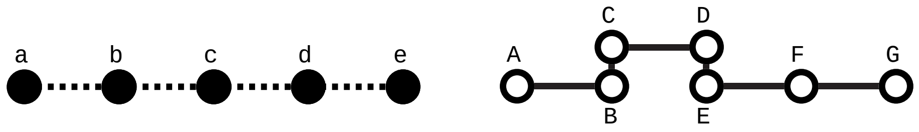
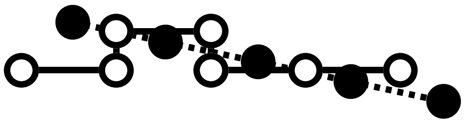
Figure 1: How Pairwise Structural Alignment Works

This schematic of our method of structural alignment is to be read from top to bottom. At TOP are two highly simplified structures (ABCDEFG and abcde) in an arbitrary, initial orientation. An initial equivalence is chosen, based on matching the ends of the two structures. Using this equivalence, we can least-squares superimpose the two molecules (giving an RMS deviation in corresponding atoms of 1.96 Å, UPPER-MIDDLE). Then based on relative positioning of the molecules determined from the fit, we calculate the distance, d_{ij}, between every atom i in the first structure and every atom j in the second structure. Each distance is transformed into a similarity value S_{ij} to form the similarity matrix shown at UPPER-MIDDLE-RIGHT, ($S_{ij} = M/[1+(d_{ij}/d_o)^2]$, where M=20 and d_o=2.24 Å). In the initial orientation atom “a” is close to atom “A” and even closer to atom “C,” and this is reflected in the S_{ij} matrix values. Dynamic programming chooses the pairs indicated by the boldface S_{ij} entries. The score for this selection is the sum of the S_{ij} values of the selected pairs less the gap penalty for each chain break (nbrk). Using a default gap penalty of 10 (M/2), the score is 7 + 12 + 12 + 13 + 13 - 10 - 10, for the S_{ij} matrix at UPPER-MIDDLE-RIGHT. The pairs chosen by dynamic programming give a new set of equivalences shown in LOWER-MIDDLE. These are used to do a second least-squares fit (giving an RMS of 0.65 Å). A new similarity matrix S_{ij} can now be calculated (shown at LOWER-MIDDLE-RIGHT), and dynamic programming again used to find new equivalences. Finally, at BOTTOM we see that these equivalences give a perfect match, so a final cycle of dynamic programming does not change the alignment. The iteration has converged on an alignment.

Figure 1 Graphic

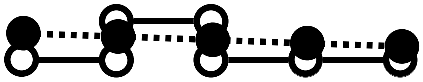


```
Initial Equivalences   - - a b c d e
                           | | | | |
                       A B C D E F G
```



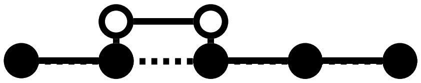
```
a - b - c d e    Score    57
|   |   | | |    Nbrk      2
A B C D E F G    RMS    1.96
```

	A	B	C	D	E	F	G
a	**7**	5	9	2	1	0	0
b	2	9	**12**	9	7	2	0
c	1	2	2	10	**12**	8	2
d	0	1	1	2	2	**13**	7
e	0	0	0	0	1	2	**13**



```
a b - - c d e    Score    91
| |     | | |    Nbrk      1
A B C D E F G    RMS    0.65
```

	A	B	C	D	E	F	G
a	**19**	4	4	1	1	0	0
b	4	**16**	16	4	4	1	0
c	1	4	4	14	**18**	4	1
d	0	1	1	4	4	**19**	4
e	0	0	0	1	1	4	**19**



```
a b - - c d e    Score   100
| |     | | |    Nbrk      1
A B C D E F G    RMS    0.23
```

	A	B	C	D	E	F	G
a	**20**	4	3	1	1	0	0
b	4	**20**	12	4	4	1	0
c	1	4	4	11	**20**	4	1
d	0	1	1	4	4	**20**	4
e	0	0	0	1	1	4	**20**

‘

Figure 2: Overall Performance on the Scop Superfamily Pairs

This figure shows the overall performance of our structural alignment algorithm on the 2107 scop superfamily pairs. Part (a) shows a plot of RMS vs. number of residues matched N for each of the pairs. A demarcation line separating good matches from bad ones is drawn as RMS = 4 (N + 135) / 225. Each pair that has some sequence similarity is indicated by an open circle. Clearly, these pairs tend to have somewhat closer structural matches. Sequence similarity was determined by doing an all-vs-all *sequence* comparison of the 941 scop domains using the FASTA program (with a k-tup value of 1) (Pearson & Lipman, 1988). An e-value for a pair less than .01 was taken to indicate significant sequence similarity with an expected false positive error rate of 1% (Pearson, 1996; Brenner et al., 1995). Note that none of the 941 domain structures in the 2107 scop superfamily pairs has sequence identity greater than 40%, so the sequence similarity found by FASTA is, by definition, somewhat marginal. Part (b) is similar to part (a) but now a plot of the normalized RMS' vs. N is shown for the same pairs (RMS’= 225RMS/(N+135)). The demarcation line is now RMS' = 4 Å.

