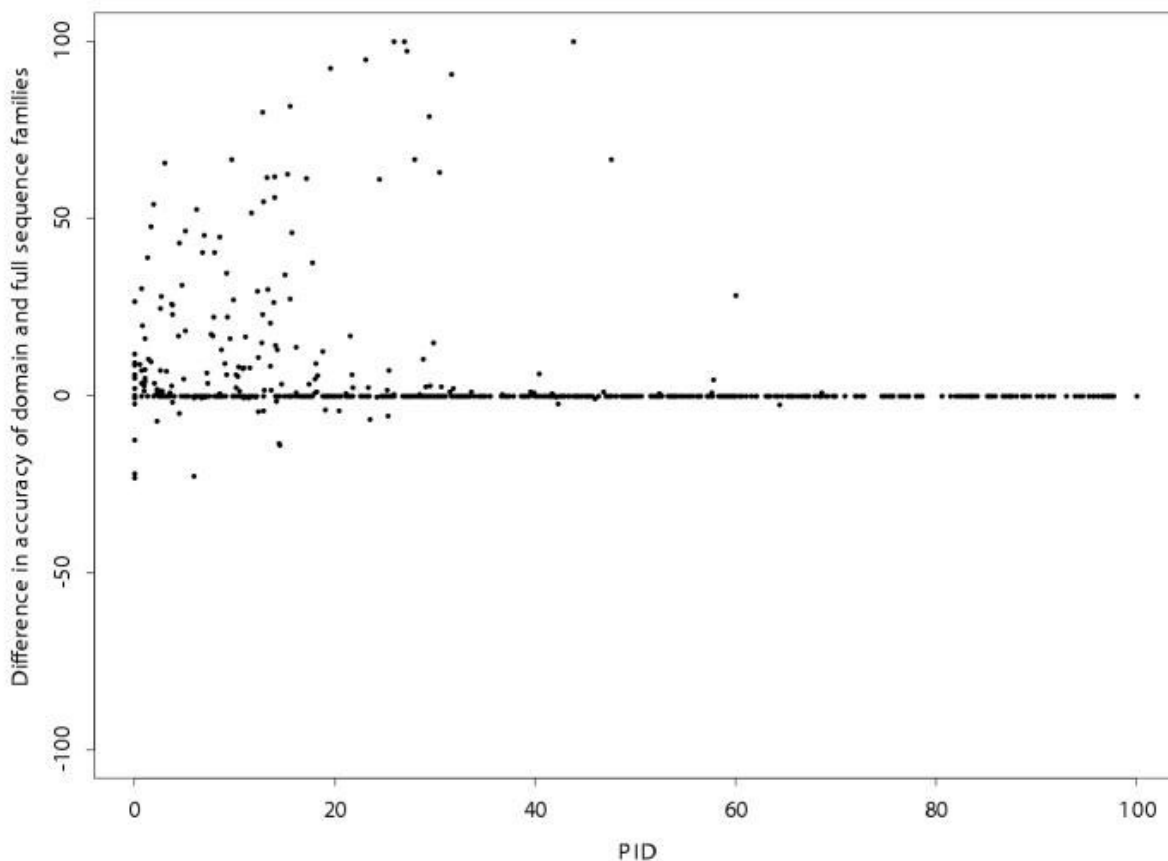
Figure 10
The difference in Acc_{SCR} for domain families and Full-length Sequence Families. The alignment was obtained by AMPS using BLOSUM75 and a penalty of 10 with NAS for clustering. Positive values show alignments that have higher accuracy when only the domains are compared relative to the full-length sequences that contain the domains. Data are plotted against the percentage sequence identity (PID).

Table 9: The performance of AMPS on the Master data set of domain families and on full-length sequences that include the domains. Parameters and column labels as in Table 7.

PID Average 1	Number of Families 2	Acc_{SCR} (Domain Families) 3	Acc_{SCR} (Full-length) 4	Difference in Accuracy (3–4)	p
0–10	22	21.1	4.2	16.9	0.000482
10–20	62	55.8	33.3	22.5	2.65e-08
20–30	70	82.6	71.2	11.4	0.000189
30–50	182	92.0	83.9	8.1	2.17e-11
50–100	269	98.2	97.9	0.3	0.466
Total	605	87.4	80.6	6.8	<2.2e-16

Table 10: The Acc$_{SCR}$ for methods on the Master data set.

