BIOINFORMATICS

Vol. 22 no. 14 2006, pages e35–e39
doi:10.1093/bioinformatics/btl218

The iRMSD: a local measure of sequence alignment accuracy using structural information

Fabrice Armougom[1], Sébastien Moretti[1], Vladimir Keduas[1] and Cedric Notredame[1,*]

[1]Laboratoire Information Génomique et Structurale, CNRS UPR2589, Institute for Structural Biology and Microbiology (IBSM), Parc Scientifique de Luminy, case 934, 163 Avenue de Luminy, FR-13288, Marseille cedex 09

ABSTRACT

Motivation: We introduce the iRMSD, a new type of RMSD, independent from any structure superposition and suitable for evaluating sequence alignments of proteins with known structures.

Results: We demonstrate that the iRMSD is equivalent to the standard RMSD although much simpler to compute and we also show that it is suitable for comparing sequence alignments and benchmarking multiple sequence alignment methods. We tested the iRMSD score on 6 established multiple sequence alignment packages and found the results to be consistent with those obtained using an established reference alignment collection like Prefab.

Availability: The iRMSD is part of the T-Coffee package and is distributed as an open source freeware (http://www.tcoffee.org/).

Contact: cedric.notredame@europe.com; cedric.notredame@igs.cnrs-mrs.fr

1 INTRODUCTION

The computation of accurate sequence alignments constitutes a pre-requisite for an ever increasing number of biological analyses. These include phylogenetic reconstruction, structure prediction, domain based analysis, function prediction and comparative genomics. In all these cases, the purpose of the alignment is to exploit evolutionary variations in order to reveal biologically meaningful patterns. The discovery and the proper analysis of these patterns depend entirely on the alignment correctness.

In many cases, an alignment is considered to be biologically correct when it accurately reflects the structural relationship between the considered sequences. This result is achieved by matching structurally equivalent residues. Assembling such an alignment is trivial when the sequences are highly similar but becomes harder for remote homologues. When considering alignments of sequences with less than 25% identity (the so-called twilight zone), standard scoring schemes like substitution matrices become uninformative and it can be difficult to determine the alignment accuracy, or even whether the sequences are truly related or not. So far, the most satisfying way of aligning remote homologues has been to use structural information whenever possible (Huang and Bystroff, 2006; Lesk and Chothia, 1980).

The use of structural information, however, carries its own peril, and while the sequence analysis community tends to consider structure based alignments as unambiguous and unquestionable gold standards, a closer look reveals a much less clear cut situation. More than 20 structure alignment packages have been developed (Goldsmith-Fischman and Honig, 2003). All these packages tend to produce different alignments because of their different underlying optimization algorithms. Furthermore, the lack of a universally accepted criterion for describing the quality of a structural alignment makes it difficult to determine the relative merits of all these packages (Kolodny, *et al*., 2005). The most common procedure to evaluate structure superpositions is to use the root mean square distance deviation (RMSD) of superposed atoms. This measure estimates the mean square distance between the equivalent alpha carbons of the two superposed structures. It can be ambiguous because of its dependence on two critical parameters: the minimization method and the procedure used to exclude structurally non equivalent regions (loops for instance).

Having several methods that deliver structure based sequence alignments and not knowing which one does best is a major issue in a context where structure-based alignments are routinely used to improve and guide the development of sequence alignment methods (Wallace, *et al*., 2005). A direct consequence of this situation has been the development of at least five collections of reference structure based sequence alignments (Edgar, 2004; Mizuguchi, *et al*., 1998; O'Sullivan, *et al*., 2004; Raghava, *et al*., 2003; Thompson, *et al*., 2005; Van Walle, *et al*., 2005). These collections are all used for a similar purpose: the benchmark of sequence alignment algorithms. Since it is virtually impossible to compare these datasets and decide whether some are more informative than others, the most common practice is to use them all, and look for common trends in the global results (Katoh, *et al*., 2005).