Figure 2(a): RMS vs N

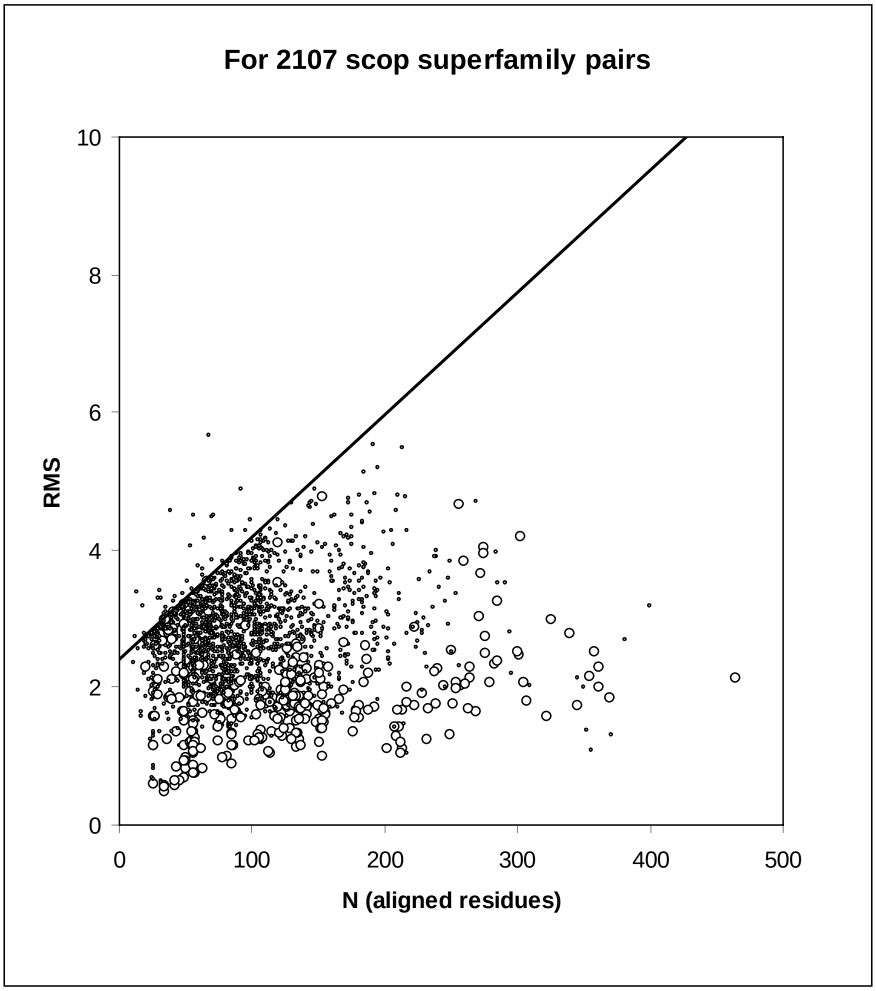


Figure 2(b): RMS’ vs N

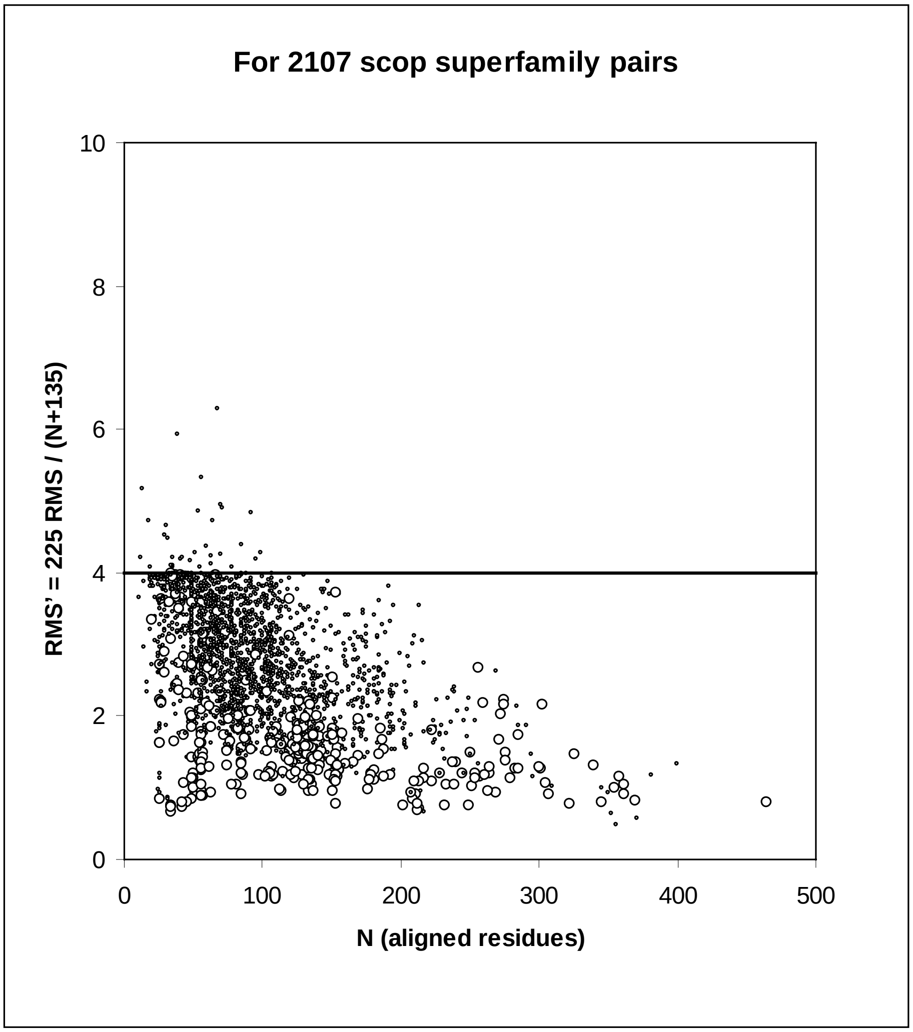


Figure 3: Distribution of RMS Values on the Scop Pairs

This figure shows the distribution of RMS and RMS' values resulting from aligning each of the 2107 Scop superfamily pairs.

