# BIOINFORMATICS

# APDB: a novel measure for benchmarking sequence alignment methods without reference alignments

*Orla O'Sullivan[1], Mark Zehnder[3], Des Higgins[1], Philipp Bucher[3], Aurelien Grosdidier[3] and Cédric Notredame[2, 3,*]*

[1]*Department of Biochemistry, University College, Cork, Ireland,* [2]*Information Génétique et Structurale, CNRS UMR-1889, 31, Chemin Joseph Aiguier, 13402 Marseille, France and* [3]*Swiss Institute of Bioinformatics, Chemin des Boveresse, 155, 1066 Epalinges, Switzerland*

## ABSTRACT

Author please check use of A and B heads is correct

**Motivation:** We describe APDB, a novel measure for evaluating the quality of a protein sequence alignment, given two or more PDB structures. This evaluation does not require a reference alignment or a structure superposition. APDB is designed to efficiently and objectively benchmark multiple sequence alignment methods.

**Results:** Using existing collections of reference multiple sequence alignments and existing alignment methods, we show that APDB gives results that are consistent with those obtained using conventional evaluations. We also show that APDB is suitable for evaluating sequence alignments that are structurally equivalent. We conclude that APDB provides an alternative to more conventional methods used for benchmarking sequence alignment packages.

**Availability:** APDB is implemented in C, its source code and its documentation are available for free on request from the authors.

**Contact:** cedric.notredame@europe.com

## INTRODUCTION

We introduce APDB (Analyze alignments with PDB), a new method for benchmarking and improving multiple sequence alignment packages with minimal human intervention. We show how it is possible to avoid the use of reference alignments when PDB structures are available for at least two homologous sequences in a test alignment. Using this method it should become possible to systematically benchmark or train multiple sequence alignment methods using all known structures, in a completely automatic manner.

There are strong justifications for improving multiple sequence alignment methods. Many sequence analysis techniques used in bioinformatics require the assembly of a multiple sequence alignment at some point. These include phylogenetic tree reconstruction, detection of remote homologues through the use of profiles or HMMs, secondary and tertiary structure prediction and more recently the identification of the nsSNPs (non synonymous Single Nucleotide Polymorphisms) that are most likely to alter a protein function. All of these important applications demonstrate the need to improve existing multiple sequence alignment methods and to establish their true limits and potential. Doing so is complicated, however, because most multiple sequence alignment methods rely on a complicated combination of greedy heuristic algorithms meant to optimize an objective function.

This objective function is an attempt to quantify the biological quality of an alignment. Almost every multiple alignment package uses a different empirical objective function of unknown biological relevance. In practice, most of these algorithms are known to perform well on some protein families and less well on others, but it is difficult to predict this in advance. It can also be very hard to establish the biological relevance of a multiple alignment of poorly characterized protein families. See Duret and Abdeddaim (2000) and Notredame (2002) for two recent reviews of the wide variety of techniques that have been used to make multiple alignments.

Given such a wide variety of methods and such poor theoretical justification for most of them, the main option for a rational comparison is systematic benchmarking. This is usually accomplished by comparing the alignments produced by various methods with 'reference' alignments of the same sequences assembled by specialists with the help of structural information. Barton and Sternberg (1987) made an early systematic attempt to validate a multiple sequence alignment method using structure based alignments of globins and immunoglobulins. Later on,

---

*To whom correspondence should be addressed.

Notredame and Higgins (1996) used another collection of such alignments assembled by Pascarella and Argos (1992). More recently, it has become common practice to use BAliBASE (Thompson *et al.*, 1999); a collection of multiple sequence alignments assembled by specialists and designed to systematically address the different types of problems that alignment programs encounter, such as the alignment of a distant homologue or long insertions and deletions. In this work, we examined two such reference collections: BaliBase and Homstrad (Mizuguchi *et al.*, 1998), a collection of high quality multiple structural alignments.