While results measured on these reference collections tend to agree for datasets with more than 30% identity, variations appear when considering sets of remote homologues (Katoh, *et al*., 2005). Aside from potential accuracy problems, the simplest explanation for these discrepancies is the possibility for alternative sequence alignments to be structurally equivalent, especially when considering remote homologues (Lackner, *et al*., 2000). In this context, setting one specific alignment as a reference becomes an arbitrary choice and therefore a bias toward specific alignment methods. In practice, the authors try to minimize that effect by specifying the core regions that should be used for the comparison, but this choice is also difficult and somehow arbitrary. We suggest in this paper that replacing the reference alignments with an RMSD measure would

*To whom correspondence should be addressed.

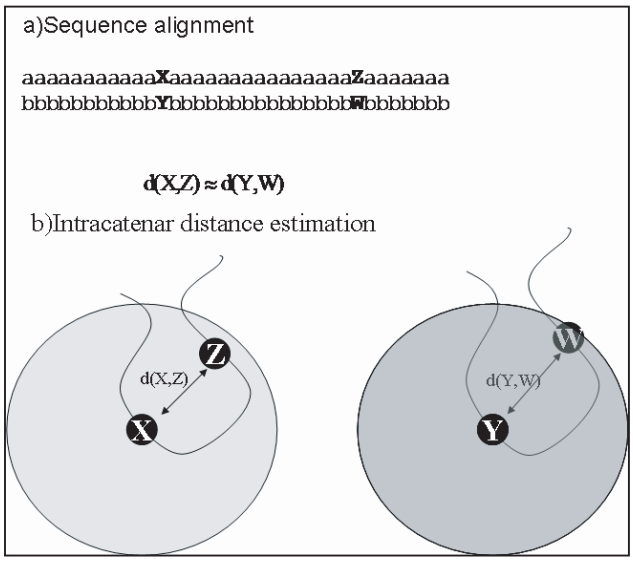

Fig 1. Basic principle of the iRMSD. Equivalences implied by the sequence alignment are tested on the structure. The assumption is that if XY and ZW are correctly aligned, then the distance between residues XZ and YW must be similar. ZW pairs are only considered if they are within a sphere of radius *R*, centered on X and Y.

be a more objective way to evaluate the sequence alignments of proteins. The RMSD has two advantages over standard methods: no dependence on a reference alignment and the possibility to quantify the structural correctness of any protein sequence alignments (provided the protein structures are known). The main drawback, however, is the reliance of the RMSD on a structure superposition strategy. This key step affords many alternative solutions whose relative merits are difficult to estimate (Kolodny, *et al*., 2005).

We redesigned the RMSD measure to make it independent from any structure superposition procedure. We named this measure iRMSD because it is an RMSD based on intra-molecular distance comparisons. The iRMSD is a follow up of the APDB measure (O'Sullivan, *et al*., 2003), designed to evaluate alignments for their compatibility with the structural superposition they imply. While APDB was a complex measure depending on three semi arbitrary parameters, the new iRMSD algorithm only requires one parameter. We show here that the iRMSD behaves just like a standard RMSD both numerically (values range) and structurally (similar structural meaning). We finally show that a straightforward normalization makes the iRMSD perfectly suitable for evaluating and comparing sequence alignment methods without the need of pre-established reference alignment collections.

2 METHODS

2.1 The iRMD measure

The iRMSD measure follows the underlying principle of APDB: given a correct alignment of two protein sequences A and B (Figure 1), if X is aligned with Y and Z with W, then the XZ distance (d(XZ)) must be similar to d(YW). The better the alignment of A and B, the smaller the average difference between all possible pairs d(XZ) and d(YW). The iRMSD associated with the aligned pair X and Y is estimated by considering every aligned pair Z and W within a sphere of radius (r) centered on X and Y that verifies the equation:

$$\mathrm{d(XW)} < r \ \textbf{AND} \ \mathrm{d(YZ)} < r \tag{1}$$

The ensemble of pairs ZW that verify equation 1 is named the neighborhood and noted N(XY). The default value of r is 10 Å (O'Sullivan, *et al*., 2003), which corresponds to a neighborhood size of 20-40 residues. The local iRMSD can be estimated as follows:

$$iRMSD(XY) = \sqrt{\frac{\sum_{ZW}(d(XZ)-d(YW))^2}{N(XY)}} \tag{2}$$

The summation is made over all the aligned ZW pairs within the neighborhood (Equation 1). Pairs XY with an empty neighborhood have their local iRMSD undefined. The global measure is obtained by summing on every pair XY and dividing by the number of pairs with a non empty neighborhood (N):

$$iRMSD = \frac{\sum_{XY} iRMSd(XY)}{N} \tag{3}$$

The iRMSD thus defined is not suitable for comparing alternative alignments, as it tends to give a better score to alignments with long gaps and few well aligned residues. In order to simultaneously take into account the superposition accuracy and the extent of the alignment (i.e. the number of matched residues), we adapted the CI formula of Kleywegt and Jones (Kleywegt and Jones, 1994) to turn the iRMSD into a Normalized iRMSD (NiRMSD):

$$NiRMSd = \frac{iRMSD^{*}MIN(L1,L2)}{N} \tag{4}$$

L1 and L2 are the respective lengths of the two sequences, and N the number of residue pairs with a non empty neighborhood. This formula amounts to incorporating a gap penalty that deals with indels and aligned pairs whose neighborhood is empty.

2.2 Validation procedure using Prefab

We used the Prefab (Edgar, 2004) collection of reference alignments to analyze the iRMSD. Prefab is an extensive collection of 1682 pairwise structural alignments obtained by combining the output of two structure alignment programs: CE (Shindyalov and Bourne, 1998) and DALI (Holm and Sander, 1993). In each of these alignments the authors have defined core regions where the DALI and the CE methods agree and have used these regions for evaluation purpose. Given one Prefab reference alignment and an alternative target alignment of the same sequences, the Qscore is defined as the fraction of core columns in the reference alignment found aligned identically in the target. In order to evaluate multiple sequence alignment packages, Prefab also includes in each dataset a collection of about 48 sequences homologous to the two structures. When evaluating an MSA package, the large dataset is aligned and the Qscore is measured on the core regions of the induced alignment of the two structures.

We evaluated the RMSD and the iRMSD of Prefab alignments. However, because of various inconsistencies between the ATOM, the SEQRES fields of the PDB entries and the sequences of the Prefab alignments, LSQMAN could only handle 587 of the original Prefab entries. This sample had roughly the same identity distribution as the entire Prefab (243 dataset having with than 20% identity (on the reference Prefab alignment), 172 between 20 and 40% identity and 171 with more than 40% identity). We believe it to be representative and large enough for the purpose of the present analysis.

2.3 Evaluation of the standard RMSD

We used the LSQMAN package (Kleywegt and Jones, 1999) to estimate the standard RMSD associated with the Prefab alignments. The local RMSD was estimated by superposing the residues contained in a window of size 21 (2∗10+1) centered on a pair of aligned residues. The superposition was

made using the Xalignment function of the LSQMAN package. The overall RMSD was obtained by sliding the window and averaging over all the windows.

2.4 Multiple sequence alignment methods

We benchmarked the iRMSD measure on the alignments produced using the public distributions of six multiple sequence alignment packages: ClustalW (Version 1.83) (Thompson, *et al*., 1994), DialignII (Version 2.2.1) (Morgenstern, 1999), Muscle (Version 3.6) (Edgar, 2004), Mafft (Version 5.6) (Katoh, *et al*., 2005), ProbCons (Version 1.10) (Do, *et al*., 2005) and T-Coffee (Version 3.75) (Notredame, *et al*., 2000).

2.5 Availability

The iRMSD package is part of the t_coffee package. It is an open source freeware that can be downloaded on http://www.tcoffee.org/. It comes along with an extensive documentation.