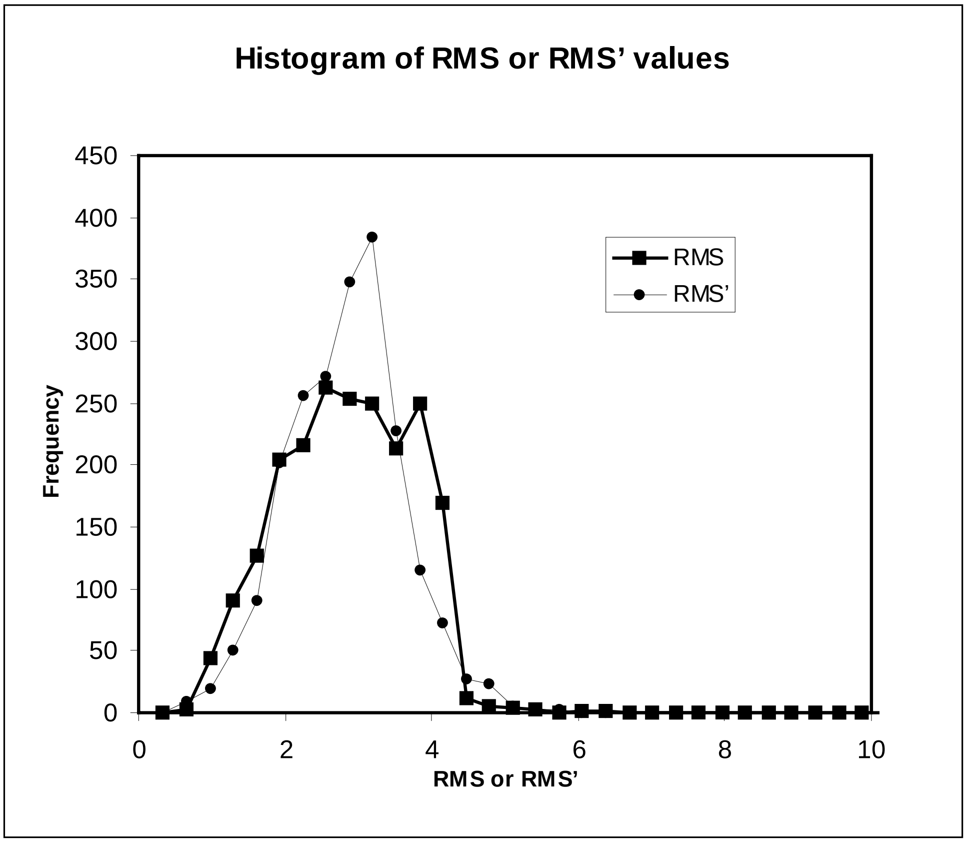


Figure 4: An Easy to Align Pair (Globins)

This figure shows a sample structural alignment of a pair of globins (d1mbd__ and d1ecd__, see methods for a discussion of the scop identifier conventions). The aligned positions are indicated by small, gray CPK spheres. This alignment is "easy" in the sense that it is obtainable from either the basic algorithm (Cα) or any variant (e.g. Cβ) and that there are very few mismatches compared to the hand alignment taken from the literature. See figure 7(b) for another view of this alignment.

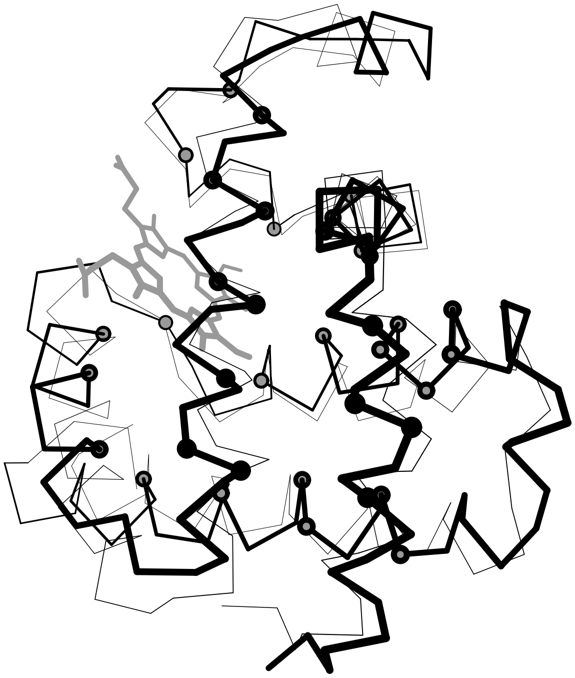


Figure 5: A Harder to Align Pair (Immunoglobulins)

This figure shows an alignment of immunoglobulin light-chain variable domain (d7fabl2) with an immunoglobulin constant domain (d1reia1). One can readily “match” this pair with the basic method (Cα) or any of the variants (in the sense that one can get a good RMS’ value). However, it is deceptively difficult to get the correct alignment in detail. The alignment from the basic method, just matching Cα atoms, is shown on the RIGHT. It gets a reasonable RMS from matching all the atoms and after elimination (see table below). However, it is clearly wrong because it misaligns the conserved disulfide (shown by the CPK spheres in the figure). In fact, comparison with the hand alignment shown in figure 7(c) indicates that every strand is slightly misaligned, giving 28 mismatches in total. It is necessary to use a variant method, which takes into account sidechain orientation and variable gap penalties, to get an alignment that gets the disulfides right. This alignment is shown at LEFT.

