Chaz Hyseni
Final Project
MCDB 452 — Fall 2006

Perspective on Protein-Protein Alignment with an Emphasis on Structural Comparison

Advances in full-genome sequencing are shed in a different light when combined with advances in structural genomics. To put novel information and data into perspective and relate them to existing knowledge, as in many other fields one must employ the strategy of comparison; only by comparison of sequences and structures, as well as grouping and classification, can new insights be formed, especially with respect to protein function (it is this function that gives meaning to the genes that encode them and may hold implications for fields such as drug discovery).

Proteins are at the core of practically every cellular activity. Life within and outside the cell is dependent on the activity of a variety of proteins. Based on function, proteins can be classified as catalytic (enzymes), regulatory, transport, storage, structural, contractile, scaffold, protective etc. (Tsai 2007). Another classification is based on shape and solubility: fibrous proteins are insoluble (e.g. collagen, fibroin, alpha-keratin), globular proteins are generally soluble (e.g. cytosolic enzymes, regulatory proteins), and membrane proteins, which are insoluble (e.g. transport proteins) (Tsai 2007).

Protein structure is essential to function, and thus structural comparison may yield information about functional similarities. Pairwise comparison and multiple alignment represent a way to analyze sequences and determine the level of homology. Similarly, a fundamental issue in analyzing protein structure is to formulate and compute a measure of similarity. Analysis of protein structure in conjunction with sequence alignments proffers more information than sequence analysis alone. For instance when Wallqvist and colleagues (2000) combined secondary structure information with sequence alignment, homology detection power increased. Strong sequence similarity alone is considered to be sufficient evidence for common ancestry but for distant homologs where there is no significant sequence similarity close structural and functional similarities provide such evidence (Reddy and Bourne 2003). Structure is more conserved than sequence, and thus structural comparison and alignment allows for identification of more distant homologs. The presence of convergence complicates structural comparison somewhat (convergence does not occur over great sequence lengths, so analogy does not affect sequence comparison as much), but homology can still be detected reliably (Sierk and Kleywegt 2004).

As of December 2006, the EMBL nucleotide sequence database contained some 83,000,000 entries (EBI website) while the UniProt protein sequence database contained some 250,000 entries (EBI website). On the other hand, the Protein Data Bank (PDB) contained approximately 37,000 protein structures or entries (PDB[a]). As evidenced by these figures, sequence entries far outnumber structure entries. Given the abundance of sequences and the relative ease with which they are obtained (as compared to the more exacting work of structure determination), there have justifiably been efforts to develop methods of predicting 3D structure from amino acid sequence. Proteins fold spontaneously into a unique three-dimensional conformation, which implies that there is a set of instructions that nature follows in attaining structure from an amino acid sequence. The next step between sequence information and being able to infer functional properties is the secondary structure of proteins. Protein chains usually fold to give secondary structures arranged in one of a few common patterns (Chothia 1984). This arrangement of patterns is classifiable, meaning that knowledge of secondary structure bears significance to classification and also functional aspects (Andersen and Rost 2003).

The first secondary structure prediction methods were based on observations of amino acid sequences and protein structures, i.e. propensities of amino acids for certain secondary structures. By dealing with it this way, the prediction problem becomes a pattern classification problem tractable by pattern recognition algorithms (Rost 2003). One such algorithm is GOR, which despite all the recent improvements is only at a level of accuracy where 64.4% secondary structures are correctly predicted (Garnier et al 1996). Rost (2003) provides a good summary of advances in secondary structure prediction methods. While neural networks have achieved some improvement by considering nonlocal interactions, and accuracy has improved further through more advanced algorithms, larger databases, and better search techniques (PSI-BLAST, hidden Markov models), prediction accuracy is still only at 77% for the best methods (as of 2003): PSIPRED, PROFphd, and SSpro (Rost 2003). Ab-initio, homology-modeling, and threading methods of computational prediction from protein sequence to structure have been improving and will likely reveal information about processes underlying protein folding. However, if the protein structures are known (this number is increasing rapidly thanks to experimental techniques that make high-throughput structure determination possible – some 6300 in 2006 (PDB[b])) the insight gained from structural alignment may actually be more useful in inferring functional properties by being able to identify homologous structures.

It is characteristic of biological systems that the objects we perceive to have a particular form arose by evolution from related objects with similar but not identical form (Lesk 2002). In terms of the macromolecular world, while some variation in structure is tolerated, certain features are more conserved than others. Although there are countless combinations possible, there are only somewhere between 1000 and 5000 protein folds (Chothia 1992). The PDB figure of 37,000 protein structures is much smaller than the 250,000 UniProt sequences. Structures are more conserved than sequences, and it follows that they should display some robustness, i.e. ability to accommodate change in amino acid sequence. Sequence divergence leads to increasing distortions in the mainchain conformation, and the fraction of residues in the core usually decreases, but these effects are limited until sequence similarity drops to below about 40-50% (Lesk 2002). The limited structural variations confer structural comparisons the potential to detect more distant homologs than sequence comparison alone.