3 RESULTS

We started by comparing the iRMSD with the standard RMSD. We did so by measuring the scores associated with the 587 Prefab alignments. The measurements were either made on core regions (Figure 2a) or on the entire Prefab Alignments (Figure 2b). Both figures indicate a very strong correlation between the two measures. The core analysis gave an r^2 correlation coefficient of 0.92 while the measure on the entire alignments gave an r^2 of 0.93. As expected, the dispersion increases with the RMSD values. The Prefab alignments are high quality structure based alignments, but we also checked the behavior of the methods when analyzing alignments of lower quality (Figure 2c). We selected the Dialign method whose alignments have an average Prefab Qscore of 0.65 on the entire dataset (0.32 in the [0-20] identity range). Figure 2c shows that the two measures remain correlated up to an RMSD of 2.5 Å ($r^2 = 0.75$), indicating a saturation of the iRMSD measure for values above 1.6 Å. This apparent saturation is a consequence of the different local substructures compared by each method (windows for the RMSD and sphere for the iRMSD) and it does not occur when measuring the standard RMSD on spheres of radius 10 Å rather than on windows. When doing so the correlation is very good ($r^2 = 0.91$ over the full range, data not shown).

We further checked the local aspect of the measures by plotting both the local iRMSD and the local RMSD against several Prefab alignments. The 1aoh_1anu example is displayed on Figure 3 and clearly shows that both measures are well coordinated all along the alignment. While the iRMSD indicates two narrow peaks not found in the RMSD, both methods agree on the final series of peaks. We used LSQMAN to superpose the two structures and were satisfied to find that the peaks showing in the iRMSD curve effectively correspond to regions poorly superposed. Although the iRMSD seems to reveal more sharply these locations, it is fair to say that the standard RMSD could probably be parameterized to yield similar results (for instance by lowering the window size).

Having established that the iRMSD behaves like a standard RMSD measure we then estimated whether that measure is suitable for evaluating the relative accuracy of multiple sequence alignment packages. For that purpose, we aligned the Prefab datasets with six MSA methods and for each of these methods we evaluated the Qscore, the Normalized iRMSD (NiRMSD, Equation 3) and estimated the fraction of alignments having a NiRMSD better or