There are two simple ways to use a reference alignment for the purpose of benchmarking Karplus and Hu (2001). One may count the number of pairs of aligned residues in the target alignment that also occur in the reference alignment and divide this number by the total number of pairs of residues in the reference. This is the Sum of Pairs Score (SPS). The main drawback is that it is not very discriminating and tends to even out differences between methods. The more popular alternative is the Column Score (CS) where one measures the percentage of columns in the target alignment that also occur in the reference alignment. This is widely used and is considered to be a stringent measure of alignment performance. In order to avoid the problem of unalignable sections of protein sequences (i.e. segments that cannot be superimposed), it is common practice to annotate the most reliable regions of a multiple structural alignment and to only consider these core regions for the evaluation. In BaliBase, the core regions make up slightly less than 50% of the total number of alignment columns.

Such use of multiple sequence alignment collections for benchmarking is very convenient because of its simplicity. However, a major problem is the heavy reliance on the correctness of the reference alignment. This is serious because, by nature, these reference alignments are at least partially arbitrary. Although structural information can be handled more objectively than sequence information, the assembly of a multiple structural alignment remains a very complex problem for which no exact solution is known. As a consequence, any reference multiple alignment based on structure will necessarily reflect some bias from the methods and the specialist who made the assembly. The second drawback is that given a set of structures there can be more than one correct alignment. This plurality results from the fact that a structural superposition does not necessarily translate unambiguously into one sequence alignment. For instance, if we consider that the residues to be aligned correspond to the residues whose alpha carbons are the closest in the 3-D superposition, it is easy to imagine that sometimes an alpha carbon can be equally close to the alpha carbons of two potential homologous residues. Most structure based sequence alignment procedures break this tie in an arbitrary fashion, leading to a reference alignment that represents only one possible arrangement of aligned residues.

This problem becomes most serious when the sequences one is considering are distantly related (less than 30% identity). Unfortunately, this is also the most interesting level of similarity where most sequence alignment methods make errors and where it is important to accurately benchmark existing algorithms. The APDB method that we describe in this work has been designed to specifically address this problem and remove, almost entirely, the need for arbitrary decisions when using structures to evaluate the quality of a multiple sequence alignment.

In APDB, a target alignment is *not* evaluated against a reference alignment. Rather, we measure the quality of the structural superposition induced by the target alignment given any structures available for the sequences it contains. By treating the alignment as the result of some sort of structure superposition, we simply measure the fraction of aligned residues whose structural neighborhoods are similar. This makes it possible to avoid the most expensive and controversial element of the MSA benchmarking methods: the reference multiple sequence alignment. APDB requires just three parameters. This is tiny if we compare it with any reference alignment where each pair of aligned residue can arguably be considered as a free parameter.

In this work we show how the APDB measure was designed and characterized on a few carefully selected pairs of structures. Among other things we explored its sensitivity to parameter settings and various sequence and structure properties, such as similarity, length, or alignment quality. Finally, APDB was used to benchmark known methods using two popular data sets: BaliBase and Homstrad. These were either used as standard reference alignments or as collections of structures suitable for APDB.

It should be noted that there are several methods for evaluating the quality of structure models and predictions using known structures. The development of these has been driven by the need to evaluate entries in the CASP protein structure prediction competition and have been reviewed by Cristobal *et al.* (2001). These all depend on generating structure superpositions between the model and the target and evaluating the quality of the match using, for example, RMSD between the two or using some measure of the number of alpha carbons that superimpose well (e.g. MaxSub by Siew *et al.* (2000)). In principle, this could also be used to benchmark alignment methods. However, one serious disadvantage is the requirement for a superposition, which is itself a difficult problem. A second disadvantage is the way RMSD measures behave with different degrees of sequence divergence and their sensitivity to local or global alignment differences. We

have carefully designed APDB so that on the one hand it remains very simple but on the other hand it is able to measure the similarity of the structural environments in a manner that lends itself to measuring alignment quality.

## SYSTEM AND METHODS

### The APDB scoring function

APDB is a measure designed to evaluate how consistent an alignment is with the structure superposition this alignment implies. Let us imagine that A and B are two homologous structures. If the structure of sequence A tells us that the residues X and Z are 9Å apart, then we expect to find a similar distance between the two residues Y and W of sequence B that are aligned with X and Z. The difference between these two distances is an indicator of the alignment quality.