Methods	Percentage Identity Range					Overall
	0–10	10–20	20–30	30–50	50–100	
AMPS	22.2	60.2	81.3	92.3	98.9	89.9
CLUSTALW	21.4	55.0	80.0	92.0	98.9	89.1
PILEUP	25.9	55.9	77.7	90.9	98.9	88.9
PRRP	20.6	56.2	78.7	91.0	98.0	88.4
PIMA	17.4	53.2	76.3	90.6	98.8	88.1
N. Family	21	57	64	175	355	672

Table 11: Difference in average performance of methods on the Master data set. A positive value indicates that the row method gives a higher accuracy than the column method. Significant differences as calculated by the Wilcoxon Signed Rank Pair test are marked: †p < 0.05; ‡p < 0.01; § p < 0.001.

	AMPS	CLUSTALW	PILEUP	PRRP
AMPS				
CLUSTALW	-0.73			
PILEUP	-0.98§	-0.25†		
PRRP	-1.50§	-0.77§	-0.52†	
PIMA	-1.73§	-1.00§	-0.75†	-0.23

HMMER [46] (data not shown), all methods gave an average Acc$_{SCR}$ of over 88% across all families. The performance of HMMER was very poor in this study since for fair comparison with the other multiple alignment methods, no seed alignment was provided to the method. Since HMMER is not strictly a multiple alignment method and its alignments were consistently worse than all other methods, they are not discussed further in this paper.

Table 11 highlights the differences in accuracy (and significance) between the methods. The order of performance based on overall average Acc$_{SCR}$ was PIMA < PRRP < PILEUP < CLUSTALW < AMPS. However, overall differences for AMPS *versus*, CLUSTALW, CLUSTALW *versus* PILEUP and PILEUP *versus* PRRP were not significant, but all other differences were significant. For the majority of methods, the most significant differences between methods occurred in the PID ranges 10–20, 20–30 and 30–50% (data not shown). For example, the 10–20% identity range has 57 alignments with an average of 45.4% of sequence in the SCRs. The difference in performance in this range between AMPS (60.2%) and CLUSTALW (55%) is 5.2% and significant at p = 0.0077. PRRP was the only method to show significant differences in the highest (50–100%) PID range. For example, in this range there are 355 alignments with on average 87.8% in the SCRs. AMPS

gave 98.9% accuracy while PRRP gave 98%, a difference of 0.9% and significant at $p = 10^{-15}$. One specific example is family 30t3 (Bacillus 1–3,1–4-beta-glucanase) for which PRRP gave only 40% accuracy despite the high PID of 59.1%. AMPS and CLUSTALW were able to align this family at 100% accuracy in the SCRs. Overall, AMPS with the BLOSUM75 matrix and gap-penalty of 10 gave the maximum average accuracy in all the PID ranges except 0–10%, where PILEUP gave the best performance. Second best was CLUSTALW which performed less well than AMPS in the lower PID ranges (PID ≤ 20) though the difference was only significant in the 10–20% range.

Evaluation of methods on families with ≤ 8 Sequences
The more computationally intensive methods are either unable to align families in the Master data set, or would take an unreasonably long time. For this reason, DIALIGN [33] MSA [33] and T-COFFEE [9] were tested on a set of small families from the master data set where the number of sequences in a family was eight or less.

Table 12 shows the order of performance based on average Acc$_{SCR}$ for all methods on the MSA set to be DIALIGN < PRRP < MSA ≤ PIMA < CLUSTALW ≤ PILEUP < AMPS < T-COFFEE. Table 13 illustrates the differences in overall accuracy between methods and their significance. It is per-

Table 12: The performance of methods on the MSA data set (families with ≤ 8 members.)

