PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

SHORT COMMUNICATION

MALIDUP: A database of manually constructed structure alignments for duplicated domain pairs

Hua Cheng,[1] Bong-Hyun Kim,[1] and Nick V. Grishin[1,2*]

[1]Howard Hughes Medical Institute, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9050
[2]Department of Biochemistry, University of Texas Southwestern Medical Center, Dallas, Texas 75390-9050

ABSTRACT

We describe MALIDUP (manual alignments of duplicated domains), a database of 241 pairwise structure alignments for homologous domains originated by internal duplication within the same polypeptide chain. Since duplicated domains within a protein frequently diverge in function and thus in sequence, this would be the first database of structurally similar homologs that is not strongly biased by sequence or functional similarity. Our manual alignments in most cases agree with the automatic structural alignments generated by several commonly used programs. This carefully constructed database could be used in studies on protein evolution and as a reference for testing structure alignment programs. The database is available at http://prodata.swmed.edu/malidup.

AQ1

Proteins 2007; 00:000–000.

Key words: duplication; homology; structure alignment.

INTRODUCTION

Protein homology is usually inferred by statistically significant sequence similarity. Since protein three-dimensional structures are generally more conserved than sequences, structural similarity can be used to find more distant homologs. Yet structural similarity does not necessarily imply homology, because it can be explained in terms of either divergent evolution or convergent evolution.[1,2] Thus fold similarity is usually supplemented by other considerations to provide convincing evidence for remote homology.[3,4] However, since internal duplications are frequently observed in molecular evolution,[5] two structurally similar domains occurring in tandem within the same peptide chain have a much greater chance to have arisen from a duplication event than from converging to the same structure independently. In other words, these domains are most likely to be homologs, even if they lack sequence or functional similarities. For instance, although the two domains in DNA helicases exhibit different binding activities and varied sequence motifs, their close resemblance in 3D structure strongly suggests that they are homologs resulting from duplication.[6–8] Therefore, looking for structural similarities between domains in the same peptide chain, one can find remote homologs while being less constrained or biased by sequence or functional considerations.

We selected cases of internal duplications from SCOP 1.69 database,[9] constructed manual alignments for the duplicated domains, and compared these alignments to those generated by three automatic structure aligners: DALI,[10,11] TM-align,[12] and FAST.[13] One of the goals of this project is to provide a library of well-constructed alignments. Manual attention to every domain pair with consideration of not only topological and spatial similarity but also sequence, structural, and functional features and other homologous proteins promises evolutionarily meaningful alignments of

Abbreviations: PDB, protein data bank; RMSD, root mean square deviation; SCOP, structural classification of proteins; 3D, three-dimensional.

Grant sponsor: NIH; Grant number: GM67165.
**Correspondence to:* Nick Grishin, HHMI/Biochemistry, University of Texas Southwestern Medical Center at Dallas, Dallas, TX 75390-9050. E-mail: grishin@chop.swmed.edu
Received 9 May 2007; Revised 6 July 2007; Accepted 17 July 2007
Published online 00 Month 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21783

higher quality than those produced by any given structure alignment program. The following general principles were used in the manual alignment construction: (1) core regions were aligned and variable loops were ignored; (2) H-bonding networks in β-sheets were followed, that is, if two residues were aligned, their respective H-bond partners were also aligned; (3) gaps were avoided as much as possible, especially in secondary structure elements; (4) two residues far from each other in the spatial superposition could be aligned (e.g. equivalent positions in two corresponding yet somewhat differently oriented helices), and two residues close in the superposition could be ignored (e.g. positions in random loops that happened to be near one another); and (5) structures were usually, but not always, treated as rigid bodies.

The alignments in MALIDUP can be used as a testing set for development of structural alignment programs, algorithms for remote homology inference using structural arguments, methods for evolutionary distance estimation from structures, and profile-based sequence similarity search tools that are seeded with structure-based alignments of remote homologs. It is also applicable in various studies of protein evolution, for example, structural and functional divergence after duplication.

MATERIALS AND METHODS

Selection of duplicated domains

From the SCOP database (version 1.69),[9] we retrieved all the domains with the word "duplication" in their annotations and grouped them by superfamilies. Some superfamilies were removed for various reasons, for example, the two repeats were too dissimilar in 3D structure to convincingly suggest homology. To avoid redundancy, we only selected one representative structure from a SCOP superfamily, mainly based on the structure's qualities (better resolution, smaller number of disordered residues). Currently, MALIDUP database contains 241 pairs of duplicated domains coming from 7 SCOP classes, 175-folds, and 209 superfamilies (some representative structures have more than one duplicated domain).