Method	Variant (Cα-Cβ + var. gap)	Basic (Cα atoms)
Mismatches vs. hand (36 aligned so /72)	6	28
RMS from all equiv. Cα's (84)	4.0 Å	3.1 Å
RMS after elimination (best 36)	1.7 Å	2.0 Å

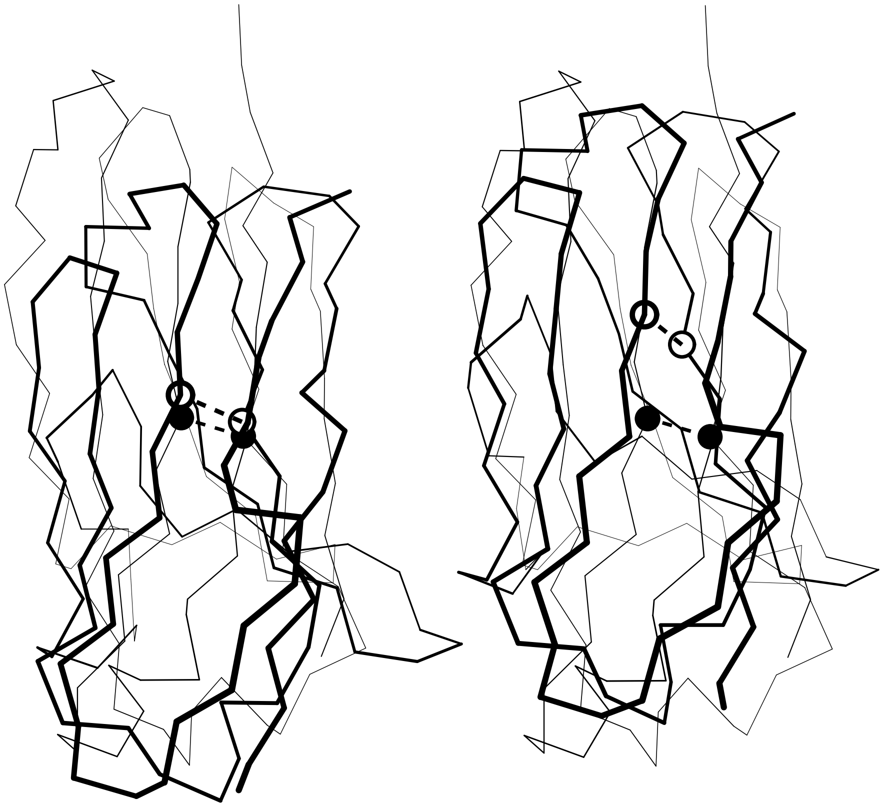


Figure 6: A Very Hard to Align Pair (G3P Dehydrogenase C-term. Domain)

This figure shows a scop pair that our program was not able to align at all. These structures (d1gd1o2 in the MIDDLE and d1gd1o2 at BOTTOM) are considered to share the fold of the C-terminal domain glyceraldehyde-3-phosphate dehydrogenase. However, they have in common only a small core region of similar topology, consisting of a four-stranded sheet with two helices packed on a face. This is highlighted in the structures and indicated in the topology diagram at TOP. The structures are grouped together in scop principally because they share an unusual type of cross-over connection, joining the strands in the sheet. This connection is highlighted by bold line in the topology diagram and a thick ribbon in the MIDDLE and BOTTOM subfigures. In both structures the crossed loops are inserted into the Rossmann-fold NAD(P)-binding domain in the same place, so they form an equivalent part of the active site. Furthermore, there is a third member of this scop superfamily family (d1dih_2) that has a pair of cross-loops equivalently inserted into a Rossmann-fold like domain.

