PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Overcoming sequence misalignments with weighted structural superposition

Nickolay A. Khazanov,[1] Kelly L. Damm-Ganamet,[2] Daniel X. Quang,[3] and Heather A. Carlson[1,2]*

[1] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109-2218

[2] Department of Medicinal Chemistry, University of Michigan, Ann Arbor, Michigan 48109-1065

[3] The Cooper Union, New York, New York 10003

ABSTRACT

An appropriate structural superposition identifies similarities and differences between homologous proteins that are not evident from sequence alignments alone. We have coupled our Gaussian-weighted RMSD (wRMSD) tool with a sequence aligner and seed extension (SE) algorithm to create a robust technique for overlaying structures and aligning sequences of homologous proteins (HwRMSD). HwRMSD overcomes errors in the initial sequence alignment that would normally propagate into a standard RMSD overlay. SE can generate a corrected sequence alignment from the improved structural superposition obtained by wRMSD. HwRMSD's robust performance and its superiority over standard RMSD are demonstrated over a range of homologous proteins. Its better overlay results in corrected sequence alignments with good agreement to HOMSTRAD. Finally, HwRMSD is compared to established structural alignment methods: FATCAT, secondary-structure matching, combinatorial extension, and Dalilite. Most methods are comparable at placing residue pairs within 2 Å, but HwRMSD places many more residue pairs within 1 Å, providing a clear advantage. Such high accuracy is essential in drug design, where small distances can have a large impact on computational predictions. This level of accuracy is also needed to correct sequence alignments in an automated fashion, especially for omics-scale analysis. HwRMSD can align homologs with low-sequence identity and large conformational differences, cases where both sequence-based and structural-based methods may fail. The HwRMSD pipeline overcomes the dependency of structural overlays on initial sequence pairing and removes the need to determine the best sequence-alignment method, substitution matrix, and gap parameters for each unique pair of homologs.

Proteins 2012; 80:2523–2535.

Key words: homolog; protein flexibility; sequence alignment; structure overlay; RMSD; structure alignment.

INTRODUCTION

Evolutionarily related proteins generally retain a tertiary fold that is more conserved than the amino acid sequence.[1,2] Structure is related to function; hence, proteins with similar structures may also share a common biological activity.[3] As a result, the identification of a homolog is a very useful means to infer the function and/or predict the structure of an uncharacterized protein. Many databases exist that classify proteins into families by their structures, including but not limited to SCOP,[4] CATH,[5] DaliDB,[6] PASS2,[7] MMDB,[8] ASTRAL,[9] HOMSTRAD,[10] and LPFC.[11] A review from Orengo and Thornton provides a very thorough discussion of protein evolution from a structural standpoint,[12] and another recent review stresses that the classification in an evolutionary context is still an open problem.[13]

An appropriate structural superposition provides a means to compare the similarity or dissimilarity between protein structures. However, to perform a structural comparison, the corresponding residues (atom pairs) between the proteins must be determined. This task can be accomplished (1) in a sequence-dependent manner

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: N.A.K. thanks the Bioinformatics Training; Grant number: GM070449; Grant sponsor: The National Institutes of Health; Grant number: GM65372; Grant sponsor: Beckman Young Investigator Award; Grant sponsor: NSF CAREER Award; Grant number: MCB 0546073; Grant sponsor: The Pharmacological Sciences Training Program; Grant number: GM007767.

*Correspondence to: Heather A. Carlson, Department of Medicinal Chemistry, 428 Church Street, University of Michigan, Ann Arbor, MI 48109-1065. E-mail: carlsonh@umich.edu.

Received 30 October 2011; Revised 5 June 2012; Accepted 10 June 2012
Published online 26 June 2012 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24134

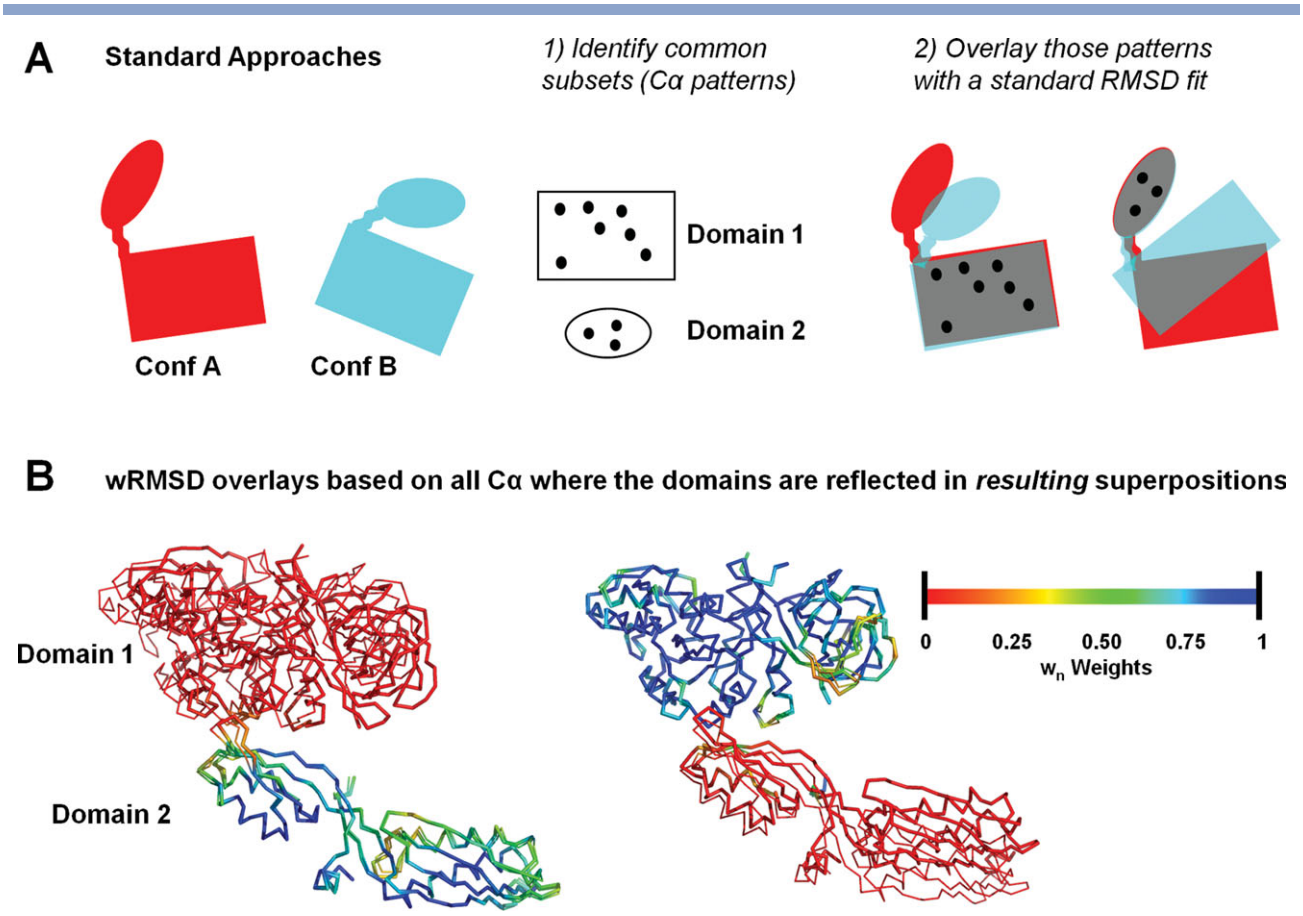

Figure 1

(**A**) Most methods for superimposing flexible proteins are based on two steps, which involve determining a subset of related atoms and overlaying the subset using a standard RMSD-fit procedure. Each technique differs in the way that it identifies the related subsets, but in the superposition step, all of the techniques designate each Cα as "in" or "out" of the calculated fit. (**B**) Our weighted superposition is based on all Cα. Multiple solutions can be found where the domains are reflected in the resulting weights and superpositions. Blue and green regions have high weights, align well, and define a domain. Red regions have low weights and poor agreement in the overlay.

using an initial sequence alignment or (2) solely through structural information in a sequence-independent manner. Sequence-based techniques can miss similarity between homologous proteins with intermediate to low-sequence identity (twilight zone). Fold-based methods can identify structural similarity, even between homologs with divergent sequences, but they can be misleading in the case of flexible proteins. A technique that combines the two approaches and overcomes limitations caused by the protein flexibility would be an ideal choice for superimposing homologs. In 2005, a thorough evaluation[14] of six structural comparison techniques—SSAP,[15] STRUCTAL,[16,17] DALI,[18] LSQMAN,[19] combinatorial extension (CE),[20] and secondary-structure matching (SSM)[21] demonstrated many strengths and limitations of current approaches and the caveats of metrics used for evaluation and comparison. A more recent review with similar scope indicated room for improvement of alignments, especially in proteins with extensive conformational variability or structural repetitions.[22] Additional reviews of the field call for combining techniques and using consensus across several methods to best define a structural comparison.[2,23,24] Here, we present a structural alignment method that accounts for protein flexibility and utilizes a superposition-driven approach to capture structural similarity in a more systematic and intuitive way.