Methods	Percentage Identity Range					Overall
	0–10	10–20	20–30	30–50	50–100	
AMPS	22.2	62.2	81.5	91.3	99.0	89.68
CLUSTALW	21.4	57.0	79.3	91.2	99.0	88.94
PILEUP	25.9	59.5	78.4	90.2	99.0	89.00
PRRP	20.6	58.2	78.6	89.7	98.1	88.14
PIMA	17.4	56.6	78.7	90.1	99.0	88.46
DIALIGN	13.5	44.4	68.3	81.9	96.3	82.91
MSA	18.3	55.2	79.4	90.3	98.5	88.24
T COFFEE	23.1	69.0	87.2	93.3	99.2	91.39
N. Family	21	49	53	142	317	582

Table 13: Overall difference in performance of methods on the MSA data set. Symbols as for Table 11.

	AMPS	CLUSTALW	PILEUP	PRRP	PIMA	DIALIGN	MSA
AMPS							
CLUSTALW	-0.73						
PILEUP	-0.67‡	0.06					
PRRP	-1.53§	-0.80§	-0.86‡				
PIMA	-1.22§	-0.48†	-0.54	0.32†			
DIALIGN	-6.77§	-6.03§	-6.10§	-5.23§	-5.55§		
MSA	-1.44§	-0.70†	-0.76	0.10‡	-0.22	5.33§	
T-COFFEE	1.71§	2.45§	2.39§	3.25§	2.93§	8.48§	3.15§

haps surprising that MSA [45] which implements a method that attempts to optimise the multiple alignment across all sequences does not perform as well on this benchmark as the hierarchical methods. However, this probably reflects the high level of development that has gone into optimising hierarchical alignment methods for biological sequence analysis in the context of protein structure and function.

The clear winner on the MSA benchmark was the newest method T-COFFEE [9]. T-COFFEE gave an average Acc_{SCR} over 582 families of 91.39% which was 1.71% better than the second best average accuracy achieved by AMPS ($p = 4.7 \times 10^{-14}$). Average differences in accuracy between T-COFFEE alignments and AMPS alignments were all positive, and significant in all but the 0–10% identity range. The largest improvement over AMPS was seen in the 10–20% identity range where the accuracy improves from 62.2% to 69.0% ($p < 0.005$). When T-COFFEE alignments were compared to CLUSTALW the improvement was even more dramatic with an increase in accuracy of 2.45% ($p = 1.5e \times 10^{-15}$) over all alignments and a 12.0% increase in accuracy in the 10–20% identity range.

Evaluation of pairwise alignment
The purpose of this study was to compare multiple sequence alignment methods, but in order to understand the methods closely, the performance of the methods were also examined on families with only two sequences. The results are summarised in Table 14. Overall, the order of performance of the methods was DIALIGN < PRRP < MSA ≤ PIMA < CLUSTALW ≤ AMPS < PILEUP < T COFFEE. The difference in performance of AMPS, PILEUP and CLUSTALW was not significant. This order of performance is very similar to that for multiple alignment on the Master data set, and indicates that the performance of multiple sequence alignment methods based on the progressive approach is proportional to their ability to align pairs of sequences. This suggests that to improve the accuracy of a multiple alignment method, one should first optimise performance for pairwise methods.

Performance of methods on full-length sequence families
Only AMPS, CLUSTALW and PRRP were able to align all 614 full-length sequence families. Table 15 shows the overall trends in accuracy to be the same for these methods, but the absolute accuracy of alignment was reduced

Table 14: Performance of methods on pairwise alignments.

Methods	Percentage Identity Range					Overall
	0–10	10–20	20–30	30–50	50–100	
AMPS	20.7	62.3	89.8	95.4	99.5	88.97
CLUSTALW	19.9	58.8	88.9	96.1	99.6	88.54
PILEUP	24.3	61.7	90.0	94.5	99.6	89.01
PRRP	17.3	58.3	85.2	93.0	99.1	87.22
PIMA	15.9	58.5	89.9	94.5	99.6	88.12
DIALIGN	14.2	43.6	75.4	88.8	97.6	82.86
MSA	17.4	55.1	89.2	94.5	99.2	87.48
T COFFEE	15.4	67.6	93.6	96.5	99.6	89.90
N. Family	15	35	22	45	156	273

Table 15: Performance of methods on families of the full-length sequence data set. Only AMPS, CLUSTALW and PRRP were able to align all families in this set.