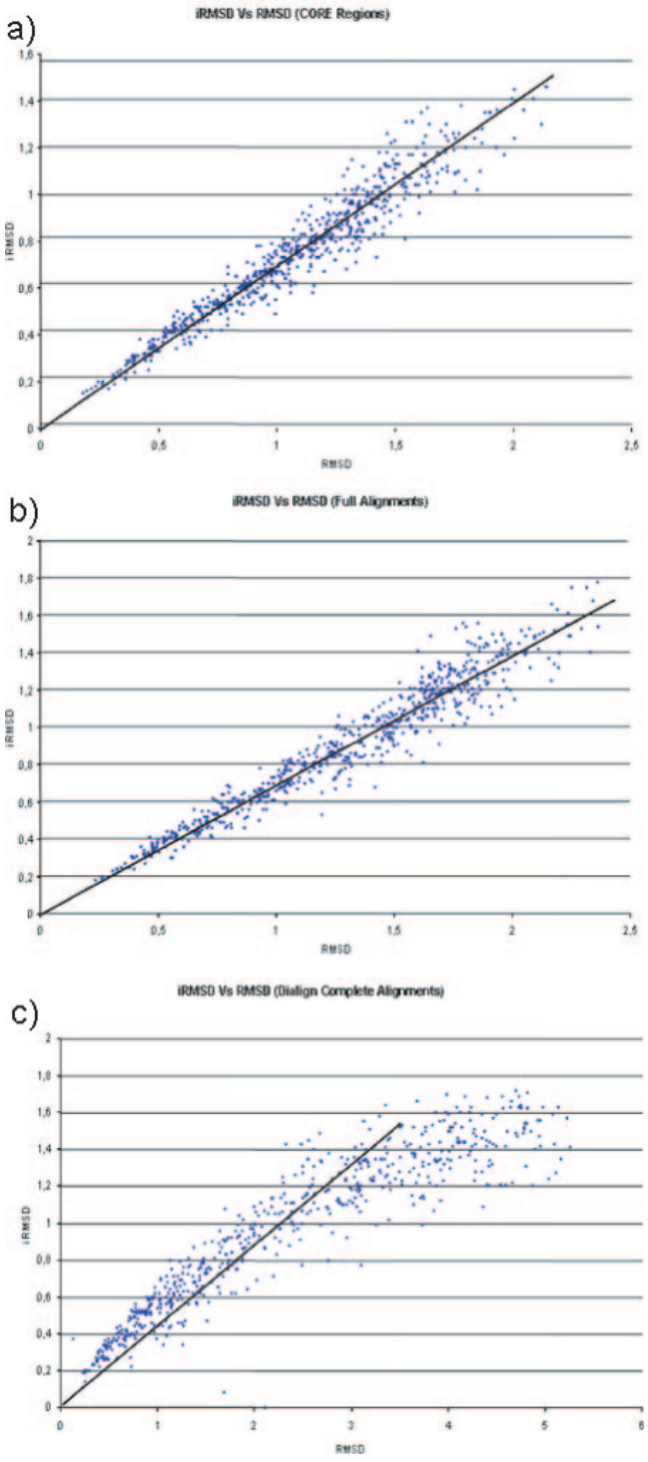

Fig 2. Correlation between the iRMSD and a standard LSQMAN RMSD. 1a) RMSD versus iRMSD of 587 of Prefab reference Alignments. The (i)RMSDs were only measured on the regions annotated as core in Prefab. The iRMSD is on the vertical axis and the regular RMSD, as obtained from LSQMAN, is on the horizontal axis. Each dot corresponds to one dataset. 2a) RMSD versus iRMSD on 587 Prefab reference Alignments. The (i)RMSDs were measured on the entire alignments. 2c) RMSD versus iRMSD on 587 Prefab datasets, aligned by Dialign. The dataset is the same as before and the (i)RMSDs were measured on the entire alignments.

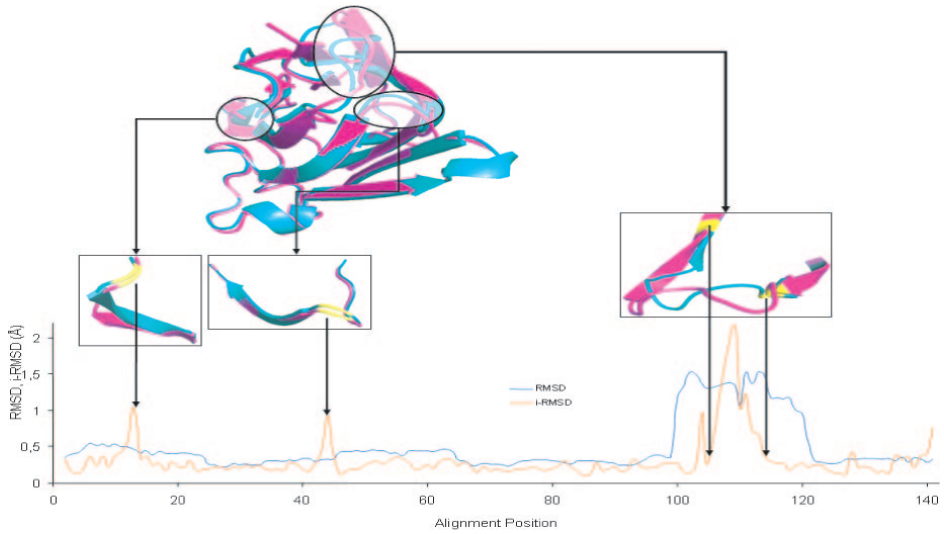

Fig 3. Local Comparison of the iRMSD against a *standard LSQMAN RMSD*. The comparison was made on the Prefab reference alignment of 1aohA_1anu. The two structures were superposed by LSQMAN (1aohA: violet, 1anu:blue). The alignment was then evaluated locally using either LSQMAN to measure the RMSD (Blue line) or T-Coffee/iRMSD to measure the local iRMSD. The (i)RMSDs values were plotted on the vertical axis against the alignment positions. Portion of the superposition corresponding to the peak were extracted and encapsulated.