Identification of homologous relationships is more than merely detection of sequence similarity. Homology is not a measure of similarity, but rather represents divergence as opposed to convergence of sequences. Sequence comparison is relatively easy when the level of similarity is higher than 50%, but in the event that two sequences should share less than 20% identity, it becomes difficult or impossible to establish whether they might have arisen via divergence or convergence (Tsai 2007). In terms of translating sequence similarity or dissimilarity into a structural and functional context, a residual identity of over 45% in an optimal alignment generally denotes similar structures, and practically similar function; a 25% sequence identity means a potentially similar general folding pattern; at 18-25% the 'twilight zone' is reached and sequence homology is marginal and unreliable (Tsai 2007).

The structural meaning of sequence alignment is not always straightforward. The alignment needs to be sufficiently long for sequence similarity to have any structural significance. To illustrate this point, Sander and

Schneider (1991) mention two extreme examples: 1) extended weak sequence similarity yet very similar structures between ras p21 protein and elongation factor TU (2.4 Ǻ rms C(α) deviation, but less than 20% sequence identity) and 2) short strong sequence similarity yet different structures in the case of octapeptides from subtilisin (2SBT) and an immunoglobulin (3FAB) (4.7 Ǻ rms C(α) deviation, but 75% sequence identity).

I attempted to provide some background information on certain methods of analyzing proteins and the motivation behind them, but intend to focus on structural comparison and alignment, as well as some algorithms developed to facilitate the process. But why is structural comparison and alignment important? As mentioned before, one reason for considering structural alignment is because it uncovers distant relationships not available from sequence alignments alone. Some other uses are in protein classification (the CATH database, where proteins are clustered at four major levels, uses structural alignment among other methods; the Scop database uses structural alignment as well), functional assessment (alignment of a protein of unknown function against proteins whose function is known), as well as for testing predicted structures (predicted through one of the previous or other similar methods) against a variety of known structures from a protein database.

A variety of methods exist that perform structural alignments and have been developed and used for years. Many of these have been compared against each other in reviews, and some of them are used by public databases, such as DALI (Holm and Sander 1993) (or DaliLite, current version 2.4.2, used for pairwise comparison) by the European Bioinformatics Institute (http://www.ebi.ac.uk/dali/) for the FSSP (Fold classification based on Structure-Structure alignment of Proteins) database, VAST (Gibrat et al 1996) used by the NCBI (http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml), or SSAP (Taylor and Orengo 1989) used by CATH (http://www.cathdb.info/cgi-bin/cath/SsapServer.pl). As reported by Bourne and Shindyalov (2003), as of May 20, 2002 there were 890 citations for DALI, 248 for SSAP, and 122 for VAST. DALI is by far the most used method.

DALI is based on the alignment of 2D distance matrices, which represent all intramolecular distances between C(α)s (residue centers) of a protein structure (Holm and Sander 1993). Protein structures change through evolution but the patterns of contacts between residues are highly conserved (if residues are in contact in one protein, they should be in contact in a related protein) (Lesk 2002). These contact patterns are represented in the 2D distance matrices. The hard part of the problem is finding the optimal submatrices. DALI uses the branch-and-bound algorithm (Lathrop and Smith 1996) to reduce the search space and thus speed up its performance (Singh and Brutlag 2000).

DALI is an efficient algorithm with a high level of accuracy, which explains its wide use in structural alignment. While the SCOP database is often cited as the gold standard because it is manually constructed by human experts, DALI is fully automated. According to Chi et al (2006), a structural alignment algorithm such as DALI, to reduce the computational effort of scanning large-scale protein databases, employs heuristics and the tradeoff may be the return of divergent results from the same query protein. In an attempt to test the performance of the different methods utilized by the EBI website, the CATH database SSAP server, and the latest SCOP v1.69 release (Chi et al 2006), I ran a DaliLite 2.4.2 pairwise comparison through the server (http://www.ebi.ac.uk/DaliLite/), one through the SSAP server (the proteins compared were 1HSA and 1DLH, described in the appendix below), and tested the new SCOP v1.69 release for both proteins. DaliLite performed comparisons of all chains automatically while these have to be selected manually on the SSAP server; in addition DaliLite performed slightly faster and reported a better rmsd (root mean square deviation) value (1.4 as compared to 2.2 Ǻ).

While such alignment of two arbitrary proteins is not possible with SCOP v1.69, comparison of a protein of choice against the database is performed relatively fast (but this is done one chain at a time). The Global-to-global alignment of the SCOP server performs the alignments one chain at a time, too, while DALI returns an automatic comparison of all chains against the database via email in under a minute (CATH never returned the results). The visual representation of structural alignment on the SCOP server is a plus, and the output is represented in a user-friendly format. Even though manual classification provides reliable results, it is labor intensive; as of May 30th, 2006, 10864 newly-discovered proteins deposited in the PDB have not been classified in SCOP v1.69 (Chi et al 2006). Despite the reported reliability of the SCOP database, the full automation of DALI is a useful feature (as will be seen shortly, the cost of this feature is not very high).