Pre-processing of coordinate files

For each pair, we defined the two duplicates' boundaries by consulting SCOP annotations, taking care to delineate the duplicates as compact structural domains. We extracted the duplicates' coordinates from the original PDB file and preprocessed these coordinate files in the following way: (1) if the two duplicates were circularly permuted relatively to each other, one of them was rearranged so that they had the same sequential order of structurally equivalent secondary structure elements; (2) the residues in every coordinate file were renumbered continuously, starting from 1; (3) the chain id in every coordinate file was changed to A regardless of the original chain id; and (4) the names of chemically modified amino acids were changed to the names of standard amino acids.

Manual and automatic alignments

We manually aligned the two duplicated domains in each pair in two steps. First, we identified corresponding secondary structure elements and superimposed the two domains in the software "Insight II." In doing so, we tried to align each pair in an evolutionarily meaningful way whenever possible. Since homologs usually preserve their core regions due to structural or functional reasons but diverge in peripheral regions,[14,15] it is reasonable to assume that structurally and topologically equivalent residues in the core regions are in most cases evolutionarily equivalent as well. For the majority of the pairs, the evolutionarily relevant, overall superposition could be easily identified by several tightly aligned loops and/or turns. For those more difficult pairs, where the structural similarity between the two domains was low and several different superpositions looked equally possible, we searched for shared sequence, structural, and functional features that were likely to have been inherited from the common ancestor.[3] Such features included conformations of loops and turns, disulphide bonds, ligand-binding residues, β-bulges, α-helix caps, residues with unusual conformations, and H-bonds. These features could be found by examining the structures carefully, comparing various members in the specific SCOP superfamily, and consulting literature. An example of using these features to align duplicated domains is described in Cheng and Grishin.[16] However, for a few most difficult pairs, the evolutionarily relevant superposition remained elusive even after these careful studies, and we provide several possible alignments for them. In the second step, we aligned the two domains' sequences according to the structural superposition made in the first step. In doing so, we followed the general principles listed in "Introduction."

The preprocessed coordinate files for every pair were submitted to three programs, DALI, TM-align, and FAST. DALI failed to output alignments for seven pairs, maybe due to the small number of secondary structural elements or low similarity. Thus, we ended up with 234 DALI alignments, 241 TM alignments, and 241 FAST alignments.

Score calculations

We used eight PSI-BLAST[17] iterations with *E*-value threshold of 0.001 against the NCBI nonredundant database to build a sequence profile for every duplicated domain in MALIDUP. The query sequence was the entire

 DOI 10.1002/prot

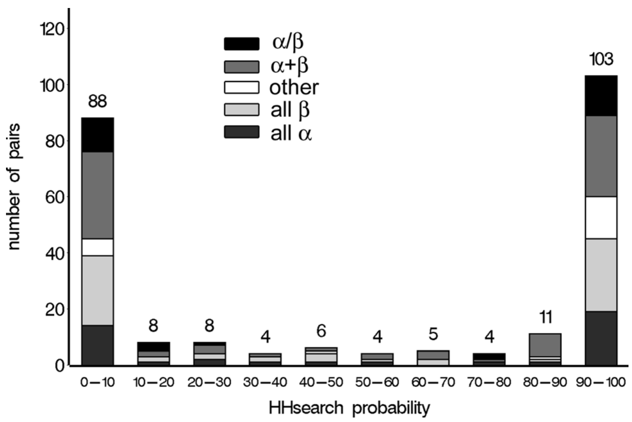


Figure 1

HHsearch "probability" distribution of the 241 pairs in MALIDUP. HHsearch "probability" ranges from 0 to 100, and is evenly divided into 10 bins. The horizontal axis shows the ranges of the bins, and the vertical axis represents the number of pairs that fall into a specific bin. Different scales of gray represent different SCOP classes (inset). The number above each bar is the total number of pairs in that bin.

PDB chain, and the part corresponding to the duplicate was extracted from the final profile. The two sequence profiles for every pair were aligned by HHsearch[18] with secondary structure prediction option.