Methods	Percentage Identity Range					Overall
	0–10	10–20	20–30	30–50	50–100	
AMPS	4.1	33.8	70.9	83.7	97.8	80.33
CLUSTALW	5.3	36.0	70.2	84.5	97.6	80.65
PILEUP	7.7	36.2	69.9	83.1	97.9	80.45
PRRP	8.3	27.6	64.6	81.3	96.2	77.63
PIMA	4.8	34.4	68.5	84.2	97.8	80.28
N. Family	23	62	71	179	269	604

Table 16: Maximum accuracy achievable over all methods on each family. AMPS (Opt): results for AMPS with BLOSUM75/10 parameters. AMPS (Max): result of taking the most accurate alignment for each family over all tested parameter combinations. All Methods (Max): result of taking the most accurate alignment over AMPS (Max) and alignments by all other methods.

Methods	Percentage Identity Range					Overall
	0–10	10–20	20–30	30–50	50–100	
AMPS (Opt)	22.17	60.25	81.27	92.33	98.93	89.85
AMPS (Max)	47.41	75.15	89.29	95.96	99.38	93.85
All Methods (Max)	52.48	77.75	90.83	96.41	99.43	94.52

on average by 9% when compared to the Master data set test. For example, while AMPS gave 89.85% (table 16) accuracy on the Master data set, this was reduced to 80.33% on the full-length sequence families.

Maximum possible accuracy
The results presented so far have focused on the accuracy of methods applied with one parameter combination

across a complete data set. Table 16 summarises a different view of the data where for each family, we recorded the maximum accuracy obtained by any of the methods on that family and for AMPS run with any of the parameter combinations applied to the Master set. The maximum Acc_{SCR} for AMPS is 93.85%, 4% higher than the value obtained with BLOSUM75/10. When all methods that can run on the Master set are combined, the accuracy

increases to 94.52% overall. The result shows that for families with >50% identity 100% accuracy was achieved by some parameter or method combination, while the maximum accuracy increased most in the lower PID ranges. For example, for the range 0–10% AMPS achieved 22.17%, but the maximum accuracy over all methods was 52.48%. This suggests that there is scope for improving the average accuracy of alignment by any one method to at least this level.

Discussion

An early surprise in applying the benchmark to several multiple alignment methods was that our own comparatively simple method AMPS [13,14] performed as well or better than many more sophisticated methods. Even with the original 1986-default pair-score matrix and gap-penalty of PAM250/8 applied with the single-order option in the method as originally published [14], AMPS achieved 88.5% Acc_{SCR} on the Master data set of domains (Table 5). With this protocol, AMPS showed no significant difference to the more complex PRRP [17] which gave 88.4% and PILEUP [6] which gave 88.9%, but was lower in accuracy than CLUSTALW with default parameters at 89.1% (Table 10.) While the original default method in AMPS followed a single alignment order, AMPS is normally run with tree-order [13]. Simply switching to tree-order and updating the pair-score matrix to one of the BLOSUM series (50 or 75) to be consistent with more modern methods, raised the accuracy of AMPS to 89.9% which showed no significant difference to CLUSTALW with default parameters (Tables 10 and 11).

The results obtained for PAM250/8 are above suspicion since these parameters were published in 1987. However, an obvious criticism of the BLOSUM results for AMPS on the Master data set is that a subset of this data set was used to select the BLOSUM matrix and gap-penalty combination. As discussed in Results, this procedure breaks the rule of separating training data from test data and so in principle invalidates any conclusions about the relative merits of AMPS compared to methods which were not "optimised" on the same subset.

Although one has to bear this in mind when examining the results, all the methods considered, except AMPS with its original options, were developed to use BLOSUM series matrices. The only change necessary to bring AMPS in line with the best of these methods, CLUSTALW, was to update the matrix to one from the BLOSUM series. While not conclusive, this strongly suggests that the improvements in accuracy obtained by CLUSTALW were due primarily to the change in pair-score matrix rather than the many refinements to the CLUSTALW algorithm that are not present in AMPS. This finding is consistent with results for secondary structure prediction, where the

addition of data rather than underlying algorithm appear to dominate improvements in accuracy [49,50].