In an evaluation of several structural alignment methods conducted by Singh and Brutlag (2000), DALI performed very well. Other algorithms evaluated were STRUCTAL (Gerstein and Levitt 1996), VAST, MINAREA (Falicov and Cohen 1996), LOCK (Singh and Brutlag 1997), and 3dSEARCH (cit. as unpubl. in Singh and Brutlag 2000). Three query structures (an immunoglobulin, a myoglobin, and a TIM) were compared against a subset of 685 structures obtained from the PDB. In their comparison of these methods DALI performed consistently well, with high sensitivity and specificity, on each of the three queries. LOCK, STRUCTAL, and VAST performed well on the myoglobin and TIM, but not so well on the immunoglobulin.

Another review in Bourne and Shindyalow (2003) looks at DALI, SSAP, VAST, CE (Shindyalov and Bourne 1998), HOMSTRAD (Mizuguchi et al 1998; cited in Bourne and Shindyalov 2003), and SARF2 (Alexandrov 1996; cited in Bourne and Shindyalov 2003). Structure alignment is an NP-hard problem that is solved heuristically by all methods (Bourne and Shindyalov 2003). There is no exact solution to the protein structure alignment problem, only the best solution for the heuristics used in the calculation (Shindyalov and Bourne 1998). In discussing the different methodologies the authors suggest that all of the heuristics basically boil down to: representing the proteins to be compared in a coordinate-independent space, comparing and optimizing the alignment between the proteins, and measuring the statistical significance of the alignment against some random set of structure comparisons. However, different methods capture different aspects of protein structure and differ in how they search for optimal structure alignments; thus, while the basic concepts behind structural alignment are similar, different paths to the objective are chosen, and thus different results obtained.

CE, similarly to DALI, uses a distance approach; however, instead of hexapeptide the fragments are octameric. SSAP, on the other hand, uses the comparison of intraprotein C(β)-C(β) vectors as an indicator of directionality. VAST represents structures as a set secondary-structure-element vectors; the type, directionality, and connectivity of these vectors represent the topology of the protein. STRUCTAL uses iterative dynamic programming to minimize the rmsd between two protein backbones (Gerstein and Levitt 1996); the matrix of pairwise distances between C(α)s here is turned into a similarity matrix.

Root mean square (rms) deviation has been a standard of measurement of structural similarity. While a common metric, it is not perfect; rms deviation is a good similarity measure for proteins of the same length, but this dependence on length renders the absolute magnitude of rms deviation meaningless (Zhang and Skolnick 2005). Another problem with rms deviation is that it weights the distances between all residue pairs equally, meaning that a small number of local structural deviations could result in a high rmsd value, even when the global topologies of the compared structures are similar (Zhang and Skolnick 2005). Levitt and Gerstein (1998) made a modification in the step of structure comparison statistics, i.e. instead of rmsd they used the alignment score S. On STRUCTAL they compared the alignment score against rmsd and the E-value statistics based on the former did much better than those based on rms.

Similarly, a TM-score exploits a variation of Levitt-Gerstein (LG) weight factor that weights the residue pairs at smaller distances relatively stronger than those at larger distances; this increases sensitivity to the global topology rather than the local structural variations (Zhang and Skolnick 2005). The value of the TM-score is also normalized in a way that the score magnitude relative to random structures is not dependent on the protein's size (Zhang and Skolnick 2005). According to the authors, in STRUCTAL the LG-score is calculated based on the Kabsch rotation matrix that was defined for minimizing the rmsd rather than maximizing the LG-score, which slows down the convergence of the iteration procedure and reduces the efficiency of the algorithm. Their new algorithm, TM-align uses the TM-score rotation matrix to speed up the process of identifying the best structure alignments (Zhang and Skolnick 2005). An improvement over STRUCTAL and SAL (Kihara et al 2003), TM-align also outperformed DALI in reaching the highest TM-score. However, DALI still did well, although the version of DaliLite (DaliLite 2.3) that was evaluated is now superseded by a newer version of DaliLite, DaliLite 2.4.2.

DALI has been shown to be a very useful method in structural comparison, and has been performing well in evaluations. One of its drawbacks may be that it bases its statistical significance score on the rmsd value, which is considered to be suboptimal. Recent advances have been made by choosing better alignment scores, such as the TM-score incorporated in the TM-align algorithm. Other approaches exist, such as the one reported in de Melo et al. (2006), a contact map matching approach; also, the FAST algorithm (Zhu and

Weng 2005) (a clustering-based algorithm) represents a new development. At the heart of every development of structural alignment methods are the advances in experimental techniques such as X-ray crystallography and NMR spectroscopy, which allow for high-throughput structure determination. As the databases grow exponentially, the comparison power is increased and advances in computer algorithms are made possible, which allow these algorithms to progressively approximate biology.

References:

Andersen, C. A. F.; and Rost, B. 2003. *Secondary structure assignment.* In *Structural bioinformatics.* Bourne, P. E.; and Weissig, H. *Eds.* Hoboken, N.J.: Wiley-Liss.

Bourne, P. E.; and Shindyalov, I. N. 2003. *Structure comparison and alignment.* In *Structural bioinformatics.* Bourne, P. E.; and Weissig, H. *Eds.* Hoboken, N.J.: Wiley-Liss.

Chi, P. H.; Shyu, C. R.; and Xu, D. 2006. *A fast SCOP fold classification system using content-based E-Predict algorithm.* BMC Bioinformatics 7, 362.