To characterize a manual or automatic alignment, we calculated several scores: aligned length, sequence identity, Cα RMSD, and GDT_TS.[19] In addition, we computed an alignment-based COMPASS[20] score in the following way: from the aforementioned sequence profiles, we extracted columns corresponding to the aligned positions in the structure alignment, the two columns for every aligned position were scored by the COMPASS scoring function, and the final alignment-based COMPASS score was calculated as the sum over all the aligned positions.

To calculate the consensus score (a score characterizing how well an alignment matches the consensus of several aligners), we first delineated the "common positions" — those positions that were aligned in the same way by at least two of the four aligners, and then we counted how many of these common positions were correctly aligned by a specific aligner and divided this number by the total number of positions aligned by this aligner.

RESULTS AND DISCUSSION

To characterize the content of the newly defined MALIDUP database, we first show that MALIDUP contains many very remote homologs by computing the HHsearch probabilities for the 241 pairs. Then we demonstrate the high quality of the manual alignments by comparing them to the automatic alignments generated by different programs.

HHsearch analysis of MALIDUP

Figure 1 shows the distribution of HHsearch[18] probabilities for the 241 domain pairs in MALIDUP. HHsearch combines sequence profile information with predicted secondary structure information and is the state-of-the-art tool for inference of remote homology from sequences. Yet about half of the pairs in MALIDUP are beyond the detection power of HHsearch, as indicated by their low probabilities. For these pairs, the homologous relationship between the two domains is mainly justified by 3D structural similarity and sharing of the same peptide chain, as argued in "Introduction." Figure 1 also suggests that our method for selecting homologous pairs is not strongly biased by sequence similarity. F1

Agreement between manual and automatic alignments

For each pair, the agreement between the manual alignment and an automatic alignment equals the number of positions aligned in the same way in the two alignments divided by the total number of aligned positions in the manual alignment. For the seven pairs where DALI failed to output any alignments, the agreements between Manual and DALI are recorded as 0. The distribution of these agreements is shown in Figure 2. It should be noted that, F2
by examining the pairs with low agreements, we found several cases where the initial manual alignment was wrong. Figure 2 was plotted after we corrected these mistakes.

This histogram suggests that manual alignments generally agree with automatic alignments quite well, especially with DALI. Yet some pairs have very low agreements between Manual and a specific program. For 195 pairs,

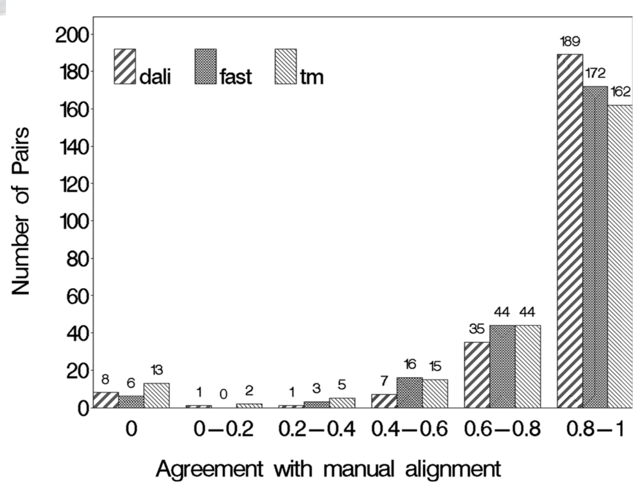
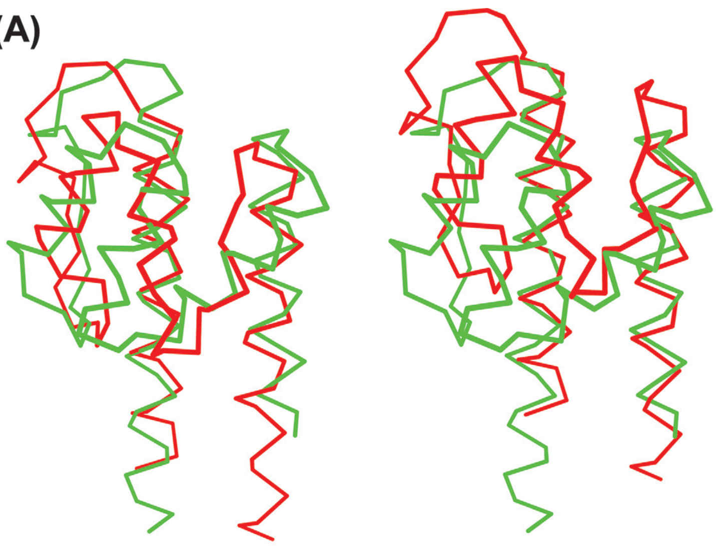


Figure 2

Agreement between manual and automatic alignments. Different patterns represent the three different programs. The horizontal axis shows the ranges of the agreement bins, and the vertical axis is the number of pairs that fall into each bin.