In 2006, we introduced a superposition technique that overcomes the limitations of protein flexibility[25] by implementing a Gaussian-weighting term into the RMSD-fit algorithm determined by Kabsch.[26] The calculated weight is directly related to the distance between two atoms in space. Consequently, atom pairs in close proximity have a greater weighting than those further apart, biasing the superposition toward the regions that remain relatively rigid between conformations. Our method is the reverse of techniques used for the last 20 years, which perform two steps: (1) identify related subsets of Cα and (2) overlay those related subsets by a

standard RMSD fit [sRMSD; Fig. 1(A)]. Using our technique, the overlay defines the domains, rather than the domains defining the overlay. The resulting weights identify the domains. As shown in Figure 1(B), the backbones of two protein conformations are well superimposed in the blue, high-weight regions but can be seen separately in the red, low-weight regions. Each solution is based on a unique domain of the protein and each is an equally valid overlay. Complete mathematical details of the weighted RMSD (wRMSD) procedure can be found in our original study,[25] and an abbreviated presentation is provided in the Supporting Information Material.

In this study, we have coupled our wRMSD technique with basic sequence alignment algorithms from the EMBOSS package[27] to provide initial alignment of homologous sequences. Previously, we showed that wRMSD is more appropriate than sRMSD for superimposing two conformations of a flexible protein. Here, we show that this approach is also superior for superimposing flexible homologs. In cases where the structures of distant homologs are available in only alternate conformations, fold-based techniques have difficulty. By only weighting regions of the protein in good structural agreement, our method is able to overcome errors from (1) the initial atom pairing resulting from low-sequence identity and (2) large conformational differences.

Once a sufficient overlay is established, the seed extension (SE) structural alignment algorithm[28] lets the wRMSD superposition dictate a new and improved sequence alignment. The SE method extends the alignment of residue pairs in very close proximity (seeds) along the protein chain. Many modern structural alignment methods combine sequence and structure data in their alignment procedure. Our method is modular, and allows the structural information to dominate the superposition, producing a consistent structural alignment solution. The SE algorithm allows us to then convert the information from the structural alignment into a corrected sequence alignment. Below, we demonstrate the robustness of the procedure with respect to initial sequence alignment, and the ability of the method to consistently outperform standard superposition. We show that the final, corrected sequence alignments are in good agreement with HOMSTRAD.[10]

We then compare HwRMSD's performance to that of several popular structure alignment programs. Based on the number of residue pairs within 2 Å, most methods produced similar results. However, HwRMSD superpositions have many more residue pairs within 1 Å. This level of accuracy is absolutely essential. This level of accuracy is needed when correcting sequence alignments in an automated fashion. A 2-Å offset in aligned α-helices can result in mis-paired residues, which misleads scientists in analyzing phylogenetic relationships and planning mutagenesis studies. Finally, drug design requires accuracy of 1 Å or better. Any skew in the alignment of the binding sites can affect the results dramatically, particularly when trying to design ligands specific to one homolog over another (i.e., specificity across human kinases).

METHODS

Homologous protein pairs were obtained from the HOMSTRAD database.[10] The protein co-ordinates were downloaded from the Protein Data Bank (PDB),[29] and the specific protein chains used were dictated by the pairings listed on the HOMSTRAD website. For this study, we chose to focus on the more difficult cases of homologous proteins with lower sequence identities (ID) (39–16%). We examined homolog pairs with <16% ID, but the sequence alignment algorithms used to obtain an initial alignment failed in many instances, giving nonsensical alignments. The alternate structural alignment programs used in this study also failed with these cases, making <16% ID a relatively universal cutoff for current methods.

Our technique is performed using Cα co-ordinates, but it is easily extended to any atom subset. The HwRMSD procedure consists of three sequential steps:

1. Use a simple sequence alignment to determine an initial list of paired residues. We call this the *initial* sequence alignment.
2. Calculate a wRMSD overlay of the two structures based on the initial sequence pairing and a wRMSD-fitting parameter of $c = 5$ Å^2.
3. Obtain a corrected sequence alignment from the structural superposition using SE. We call this the final or *corrected* sequence alignment.

The program *needle*, an implementation of the Needleman–Wunsch (NW) global alignment from the EMBOSS[27] package, was used to generate the pair-wise sequence alignment. This alignment determines the residue correspondence between the two proteins, which is then used to guide the initial structural superposition.

To obtain a sequence alignment from the structural superposition, SE is used with default parameters.[28] Briefly, SE finds "seed" pairs of structurally equivalent residues from overlaid structures based on their physical proximity and chemical similarity. Consecutive triplets of seeds are then extended along the alignment matrix in both directions using distance and amino acid similarity to resolve conflicts that arise during the extension of more than one diagonal.

Robustness

To test the robustness of the method with respect to initial sequence alignments, *water*, an implementation of the local Smith–Waterman (SW) sequence alignment from EMBOSS,[28] was also used. The four different scoring matrices used were BLOSUM50, BLOSUM62,

PAM120, and PAM250; each employed its optimal gap-open and gap-extension penalty parameters: (−10,−2), (−7,−1), (−16,−4), and (−10,−2), respectively. The optimal parameters for each scoring matrix were recommended by the European Bioinformatics Institute.[30]

For each protein pair, the pair-wise distances between the superpositions obtained with the same sequence alignment algorithm and the different parameters were calculated, and the median of these values was chosen to represent the similarity of the solutions. The distances between the superpositions were calculated using a simple all-atom RMSD measure (not superposition in this case). In cases such as the PHBD-like proteins (1FOH[31] and 1PBE[32]) and several others, some initial sequence alignments were too small to be considered reasonable solution. To avoid such invalid outliers, any solution with an alignment length of <10 residue pairs was discarded, and the median distance was calculated between the two or three remaining solutions. Such cases were mostly in the range of low-sequence identity (<20% ID), and they illustrate the practical limits of current approaches.

Comparison

To compare the results of HwRMSD to other tools, we used the EMBOSS-wrapped implementations of CE,[20] FATCAT (flex),[33] and the native SSM[21] and DaliLite[6] servers to perform structural alignments on each structure pair. Default parameters were used for all structural alignment methods. For SSM and DaliLite, only the best alignment solution was considered, based on highest *Q*-Score or *Z*-Score, respectively.

RMSD is a standard way to compare the "goodness" of a structural superposition, but it can be misleading. In cases where the aligned proteins have large structural variation among multiple domains, the solution with the lowest RMSD does not necessarily provide the best overlay between the most structurally similar domains. The best superposition of homologous structures should have the largest number of sequence-aligned residues in close proximity, even if it is at expense of a large distance among a few pairs of residues in a flexible domain. Therefore, we calculated the number of sequence-aligned residues with Cα within 1 Å to measure the quality of a structural superposition. Rather than expressing these as a raw number of residue pairs, we used the percentage of residue pairs (%Cα $\leq$ 1 Å) for an even comparison across proteins of various lengths. The percentage of pairs within 2 Å (%Cα $\leq$ 2 Å) was also used as a secondary, looser criterion.