Chivian, D.; Robertson, T.; Bonneau, R.; and Baker, T. 2003. *Ab initio methods.* In *Structural bioinformatics.* Bourne, P. E.; and Weissig, H. *Eds.* Hoboken, N.J.: Wiley-Liss.

Chothia, C. 1984. *Principles that determine the structure of proteins.* Ann Rev. Biochem. 53, 537–72.

Chothia, C. 1992. *Proteins. One thousand families for the molecular biologist.* Nature 357, 543–544.

de Melo, R. C.; Lopes, C. E. R.; Fernandes, F. A.; da Silveira, C. H.; Santoro, M. M.; Carceroni, R. L.; Meira, W.; and de A. Araujo, A. 2006. *A contact map matching approach to protein structure similarity analysis.* Genet. Mol. Res. 5 (2), 284-308

Falicov, A.; and Cohen, F. E. 1996. *A surface of minimum area metric for the structural comparison of proteins.* J. Mol. Biol. 258, 871-892.

Gerstein, M.; and Levitt, M. 1996. *Using iterative dynamic programming to obtain accurate pair-wise and multiple alignments of protein structures.* In *Proc. Fourth Int. Conf. on Intell. Sys. for Mol. Biol.* Menlo Park, CA: AAAI Press.

Gibrat, J. F.; Madej, T.; and Bryant, S. H. 1996. *Surprising similarities in structure comparison.* Curr. Opin. Struct. Biol. 6 (3), 377–85.

Godzik. A. 2003. *Fold recognition methods.* In *Structural bioinformatics.* Bourne, P. E.; and Weissig, H. *Eds.* Hoboken, N.J.: Wiley-Liss.

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

Kihara, D.; and Skolnick, J. 2003. *The PDB is a covering set of small protein structures.* J. Mol. Biol. 334, 793–802.

Krieger, E.; Nabuurs, S. B.; and Vriend, G. 2003. *Homology modeling*. In *Structural bioinformatics*. Bourne, P. E.; and Weissig, H. *Eds*. Hoboken, N.J.: Wiley-Liss.

Lathrop, R. H.; and Smith, T. F. 1996. *Global optimal protein threading with gapped alignment and empirical pair potentials*. J. Mol. Biol. 255, 641–665.

Lesk, A. M. 2002. *Introduction to bioinformatics*. Oxford; New York: Oxford University Press.

Levitt, M.; and Gerstein, M. 1998. *A unified statistical framework for sequence comparison and structure comparison*. Proc. Natl Acad. Sci. USA 95, 5913–5920.

Reddy, B. V. B.; and Bourne, P. E. 2003. *Protein structure evolution and the Scop database*. In *Structural bioinformatics*. Bourne, P. E.; and Weissig, H. *Eds*. Hoboken, N.J.: Wiley-Liss.

Rost, B. 2003. *Prediction in 1D: Secondary structure, membrane helices, and accessibility*. In *Structural bioinformatics*. Bourne, P. E.; and Weissig, H. *Eds*. Hoboken, N.J.: Wiley-Liss.

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.

Sierk, M. L; and Kleywegt, G. J. 2004. *Déjà vu all over again: Finding and analyzing protein structure similarities*. Structure 12, 2103–2111.

Singh, A. P.; and Brutlag, D. L. 1997. *Hierarchical protein structure superposition using both secondary structure and atomic representations*. In *Proc. Fifth Int. Conf. on Intell. Sys. for Mol. Biol.* Menlo Park, CA: AAAI Press.

Singh, A. P.; and Brutlag, D. L. 2000 *Protein structure alignment: A comparison of methods*. (http://cmgm.stanford.edu/~brutlag/Papers/singh00.pdf)

Taylor, W. R.; and Orengo, C. A. 1989. *Protein structure alignment*. J. Mol. Biol. 208 (1), 1–22.

Tsai, C. S. 2007. *Biomacromolecules: Introduction to structure, function and informatics*. Hoboken, N.J.: Wiley-Liss.

Wallqvist, A.; Fukunishi, Y.; Murphy, L. R.; Fadel, A.; and Levy, R. M. 2000. *Iterative sequence/secondary structure search for protein homologs: comparison with amino acid sequence alignments and application to fold recognition in genome databases***.** Bioinformatics 16 (11), 988–1002.

Zhang, Y.; and Skolnick, J. 2005. *TM-align: A protein structure alignment algorithm based on the TM-score*. Nucleic Acids Res. 33 (7), 2302–2309.

Zhu, J. H.; and Weng, Z. P. 2005. *FAST: A novel protein structure alignment algorithm*. Proteins 58 (3), 618-627.

Further Reading:

Bartlett, G. J.; Todd, A. E.; and Thornton, J. M. 2003. *Inferring protein function from structure*. In *Structural bioinformatics*. Bourne, P. E.; and Weissig, H. *Eds*. Hoboken, N.J.: Wiley-Liss.