*Author please resupply fig as not reproducing correctly, thanks*

```
                    _____9Å_____
                         _             _
  A  aaaaaaaaaaaXaaaaaaaaaaaaaaaaaZaaaaaaa

  B  bbbbbbbbbbbYbbbbbbbbbbbbbbbbbWbbbbbbb
                    _____
                         9Å?
```

In APDB we take this idea further by measuring the differences of distances between X:Y (X *aligned with* Y) and Z:W within a bubble of fixed radius centered around X and Y. The bubble makes APDB a local measure, less sensitive than a classic RMSD measure to the existence of non-superposable parts in the structures being considered. Furthermore it ensures that a bad portion of the alignment does not dramatically affect the overall alignment evaluation. The typical radius of this bubble is 10Å, and it contains 20 to 40 amino acids. We consider two residues to be properly aligned if the distances from these two residues to the majority of their neighbors within the bubble are consistent between the two structures. In other words, we check whether a structural neighborhood is supportive of the alignment of the two residues that sit at its center. This can be formalized as follows:

**X : Y** and is a pair of aligned residues in the alignment
**N** Number of aligned pairs of residues
**d(X, Z)** is the distance between the C$\alpha$ of the two residues $X$ and $Z$ within one structure.
**Brad** is the radius of the bubble set around residues $X$ and $Y$ (Brad = 10 Å by default).
**T1** is the maximum difference of distance between $d(X, Z)$ and $d(Y, W)$ ($T1 = 1$ Å by default).
**T2** is the minimal percentage of residues that must respect the criterion set by $T1$ for $X$ and $Y$ to be considered correctly aligned (70% by default).
**considered**$_{X:Y}$(**Z : W**) is equal to 1 if the pair $Z : W$ is in the bubble defined by pair $X : Y$

**correct**$_{X:Y}$(**Z : W**) is equal to 1 if $d(X, Z)$ and d(Y,W) are sufficiently similar as set by $T1$.
**aligned**(**X : Y**) is equal to 1 if most pairs $Z : W$ in the $X : Y$ bubble are correct as set by $T2$.

$$considered_{X:Y}(Z : W) = 1$$
$$\text{if } d(X, Z) < \text{Brad and } d(Y, W) < \text{Brad} \quad (1)$$
$$correct_{X:Y}(Z : W) = 1$$
$$\text{if } d(X, Z) < \text{Brad and } d(Y, W) < \text{Brad} \quad (2)$$
$$\text{and } |d(X, Z) - d(Y, W)| < T1$$

$$aligned(X : Y) = 1$$
$$\text{if } \frac{\sum_{Z:W} Correct_{X:Y}(Z : W)}{\sum_{Z:W} Considered_{X:Y}(Z : W)} \times 100 > T2 \quad (3)$$

Finally, the APDB measure for the entire alignment is defined as:

$$APDB \; Score = \frac{\sum_{X:Y} Aligned(X : Y)}{N} \quad (4)$$

Given a multiple alignment of sequences with known structures, the APDB score can easily be turned into a sum of pairs score by summing the APDB score of each pair of structures and dividing it by the total number of sequence pairs considered.