Another possible explanation of the accuracy seen for AMPS, even before changing to the BLOSUM matrix, is that we unconsciously selected test data and evaluation measures that favour the type of alignments that AMPS produces. If we had stopped with the methods compared in Table 10 it would have been difficult to refute this charge. However, after developing the benchmark and running most of the tests, the T-COFFEE [9] algorithm became available. When tested by its authors, T-COFFEE was reported as out-performing other methods, so by running T-COFFEE alignments through our benchmark we could test both the method and to some extent the validity of the benchmark. As the data discussed in Results and shown in Tables 12,13,14 indicate, T-COFFEE outperforms all other methods on our benchmark. Since our result agrees with the authors of T-COFFEE but on a different evaluation method, it suggests that the test data and evaluation measures applied in this work are effective and likely to be free of bias towards any single alignment method.

When applying any benchmark to a method, it is important to ensure that there is a separation between the data used to develop the method and optimise parameters, and those data being used to test. Unfortunately, it is rarely possible to make a totally clean separation. Even when test data appear to have been explicitly excluded from training data there may be implicit data and assumptions that are built into the algorithm under test. This is a particular problem when testing sequence alignment algorithms since nearly all make use of amino acid substitution matrices that are derived from the analysis of sequence alignments. The BLOSUM series of matrices in particular, are derived from alignments of protein families that will almost certainly include families that are featured in any reasonable benchmark data set.

A further problem arises when a benchmark is applied to a range of methods from different laboratories, since one does not know exactly which data were used to refine each method. For example, Cuff & Barton [49] faced this problem for the comparison of secondary structure prediction algorithms where often, even the authors of the methods were unsure of which data had been used in training! A fair test of the different prediction methods was only possible when a completely new test set was generated that could not have been used in the development of any of the methods [50].

Given these concerns, the good performance of T-COFFEE on our benchmark, may in part be due to overlap between the families of alignments used to optimise T-COFFEE

and those in this benchmark. Such overlaps are a problem for those trying to tease out which features of an alignment algorithm are most important. However, an end-user of an alignment algorithm would like it to give the most accurate alignment possible on any presented set of sequences. Developing a method on large data sets such as those presented here is likely to lead to an overall improvement in accuracy and so provide better quality alignments for most users.

Although useful in its current form, OXBench could be improved in a number of ways. OXBench contains alignments that range from 0 to 98.9% percentage identity across the complete alignment (PID_w in Table 1), but below 20% identity, the data sets have only 28 alignments of more than two sequences. With so few alignments in this range it is difficult to see significant differences between methods, though general trends are apparent. The deficit in low percentage identity alignments is partly due to the automatic sequence/structure clustering used to construct the 3Dee domains database [23,24] and the need to avoid incorrect clusters. In the 3Dee structural hierarchy, the domain sequences are first clustered on sequence similarity, then representatives taken and clustered by structure comparison [23,24]. As structural similarity reduces, the possibility of incorrectly clustering structures increases, so we took a conservative threshold of structural similarity when choosing the families to include in OXBench to avoid polluting the reference data. A further reason for the relatively low numbers of low similarity domains in the benchmark is the general difficulty of constructing multiple structure alignments where the proteins only have a basic topology and a few key positions in common. Again, we erred on the side of caution in selecting domain families for which the STAMP algorithm [25] will give consistent multiple structure alignments.

Structural hierarchies such as SCOP [51] that are curated by humans rather than generated entirely automatically provide a more reliable grouping for structures that share remote similarities. A benchmark data set derived from alignments built at the SCOP superfamily or fold level may extend the number of examples in the low PID range. In addition, SCOP is regularly updated and so the growth in the structural database would naturally lead to an expansion in the number of low identity examples described here. However, even with this approach, the problem still remains to generate reliable multiple structure alignments of such families to use as reference alignments for benchmarking. One way to approach this problem would be to be less conservative in selecting STAMP aligned subsets, then inspect each alignment and structure superposition by hand with reference to hydrogen-bonding diagrams and known functional sites. The

automatically identified STAMP SCRs could then be extended or contracted as appropriate in order to separate out clear regions of structural similarity. The resulting hand-curated multiple structure alignments should provide a valuable reference resource for the improvement of multiple sequence alignment methods on hard-to-align examples.