Cymerman I. A.; Feder M.; Pawlowski M.; Kurowski M. A.; and Bujnicki J. M. 2004. *Computational methods for protein structure prediction and fold-recognition*. In Nucleic Acids and Molecular Biology series, *Practical Bioinformatics*. Bujnicki J. M. *Ed*. Berlin; Heidelberg: Springer-Verlag

Gerstein, M; and Levitt, M. 1998. *Comprehensive assessment of automatic structural alignment against a manual standard, the Scop classification of proteins*. Protein Sci. 7 (2), 445–456.

Grant, R. P. *Ed.* 2004. *Computational genomics: Theory and application*. Wymondham: Horizon Bioscience.

Orengo, C. A.; Pearl, F. M. G.; and Thornton, J. M. 2003. *The CATH domain structure database*. In *Structural bioinformatics*. Bourne, P. E.; and Weissig, H. *Eds*. Hoboken, N.J.: Wiley-Liss.

Raychaudhuri, S. 2006. *Computational text analysis for functional genomics and bioinformatics*. Oxford; New York: Oxford University Press.

Seckbach, J.; and Rubin, E. *Eds*. 2004. *The new avenues in bioinformatics*. Dordrecht; Boston: Kluwer Academic Publishers.

Wang, J. T. L.; Zaki, M. J.; Toivonen, H. T. T.; Shasha, D. *Eds*. 2005. *Data mining in bioinformatics*. London: Springer.

Westhead, D. R.; Parish, J. H.; and Twyman, R. M. 2002. *Bioinformatics*. Oxford: BIOS.

HTML content:

Nucleotide Sequence Database. European Bioinfornatics Institute (EBI). http://www.ebi.ac.uk/embl/

Protein Sequence Database. European Bioinfornatics Institute (EBI). http://www.ebi.ac.uk/swissprot/

Protein Data Bank (PDB[a]). http://www.rcsb.org/pdb/statistics/holdings.do

Protein Data Bank (PDB[b]).
http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100

Appendix

DaliLite Results

SUBMISSION PARAMETERS

Structure 1 1HSA Structure 2 1DLH

Results of Structure Comparison

Each chain of mol1 is compared structurally to each chain of mol2 using the DaliLite program. The Dali method optimises a weighted sum of similarities of intramolecular distances. Sequence identity and the root-mean-square deviation of C-alpha atoms after rigid-body superimposition are reported for your information only, they are ignored by the structural alignment method. Suboptimal alignments do not overlap the optimal alignment or each other. Suboptimal alignments detected by the program are reported if the Z-score is above 2; they may be of interest if there are internal repeats in either structure. In the C-alpha traces, the chains of the first and second structure are renamed 'Q' and 'S', respectively. The best match to each chain in the second structure is highlighted in the table below. Z-Scores below 2 are not significant.

First Structure & Chain: mol1A

No.	Second Structure & Chain	Z-Score	Aligned Residues	RMSD [Å]	Seq. Identity [%]	Structural Alignment	Superimposed C-alpha Traces	PDB Files: mol2 is rotated / translated to mol1 position
1	mol2B	14.6	162	3.9	24	click here	CA_1.pdb	mol1_original.pdb mol2_1.pdb
2	mol2A	14.2	115	3.3	23	click here	CA_2.pdb	mol1_original.pdb mol2_2.pdb
3	mol2D	14.2	114	3.2	23	click here	CA_3.pdb	mol1_original.pdb mol2_3.pdb
4	mol2E	14.1	181	4.4	24	click here	CA_4.pdb	mol1_original.pdb mol2_4.pdb
5	mol2D	7.2	84	3.8	11	click here	CA_5.pdb	mol1_original.pdb mol2_5.pdb
6	mol2A	7.2	84	3.8	11	click here	CA_6.pdb	mol1_original.pdb mol2_6.pdb
7	mol2B	6.8	82	2.8	11	click here	CA_7.pdb	mol1_original.pdb mol2_7.pdb
8	mol2E	6.3	80	2.8	13	click here	CA_8.pdb	mol1_original.pdb mol2_8.pdb
9	mol2B	5.5	78	3.0	10	click here	CA_9.pdb	mol1_original.pdb mol2_9.pdb
10	mol2E	5.1	72	2.7	8	click here	CA_10.pdb	mol1_original.pdb mol2_10.pdb
11	mol2A	4.9	78	2.7	9	click here	CA_11.pdb	mol1_original.pdb mol2_11.pdb
12	mol2D	4.8	78	2.7	9	click here	CA_12.pdb	mol1_original.pdb mol2_12.pdb
13	mol2B	4.6	69	2.7	7	click here	CA_13.pdb	mol1_original.pdb mol2_13.pdb
14	mol2B	4.4	83	5.3	8	click here	CA_14.pdb	mol1_original.pdb mol2_14.pdb
15	mol2E	4.2	83	5.7	8	click here	CA_15.pdb	mol1_original.pdb mol2_15.pdb
16	mol2E	2.3	67	3.4	4	click here	CA_16.pdb	mol1_original.pdb mol2_16.pdb