### Design of a benchmark system for apdb

In order to study the behavior of APDB, we used two established collections of reference alignments: BAliBASE (Thompson *et al.*, 1999) and HOMSTRAD (Mizuguchi *et al.*, 1998). First we extracted 9 structure based pair-wise sequence alignments from HOMSTRAD, which we refer to as HOM_9. These reference alignments were chosen so that their sequence identities (as measured on the HOMSTRAD reference alignments) evenly cover the range 17 to 90%. These alignments are between 200 and 300 residues long and are used for detailed analysis and parameterization of APDB. The PDB names of the pairs of structures are given in the figure legend for Figure 2. Next, in order to assemble a discriminating test set, we selected the most difficult alignments from HOMSTRAD. We chose alignments which had at least 4 sequences and where the average percent identity was 25% or less. This resulted in a selection of 43 alignments, which we refer to as HOM_43. BAliBASE version 1 has 141 alignments divided into 5 reference groups. We chose all alignments where 2 or more of the sequences had a known structure. This resulted in a subset of 91 alignments from the first 4 reference groups of BAliBASE which we refer to as BALI_91. Minor adjustments had to be made to ensure consistency between BAliBASE sequences and the corresponding PDB files. HOM_43 and BALI_91 test sets are available in the APDB distribution.
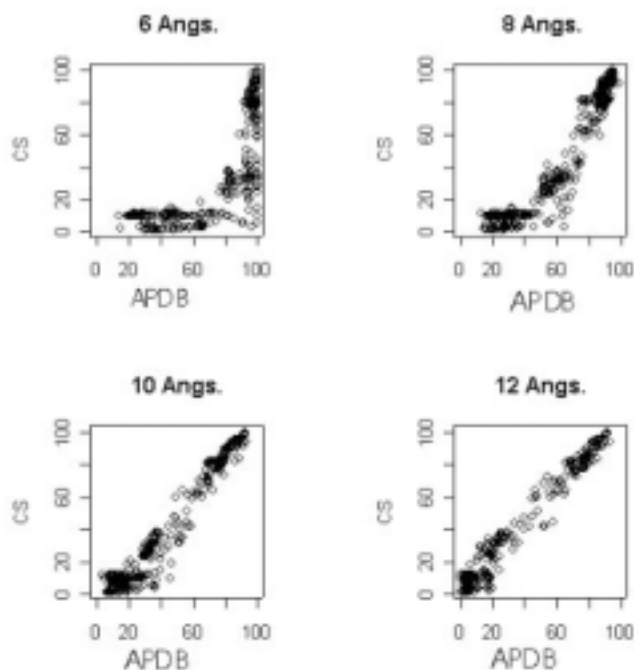
**Fig. 1. Tuning of Brad, the bubble radius using sub-optimal alignments of two sequences from HOM_9** Each graph represents the correlation between CS and APDB for 4 different Bubble Radius values (Brad of 6, 8, 10 and 12Å). In each graph, each dot represents a sub-optimal alignment from HOM_9, sampled from the genetic algorithm.

### Generation of multiple alignments

We compared the performance of APDB on two different multiple alignment methods. We tested the widely used ClustalW version 1.80 (Thompson *et al.*, 1994). We also tested the more recent T-Coffee version 1.37 (Notredame *et al.*, 2000) using default parameters.

### Generation of suboptimal alignments

In order to evaluate the sensitivity of APDB to the quality of an alignment, we used an improved version of the genetic algorithm SAGA (Notredame and Higgins, 1996) in order to generate populations of sub-optimal alignments. In each case a pair of sequences was chosen in HOM_9 and 50 random alignments were generated and allowed to evolve within SAGA so that their quality gradually improved (as measured by their similarity with the HOMSTRAD reference alignment). Ten alignments were sampled at each generation in order to build a collection of alternative alignments with varying degrees of quality. This algorithm was stopped when optimality was reached, thus typically yielding collections of a few hundred alignments.

A second method for generating sub-optimal alignments was based on the PROSUP package (Lackner *et al.*, 2000). PROSUP takes two structures, makes a rigid body superposition and generates all the sequence alignments that are consistent with this superposition, thus producing alternative alignments that are equivalent from a structural point of view. Typically PROSUP yields 5 to 25 alternative alignments within a very narrow range of RMSDs.

### Comparison of apdb with other standard measures

In order to compare the APDB measure with more conventional measures, we used the Column Score (CS) measure as provided by the aln_compare package (Notredame *et al.*, 2000). CS measures the percentage of columns in a test alignment that also occur in the reference alignment. In BAliBASE this measure is restricted to those columns annotated as core region in the reference. Although alternative measures have recently been introduced (Karplus and Hu, 2001), CS has the advantage of being one of the most widely used and the simplest method available today.