The accuracy of the HwRMSD-corrected sequence alignments was evaluated through comparison to the alignments provided by HOMSTRAD,[10] a "gold standard" for the field. As a first metric, we simply counted the number of unique residue pairings in common between the two alignments. This overlap quantified the extent to which the HwRMSD-corrected sequence alignment recapitulates the alignment considered optimal by HOMSTRAD. However, some disagreement between the two alignments may come from small errors in HOMSTRAD. To evaluate that possibility, we turned to our definition of a good structural overlay as a second metric. If the sequence alignment from HOMSTRAD was the best possible pairing of the structurally similar residues, it would produce more sequence pairs within 1 or 2 Å in a superposition generated by wRMSD (i.e., higher %Cα $\leq$ 1 Å and/or %Cα $\leq$ 2 Å values). Finally, sequence alignments cannot be strictly defined, and more than one solution may be equally valid. Equivalent solutions might differ slightly in their residue pairings, but result in the same %Cα $\leq$ 1 Å or %Cα $\leq$ 2 Å.

When comparing the HwRMSD method to the THESEUS structural superposition algorithm,[34] we ran the THESEUS program with the standard sequence alignment used by HwRMSD (needle with BLOSUM50 parameter set). As THESEUS, like the stand-alone wRMSD method, does not return its own structure-based sequence alignment, we ran SE on the THESEUS superposition to generate a THESEUS-corrected sequence alignment. By "plugging in" THESEUS in this way into our pipeline, we can directly compare the effect of the wRMSD structural superposition with that of an alternate superposition from THESEUS.

PyMOL[35] was used for visualization and the creation of figures for this manuscript.

RESULTS AND DISCUSSION

HwRMSD for superimposing homologous proteins and correcting their sequence alignments

Aligning proteins with high-sequence identity is straightforward, so for this study, we chose to focus on the more difficult cases of homologous proteins in the low to intermediate range (39–16% ID), as listed in Table I (for more details, see Supporting Information Table 1). In our previous study,[25] we were able to show an improved superposition of two conformations of the chaperonin protein GroEL, which undergoes a large conformational change between the bound and the apo forms (PDB[29] codes 1AON[36] and 1OEL,[37] respectively). In Figure 2, we use this system again to demonstrate the potential difficulties of fitting homologous, flexible proteins. With our technique, either conformation of GroEL can be appropriately superimposed to the bound form of its archaeal homolog, the thermosome (1A6E[38]). The easier case of fitting the two bound conformations is shown in Figure 2(A,B) which indicates the more difficult comparison of the bound form of the thermosome

Table I
Homologous Protein Pairs Obtained from the HOMSTRAD Database and the Performance of sRMSD and wRMSD Superposition to Bring Paired Residues into Close Proximity[a]

		%Cα ≤ 2 Å		%Cα ≤ 1 Å	
Homolog proteins and PDB IDs	%ID	sRMSD (%)	wRMSD (%)	sRMSD (%)	wRMSD (%)
Serine/threonine phosphatase FJM and 1TCO	39	80.00	84.4	40.34	66.1
Glutathione synthase 1M0W and 2HGS	37	64.48	65.8	39.11	43.1
Interferon 1AU1 and 1ITF	35	35.93	60.5	6.59	25.7
Adenosylmethionine decarboxylase 1I7B and 1MHM	33	74.03	76.0	40.91	47.1
Clostridial neurotoxin zinc protease 1EPW and 3BTA	31	52.74	60.7	12.48	34.6
Sulfatase 1AUK and 1FSU	29	60.8	60.8	35.6	40.9
Translation initiation factor 1AP8 and 1EJH	29	23.76	35.4	3.31	13.3
Protocatechuate-3,4-dioxygenase 3PCG (chain A) and 3PCG (chain M)	28	24.88	62.2	5.97	46.3
Aminotransferase 1A3G and 5DAA	27	82.01	85.6	39.93	51.8
SpoU rRNA methylase 1IPA and 1GZ0	26	15.23	48.6	2.47	27.6
FMN oxidoreductase 1OYC and 2TMD	25	35.78	59.5	9.97	29.0
Queuine tRNA-ribosyltransferase 1IQ8 and 1K4G	25	63.99	69.3	16.96	25.6
tRNA synthestase 1GLN and 1QTQ	24	12.38	51.7	1.86	23.2
DNA methylase 1BOO and 2ENT	23	3.32	63.5	0.00	41.0
DNA topoisomerase 1AB4 and 1BJT	22	29.72	40.6	4.88	14.0
Pyridoxal-phosphate enzymes 1TDJ and 2TYS	21	33.63	49.5	4.20	14.7
Iron/ascorbate oxidoreductase 1BK0 and 1DCS	20	34.29	47.1	8.93	15.7
Molybdopterin dehydrogenase 1FFV and 1FO4	19	51.39	61.5	13.19	28.8
Spliceosomal Protein, Internalin B 1A9N and 1D0B	19	24.54	40.5	0.61	23.3
Asp/Glu/Hydantoin racemase 1B74 and 1JFL	18	17.47	41.0	1.75	15.7
Polysaccharide lyase 1CB8 and 1EGU	18	36.01	50.5	7.40	18.6
PHBH-like proteins 1FOH and 1PBE	17	3.32	43.6	0.26	12.8
Adaptin, clathrin appendage domain 1E42 and 1QTS	16	6.41	29.1	0.43	9.8

[a]Sequence pairings for the %Cα calculations were obtained from SE after the superposition step.

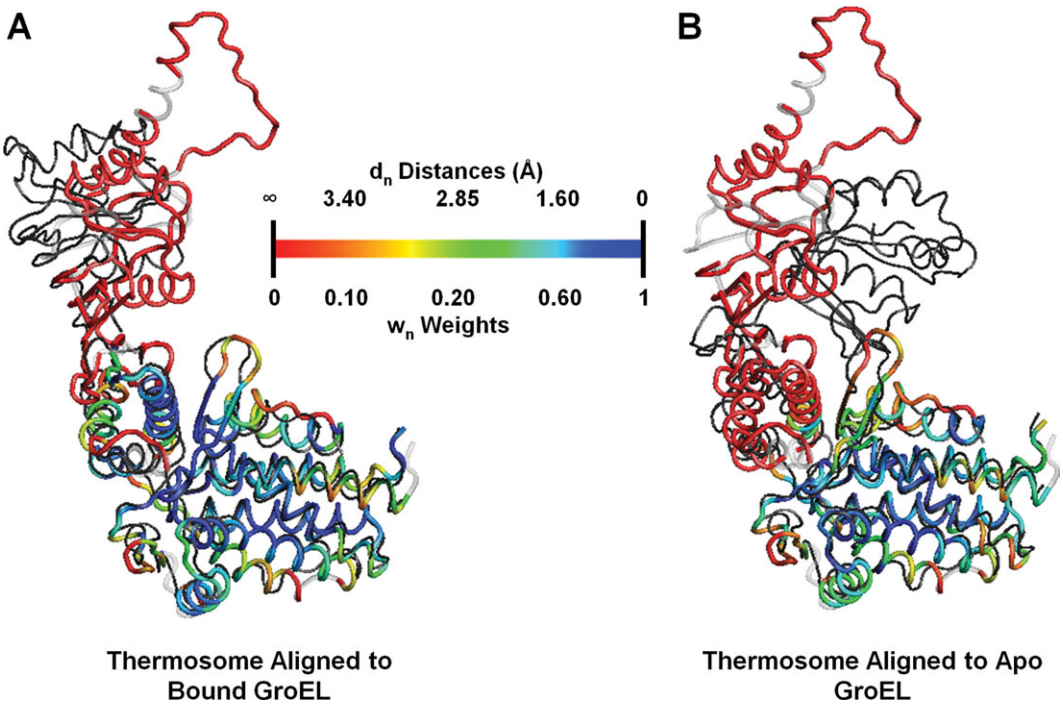

Figure 2
Chaperonin family (20.8% ID). Most techniques would readily identify the similarity between the thermosome and the GroEL in the similar bound conformation, but they may not identify its similarity with the apo conformation of GroEL. (**A**) wRMSD superposition of the bound conformation of GroEL[36] (thick, colored lines) onto the homologous thermosome[38] (thin, black lines). Light gray regions of GroEL indicate residues within gaps in the alignment. (**B**) wRMSD fit of the apo conformation of GroEL[37] (thick, colored lines) onto its homolog thermosome (thin, gray lines).