First Structure & Chain: mol1B

No.	Second Structure & Chain	Z-Score	Aligned Residues	RMSD [Å]	Seq. Identity [%]	Structural Alignment	Superimposed C-alpha Traces	PDB Files: mol2 is rotated / translated to mol1 position
17	mol2A	16.4	98	1.4	30	click here	CA_17.pdb	mol1_original.pdb mol2_17.pdb
18	mol2D	16.4	98	1.4	30	click here	CA_18.pdb	mol1_original.pdb mol2_18.pdb
19	mol2B	16.2	96	1.2	33	click here	CA_19.pdb	mol1_original.pdb mol2_19.pdb
20	mol2E	15.7	96	1.3	33	click here	CA_20.pdb	mol1_original.pdb mol2_20.pdb

First Structure & Chain: mol1D

No.	Second Structure & Chain	Z-Score	Aligned Residues	RMSD [Å]	Seq. Identity [%]	Structural Alignment	Superimposed C-alpha Traces	PDB Files: mol2 is rotated / translated to mol1 position

21	mol2B	14.7	182	3.9	24	click here	CA_21.pdb	mol1_original.pdb mol2_21.pdb
22	mol2A	14.3	115	3.3	22	click here	CA_22.pdb	mol1_original.pdb mol2_22.pdb
23	mol2D	14.2	114	3.1	23	click here	CA_23.pdb	mol1_original.pdb mol2_23.pdb
24	mol2E	14.2	181	4.3	24	click here	CA_24.pdb	mol1_original.pdb mol2_24.pdb
25	mol2A	7.1	87	6.8	10	click here	CA_25.pdb	mol1_original.pdb mol2_25.pdb
26	mol2D	7.1	84	6.9	11	click here	CA_26.pdb	mol1_original.pdb mol2_26.pdb
27	mol2B	6.8	82	2.8	11	click here	CA_27.pdb	mol1_original.pdb mol2_27.pdb
28	mol2E	6.3	89	6.3	12	click here	CA_28.pdb	mol1_original.pdb mol2_28.pdb
29	mol2B	5.7	83	3.2	10	click here	CA_29.pdb	mol1_original.pdb mol2_29.pdb
30	mol2E	5.6	89	3.2	9	click here	CA_30.pdb	mol1_original.pdb mol2_30.pdb
31	mol2B	5.2	72	2.8	8	click here	CA_31.pdb	mol1_original.pdb mol2_31.pdb
32	mol2E	5.1	71	2.7	8	click here	CA_32.pdb	mol1_original.pdb mol2_32.pdb
33	mol2D	4.9	77	2.6	10	click here	CA_33.pdb	mol1_original.pdb mol2_33.pdb
34	mol2A	4.8	102	11.1	8	click here	CA_34.pdb	mol1_original.pdb mol2_34.pdb
35	mol2E	2.3	67	3.4	4	click here	CA_35.pdb	mol1_original.pdb mol2_35.pdb

First Structure & Chain: mol1E

No.	Second Structure & Chain	Z-Score	Aligned Residues	RMSD [Å]	Seq. Identity [%]	Structural Alignment	Superimposed C-alpha Traces	PDB Files: mol2 is rotated / translated to mol1 position
36	mol2D	16.5	98	1.4	30	click here	CA_36.pdb	mol1_original.pdb mol2_36.pdb
37	mol2A	16.4	98	1.4	30	click here	CA_37.pdb	mol1_original.pdb mol2_37.pdb
38	mol2B	16.2	96	1.3	33	click here	CA_38.pdb	mol1_original.pdb mol2_38.pdb
39	mol2E	15.8	96	1.3	33	click here	CA_39.pdb	mol1_original.pdb mol2_39.pdb
40	mol2B	5.5	75	2.9	4	click here	CA_40.pdb	mol1_original.pdb mol2_40.pdb
41	mol2A	4.9	74	3.0	3	click here	CA_41.pdb	mol1_original.pdb mol2_41.pdb

Additional data

- Rotation-translation matrices for superimposition
- Listing of structurally equivalent residue ranges
- View the log - this is only informative to experts

Inputs
Here you can check that your PDB structures have been uploaded and parsed successfully. Below, the HEADER, TITLE, COMPND and AUTHOR records are echoed from the input PDB files, and the number of residues and secondary structure elements are listed for each chain.

First structure = mol1
HEADER   HISTOCOMPATIBILITY ANTIGEN   11-AUG-92   1HSA
COMPND   HUMAN CLASS I HISTOCOMPATIBILITY ANTIGEN/HLA-B
AUTHOR   D.R.MADDEN,J.C.GORGA,J.L.STROMINGER,D.C.WILEY

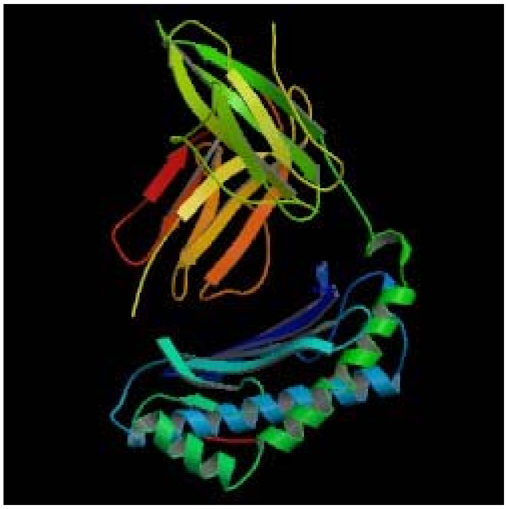
Chains found in mol1:
mol1A: 276 residues and 20 secondary structure elements
mol1B: 99 residues and 9 secondary structure elements
mol1D: 276 residues and 20 secondary structure elements
mol1E: 99 residues and 9 secondary structure elements