## RESULTS AND DISCUSSION

### Fine tuning of apdb

Three parameters control the behaviour of APDB: Brad (the bubble radius), T1 (the difference of distance threshold) and T2 (the fraction of the bubble neighbourhood that must support the alignment of two residues). We exhaustively studied the tuning effect of each of these parameters using HOM_9 and parameterised APDB so that its behaviour is as consistent as possible with the behaviour of CS on HOM_9.

In Figure 1 we show the relationship between CS and APDB for 250 sub-optimal alignments generated by genetic algorithm for one of the 9 test cases from HOM_9 over 4 different settings of Brad, the Bubble Radius. While the two scoring schemes are in broad agreement, the correlation improves dramatically as Brad increases. This trend can be summarised using the correlation coefficient measured on each of the graphs similar to those shown in Figure 1. The overall results for all nine HOM_9 test cases are shown in Figure 2. These results clearly show that the behaviour of APDB is best for values of Brad of 10 Å or above. With these values the level of correlation between CS and APDB increases and so does the agreement across all 9 test cases. We chose 10 Å as the default value in order to ensure a proper behaviour while retaining as much as possible the local character of the measure. Given the default value of 10 Å for Brad, we examined T1 and T2 in a similar fashion and found the most appropriate values as 1 Å for T1 and 70% for T2.
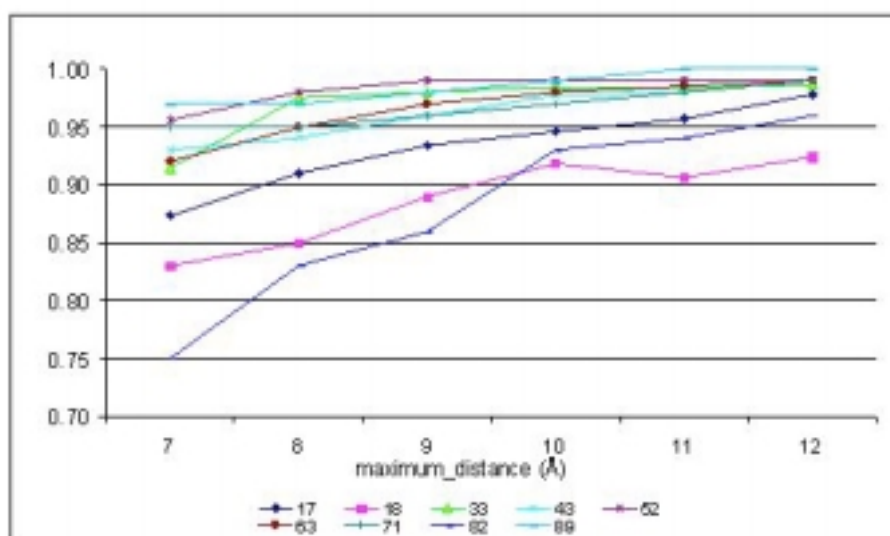
**Fig. 2. Correlation between the Column Score measure (CS) and APDB on HOM_9** Each HOM_9 test set is labelled according to its average percent sequence identity as measured on the HOMSTRAD reference. The horizontal axis indicates the value of Brad. The vertical axis indicates the correlation coefficient between CS and APDB as measured on a population of sub-optimal alignments similar to the ones in Figure 1. Each dot indicates a correlation coefficient measured on one HOM_9 test set, using the indicated value of Brad. Each HOM_9 test set is an alignment between two sequences whose PDB names are as follows: 17: 2gar versus 1fmt, 18: ljfl versus lb74, 33: 1isi versus 11be, 43: 2cev versus 1d3v, 52: 1aq0 versus 1ghs, 63: 2gnk versus 2pii, 71: 1hcz versus 1cfm, 82: 1dvg versus 1qq8, 89: 1k25 versus 1qme.

## Sensitivity of apdb to sequence and structure similarity

It is important to verify that the behaviour of APDB remains consistent across a wide range of sequence similarity levels. It is especially important to make sure that when two different alignments of the same sequences are evaluated, the best one (as judged by comparison with the HOMSTRAD reference) always gets the best APDB score. In order to check for this, we used the genetic algorithm to generate sub-optimal alignments for each test case in HOM_9. In each case, we gathered a collection of 250 sub-optimal alignments with CS scores of 0–40%, 41–60%, 61–80% and 81–100%. The CS score measures the agreement between an alignments and its reference in HOMSTRAD. We then measured the average APDB score in each of these collections. Each of these measures corresponds to a dot in Figure 3 where vertically aligned series of dots correspond to different measures made on the same HOM_9 test set.