In this paper we have considered a range of different, commonly applied alignment evaluation measures. While these measures are not a comprehensive survey, the OXBench code has been written in a modular way to allow extension that may include different measures as needed. For example, some groups when developing techniques that aim to identify the reliably aligned regions of an alignment (e.g. see: [52–54]) prefer an evaluation measure such as the "shift score" proposed by Cline *et al.* [54] that attempt to penalise alignments that delete large parts of the structure as well as alignments of regions that should not be aligned. Measures such as this and other evaluation variants [22] would be straightforward to add as additional modules to the evaluation code.

Ideally, one would dispense with reference alignments and simply judge the quality of the three-dimensional model inferred from alignment as in the *independent* measures considered above. Although conceptually harder to understand, the S_c independent measure that considers both distance and local conformation performed as well as the Acc_{SCR} and so warrants further exploration as a ranking method. More sophisticated model-building approaches followed by 3D-structure evaluation may in future provide even more effective guides to improving alignment methods.

Conclusions

In this paper we have developed sets of protein multiple sequence alignments derived from tertiary structure comparison that can be used to test sequence alignment methods. The sets include those designed to test the ability of a method to align residues in known SCRs, to test the effect of aligning domains within full-length sequences, to test accuracy on pairwise and multiple alignment, and to test the effect of adding additional sequences to an alignment.

We have evaluated a range of different measures of alignment accuracy. The methods included those *dependent* on comparison to a reference alignment, and those that worked by comparison of the three-dimensional structure inferred from the sequence alignment (*independent* measures). We tested these measures on one alignment method, AMPS [14], then applied one measure that gave good discrimination, the Acc_{SCR}, to eight further methods. In order to allow new methods to be tested with the same

data and tools and compared to the work described here, the OXBench software and the result of all the analyses has been made available *via* a searchable database on the WWW. The main conclusions of applying the benchmark in this study were:

1. All measures of alignment accuracy provided some degree of discrimination of alignment quality but measurement of accuracy within structurally conserved regions (Acc_{SCR}) was preferred since it focused on the most reliable structurally aligned regions in the reference alignment. A structural similarity score that takes account of distance and local conformation (S_c) was the most discriminating measure that was independent of a reference alignment.

2. The pair-score matrix (BLOSUM, PAM, etc.) has a bigger effect on alignment accuracy than other parameters such as clustering method. For AMPS an overall improvement of 1.25% was seen on moving from PAM250 to BLOSUM75 matrix for tree-based multiple alignment from Z-scores.

3. Adding more sequences to an alignment improves the accuracy of multiple alignments for families with low sequence identity. For AMPS the largest effect was in the 0–10 PID bin where an average improvement of 13.6% was seen ($p = 0.0045$; Tables 7,8).

4. Aligning full-length sequences degraded the accuracy of alignment for embedded domains by all methods considered. For example, for AMPS the accuracy drops by 6.8% on average ($p \leq 2.2 \times 10^{-16}$; Table 9).

5. The order of performance for methods on pairwise alignment (Table 14) was very similar to that on the Master data set for multiple alignment. This suggests that the performance of multiple sequence alignment methods based on the progressive approach is proportional to their ability to align pairs of sequences.

6. On the MSA data set of alignments that contain ≤ 8 sequences, the best hierarchical alignment methods outperformed more rigorous methods that optimise sum-of-pairs and similar measures. For example, MSA gave alignments that were on average 1.4% worse than AMPS ($p < 0.001$; Table 13).

7. T-COFFEE [9] out-performed all other methods on the MSA data set benchmark by giving 91.39% accuracy compared to 88.94% for CLUSTALW a difference significant at $p < 0.001$ (Tables 12,13).

8. The maximum accuracy achieved by taking the maximum accuracy alignment for each family over all methods and parameter combinations was 94.52% (Table 16).

Authors' contributions
GPSR developed the initial database and analysis code, web site and database, he also developed and tested the different evaluation measures. SMJS refined the database, software and included statistical analysis. PCA and JB installed and debugged the database, evaluation software and web portal. PCA developed the modular OXBench software. GJB conceived the project, coordinated it and refined the manuscript drafted by GPSR and SMJS.