Table 1. Average Qscore

Range	N	Dialign	Clustal	Muscle	TCoffee	ProbC.	MAFFT	PREFAB
0-20	243	0.32	0.34	0.43	0.44	0.48	**0.49**	----
20-40	171	0.80	0.83	0.86	0.87	**0.88**	**0.88**	----
40-100	173	0.96	0.96	0.97	**0.98**	0.97	**0.98**	----
Total	587	0.65	0.67	0.71	0.73	0.74	**0.75**	----

a) **Average Qscore**: Range is the range of identity of the considered Prefab datasets, as measured on the reference alignments. N is the number of Prefab datasets in each range. Dialign, ClustalW, Muscle, TCoffee, ProbCons and Mafft are the average Qscores as measured on the alignments produced by these packages. The entries corresponding to the best performance for each category are underlined and in bold. The best Qscore are the highest.

Range	N	Dialign	Clustal	Muscle	TCoffee	ProbC.	MAFFT	PREFAB
0-20	243	3.46	2.10	1.82	2.16	1.85	**1.76**	0.85
20-40	171	0.91	0.82	0.80	0.79	**0.77**	**0.77**	0.67
40-100	173	0.44	0.58	0.44	0.44	0.44	**0.43**	0.43
Total	587	1.83	1.28	1.11	1.25	1.12	**1.08**	0.67

b) **Average NiRMSD**: The labels are the same. The measure is the average NiRMSD as measured on the core regions of the alignments. The Prefab column corresponds to the evaluation of the Prefab reference alignments. The best NiRMSD scores are the lowest.

Range	N	Dialign	Clustal	Muscle	TCoffee	ProbC.	MAFFT	PREFAB
0-20	243	0.02	0.10	0.05	0.09	0.06	**0.10**	----
20-40	171	0.36	0.36	0.46	0.56	**0.57**	**0.54**	----
40-100	173	0.86	0.89	0.89	**0.92**	0.89	0.91	----
Total	587	0.36	0.40	0.42	0.47	0.45	**0.47**	----

c) **Best NirRMSD** Fraction: fraction of alignments having a NiRMSD better or equal to the Prefab reference as measured on the core regions. The labels are the same.

Table 2. Consistency between the NiRMSD and the Qscore

Range	Npair	Consistent	Inconsistent
0-20	7290	0.86*	0.14*
20-40	5130	0.90*	0.10*
40-100	5190	0.94*	0.06*
Total	17610	0.90*	0.10*

a) Core Regions: Range is the range of identity of the considered Prefab datasets, as measured on the reference alignments, Np is the number of pairs on which the comparison was carried out. Consistent is the fraction of pairs for which the Qscore and the NiRMSD score were consistent. For the purpose of this table, two pairs were considered consistent whenever their Qscore differed by less than 1 point percent and their NiRMSD by less than 0.05 Å. A binomial test was carried out on the results and entries marked with * indicate results whose p-value is lower than 0.000001.

Range	Npair	Consistent	Inconsistent
0-20	7290	0.79*	0.21*
20-40	5130	0.84*	0.16*
40-100	5190	0.84*	0.16*
Total	17610	0.82*	0.18*

b) Same as a) but with the NiRMSDs measured on the entire alignments.

equal to the Prefab reference (Best NiRMD fraction), as measured on the core regions.

The results (Table 1a,b and c) are unambiguous and clearly show a high correlation between the Qscore, the average NiRMSD and the Best NiRMSD fraction. As expected, the Prefab reference alignments outperform every other method (Table1b, Prefab), with a NiRMSD always lower than the rest, especially in the distant homologue category (Table 1b, Prefab, [0-20]). The rankings suggested by each score are in broad agreement when considering equivalent lines in each table. We looked at the statistical signifi-

cance of all these analyses. For doing so we considered every dataset individually and estimated the consistency between the Qscore and the NiRMSD measured on two alternative alignments. For instance, given a dataset and two alignments (aln1 and aln2) generated by two different methods, the Qscore and the NiRMSD are consistent if they indicate the same relationship between the two alignments (*e.g.* aln1 better than aln2 according to Qscore AND NiRMSD).

This measure was used to analyze every possible pair of methods (Table 2a,b). The results show that Qscore and NiRMSD are highly correlated with 90% consistency between the two measures on core regions and 82% when considering entire alignments. The correlation is not affected by the level of identity between the considered sequences. These figures were measured on more than 17000 pairs of alignments. We checked these results for statistical significance, using a binomial test and assuming an equal probability of 0.5 for consistency and inconsistency. The results are highly significant on each category, with P-Values systematically lower than 10^{-6}. These results confirm that the NiRMSD measure is at least as discriminative as Prefab.