Figure 3 clearly shows that regardless of the percent identity within the HOM_9 test set being considered, alignments with better CS scores always correspond to a better APDB score (this results in the lines never crossing one another on Fig. 3). We did a similar analysis using the RMSD as measured on the HOMSTRAD alignment in place of sequence identity. The behaviour was the

same and clearly indicates that APDB gives consistent results regardless of the structural similarity between the structures being considered.

## Suitability of apdb for analysing sub-optimal alignments

Collections of sub-optimal alignments for each of the nine HOM_9 test sets were generated using SAGA and evaluated for their CS scores and APDB scores. These results were pooled and are displayed on the graph shown on Figure 4. This Figure indicates good agreement between the CS and the APDB score regardless of the level of optimality within the alignment being considered. It suggests that APDB is informative over the complete range of CS values. It also confirms that APDB is not 'too generous' with sub-optimal alignments

We also checked whether sequence alignments that are structurally equivalent obtain similar APDB scores even if they are different at the sequence level. For this purpose, we used PROSUP (Lackner *et al.*, 2000). Given a pair of structures, PROSUP generates several alignments that are equally good from a structure point of view (similar RMSD), but can be very different at the sequence level (different Column Score). We manually identified two such test sets in HOMSTRAD and the results are summarized in Table 1. For each of these two test sets, we
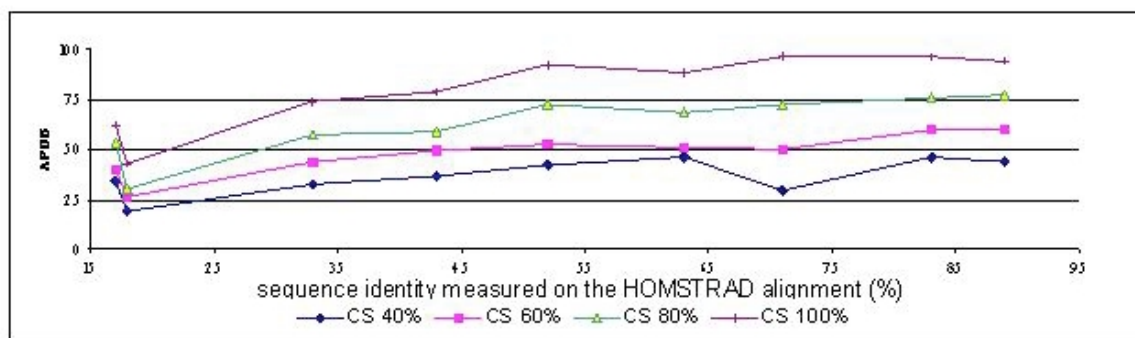
**Fig. 3. Estimation of the sensitivity of APDB to sequence identity** On this graph, each set of vertically aligned dots corresponds to a single HOM_9 test set. The 9 HOM_9 test sets are arranged according to their average identity (17–89%, see Figure 2). Each dot represents the average APDB score of a population of 250 sub-optimal alignments (generated by genetic algorithm) with a similar CS score (binned in four groups representing CS of <40%, 41–60%, 61–80% and 81–100%) generated for one of the 9 HOM_9 test sets.
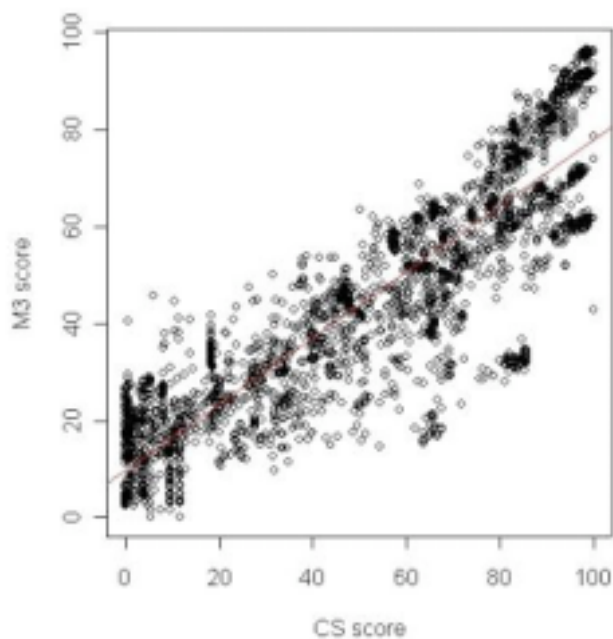