Acknowledgements

References
1. Taylor WR: **Identification of protein sequence homology by consensus template alignment.** *J Mol Biol* 1986, **188:**233-258.
2. Barton GJ: **Protein sequence alignment and database scanning.** In *Protein structure prediction: A practical approach* Edited by: Sternberg MJE. Oxford: IRL Press at Oxford University Press; 1996:31-63.
3. Livingstone CD and Barton GJ: **Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation.** *Comput Appl Biosci* 1993, **9:**745-756.
4. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F and Sali A: **Comparative protein structure modeling of genes and genomes.** *Annu Rev Biophys Biomol Struct* 2000, **29:**291-325.
5. Barton GJ and Sternberg MJ: **Evaluation and improvements in the automatic alignment of protein sequences.** *Protein Eng* 1987, **1:**89-94.
6. Feng DF and Doolittle RF: **Progressive sequence alignment as a prerequisite to correct phylogenetic trees.** *J Mol Evol* 1987, **25:**351-360.
7. Lipman DJ, Altschul SF and Kececioglu JD: **A tool for multiple sequence alignment.** *Proc Natl Acad Sci USA* 1989, **86:**4412-4415.
8. Thompson JD, Higgins DG and Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22:**4673-4680.
9. Notredame C, Higgins DG and Heringa J: **T-coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2000, **302:**205-17.
10. Needleman SB and Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *J Mol Biol* 1970, **48:**443-453.
11. Sellers PH: **On the theory and computation of evolutionary distances.** *J App Math* 1974, **26:**787-793.
12. Murata M, Richardson JS and Sussman JL: **Simultaneous comparison of three protein sequences.** *Proc Natl Acad Sci USA* 1985, **82:**3073-3077.
13. Barton GJ: **Protein multiple sequence alignment and flexible pattern matching.** *Methods Enzymol* 1990, **183:**403-28.
14. Barton GJ and Sternberg MJ: **A strategy for the rapid multiple alignment of protein sequences, confidence levels from tertiary structure comparisons.** *J Mol Biol* 1987, **198:**327-337.