(A)



(B)

```
REAVRLLLLRNDLHNLQALLRAKATGRPFEEVLLLPGTLREEVWRQAYEAQDPAGMAQVLAVPGHPLARALRAVLRETQDLARVEALLAKRFFEDVAK domain1
PALRDYLALEVDAENLRTAFKLQGSGLAPDA FFLK FVDRVRFARLMEG    YAVLDELS  GTPFS          VRDLKALERGLRCVLLKEAKK  Manual
    PALRDYLALEVDAENLRTAFKL PDAFFLKG           DRVRFAR VLDE            GTPFS GLSGV ALERGLRCVLLKEAKKGVQ     DALI
PALRDYLALEVDAENLRTAFKLQGSGLAPDAFFLKGGRFVDRVRFARLMEGDYAVLDELSGTPFSGLSGVRDLKALERGLRCVLLKEAKKGVQDP     domain2
```


Figure 3

Comparison of Manual and DALI alignments. ***A:*** *Manual (left) and DALI (right) superpositions for domain 1 and domain 2 in Thermus thermophilus V-type ATP synthase subunit C*[22] *(PDB 1v9m). Compared to the manual superposition, the domain in red is shifted one-turn upwards in the DALI superposition. Diagrams are generated by MOLSCRIPT.*[23] ***B:*** *"Multiple alignment" format*[24] *of the Manual and the DALI alignments. The intact sequences of domain 1 and domain 2 are shown as the top and the bottom line, respectively. The middle lines are the aligned positions of domain 2, aligned to their corresponding positions in domain 1 (top line). The shifts between Manual and DALI alignments are indicated by slashes and backslashes.*

all of the three agreements are above 0.5; for 37 pairs, two of the three agreements are above 0.5; and for 9 pairs, one of the three agreements is above 0.5.

An example

Thermus thermophilus V-type ATP synthase subunit C has three structural domains.[21] Figure 3(A) shows the Manual and DALI superpositions of domain 1 and domain 2. These two superpositions differ by a one-turn shift of the mutual positions of the corresponding helices, resulting in 3- or 4-residue shifts in the sequence alignments shown in Figure 3(B). Thus the agreement between Manual and DALI alignments is only 1%. A detailed inspection of the two domains reveals several structural features in support of the manual alignment, for example, in domain 1, the side chain of Asn95 forms H-bonds with the backbone of Leu115, and in domain 2, their respective equivalent residues, Asn200 and Leu219, form H-bonds in the same fashion. Furthermore, with a high probability (96.2%), HHsearch independently arrives at an alignment that agrees with the manual alignment in most parts. Therefore, we are confident that the manual alignment is evolutionarily meaningful. For this pair, the agreement between Manual alignment and TM or FAST alignment is 61 or 73%, respectively.

Comparison of alignment-based scores

Six scores, namely aligned length, sequence identity, RMSD, GDT_TS, COMPASS, and consensus, were calculated for every pair based on alignments generated by the four aligners (DALI, TM-align, FAST, and Manual) as described in "Materials and Methods." The results are shown in Table I.

Compared with DALI and TM-align, FAST and Manual alignments are generally shorter but have better sequence identity, RMSD, and COMPASS score. DALI and TM-align appear less conservative and align more residues in the peripheral regions.

The consensus of individual programs has been shown to deliver better performance in structure predictions and multiple sequence alignments.[25,26] In the same spirit, we calculate a consensus score for each of the four aligners as described in "Materials and Methods." This consensus score equals the percent of an alignment that

DOI 10.1002/prot

Table I
Mean and Standard Error of Various Scores for Each Aligner

	DALI	TM-align	FAST	Manual
Aligned length (a.a.)	86.26 ± 2.78	**87.26** ± 2.87	75.14 ± 2.54	78.16 ± 2.54
Sequence identity (%)	16.95 ± 0.73	16.42 ± 0.72	17.88 ± 0.74	**18.00** ± 0.72
RMSD (Å)	2.74 ± 0.07	2.63 ± 0.05	2.56 ± 0.07	**2.49** ± 0.06
GDT_TS (%)	65.96 ± 0.85	67.72 ± 0.78	67.12 ± 0.94	**68.53** ± 0.86
COMPASS	3.65 ± 0.86	2.53 ± 0.86	**5.37** ± 0.87	5.23 ± 0.87
Consensus (%)	82.57 ± 1.29	74.68 ± 1.54	87.38 ± 1.24	**91.48** ± 0.63