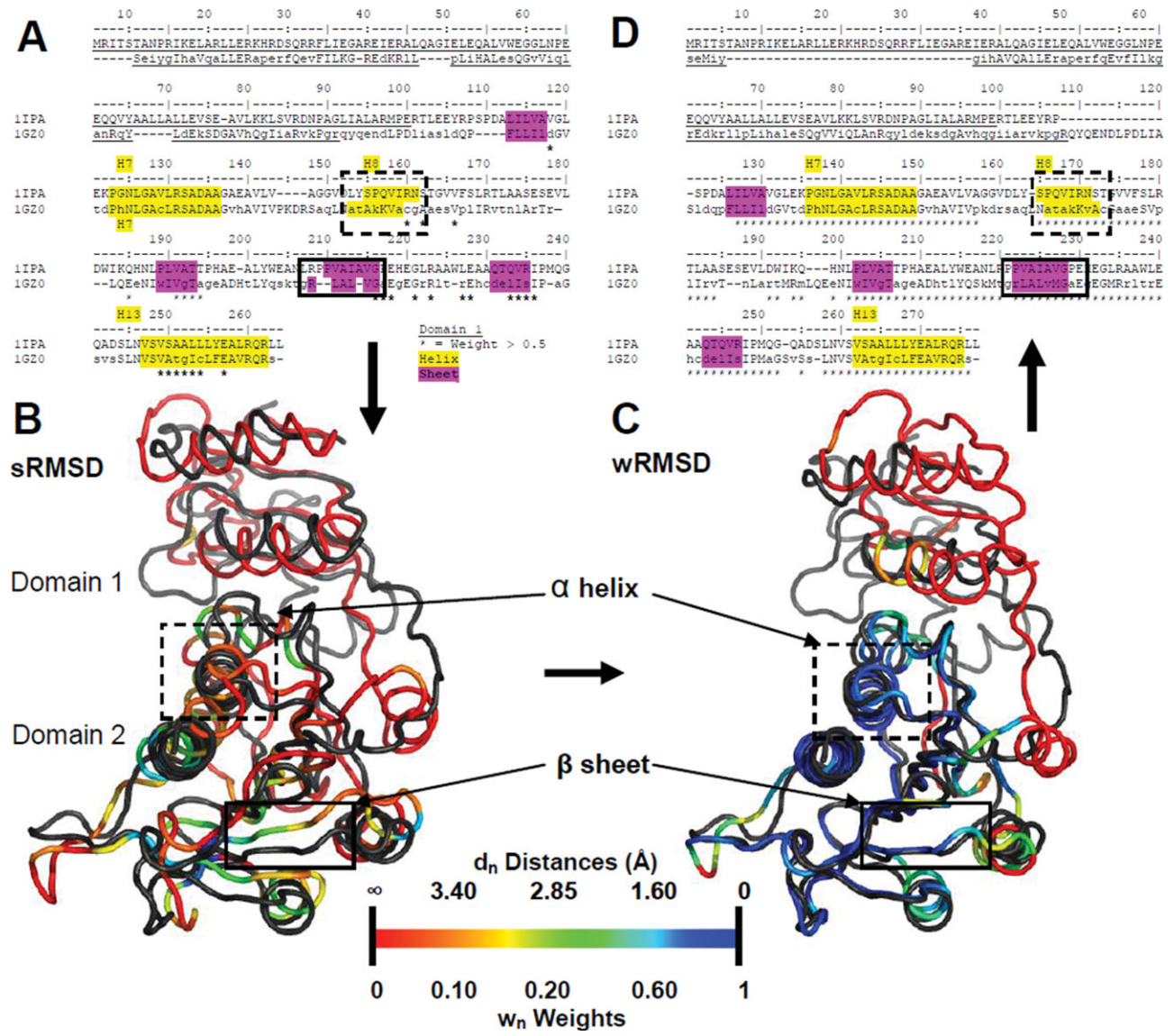

Figure 3

SpoU rRNA methylase family (26% ID). (**A**) NW sequence alignment of 1IPA[40] and 1GZ0[39] using default parameters. Lower case represents sequence dis-similarity, and gaps are shown with dashes. The underlined region notes domain 1, yellow represents α-helices, purple represents β-sheets, and boxes represent misaligned residues corresponding to the labeled α-helix and β-sheet in (**B**) and (**C**). Atom pairs with a weighting of 50% or greater in the wRMSD calculation are noted with asterisks. (B) Standard superposition superpositions of 1GZ0 (colored ribbon) onto 1IPA (black ribbon) obtained using the initial sequence alignment (from A), colored by weight. (C) Weighted superposition obtained from the same initial sequence alignment. (**D**) SE sequence alignment based on the wRMSD superposition, which now corrects the alignment of the secondary structure elements based on their spatial proximity in (C).

to the apo form of GroEL. The superpositions are colored by weight of the aligned pairs of Cα atoms, with higher weights indicating closer proximity and stronger structural similarity. It is obvious that a sRMSD overlay would be badly skewed in the case of Figure 2(B).

Figure 3 provides a thorough illustration of how specific errors in an initial sequence alignment are overcome by weighted superposition and corrected by HwRMSD owing to structural similarity. This example is based on homologs from the SpoU rRNA methylase family with 26% ID (1GZ0[39] and 1IPA[40]). Figure 3(A) shows the initial global NW sequence alignment using default parameters (BLOSUM50), and the resulting standard and weighted superpositions are shown in Figure 3(B,C). The final, corrected sequence alignment generated by SE is shown in Figure 3(D). Any residue pairs that received a weight of 0.5 or greater from the wRMSD calculation are noted with an asterisk. Residue pairings in regions of

good structural agreement will be heavily weighted in the wRMSD calculation and drive the superposition.

Protein regions that have been brought into close spatial proximity, despite low weights, indicate potentially incorrect pairings in the sequence. The underlined region of the sequence alignment in Figure 3(A,D) corresponds to a flexible domain between the proteins; as would be expected, none of these residues was significantly weighted to contribute to the superposition. The black boxes in Figure 3(A) indicate two regions of incorrect atom pairing. The first is owing to a gap placement (in 1IPA) and corresponds to the mis-aligned residues of the denoted H8 α-helix in Figure 3(B,C). However, after the weighted superposition, they are brought into close spatial proximity, and the final sequence alignment obtained by SE eliminates the gap to produce a correct pairing as evidenced by the high weights [Fig. 3(C)]. The β-sheet, shown in Figure 3(C,D), is also a misalignment that is overcome by the wRMSD superposition.

HwRMSD was developed to overcome errors in sequence alignment, but it can also overcome unexpected programming bugs. Some of the errors in Figure 3 are caused by the default behavior of the Biopython parser[41] used to pull sequence information from the coordinates in the PDB files, which omits modified methionine residues. Some parsers ignore nonstandard amino acids (listed as HETATOMs), and in the 1GZ0 structure, seven methionines have been replaced with selenomethionine to aid in solving the structure. The wRMSD superposition overrides the ambiguity, and the final alignment correctly pairs the β-sheet residues (with selenomethionine present in the sequence this time, thanks in part to SE parsing the structure correctly). Although this is easily rectified programmatically, we have used the omission to serve as an example of additional and unexpected sources of error.

Correction of an alignment is made possible by the powerful combination of wRMSD-generated superposition and the "SE" algorithm used by SE to obtain a sequence alignment from a pair of protein structures. The SE method makes no inference about secondary structure elements of the aligned structures and considers residue similarity only for tie-breaking. Additionally, the algorithm extends from a number of small "seed" pairings, and hence there are no gap penalties and no global cost optimization—two factors that are present in many structural alignment algorithms. The absence of these heuristics makes the SE algorithm a true What-You-See-Is-What-You-Get method for translating a structural superposition into a sequence alignment, and thus, it is a perfect fit for the HwRMSD protocol.

HwRMSD overlays are more robust than standard superpositions

Standard superposition is sensitive to the initial alignment, and incorrectly paired residues skew the result even in cases where large structure similarity in a protein domain is visually obvious. The weighted superposition converges to a consistent result even when a wide variety of initial alignments are used, allowing regions of the structure in closest proximity—such as large similar domains—to drive the superposition. To demonstrate the limitation of sRMSD superposition, we examined the dependence of the resulting overlay on the initial sequence alignment by varying the initial alignments used to generate the superposition. We chose the canonical implementation of global NW and local SW algorithms to perform the initial sequence alignments, and varied their parameters by using four different scoring matrices and gap penalties. For each test case, each of the sequence alignments was used to generate a standard and weighted superposition.