15. Karplus K and Hu B: **Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set.** *Bioinformatics* 2001, **17:**713-20.
16. McClure M, Vasi T and Fitch W: **Comparative analysis of multiple protein-sequence alignment methods.** *Mol Biol Evol* 1994, **11:**571-592.
17. 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.
18. Gotoh O: **Optimal alignment between groups of sequences and its application to multiple sequence alignment.** *Comp App Biosci* 1993, **9:**361-370.
19. Gotoh O: **Further improvement in methods of group-to-group sequence alignment with generalized profile operations.** *Comp App Biosci* 1993, **10:**379-387.
20. Gotoh O: **A weighting system and algorithm for aligning many phylogenetically related sequences.** *Comp App Biosci* 1995, **11:**543-551.
21. Sali A and Overington JP: **Derivation of rules for comparative protein modeling from a database of protein structure alignments.** *Protien Sci* 1994, **3:**1582-1596.
22. Thompson J, Plewniak F and Poch O: **BAliBASE: a benchmark alignment database for the evaluation of multiple sequence alignment programs.** *Bioinformatics* 1999, **15:**87-88.
23. Dengler U, Siddiqui AS and Barton GJ: **Protein structural domains: analysis of the 3Dee domains database.** *Proteins* 2001, **42:**332-344.
24. Siddiqui AS, Dengler U and Barton GJ: **3Dee: a database of protein structural domains.** *Bioinformatics* 2001, **17:**200-201.
25. Russell RB and Barton GJ: **Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels.** *Proteins* 1992, **14:**309-323.
26. Kabsch W and Sander C: **Dictionary of protein secondary structure: pattern recognition of hydrogne-bonded and geometrical features.** *Biopolymers* 1983, **22:**2577-2637.
27. Laskowski RA, Mac Arthur MW, Moss DS and Thornton JM: **PROCHECK: a program to check the stereochemical quality of protein structures.** *J Appl Cryst* 1993, **26:**283-291.
28. Godzik A: **The structural alignment between two protein: Is there a unique answer?** *Protien Sci* 1996, **5:**1325-1338.
29. Feng ZK and Sippl MJ: **Optimum superimposition of protein structures: ambiguities and implications.** *Fold Des* 1996, **1:**123-132.
30. Barton G: **OC – A cluster analysis program.** 1993 [http://www.compbio.dundee.ac.uk/Software/OC/oc.html].
31. Hermjakob H, Lang F and Apweiler R: **SPTR – A comprehensive, non-redundant and up-to-date view of the protein sequence world.** *CCP11 Newsletter* 1998, **2:**.
32. Bairoch A and Boeckmann B: **The SWISS-PROT protein sequence data bank.** *Nucleic Acids Res* 1991, **19 Suppl:**2247-2249.
33. Morgenstern B, Dress A and Werner T: **Multiple DNA and protein sequence alignment based on segment-to-segment comparison.** *Proc Natl Acad Sci USA* 1996, **93:**12098-12103.
34. Moult J, Hubbard T, Bryant SH, Fidelis K and Pedersen JT: **Critical assessment of methods of protein structure prediction (CASP): round II.** *Proteins* 1997, **Suppl 1:**2-6.
35. Mclachlan AD: **A mathematical procedure for superimposing atomic coordinates of proteins.** *Acta Crystallogr A* 1972, **A28:**656-657.
36. Diamond R: **On the comparison of conformations using linear and quadratic transformations.** *Acta Crystallogr A* 1976, **A32:**1-10.
37. Rossmann MG, Liljas A, Branden CI and Banaszak LJ: **Evolutionary and structural relationships among the dehydrogenases.** *The Enzymes* 1975, **11:**61-102.
38. Vogt G, Etzold T and Argos P: **An assessment of amino acid exchange matrices in aligning protein sequences: The twilight zone revisited.** *J Mol Biol* 1995, **249:**816-831.
39. Ihaka R and Gentleman R: **R: A language for data analysis and graphics.** *Journal of Computational and Graphical Statistics* 1996, **5:**299-314.
40. Dayhoff MO, Schwartz RM and Orcutt BC: **A model of evolutionary change in proteins. Matrices for detecting distant relationships.** In *Atlas of protein sequence and structure Volume 5*. Edited by: *Dayhoff MO. Washington DC: National biomedical research foundation*; 1978:345-358.
41. Barton GJ: **ALSCRIPT: a tool to format multiple sequence alignments.** *Protein Eng* 1993, **6:**37-40.
42. Henikoff S and Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci USA* 1992, **89:**10915-10919.
43. Gonnet GH, Cohen MA and Benner SA: **Exhaustive matching of the entire protein sequence database.** *Science* 1992, **256:**1443-1445.
44. Webber C and Barton GJ: **Estimation of P-values for global alignments of protein sequences.** *Bioinformatics* 2001, **17:**1158-67.
45. Gupta SK, Kececioglu J and Schaffer AA: **Improving the practical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment.** *J Comput Biol* 1995, **2:**459-472.
46. Krogh A, Brown M, Mian IS, Sjolander K and Haussler D: **Hidden markov models in computational biology: Applications to protein modelling.** *J Mol Biol* 1994, **235:**1501-1531.
47. Smith RF and Smith TF: **Automatic generation of primary sequence patterns from sets of related protein sequences.** *Proc Natl Acad Sci USA* 1990, **87:**118-122.
48. Smith RF and Smith TF: **Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparitive protein modelling.** *Proteins* 1992, **5:**35-41.
49. Cuff JA and Barton GJ: **Evaluation and improvement of multiple sequence methods for protein secondary structure prediction.** *Proteins* 1999, **34:**508-519.
50. Cuff JA and Barton GJ: **Application of multiple sequence alignment profiles to improve protein secondary structure prediction.** *Proteins* 2000, **40:**502-511.
51. Murzin A, Brenner SE, Hubbard T and Chothia C: **SCOP: a structural classification of proteins database and the investigation of sequences and structures.** *J Mol Biol* 1995, **247:**536-540.
52. Vingron M and Argos P: **Determination of reliable regions in protein sequence alignments.** *Protien Eng* 1990, **3:**565-569.
53. Holmes I and Durbin R: **Dynamic programming alignment accuracy.** *J Comput Biol* 1998, **5:**493-504.
54. Cline M, Hughey R and Karplus K: **Predicting reliable regions in protein sequence alignments.** *Bioinformatics* 2002, **18:**306-314.