The mean and the standard error of the mean for each score and each aligner. For RMSD, a smaller value is better; for all other scores, a larger value is better. The best mean in each row is bolded.

is aligned in the same way by at least two of the four aligners, assuming that similarities captured by different aligners are more likely to be true. Manual alignments have the best average consensus score, as well as GDT_TS, RMSD, and sequence identity, suggesting that manual alignments have the highest overall quality. The average length of manual alignments lies between FAST and DALI alignments, indicating a reasonable compromise in the number of aligned residues.

Web interface

The website for MALIDUP (http://prodata.swmed.edu/malidup) lists all the pairs in this database. Clicking on a pair name redirects the browser to that pair's specific page, which displays the basic information about the two duplicated domains, the alignment-based scores, and the manual and automatic structure alignments. The structural superpositions can be downloaded in PDB format or can be viewed in PyMol (http://pymol.sourceforge.net/). In addition, the whole database can be downloaded as a compressed file from ftp://iole.swmed.edu/pub/cheng/duplication/dup.tar.

ACKNOWLEDGMENTS

We thank Lisa Kinch for critical reading of the manuscript. We thank S Sri Krishna for helpful discussions. This work was supported by

REFERENCES

1. Orengo CA, Sillitoe I, Reeves G, Pearl FM. Review: what can structural classifications reveal about protein evolution? J Struct Biol 2001;134:145–165.
2. Krishna SS, Grishin NV. Structurally analogous proteins do exist. Structure (Camb) 2004;12:1125–1127.
3. Murzin AG. How far divergent evolution goes in proteins. Curr Opin Struct Biol 1998;8:380–387.
4. Kinch LN, Grishin NV. Evolution of protein structures and functions. Curr Opin Struct Biol 2002;12:400–408.
5. Heringa J. Detection of internal repeats: how common are they? Curr Opin Struct Biol 1998;8:338–345.
6. Subramanya HS, Bird LE, Brannigan JA, Wigley DB. Crystal structure of a DExx box DNA helicase. Nature 1996;384:379–383.
7. Gorbalenya AE, Koonin EV. One more conserved sequence motif in helicases. Nucleic Acids Res 1988;16:7734.
8. Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM. A conserved NTP-motif in putative helicases. Nature 1988;333:22.
9. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
10. Holm L, Sander C. Mapping the protein universe. Science 1996;273:595–603.
11. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
12. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
13. Zhu J, Weng Z. FAST: a novel protein structure alignment algorithm. Proteins 2005;58:618–627.
14. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J 1986;5:823–826.
15. Matsuo Y, Bryant SH. Identification of homologous core structures. Proteins 1999;35:70–79.
16. Cheng H, Grishin NV. DOM-fold: a structure with crossing loops found in DmpA, ornithine acetyltransferase, and molybdenum cofactor-binding domain. Protein Sci 2005;14:1902–1910.
17. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
18. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics (Oxford, England) 2005;21:951–960.
19. Zemla A. LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res 2003;31:3370–3374.
20. Sadreyev R, Grishin N. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J Mol Biol 2003;326:317–336.
21. Iwata M, Imamura H, Stambouli E, Ikeda C, Tamakoshi M, Nagata K, Makyio H, Hankamer B, Barber J, Yoshida M, Yokoyama K, Iwata S. Crystal structure of a central stalk subunit C and reversible association/dissociation of vacuole-type ATPase. Proce Natl Acad Sci USA 2004;101:59–64.
22. Numoto N, Kita A, Miki K. Structure of the C subunit of V-type ATPase from Thermus thermophilus at 1.85 A resolution. Acta crystallogr 2004;60 (Part 5):810–815.
23. Kraulis PJ. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Cryst 1991;24:946–950.
24. Godzik A. The structural alignment between two proteins: is there a unique answer? Protein Sci 1996;5:1325–1338.
25. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics (Oxford, England) 2003;19:1015–1018.
26. Wallace IM, O'Sullivan O, Higgins DG, Notredame C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res 2006;34:1692–1699.