Figure 6 Graphic

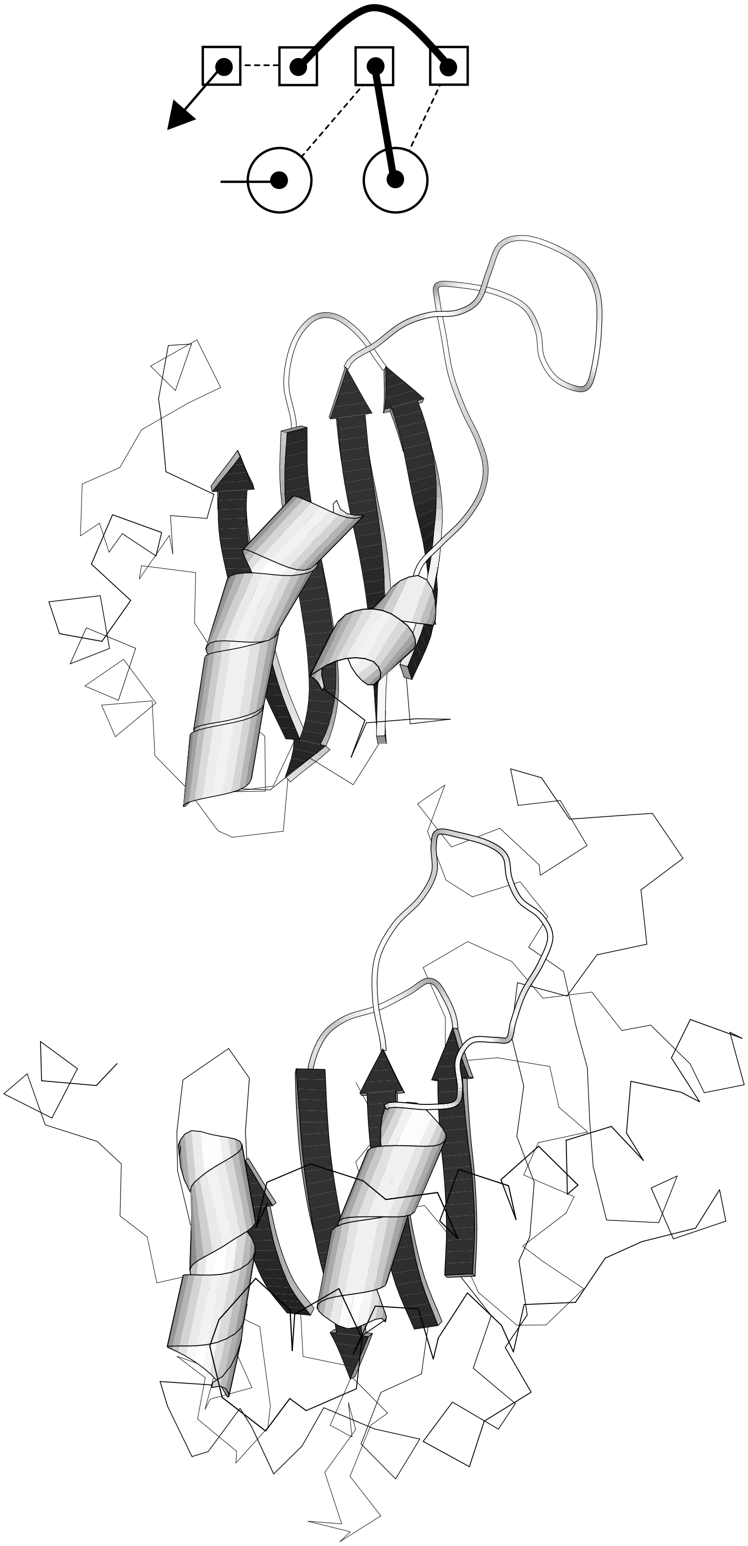


Figure 7: Sample Multiple Alignments

This figure shows sample multiple alignments for three protein families. Part (a) shows one for the dihydrofolate reductase (DHFR) family; part (b), for the globin family; and part (c), for two immunoglobulins. For each family, in turn, two separate multiple alignments are shown: the one marked "HAND" is a manually constructed "gold-standard" taken directly from the literature and the one marked "AUTO" is automatically generated by our program. The hand alignments were taken from Lesk & Chothia (1982) for the immunoglobulins, Gerstein et al. (1994) for the dihydrofolate reductases, and Lesk & Chothia (1980) for the globins. The HAND and AUTO alignments were aligned as blocks so that there are the fewest possible mismatches between them. Mismatches are scored only in the core alignable regions, marked by a “*” character in the "CORE" row. They are highlighted in the automatically generated alignment (by inverted text, changing case, and substituting "-" for "."). The DHFR alignment has 1 mismatch in total with d1dhfa_ as the central structure to which everything is aligned. The globin alignment has 18 mismatches with d1mbd__ as the central structure. For the immunoglobulins a third alignment, beyond the HAND and AUTO ones, marked SIMP is also shown. This is result of using the basic method (Cα). It clearly gets the alignment wrong and a more complex method is necessary to get the correct alignment (Cα-Cβ + var. gap). See figure 5 and the text for more details.

Figure 7(a) Dihydrofolate Reductase Alignment (Very Easy)

```
CORE 1       *********  **********  ************          *************
HAND d1dhfa_ LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
HAND d8dfr__ LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
HAND d4dfra_ ISLIAALAVDRVIGMENAMPWN-LPADLAWFKRNTL--------NKPVIMGRHTWESI
HAND d3dfr__ TAFLWAQDRDGLIGKDGHLPWH-LPDDLHYFRAQTV--------GKIMVVGRRTYESF

AUTO d1dhfa_ LNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQ-NLVIMGKKTWFSI
AUTO d8dfr__ LNSIVAVCQNMGIGKDGNLPWPPLRNEYKYFQRMTSTSHVEGKQ-NAVIMGKKTWFSI
AUTO d4dfra_ ISLIAALAVDRVIGMENAMPW-NLPADLAWFKRNTLD--------KPVIMGRHTWESI
AUTO d3dfr__ TAFLWAQDRNGLIGKDGHLPW-HLPDDLHYFRAQTVG--------KIMVVGRRTYESF

CORE 2              **********      **** **************         ***************
HAND d1dhfa_ VPEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHP
HAND d8dfr__ VPEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAVYKAAMEKP
HAND d4dfra_ ---G-RPLPGRKNIILS-SQPGTDDRV-TWVKSVDEAIAACGDVP------EIMVIGGGRVYEQFLPKA
HAND d3dfr__ ---PKRPLPERTNVVLTHQEDYQAQGA-VVVHDVAAVFAYAKQHLDQ----ELVIAGGAQIFTAFKDDV

AUTO d1dhfa_ -PEKNRPLKGRINLVLSRELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMVWIVGGSSVYKEAMNHP
AUTO d8dfr__ -PEKNRPLKDRINIVLSRELKEAPKGAHYLSKSLDDALALLDSPELKSKVDMVWIVGGTAVYKAAMEKP
AUTO d4dfra_ -G---RPLPGRKNIILSSSQPGTDDRV-TWVKSVDEAIAACGDVPE-----.IMVIGGGRVYEQFLPKA
AUTO d3dfr__ -P--KRPLPERTNVVLTHQEDYQAQGA-VVVHDVAAVFAYAKQHLD----QELVIAGGAQIFTAFKDDV

CORE 3         **********       *       **            *            ********
HAND d1dhfa_ GHLKLFVTRIMQDFESDTFFPEIDLEKYKLLPEYPGVLSDVQEEKGIK------YKFEVYEKND---
HAND d8dfr__ INHRLFVTRILHEFESDTFFPEIDYKDFKLLTEYPGVPADIQEEDGIQ------YKFEVYQKSVLAQ
HAND d4dfra_ --QKLYLTHIDAEVEGDTHFPDYEPDDWE---SVFSEF---HDADAQNSHS---YCFEILERR----
HAND d3dfr__ --DTLLVTRLAGSFEGDTKMIPLNWDDFT---KVSSRT---VEDTNPALT----HTYEVWQKKA---

AUTO d1dhfa_ GHLKLFVTRIMQDFESDTFFPEIDLEKYKLLPEYPGVLSDVQEEKG--I----KYKFEVYEK-N---
AUTO d8dfr__ INHRLFVTRILHEFESDTFFPEIDYKDFKLLTEYPGVPADIQEEDG--I----QYKFEVYQK-SV--
AUTO d4dfra_ --QKLYLTHIDAEVEGDTHFPDYEPDDWESVFSE------FHDADA--QNSHSSYCFEILER-R---
AUTO d3dfr__ --DTLLVTRLAGSFEGDTKMIPLNWDDFTKVSSR------TVEDTNPAL----THTYEVWQKKA---
```