**Fig. 4. Correlation between CS and APDB on the complete HOM_9 test set** Each dot corresponds to a sub-optimal alignment of one of the HOM_9 test cases, generated by genetic algorithm. For each alignment the graph plots the APDB score against its CS counterpart.

**Table 1.** Evaluating PROSUP suboptimal alignments with APDB

| Set | St1 | St2 | ALN | RMSD | CS | APDB |
|---|---|---|---|---|---|---|
| 1 | 1e96B | 1a17 | aln1 | 1.45Å | 100.0 | 80.2 |
|  | 1e96B | 1a17 | aln2 | 1.50Å | 65.6 | 80.7 |
| 2 | 1cd8 | 1qfpa | aln1 | 2.95Å | 100.0 | 18.7 |
|  | 1cd8 | 1qfpa | aln2 | 2.95Å | 55.1 | 17.9 |

Set indicates the test set index, St1 and St2 indicate the two structures being aligned by PROSUP, ALN indicates the alignment being considered, RMSD shows the RMSD associated with this alignment, CS indicates its CS score, with the CS score of aln1 alignments being set to 100 because they are used as references. APDB indicates the APDB score.

In both test sets, using *aln1* as a reference for the CS measure leads to the conclusion that *aln2* is mostly incorrect (cf. CS column of Table 1). This is not true since these alignments are structurally equivalent as indicated by their RMSDs. In such a situation, APDB behaves much more appropriately and gives to each couple *aln1/aln2* scores that are nicely consistent with their RMSD, thus indicating that APDB can equally well reward two sub-optimal alignments when these are equivalent from a structural point of view.

**Using apdb to benchmark alignment methods**

Table 2 shows the average CS and APDB scores for the test sets in each of the four Bali_91 categories being considered here and in HOM_43. The highest scores in all cases, for both measures, come from the reference column (the last column). This is desirable providing the reference alignments really are consistent with the

selected in the output of PROSUP two alignments (*aln1* and *aln2*) to which PROSUP assigns similar RMSDs. *aln1* is used as a reference and therefore gets a CS score of 100 while the CS score of the second alignment (*aln2*) is computed by direct comparison with its *aln1* counterpart.

**Table 2.** Correlation between APDB and CS on BaliBase and Homstrad

| Set | N | ClustalW | | T-Coffee | | Reference | |
| --- | --- | CS | **APDB** | CS | **APDB** | CS | **APDB** |
| B91 R1 | 35 | 70.1 | **59.9** | 67.7 | **58.3** | 100 | **64.7** |
| B91 R2 | 23 | 32.7 | **26.6** | 33.9 | **47.1** | 100 | **55.2** |
| B91 R3 | 22 | 46.4 | **38.5** | 48.6 | **46.9** | 100 | **53.2** |
| B91 R4 | 11 | 52.0 | **59.5** | 52.5 | **64.5** | 100 | **65.7** |
| H43 | 43 | 35.4 | **60.2** | 38.9 | **61.6** | 100 | **72.9** |

Test Set: indicates the test set being considered, either one of the BaliBase_91 references (B91R#) or HOM_43(H43), a subset of HOMSTRAD. *N* indicates the number of test alignments in this category. *ClustalW* indicates a set of measures made on alignments generated with ClustalW. *T-Coffee* indicates similar measures made on T-Coffee generated alignments. Reference indicates measures made on the reference alignments as provided in BaliBase or in Homstrad. *CS* columns are the Column Score measures while APDB indicates similar measures made using APDB.

underlying structures. If we now compare the columns two by two, we find that every variation on CS from one column to another agrees with the corresponding variation of APDB. For instance in row 1 (Bali_91 Ref1), when T-Coffee/CS is lower than ClustalW/CS, T-Coffee/APDB is also lower. This observation is true for the whole table, regardless of the pair of results being considered. When considering the 134 alignments one by one, this observation remains true in more than 70 % of the cases.