Figure 4 uses DNA methylase homologs[38,39] (23% ID) to visually show how the sRMSD superpositions are noticeably different when varying the parameters for sequence alignment [Fig. 4(A)]. As immediately visible, a standard overlay is extremely sensitive to the residue correspondence established by the initial sequence alignment (average RMSD distance among solutions ~3 Å). Conversely, the multiple wRMSD-superposition solutions are indistinguishable by eye [Fig. 4(B)] irrespective of the initial sequence alignment parameters (average RMSD distances among solutions ~0.25 Å). Although variation in the sRMSD solutions would result in different structure-based sequence alignments from SE, all the wRMSD solutions generate the same corrected SE sequence alignment.

Overall, we found that wRMSD is relatively insensitive to small variations in the initial sequence alignment. Across our entire test set, the use of weighted superposition overcomes the variation in atom-pairing arising from the different sequence-alignment parameters to give a more consistent structural comparison for each test case except one (Supporting Information Table 2). This demonstrates the robustness of the wRMSD superposition algorithm as a tool able to generate consistent structural superpositions as part of the HwRMSD protocol, regardless of the chosen proteins.

The NW alignment with a BLOSUM50 matrix produced the most appropriate alignments over all the test cases (data not shown). Although similar alignments were also obtained from BLOSUM62 and PAM250 with both NW and SW, the cases where the alignments failed or were too short for an initial alignment were more frequent. *Given its reliable performance, the NW BLOSUM50 sequence alignment has been defined as the default initial alignment for HwRMSD.*

The greatest difficulties were found when using PAM120 in both SW and NW alignments, which was expected because this matrix is the least appropriate choice for distant homologs. Furthermore, the PAM120 matrix is coupled with the most severe gap penalty

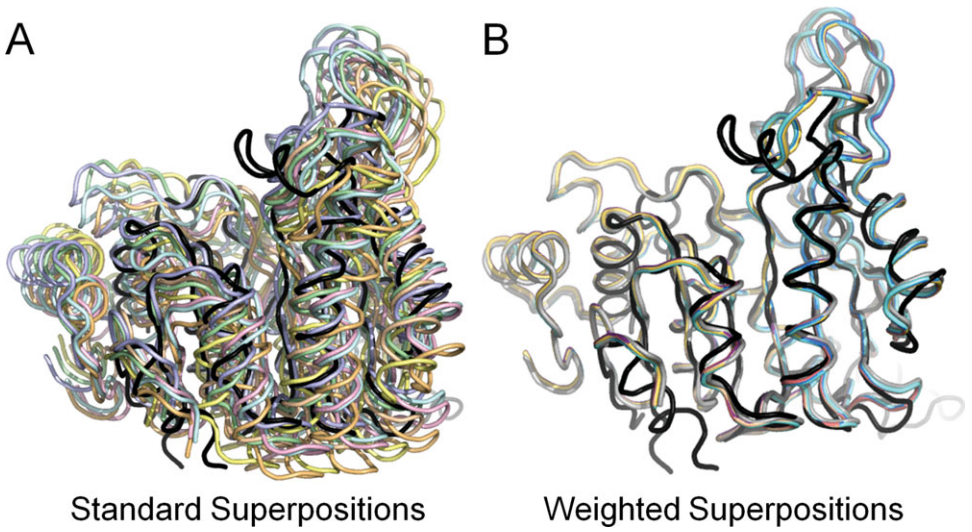

Figure 4

DNA methylase family (23% ID). Weighted structural superpositions are nearly independent of the sequence alignment method, but standard superpositions are greatly affected. (**A**) Overlays of 2ENT (black ribbon) to 1BOO[49] (colored ribbons) from standard superpositions based on seven different sequence alignments. (**B**) The seven weighted superpositions of 2ENT to 1BOO, based on the same sequence alignments converge on the same solution.

(−16) of the four parameter sets tested; it produced alignments with very few gaps when used with the global NW method and extremely short alignments with the local SW method. There were a few cases where the wRMSD superposition was not able to overcome the sequence alignment errors and the superposition was not an improvement over a sRMSD overlay. For example, in the case of spliceosomal protein[42] and internalin B[43] (1A9N and 1D0B) the NW PAM120 alignment produced a significantly different superposition. Similarly, in the case of the adaptin[44] and clathrin[45] appendage domains (1E42 and 1QTS), the SW PAM120 alignment produced an outlier superposition (a distance of ∼7 Å, RMSD, from other solutions). These poorly aligned protein pairs are in the <20% sequence identity range.

HwRMSD overlays provide better structural comparisons than standard superpositions

A consistent superposition is useful only if it is also an improved superposition. In Figure 4, the fraction of aligned Cα pairs closer than 2 Å is 3–4% for the sRMSD superpositions, and no pairs are within 1 Å. Conversely, 64% of the pairs from the wRMSD superposition are within 2 Å, and 41% are within 1 Å. This improvement is particularly prominent in the core region, which is structurally conserved between the homologous proteins.

To compare the quality of sRMSD versus wRMSD superpositions directly, the numbers of aligned residues within 2 and 1 Å were tallied for all protein sets based on the best sequence alignment (NW with the BLOSUM50 parameter set). In all cases, the percentage of sequence-aligned Cα pairs in close proximity is greater for the wRMSD superpositions (Table I). Like Figure 3, the results in Table I suggest that the wRMSD-based solutions achieve better alignments of structurally similar protein regions. The solutions have also been examined visually to confirm the improvements.

To test whether the use of wRMSD within our three-step pipeline is comparable to other weighted superposition methods, we compared our results to those generated using THESEUS[34] for the superposition. The THESEUS results were generated by plugging in THESEUS instead of wRMSD into the HwRMSD protocol. This focused the comparison on the performance of the superposition step, all other steps being equal. THESEUS does not include a sequence alignment step, nor is it specifically for homologs, and hence its incorporation into our pipeline is appropriate and consistent with its development as a superposition tool.[34] The results were almost identical with perhaps a slight advantage for wRMSD (first versus last columns in Supporting Information Table 3). The wRMSD-based solutions resulted in higher %Cα<2Å in 21 out of the 23 cases, but the results of the two approaches never differed by more than 4%. For the %Cα<1Å metric, the differences between the two methods became even more slight, with roughly a half-and-half split between cases being fit slightly better by wRMSD or THESEUS. The one exception was the test pair of 1FOH and 1PB3, for which the THESEUS algorithm failed to execute. The similarity of results suggests that the superpositions resulting from wRMSD compare

Table II
Correspondence Between HOMSTRAD and HwRMSD Sequence Alignments with the Percentage Calculated with Respect to HOMSTRAD Alignment

Homolog proteins and PDB IDs	%ID	# Pairs ∩ HOMSTRAD	% Pairs ∩ HOMSTRAD
1FJM and 1TCO	39	271	95.09
1M0W and 2HGS	37	422	92.14
1AU1 and 1ITF	35	144	89.44
1I7B and 1MHM	33	274	94.48
1EPW and 3BTA	31	461	88.82
1AUK and 1FSU	29	378	84.19
1AP8 and 1EJH	29	129	73.71
3PCG (chain A) and 3PCG (chain M)	28	143	78.14
1A3G and 5DAA	27	268	100.00
1IPA and 1GZ0	26	NA[a]	NA[a]
1OYC and 2TMD	25	292	87.95
1IQ8 and 1K4G	25	312	93.98
1GLN and 1QTQ	24	251	92.96
1BOO and 2ENT	23	203	84.23
1AB4 and 1BJT	22	355	82.18
1TDJ and 2TYS	21	265	82.04
1BK0 and 1DCS	20	224	86.49
1FFV and 1FO4	19	274	96.14
1A9N and 1D0B	19	0[b]	0.00[b]
1B74 and 1JFL	18	144	68.57
1CB8 and 1EGU	18	256	84.49
1FOH and 1PBE	17	297	80.05
1E42 and 1QTS	16	91	41.74

[a]The HOMSTRAD alignment for this protein pair was missing the multiple MSE residues present in the 1GZ0 structure, and hence we could not automatically and directly compare the two alignments without introducing modifications into the structure or alignment (see discussion of Fig. 3 in the text).
[b]In the case of Spliceosomal protein and internalin B (ID 19%), a repeating sheet-helix domain in both proteins allowed for alternate superpositions, where the two proteins were shifted relative to one another by one repeating segment of the sheet-helix coil. Both superpositions maintained a high structural correspondence although their alignments were distinct.

well to those of THESEUS, and that in general, any good weighted-superposition method can lead to robust structural superposition solutions and corrected sequence alignments in our pipeline.