Second structure = mol2

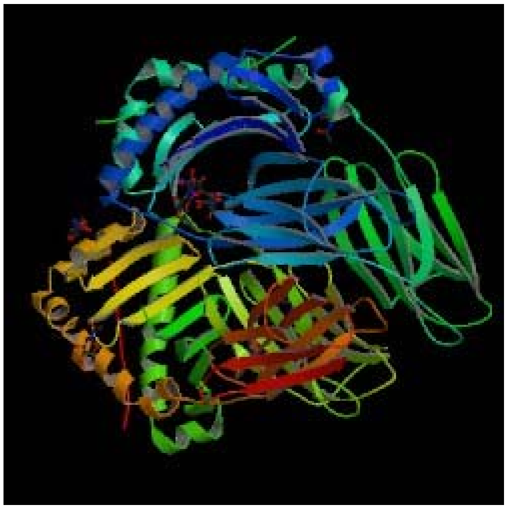
```
HEADER      HISTOCOMPATIBILITY ANTIGEN          15-FEB-94      1DLH
COMPND      HLA-DR1 (DRA, DRB1 0101) HUMAN CLASS II HISTOCOMPATIBILITY PROTEIN
             (EXTRACELLULAR DOMAIN) COMPLEXED WITH ANTIGENIC PEPTIDE
AUTHOR      L.J.STERN
```

Chains found in mol2:
mol2A: 180 residues and 14 secondary structure elements
mol2B: 188 residues and 16 secondary structure elements
mol2D: 180 residues and 14 secondary structure elements
mol2E: 187 residues and 13 secondary structure elements

Reference
Holm L, Park J (2000) DaliLite workbench for protein structure comparison.
Bioinformatics 16, 566-567.



1HSA (source: PDB)



1DLH (source: PDB)

	Domain1	Length	Domain2	Length	Equiv. Res.	Overlap (%)	Seq. id (%)	Score (0-100)	RMSD
	1hsaB0	99	1dlhA0	180	99	55	29	80.46	2.02

```
1hsaB0:pdbno   0         0         0         0
1hsaB0         |         |         |         |
1hsaB0:aa      ----------------------------------------
1hsaB0:ss

1dlhA0:pdbno   3         13        23        33
1dlhA0         |         |         |         |
1dlhA0:aa      EEHVIIQAEFYLNPDQSGEFMFDFDGDEIFHVDMAKKETV
1dlhA0:ss          SSSSSSSS    SSSSSSSS   SSSSSS    SSS


1hsaB0:pdbno   0         0         0         0
1hsaB0         |         |         |         |
1hsaB0:aa      ----------------------------------------
1hsaB0:ss

1dlhA0:pdbno   43        53        63        73
1dlhA0         |         |         |         |
1dlhA0:aa      WRLEEFGRFASFEAQGALANIAVDKANLEIMTKRSNYTPI
1dlhA0:ss      S            HHHHHHHHHHHHHHHHHHHHHH


1hsaB0:pdbno   1         11        21        31
1hsaB0         |         |         |         |
1hsaB0:aa      IQRTPKIQVYSRHPAENGKSNFLNCYVSGFHPSDIEVDLL
1hsaB0:ss           SSSSSS         SSSSSSSSSS     SSSSS

1dlhA0:pdbno   83        93        103       113
1dlhA0         |         |         |         |
1dlhA0:aa      TNVPPEVTVLTNSPVELREPNVLICFIDKFTPPVVNVTWL
1dlhA0:ss           SSSSSS         SSSSSSSSSS     SSSSS


1hsaB0:pdbno   41        50        60        70
1hsaB0         |         |         |         |
1hsaB0:aa      KNGERIE-KVEHSDLSFSKDWSFYLLYYTEFTPTEKDEYA
1hsaB0:ss      S SSSS   SSSS SSSS    SSSSSSSSS       SS

1dlhA0:pdbno   123       133       143       153
1dlhA0         |         |         |         |
1dlhA0:aa      RNGKPVTTGVSETVFLPREDHLFRKFHYLPFLPSTEDVYD
1dlhA0:ss      S SSSS   SSSS SSSS    SSSSSSSSS      SSS


1hsaB0:pdbno   80        90
1hsaB0         |         |
1hsaB0:aa      CRVNHVTLSQPKIVKWDRDM
1hsaB0:ss      SSSS       SSSS

1dlhA0:pdbno   163       173
1dlhA0         |         |
1dlhA0:aa      CRVEHWGLDEPLLKHWEFDA
1dlhA0:ss      SSSS       SSSSSS
```

Tables from Singh and Brutlag (2000) (comparison of structural alignment algorithms):