Figure 7(b) Globin Alignment (Easy)

```
CORE 1                  ****************          ********************* *
HAND d2hhba_ ---------VLSPADKTNVKAAWGKVGA----HAGEYGAEALERMFLSFPTTKTYFPHF
HAND d2hhbb_ --------VHLTPEEKSAVTALWGKV------NVDEVGGEALGRLLVVYPWTQRFFESF
HAND d2lhb__ PIVDTGSVAPLSAAEKTKIRSAWAPVYS----TYETSGVDILVKFFTSTPAAQEFFPKF
HAND d1mbd__ ---------VLSEGEWQLVLHVWAKVEA----DVAGHGQDILIRLFKSHPETLEKFDRF
HAND d2hbg__ ---------GLSAAQRQVIAATWKDIAG--ADNGAGVGKDCLIKFLSAHPQMAAVFG-F
HAND d1mba__ ---------SLSAAEADLAGKSWAPVFA----NKNANGLDFLVALFEKFPDSANFFADF
HAND d1ecd__ ----------LSADQISTVQASFDKVKG--------DPVGILYAVFKADPSIMAKFTQF

AUTO d2hhba_ ---------VLSPADKTNVKAAWGKVGA-H---AGEYGAEALERMFLSFPTTKTYFPHF
AUTO d2hhbb_ ---------HLTPEEKSAVTALWGKV---N---VDEVGGEALGRLLVVYPWTQRFFESF
AUTO d2lhb__ ---------PLSAAEKTKIRSAWAPVYSTT---YETSGVDILVKFFTSTPAAQEFFPKF
AUTO d1mbd__ ---------VLSEGEWQLVLHVWAKVEA-D---VAGHGQDILIRLFKSHPETLEKFDRF
AUTO d2hbg__ ---------GLSAAQRQVIAATWKDIAG-A-DNGAGVGKDCLIKFLSAHPQMAAVFG-F
AUTO d1mba__ ---------SLSAAEADLAGKSWAPVFA-N---KNANGLDFLVALFEKFPDSANFFADF
AUTO d1ecd__ ----------LSADQISTVQASFDKVKG--------DPVGILYAVFKADPSIMAKFTQF

CORE 2                        *********************          *****************
HAND d2hhba_ --DLS--------HGSAQVKGHGKKVADALTNAVAHV-------D--DMPNALSALSDLHAHKL-
HAND d2hhbb_ -GDLSTP---DAVMGNPKVKAHGKKVLGAFSDGLAHL-------D--NLKGTFATLSELHCDKL-
HAND d2lhb__ KGLTTA----DQLKKSADVRWHAERIINAVNDAVASM-----DDT-EKMSMKLRDLSGKHAKSF-
HAND d1mbd__ -KHLKTE---AEMKASEDLKKHGVTVLTALGAILKK--------K-GHHEAELKPLAQSHATKH-
HAND d2hbg__ SGA-----------SDPGVAALGAKVLAQIGVAVSHL-----GDE-GKMVAQMKAVGVRHKGYGN
HAND d1mba__ KGKSVA-----DIKASPKLRDVSSRIFTRLNEFVNNA-----ANA-GKMSAMLSQFAKEHVGFG-
HAND d1ecd__ -AG-KDL---ESIKGTAPFETHANRIVGFFSKIIGEL------P---NIEADVNTFVASHKPRG-

AUTO d2hhba_ DLS----------HGSAQVKGHGKKVADALTNAVAHVD---D-----.MPNALSALSDLHAHKLR
AUTO d2hhbb_ GDL----STPDAVMGNPKVKAHGKKVLGAFSDGLAHLD---N-----.LKGTFATLSELHCDKLH
AUTO d2lhb__ KGL----TTADELKKSADVRWHAERIINAVNDAVASMD---D---TEKMSMKLRDLSGKHAKSFQ
AUTO d1mbd__ KHL----KTEAEMKASEDLKKHGVTVLTALGAILKKkG---H-----.HEAELKPLAQSHATKHK
AUTO d2hbg__ SGA----SDPG-----..VAALGAKVLAQIGVAVSHLGDEGK-----.MVAQMKAVGVRH.kgyG
AUTO d1mba__ KGK----S-VADIKASPKLRDVSSRIFTRLNEFVNNAA---N---AGKMSAMLSQFAKEHVG.fG
AUTO d1ecd__ AGK-----DLESIKGTAPFETHANRIVGFFSKIIGELP---N-----.IEADVNTFVASHK.prG

CORE 3          ********************            *******************
HAND d2hhba_ --RVDPVNFKLLSHCLLVTLAAHLP-A--EFTPAVHASLDKFLASVSTVLTSKYR------
HAND d2hhbb_ --HVDPENFRLLGNVLVCVLAHHFG-K--EFTPPVQAAYQKVVAGVANALAHKYH------
HAND d2lhb__ --QVDPQYFKVLAAVIADTVAAG------------DAGFEKLMSMICILLRSAY-------
HAND d1mbd__ --KIPIKYLEFISEAIIHVLHSRHP-G--DFGADAQGAMNKALELFRKDIAAKYKELGYQG
HAND d2hbg__ -KHIKAQYFEPLGASLLSAMEHRIGGKM---NAAAKDAWAAAYADISGALISGLQS-----
HAND d1mba__ ---VGSAQFENVRSMFPGFVASVAAPP-----AGADAAWTKLFGLIIDALKAAGA------
HAND d1ecd__ ---VTHDQLNNFRAGFVSYMKAHT------DFAGAEAAWGATLDTFFGMIFSKM-------

AUTO d2hhba_ ---VDPVNFKLLSHCLLVTLAAHLPAEFTPA---VHASLDKFLASVSTVLTSKYR------
AUTO d2hhbb_ ---VDPENFRLLGNVLVCVLAHHFGKEFTPP---VQAAYQKVVAGVANALAHKY------H
AUTO d2lhb__ ---VDPQYFKVLAAVIADTVAAG------------DAGFEKLMSMICILLRSA.------Y
AUTO d1mbd__ ---IPIKYLEFISEAIIHVLHSRHPGDFGAD---AQGAMNKALELFRKDIAAKYKELGYQG
AUTO d2hbg__ NKHIKAQYFEPLGASLLSAMEHRIGGKMNAA---AKDAWAAAYADISGALISGLQS-----
AUTO d1mba__ ---VGSAQFENVRSMFPGFVASVAA--PPAG---ADAAWTKLFGLIIDALKAAG------A
AUTO d1ecd__ ---VTHDQLNNFRAGFVSYMKAHTD---FAG---AEAAWGATLDTFFGMIFSKM-------
```