## CONCLUSION

This work introduces APDB, a novel method that makes it possible to evaluate the quality of a sequence alignment when two or more tertiary structures of the sequences it contains are available. This method does not require a reference alignment and it does not depend on any complex procedure such as structure superposition or sequence alignment. We show here that APDB sensitivity is comparable with that of CS, a well-established measure that compares a target alignment with a reference alignment. Our results also indicate that APDB can discriminate better than CS between structurally correct sub-optimal sequence alignments and structurally incorrect sequence alignments, even when the structures being considered are distantly related.

Apart from the cost associated with their assembly, a serious problem with reference alignments is that they need to be annotated to remove from the evaluation regions that correspond to non-superposable portions of the structures. This is necessary because otherwise these regions (whose alignment cannot be trusted) will bias a CS evaluation toward rewarding the arbitrary alignment conformation displayed in the reference. Table 2 illustrates well the fact that such an annotation is not necessary in APDB. In our measure, thanks to the combination of local evaluation and the absence of a reference alignment, the only possible effect of non-superposable regions is to decrease the proportion of residues found aligned in a structurally optimal sequence alignment, thus yielding scores lower than 100 in the case of distantly related structures.

A key advantage of APDB is its simplicity. It only requires three parameters and a few PDB files. Most importantly, APDB does not require any arbitrary manual intervention such as the assembly of a reference alignment. In the short term, all the existing collections of reference alignment could easily be integrated and extended with APDB. In the longer term, APDB could also be used to evaluate and compare existing collections of alignments such as profiles, when structures are available.

## REFERENCES

Barton,G.J. and Sternberg,M.J.E. (1987) A strategy for the rapid multiple alignment of protein sequences: confidence levels from tertiary structure comparisons. *J. Mol. Biol.*, **198**, 327–337.

Cristobal,S., Zemla,A., Fischer,D., Rychlewski,L. and Elofsson,A. (2001) A study of quality measures for protein threading models. *BMC Bioinformatics*, **2**, 5.

Duret,L. and Abdeddaim,S. (2000) Multiple alignment for structural, functional, or phylogenetic analyses of homologous sequences. *Bioinformatics, Sequence, Structure and Databanks*. Higgins,D. and Taylor,W. (eds), Oxford University Press, Oxford.

Karplus,K. and Hu,B. (2001) Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. *Bioinformatics*, **17**, 713–720.

Lackner,P., Koppensteiner,W.A., Sippl,M.J. and Domingues,F.S. (2000) ProSup: a refined tool for protein structure alignment. *Protein Eng.*, **13**, 745–752.

Mizuguchi,K., Deane,C.M., Blundell,T.L. and Overington,J.P. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, **7**, 2469–2471.

Notredame,C. (2002) Recent progress in multiple sequence alignments. *Pharmacogenomics*, **3**, 131–144.

Notredame,C. and Higgins,D.G. (1996) SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res.*, **24**, 1515–1524.

Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel algorithm for multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.

Pascarella,S. and Argos,P. (1992) A data bank merging related protein structures and sequences. *Protein Eng.*, **5**, 121–137.

Siew,N., Elofsson,A., Rychlewski,L. and Fischer,D. (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, **16**, 776–785.

Thompson,J., Higgins,D. and Gibson,T. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4690.

Thompson,J., Plewniak,F. and Poch,O. (1999) BaliBase: a benchmark alignment database for the evaluation of multiple sequence alignment programs. *Bioinformatics*, **15(1)**, 87–88.