Validation of HwRMSD-corrected sequence alignments

HwRMSD provided superior structural superpositions, and it was important to demonstrate that it resulted in improved sequence alignments from SE. We compared the final, corrected HwRMSD sequence alignments to those provided by HOMSTRAD, the database from which our homolog pairs were sourced. Table II lists the number of residue pairs in common between the HwRMSD (NW with the BLOSUM50 parameter set) and the HOMSTRAD sequence alignments. Only four cases resulted in the fraction of aligned residue pairs being <80%. In the case of Spliceosomal Protein and Internalin B (ID 19%), there were no residue pairs in common in the alignment, yet both alignments demonstrated structural correspondence, indicating a case where multiple yet distinct solutions are possible. The repeating sheet-helix domains of these proteins demonstrate a case where an initial alignment off by a unit of the repeating region may generate a distinct solution with a reasonable structural correspondence and RMSD, a case where the single "correct" alignment may not exist.

Considering that the HOMSTRAD structural alignments were extensively curated by hand, HwRMSD alignments do very well to recapitulate the structural correspondence in a completely automated manner without parameter tweaking. Furthermore, Table III summarizes that the HOMSTRAD and HwRMSD-corrected sequence alignments result in the same structural superpositions with wRMSD. In many cases, the HwRMSD-corrected sequences result in very minor improvements, typically $\leq$5 additional residue pairs within 2 Å. In the case of 1M0W and 2HGS (36% ID), HOMSTRAD has an advantage of nine more residue pairs within 2 Å, and in the case of 1E42 and 1QTS (16% ID), HwRMSD has an advantage of 13 more residue pairs. As discussed in the **METHODS** section, sequence alignments cannot be strictly defined, and more than one solution may be equally valid. We show that the two alignments are equivalent solutions with nearly the same %C$\alpha \leq 1$ Å and %C$\alpha \leq 2$ Å. Of course, this focuses on the regions where the greatest structural similarity is seen, and HOMSTRAD may have an advantage in other regions (**Local HwRMSD alignments** section).

Of course, there may be situations where it is difficult to obtain an appropriate superposition with the weighted fitting, for example, when a protein is large and has multiple domains. If two different initial sequence alignments obtain residue pairings each focused on a different domain, rather than spanning entire protein structure, then the weighted superpositions may not converge to the same solution (**Local HwRMSD alignments** section). Another such case is when there is too little sequence or structural similarity, but this is when most comparison methods breakdown. For the test cases employed in this study, the sequence alignment tools broke down at ~16% ID, returning sporadic aligned segments that were too short and too infrequent. Homologs with so little sequence similarity are notoriously difficult to align,[46] but it may be possible in some cases to compare them using methods based on structural information such as geometric comparisons of folds.[47] However, these techniques would be successful only when there is little structural variation or flexibility. Techniques such as wRMSD are absolutely required for large structural variation.

Comparison of HwRMSD to other structural alignment solutions

Finally, we compared HwRMSD to other popular structural alignment methods. The structural alignments from CE, FATCAT, SSM, and Dalilite were used for comparison, all generated using the same set of test pairs and

Table III
Comparison of HOMSTRAD Sequence Alignment and HwRMSD Correct Sequence Alignment to Bring Paired Residues into Close Proximity with wRMSD Superpositions

Homolog proteins and PDB IDs	%ID	HOMSTRAD %Cα ≤ 2 Å (%)	HwRMSD %Cα ≤ 2 Å (%)	HOMSTRAD %Cα ≤ 1 Å (%)	HwRMSD %Cα ≤ 1 Å (%)
1FJM & 1TCO	39	84.1	84.4	66.1	66.1
1M0W & 2HGS	37	67.7	65.8	44.8	43.1
1AU1 & 1ITF	35	59.9	60.5	25.7	25.7
1I7B & 1MHM	33	75.3	76.0	47.1	47.1
1EPW & 3BTA	31	59.9	60.7	34.2	34.6
1AUK & 1FSU	29	60.4	60.8	40.7	40.9
1AP8 & 1EJH	29	35.4	35.4	13.3	13.3
3PCG (chain A) & 3PCG (chain M)	28	60.2	62.2	45.77	46.3
1A3G & 5DAA	27	85.6	85.6	51.8	51.8
1IPA & 1GZ0	26	NA[a]	48.6	NA[a]	27.6
1OYC & 2TMD	25	59.5	59.5	29.0	29.0
1IQ8 & 1K4G	25	69.3	69.3	25.6	25.6
1GLN & 1QTQ	24	51.7	51.7	23.2	23.2
1BOO & 2ENT	23	62.0	63.5	40.6	41.0
1AB4 & 1BJT	22	39.7	40.6	14.0	14.0
1TDJ & 2TYS	21	49.5	49.5	14.7	14.7
1BK0 & 1DCS	20	45.4	47.1	15.7	15.7
1FFV & 1FO4	19	61.8	61.5	28.8	28.8
1A9N & 1D0B	19	0.0[b]	40.5	0.0[b]	23.3
1B74 & 1JFL	18	40.2	41.0	15.3	15.7
1CB8 & 1EGU	18	50.2	50.5	18.6	18.6
1FOH & 1PBE	17	43.6	43.6	12.8	12.8
1E42 & 1QTS	16	23.5	29.1	8.5	9.8

[a]The HOMSTRAD alignment for this protein pair was missing the multiple MSE residues present in the 1GZ0 structure, and so we could not automatically and directly compare the two alignments without introducing modifications into the structure or alignment (see discussion of Figure 3 in the text).
[b]In the case of Splicesosomal Protein and Internalin B (ID 19%), a repeating sheet-helix domain in both proteins allowed for alternate superpositions, where the two proteins were shifted relative to one another by one repeating segment of the sheet-helix coil. The HOMSTRAD alignment was based on one superposition, while the wRMSD superposition was a distinctly different solution, shifted by one repeating sheet-helix segment. The number of atom pairs in close structural proximity was thus very limited (most pairs > 4 Å apart).

default parameters. Some methods are based on minimizing an RMSD measure; ours is made to maximize a weighting metric. The number of atoms in close proximity is the most intuitive metric for comparing structural alignments, resulting from a diverse set of algorithms. Figure 5 shows the performance of HwRMSD against the four methods, as evaluated by %Cα within 1 and 2 Å (more details are found in Supporting Information Table 3). Note that the percentages were calculated based on the shortest protein because that is the maximum number of actual matched pairs possible for each homologous pair.

In a prevailing trend, the performance of all algorithms decreases with diminishing sequence identity, indicating a general reduction in structural similarity [Fig. 5(A,B)]. Several pairs of homologs, such as the phosphatases 1FJM and 1TCO (39% ID) and the two chains from the aminotransferases structure 1A3G and 5DAA (27% ID) are aligned well by all methods, with ~86 and ~89% of the aligned residues within 2 Å, respectively [Fig. 5(A)]. Conversely, some protein pairs have few residue pairs within 2 Å (<50%), regardless of the method. Note that the percentage of closely aligned pairs is measured with respect to the number of residues in the shorter protein, but the alignment length in each system is shown in Figure 5(C) so that the reader can compare details in the performance of the various methods with the various systems.