CONCLUSION

We describe the iRMSD, a measure with all the advantages and properties of a standard RMSD without requiring any structure superposition. A simple normalization makes it possible to use the iRMSD for evaluating the accuracy of structure based sequence alignments. This measure, named NiRMSD, was applied on the alignments produced by 6 popular multiple sequence alignment packages. In 90 % of the cases the NiRMSD measure was in agreement with the Prefab ranking (Qscore). These findings, highly significant from a statistical point of view, suggest the suitability of this new measure for evaluating sequence alignments accuracy whenever structural information is available. We also expect that the method can easily be extended to sequences having a close homologue with a known structure.

Future developments will involve applying the iRMSD to Multiple Structure Alignment analysis. We are also planning to use the NiRMSD measure to compare structure alignment packages and check whether some methods clearly outperform the others or whether some structure alignment meta-method should be designed instead. Further refinement could also involve exploring the capacity of the iRMSD measure to automatically identify and exclude unalignable positions.

ACKNOWLEDGEMENTS

The development of this project was supported by CNRS (Centre National de la Recherche Scientifique), Sanofi-Aventis Pharma SA., Marseille-Nice Génopole and the French National Genomic Network (RNG). We thank Prof. Jean-Michel Claverie (head of IGS) for useful discussions and material support. We also thank Dr Phillip Bucher who provided many of the original ideas through useful discussions.

REFERENCES

Do,C.B., Mahabhashyam,M.S., Brudno,M. and Batzoglou,S. (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Res*, **15**, 330–340.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**, 1792–1797. Print 2004.

Goldsmith-Fischman,S. and Honig,B. (2003) Structural genomics: computational methods for structure analysis. *Protein Sci*, **12**, 1813–1821.

Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J Mol Biol*, **233**, 123–138.

Huang,Y.M. and Bystroff,C. (2006) Improved pairwise alignments of proteins in the Twilight Zone using local structure predictions. *Bioinformatics*, **22**, 413–422.

Katoh,K., Kuma,K., Toh,H. and Miyata,T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, **33**, 511–518.

Kleywegt,G.J. and Jones,T.A. (1994) Superposition. *CCP4/ESF-EACBM Newsletter Protein Crystallog*., **31**, 9–14.

Kleywegt,G.J. and Jones,T.A. (1999) Software for handling macromolecular envelopes. *Acta Crystallogr D Biol Crystallogr*, **55** (**Pt 4**), 941–944.

Kolodny,R., Koehl,P. and Levitt,M. (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J Mol Biol*, **346**, 1173–1188.

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.

Lesk,A.M. and Chothia,C. (1980) How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Biol*, **136**, 225–270.

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.

Morgenstern,B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment [In Process Citation]. *Bioinformatics*, **15**, 211–218.

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

O'Sullivan,O., Suhre,K., Abergel,C., Higgins,D.G. and Notredame,C. (2004) 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J Mol Biol*, **340**, 385–395.

O'Sullivan,O., Zehnder,M., Higgins,D., Bucher,P., Grosdidier,A. and Notredame,C. (2003) APDB: a novel measure for benchmarking sequence alignment methods without reference alignments. *Bioinformatics*, **19 Suppl 1**, i215–221.

Raghava,G.P., Searle,S.M., Audley,P.C., Barber,J.D. and Barton,G.J. (2003) OXBench: A benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, **4**, 47.

Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*, **11**, 739–747.

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.D., Koehl,P., Ripp,R. and Poch,O. (2005) BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins*, **61**, 127–136.

Van Walle,I., Lasters,I. and Wyns,L. (2005) SABmark--a benchmark for sequence alignment that covers the entire known fold space. *Bioinformatics*, **21**, 1267–1268.

Wallace,I.M., Blackshields,G. and Higgins,D.G. (2005) Multiple sequence alignments. *Curr Opin Struct Biol*, **15**, 261–266.