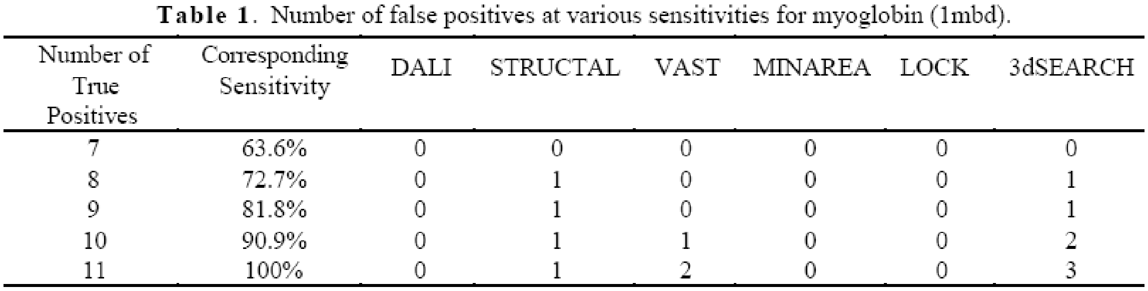


Table 1. Number of false positives at various sensitivities for myoglobin (1mbd).

Number of True Positives	Corresponding Sensitivity	DALI	STRUCTAL	VAST	MINAREA	LOCK	3dSEARCH
7	63.6%	0	0	0	0	0	0
8	72.7%	0	1	0	0	0	1
9	81.8%	0	1	0	0	0	1
10	90.9%	0	1	1	0	0	2
11	100%	0	1	2	0	0	3

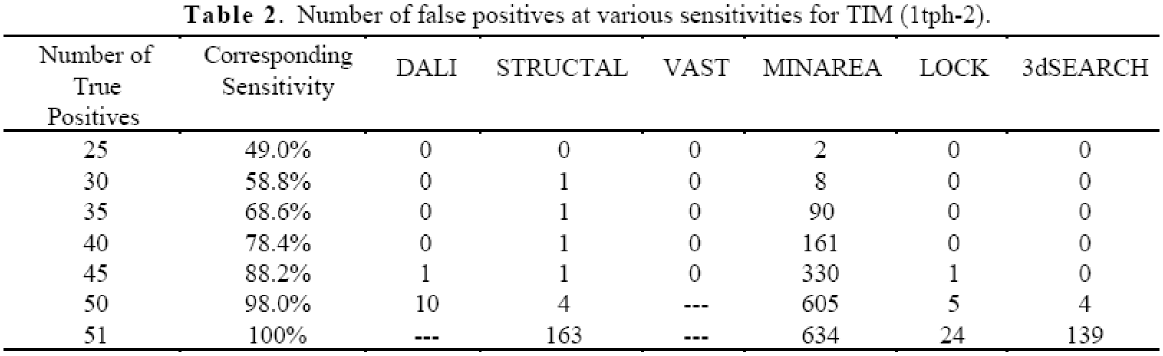


Table 2. Number of false positives at various sensitivities for TIM (1tph-2).

Number of True Positives	Corresponding Sensitivity	DALI	STRUCTAL	VAST	MINAREA	LOCK	3dSEARCH
25	49.0%	0	0	0	2	0	0
30	58.8%	0	1	0	8	0	0
35	68.6%	0	1	0	90	0	0
40	78.4%	0	1	0	161	0	0
45	88.2%	1	1	0	330	1	0
50	98.0%	10	4	---	605	5	4
51	100%	---	163	---	634	24	139

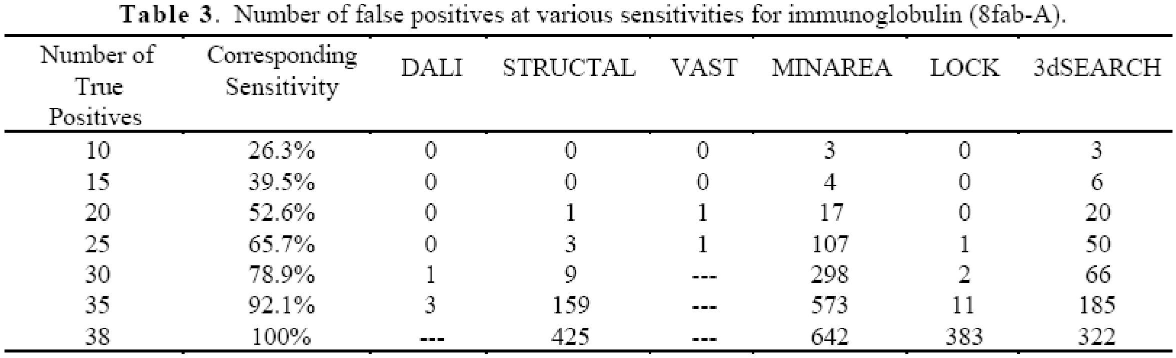


Table 3. Number of false positives at various sensitivities for immunoglobulin (8fab-A).

Number of True Positives	Corresponding Sensitivity	DALI	STRUCTAL	VAST	MINAREA	LOCK	3dSEARCH
10	26.3%	0	0	0	3	0	3
15	39.5%	0	0	0	4	0	6
20	52.6%	0	1	1	17	0	20
25	65.7%	0	3	1	107	1	50
30	78.9%	1	9	---	298	2	66
35	92.1%	3	159	---	573	11	185
38	100%	---	425	---	642	383	322