HwRMSD showed great versatility with a solid performance across all of the different test cases [Fig. 5(A)]. The %Cα < 2Å for HwRMSD was always close to the mean of the other solutions, and it has the second lowest standard deviation across the test cases (after FATCAT), signifying a lack of particularly poor solutions. In fact, HwRMSD performed best or second best in three of the more challenging cases in this test set—1AUK and 1FSU (29% ID), 1GZ0 and 1IPA (26% ID) and 1E42 and 1QTS (16%ID). On average, FATCAT performed best in terms of the %Cα < 2Å metric, with an average of 59.76% pairs in proximity among the 23 test cases. HwRMSD, CE, and SSM having comparably high averages of 58.87, 57.74, and 58.6%, respectively [Fig. 5(A)]. Dalilite tended to produce the solutions with the lowest %Cα < 2Å, 50.85% on average. Several test cases were difficult for certain methods, with the SSM algorithm performing poorly on the 1AU1 and 1ITF interferon case, and Dalilite on the dioxygenase 3CPG test case. Below 20% sequence identity, the results of the algorithms varied more widely, indicating that perhaps no one solution was an obvious "right" answer for the structural alignment.

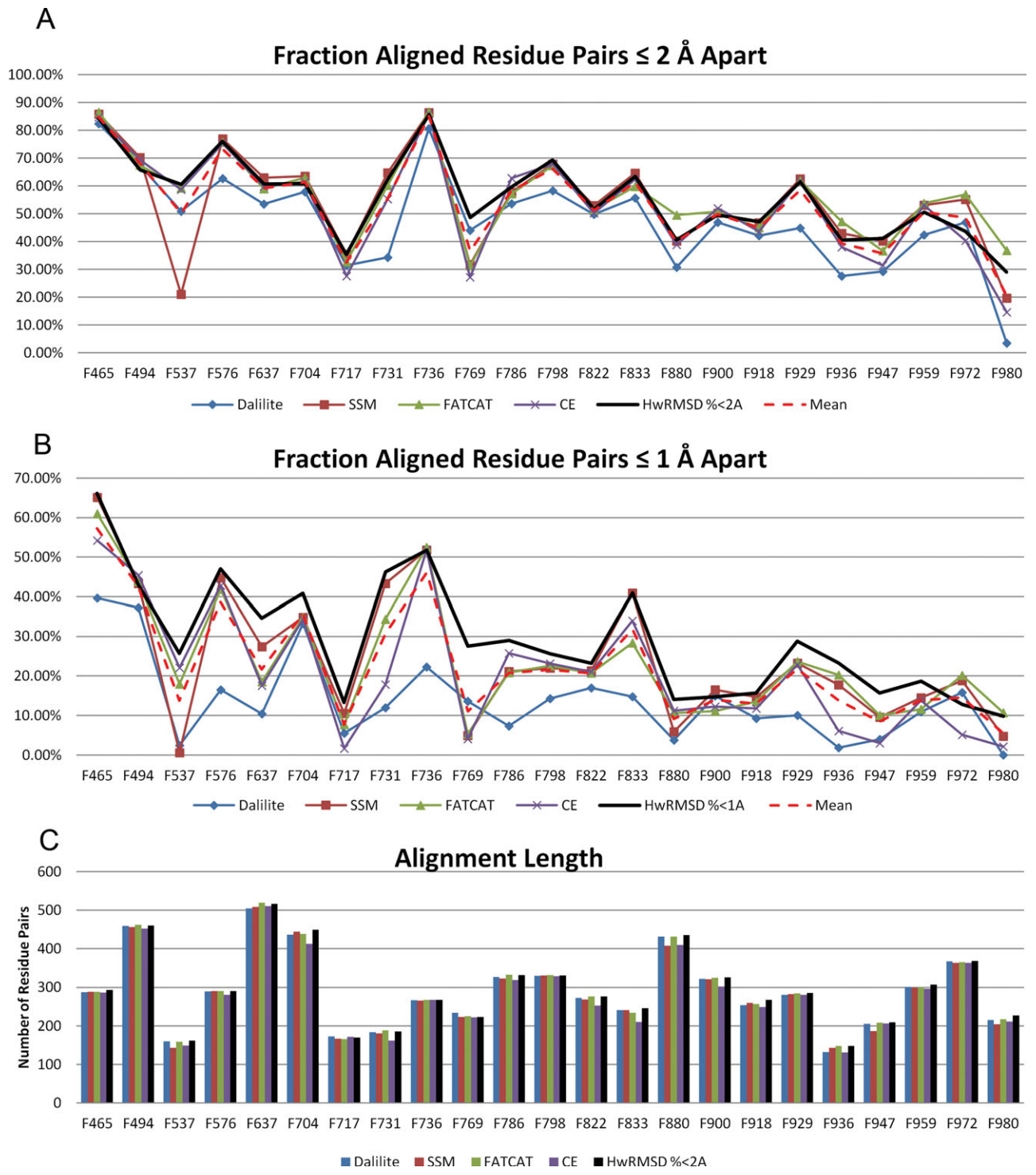

Figure 5

(**A**) Alignment results of HwRMSD (using BLOSUM50 global NW alignment) compared to other structural alignment programs using the percentage of aligned residues within 2 Å. (**B**) HwRMSD results compared to other structural alignment programs using the percentage of aligned residues within 1 Å. The fraction of pairs was calculated with respect to the shortest protein chain of the pair. (**C**) The number of residue pairs aligned in each structural alignment solution for each test case.

The particular strength of the HwRMSD method can best be demonstrated with the more stringent %Cα < 1Å metric. HwRMSD provides the highest fraction of residue pairs within 1 Å across the majority of the test cases [Fig. 5(B)]. This is especially significant considering the relative variation in performance seen in other methods,

with none consistently outperforming the others. Although most methods look roughly equivalent in Figure 5(A), Figure 5(B) shows the kind of precision that HwRMSD provides. *This superior performance in %Cα<1Å is particularly important when the structural superposition is used to correct the sequence alignment in an automated fashion.* As SE, as part of the HwRMSD pipeline, relies on proximity of paired residue to generate a corrected sequence alignment, residues that are 2 Å away or farther may still get inappropriately paired in the final sequence alignment. This chance is significantly reduced for paired residues 1 Å apart. HwRMSD produces the greatest fraction of such pairs; therefore, it is more likely to generate a corrected sequence alignment that represents the true structural correspondence.

Local HwRMSD alignments

The alignment and superposition of some homologous pairs may have multiple solutions, especially if there are multiple similar domains that move relative to one another (such as in the case of an apo versus holo structure). The wRMSD algorithm can explore alternate multiple alignments by using a "local" superposition where subsets of the initial sequence alignment are used to produce multiple, weighted superpositions. The local alignment option is built into the current implementation of wRMSD, and the local alignment functionality is extensively described in the previous wRMSD publication.[25]

By using only small segments of the sequence alignment, the weighing is initially focused upon a portion of the structure, allowing structural similarities that would have been washed out in the global alignment to possibly drive the weighted superposition. The alternate solutions can then be ranked by weights to choose the best of these "local" alignments. Use of this application might help to create better agreement with HOMSTRAD in regions outside the largest structural similarities. We plan to pursue this in the future. It is more difficult to stitch together multiple superpositions and multiple sequence alignments to create a consensus view of comparing homologs. At this point, it is most important to show that the primary solution is valid and an improvement over existing techniques.

CONCLUSIONS

We have now coupled our wRMSD method into a three-step pipeline with a sequence alignment and SE algorithm. Our method is capable of preferentially selecting out the regions with the best structural agreement between homologous proteins and generating a superposition that can identify significant similarities and differences. The SE algorithm then generates a "corrected" sequence alignment based on the improved superposition. This algorithm combination, referred to as HwRMSD, provides a flexible and transparent structure alignment method. The HwRMSD technique can be used to superimpose homologs with low-sequence identity and large conformational differences, an area where both sequence-based and structure-based methods may fail.

Employing homologs in the range of intermediate to low-sequence identity, we have shown that applying a weighting term can generate a better, more consistent overlay than standard superposition. Unlike sRMSD, HwRMSD can overcome the dependence of a structural superposition on the initial sequence alignment used to determine the appropriate Cα pairs and results in a robust structural alignment. Conserved regions of the structures are heavily weighted; thus, errors made in the initial sequence alignment are relatively discounted. The corrected alignments are not only robust but also correspond well to structure/sequence similarity, as determined by comparison to the curated alignments in the HOMSTRAD database. The wRMSD technique does not require prior knowledge of any protein system, and it removes the need to determine the best alignment method or parameters for each application. HwRMSD performs better than other structural alignment tools in closely superimposing protein regions of high structural similarity and generating structural alignments that represent that similarity. Our technique is also modular, allowing the user to control and interpret the superposition and alignment results at the individual steps of the structural alignment process—the initial alignment, the superposition, and the final structural alignment.