Figure 7(c) Immunoglobulin Alignment (Harder)

```
CORE 1             ***             ******          *****
HAND d7fabl2 PKAAPSVTLFPPSSEELQANKATLVCLISDFYPG--AVTVAWKAD----------------
HAND d1reia_ ---DIQMTQSPSSLSASVGDRVTITCQASQDI----IKYLNWYQQTPGKAPKLLIYEASNL

AUTO d7fabl2 PKAAPSVTLFPPSSEELQANKATLVCLISDFYPG--AVTVAWKAD-----GSPV-------
AUTO d1reia_ ---DIQMTQSPSSLSASVGDRVTITCQASQ--DI--IKYLNWYQQTPGKAPKLLIYEASNL

SIMP d7fabl2 PKAAPSVTLFPPSSEELQANKATLVCLISDFYPG--AVTVAWKADGSP-------------
SIMP d1reia_ -DIQMTqspSSLSA----SVGDrvtitcQASQDIIKYLnwyqqTPGKA----PKLLIYEAS


CORE 2               ***            ******            ******        *******
HAND d7fabl2 --GSPVKAGVETTTPSKQSNNKYAASSYLSLTPEQWKSHKSYSCQVTHE----GSTVEKTVAP----
HAND d1reia_ QAGVPSRFSGS---------GSGTDYTFT-ISSLQPEDIATYYCQQYQS----LPYTFGQGTKLQIT

AUTO d7fabl2 ------KAGVETTTPSKQSNNKYAASSYLSLTPEQWKSHKSYSCQVTHE----GSTVEKTVAP----
AUTO d1reia_ QAGVPSRFSGS----------GSgtdytfTISSLQPEDIATYYCQQYQS----LPYTFGQGTKLQIT

SIMP d7fabl2 ----VKA-GVETTTPSKQSNNKYAASSYLSLTPEQWKSHKSYSCQVTHE----GSTVEKTVAP----
SIMP d1reia_ NLQAGVPSrfsGSGSG------TdytftiSSLQPE----DIatyycqQYQSLPYTfgqgtklQIT--
```

Figure 8: Median Structure and Multiple-Alignment Quality

This table shows the how the quality of a multiple structural alignment decreases as one moves away from using the median structure as a basis for the alignment. Two families of structures are shown: immunoglobulin VL domains (all-β) and globins (all-α). For each family all possible pairwise alignments were done and then used to calculate the average distance (i.e. average RMS) between each structure and all the other structures. Because this distance will be smallest for structures near the cluster center, it can be used to rank each structure in terms of its proximity to the cluster center. Next, a multiple alignment was automatically generated based on a aligning all the structures in the family to a particular target structure. Every structure, in turn, was considered as the target. As described in the text, our automatically generated alignments were compared with manually generated "gold-standard" alignments, and the total number of comparisons and mismatches at core positions were tabulated. As we consider target structures farther away from the "center of the structure cluster" (in the RMS sense discussed above) the number of mismatches increases. This is true for both the highly diverged globin alignment and the less-diverged immunoglobulin alignment.