Of course, we must note that our tool, like any other, will breakdown when sequence or structural similarity is too low. We next aim to use this technique to align protein structures in our BindingMOAD database[48] to characterize ligand recognition across homologous families of protein structures.

ACKNOWLEDGMENTS

K.L.D. is grateful for receiving a Rackham Predoctoral Fellowship and a fellowship from the American Foundation for Pharmaceutical Education. Daniel Quang thanks the University of Michigan's Interdisciplinary REU Program in Structure and Function of Proteins (NSF DBI 0851723).

REFERENCES

1. Holm L, Sander C. Mapping the protein universe. Science 1996;273:595–602.
2. Watson JD, Laskowski RA, Thornton JM. Predicting protein function from sequence and structural data. Curr Opin Struct Biol 2005;15:275–284.
3. Marsden RL, Ranea JA, Sillero A, Redfern O, Yeats C, Maibaum M, Lee D, Addou S, Reeves GA, Dallman TJ, Orengo CA. Exploiting protein structure data to explore the evolution of protein function

and biological complexity. Philos Trans R Soc Lond B Biol Sci 2006;361:425–440.
4. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 2008;36:D419–D425.
5. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 2007;35:D291–D297.
6. Holm L, Rosenstrom P. Dali server: conservation mapping in 3D. Nucleic Acids Res 2010;38:W545–W549.
7. Bhaduri A, Pugalenthi G, Sowdhamini R. PASS2: an automated database of protein alignments organised as structural superfamilies. BMC Bioinformatics 2004;5:35.
8. Wang Y, Addess KJ, Chen J, Geer LY, He J, He S, Lu S, Madej T, Marchler-Bauer A, Thiessen PA, Zhang N, Bryant SH. MMDB: annotating protein sequences with Entrez's 3D-structure database. Nucleic Acids Research 2007;35:D298–D300.
9. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL Compendium in 2004. Nucleic Acids Research 2004;32:D189–D192.
10. Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: a database of protein structure alignments for homologous families. Protein Sci 1998;7:2469–2471.
11. Schmidt R, Altman RB, Gerstein M. LPFC: an internet library of protein family core structures. Protein Sci 1997;6:246–248.
12. Orengo CA, Thornton JM. Protein families and their evolution—a structural perspective. Annu Rev Biochem 2005;74:867–900.
13. Valas R, Yang S, Bourne P. Nothing about protein structure classification makes sense except in the light of evolution. Curr Opini Struct Biol 2009;19:329–334.
14. Kolodny R, Koehl P, Levitt M. Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. J Mol Biol 2005;346:1173–1188.
15. Taylor WR, Orengo CA. Protein structure alignment. J Mol Biol 1989;208:1–22.
16. Gerstein M, Levitt M. Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. Protein Sci 1998;7:445–456.
17. Subbiah S, Laurents DV, Levitt M. Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. Curr Biol 1993;3:141–148.
18. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
19. Kleywegt G. Use of Non-crystallographic symmetry in protein structure refinement. Acta Crystallogr D 1996;52:842–857.
20. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739–747.
21. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D 2004;60:2256–2268.
22. Mayr G, Domingues F, Lackner P. Comparative analysis of protein structure alignments. Biomed Chromatogr Struct Biol 2007;7:50.
23. Roland L D, Jr. Sequence comparison and protein structure prediction. Curr Opin Struct Biol 2006;16:374–384.
24. Sam V, Tai C, Garnier J, Gibrat J, Lee B, Munson P. Towards an automatic classification of protein structural domains based on structural similarity. Biomed Chromatogr Bioinformatics 2008;9:74.
25. Damm KL, Carlson HA. Gaussian-weighted RMSD superposition of proteins: a structural comparison for flexible proteins and predicted protein structures. Biophys J 2006;90:4558–4573.
26. Kabsch W. A solution for the best rotation to relate two sets of vectors. Acta Crystallogr A 1976;32:922–923.
27. Rice P, Longden I, Bleasby A. EMBOSS: The European molecular biology open software suite. Trends Genet 2000;16:276–277.
28. Tai C-H, Vincent J, Kim C, Lee B. SE: an algorithm for deriving sequence alignment from a pair of superimposed structures. Biomed Chromatogr Bioinformatics 2009;10:S4.
29. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
30. EMBL-EBI. Help-About Scoring Matrices, http://www.ebi.ac.uk/help/matrix.html; 2010.
31. Enroth C, Neujahr H, Schneider G, Lindqvist Y. The crystal structure of phenol hydroxylase in complex with FAD and phenol provides evidence for a concerted conformational change in the enzyme and its cofactor during catalysis. Structure 1998;6:605–617.
32. Mesecar AD, Koshland DE. Sites of binding and orientation in a four-location model for protein stereospecificity. IUBMB Life 2000;49:457–466.
33. Ye Y, Godzik A. Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics 2003;19:ii246–ii255.
34. Theobald DL, Wuttke DS. THESEUS: maximum likelihood superpositioning and analysis of macromolecular structures. Bioinformatics 2006;22:2171–2172.
35. Delano W. The PyMOL molecular graphics system, San Carlos, CA: DeLano Scientific; 2002.
36. Xu Z, Horwich AL, Sigler PB. The crystal structure of the asymmetric GroEL-GroES-(ADP)7 chaperonin complex. Nature 1997;388:741–750.
37. Braig K, Adams PD, Brünger AT. Conformational variability in the refined structure of the chaperonin GroEL at 2.8 A resolution. Nat Struct Mol Biol 1995;2:1083–1094.
38. Ditzel L, Löwe J, Stock D, Stetter K-O, Huber H, Huber R, Steinbacher S. Crystal structure of the thermosome, the archaeal chaperonin and homolog of CCT. Cell 1998;93:125–138.
39. Michel G, Sauve V, Larocque R, Li Y, Matte A, Cygler M. The structure of the RlmB 23S rRNA methyltransferase reveals a new methyltransferase fold with a unique knot. Structure 2002;10:1303–1315.
40. Nureki O, Shirouzu M, Hashimoto K, Ishitani R, Terada T, Tamakoshi M, Oshima T, Chijimatsu M, Takio K, Vassylyev DG, Shibata T, Inoue Y, Kuramitsu S, Yokoyama S. An enzyme with a deep trefoil knot for the active-site architecture. Acta Crystallogr D Biol Crystallogr 2002;58:1129–1137.
41. Biopython, version 1.42, http://biopython.org; 2006.
42. Price SR, Evans PR, Nagai K. Crystal structure of the spliceosomal U2B"-U2A′ protein complex bound to a fragment of U2 small nuclear RNA. Nature 1998;394:645–650.
43. Marino M, Braun L, Cossart P, Ghosh P. Structure of the lnlB leucine-rich repeats, a domain that triggers host cell invasion by the bacterial pathogen L. monocytogenes. Mol Cell 1999;4:1063–1072.
44. Owen DJ, Vallis Y, Pearse BM, McMahon HT, Evans PR. The structure and function of the beta 2-adaptin appendage domain. EMBO J 2000;19:4216–4227.
45. Traub LM, Downs MA, Westrich JL, Fremont DH. Crystal structure of the alpha appendage of AP-2 reveals a recruitment platform for clathrin-coat assembly. Proc Natl Acad Sci USA 1999;96:8907–8912.
46. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
47. Elofsson A. A study on protein sequence alignment quality. Proteins Struct Funct Bioinformatics 2002;46:330–339.
48. Benson ML, Smith RD, Khazanov NA, Dimcheff B, Beaver J, Dresslar P, Nerothin J, Carlson HA. Binding MOAD, a high-quality protein–ligand database. Nucleic Acids Res 2008;36:D674–D678.
49. Gong W, O'Gara M, Blumenthal RM, Cheng X. Structure of pvu II DNA-(cytosine N4) methyltransferase, an example of domain permutation and protein fold assignment. Nucleic Acids Res 1997;25:2702–2715.