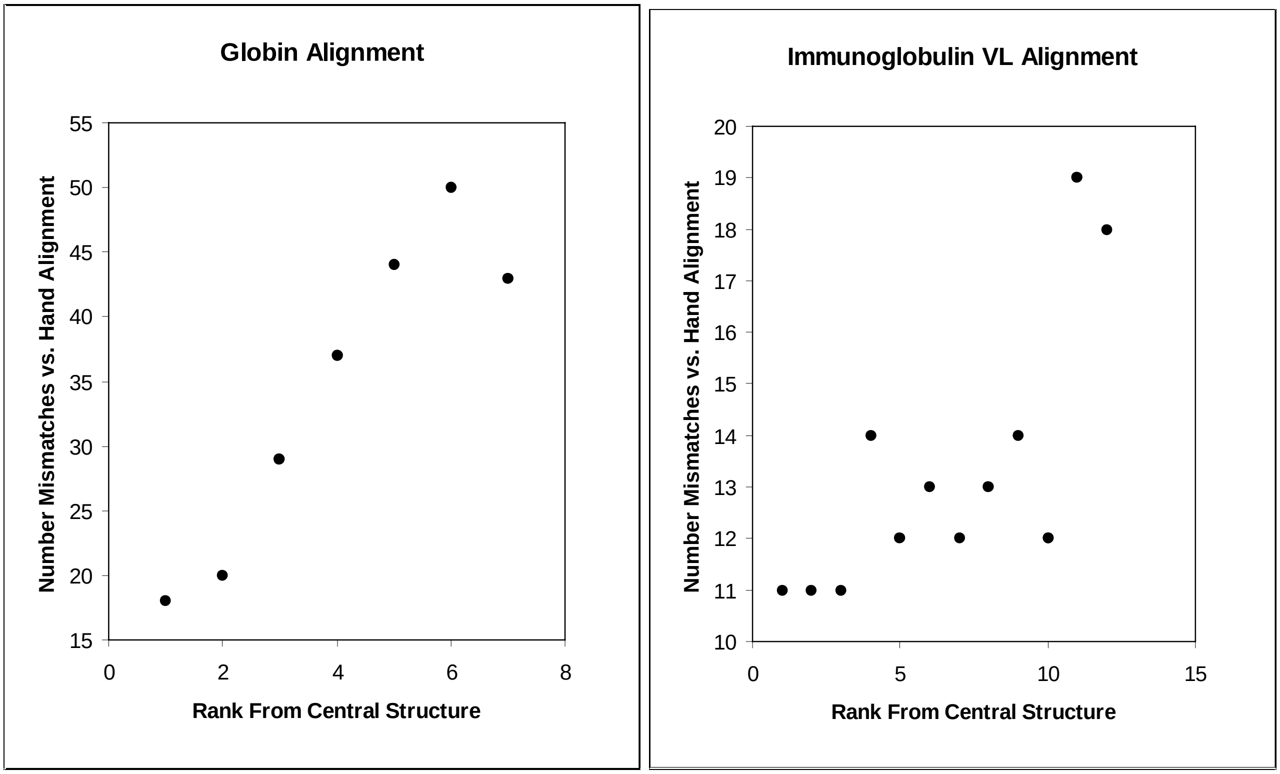


Table 1: Comparison of Automatically Generated Multiple Alignments vs. Manual "Gold-Standards"

The table shows summary statistics derived from comparing nine automatically generated alignments to manual, "gold-standard" alignments culled from the literature. These alignments are meant to correspond to as varied a selection of scop superfamilies as possible, given the limitations of the data in the literature. A detailed explanation of the statistics follows: Column "Num. Struct." gives the number of structures involved in the alignment. Column "Num. Comp." gives the number of comparisons done in comparing to the manual alignment. This is just the number of core positions times the number of structures. Column "Mismatches" gives the number of mismatches as compared to the hand alignment (which should be considered relative to the total number of comparisons). Column "Scop S.fam." gives the scop superfamily that the alignment was generated from. Column "Method" tells whether the basic method (Cα) or a variant was used in generating the alignment. Alignment 1 is from Chothia & Lesk (1982); 2, Lesk & Chothia (1982); 3, Joshua-Tor et al. (1995); 4, Graves et al. (1994); 5, Gerstein et al. (1993); 6, Harpaz & Chothia (1994) and Leahy et al. (1992); 7, Gerstein et al. (1994); 8, Lesk & Chothia (1980); 9, Chothia & Lesk (1987). All of the "gold-standard" alignments were done truly manually (i.e., not by using a different computer algorithm).

	Protein Family	Num. Struct.	Num. Comp.	Mis-matches	Scop S.fam.	Comment on structures	Method
1	Plastocyanin/azurin	2	118	2	2.05.1	all-β	Cα
2	Immunoglobulin VL-Fc (V-set + C1-set)	2	72	6	2.01.1	all-β	Cα-Cβ + var. gap
3	Cysteine proteinases (Gal6-Papain)	2	214	2	4.03.1	α+β with large insertions	Cα
4	C-type Lectins	2	212	0	4.77.1	α+β (mostly β)	Cα
5	P-loop containing NTP hydrolases (ADK)	3	534	0	3.21.1	α/β with a large conf. change	Cα
6	Immunoglobulin V-frame (V-set + I-set)	4	184	4	2.01.1	all-β (includes telokin)	Cβ + var. gap
7	Dihydrofolate Reductases	4	436	1	3.46.1	α/β	Cα
8	Globins	8	805	18	1.01.1	all-α	Cα + var. gap
9	Immunoglobulin V-set (just VL domains)	13	1183	11	2.01.1	all-β	Cβ