Improving Structural Superpositions for Complex Protein Structures by Limiting Query to Single Domains

Joanna E Boerner
Dept. Microbiology & Immunology
Stanford University School Of Medicine

Final Project for BIOCHEM 218: Computational Molecular Biology
06 June, 2002

Introduction

One limitation in deriving biologically significant information from structural superpositions is the large number of false positives resulting from alignments of small portions of the overall structures of functionally unrelated proteins. This weakness could be addresses by requiring that a reported hit and the query be superimposable over most points within each structure. For example, two proteins of high sequence identity (>80%) would be virtually superimposable. However, such a stringent algorithm would fail to report structural homology in the case where two protein structures share a common domain, but differ at other domains in their respective structures. For example, imagine there are two hypothetical transcription factors, TFA and TFB. Both transcription factors bear the same DNA-binding domain (e.g. a helix-turn-helix motif). Yet, their activation domains differ, TFA having a domain that binds another protein and TFB having a domain that binds a steroid. The structural superposition algorithm should be able to identify these two structures as similar. Current algorithms, such as DALI (*d*istance *ali*gnment tool) (Holm and Sander, 1993), are capable of doing this. Unfortunately, algorithms that can recognize alignment of single domains will necessarily also report alignments of smaller superpositions, e.g. the alignment of a few helices that bear no more than a structural role. Thus, the number of reported hits can reach into the hundreds. Sorting through all these for true positives is extremely time consuming, especially considering the search may not provide any new information (i.e. all true positives from the search had previously been recognized as structural homologs of the query). What is needed to resolve this difficulty is a method to enrich for hits to a particular portion of the query structure. One approach would be to parse the query structural file into separate files for each domain of the protein of interest, and use these individual files as queries for structural homolog searches. In this paper, I report the results from application of this method to four, 2 domain structures and the poliovirus RNA-dependent RNA polymerase.

Methods

The primary goal of this project was to demonstrate that productive structural searches could be performed with a query of a single domain or structural motif from a complex multi-domain protein. As a first step, I wished to confirm that a search using a single domain of a two domain protein would yield the expected true positives. Implicit in this step is the ability to modify manually a structural file (e.g. a pdb file) to include atom coordinates for only the domain of interest (see below). Three structures were selected from the SCOP (*S*tructural *C*lassification *o*f *P*roteins) database, 1.59 release 15 May 2002 (Murzin *et al*, 1995) with the following criteria: 1) the structures must contain a single polypeptide chain with only two domains, 2) these domains must be self contained (e.g. they visually form discrete structural units), and 3) the domains must be connected by one loop, such that a single cut in the amino acid chain would separate the domains. The selected structures, SCOP classifications, and sequence fragments are:

1daa:a D-amino acid aminotransferase, chain A; 277 residues; Class: multi-domain proteins (α and _); Fold: D-amino acid aminotransferase-like PLP-dependent enzymes, *2 domains: (1)* α+_*: _3-α2-_2; (2)* α/_*, a part of its mixed sheet forms barrel: n=6, S=8*

1daa-D1 Domain 1 fragment, residues Gly1-Arg120
1daa-D2 Domain 2 fragment, residues Pro121- Pro277

1dqs:a Dehydroquinate synthase, chain A; 381 residues; Class: multi-domain proteins (α and _); Fold: Dehydroquinate synthase-like, *2 domains: (1) α/_ of a Rossmann-fold topology, binds NAD; (2) multihelical array*

1dqs-D1 Domain 1 fragment, residues Pro1-Leu183
1dqs-D2 Domain 2 fragment, residues Pro184-Leu391

1dkz:a DnaK, C-terminal substrate binding fragment, chain A; 215 residues; Class: multi-domain proteins (α and _); Fold: Heat shock 70kD (HSP70), C-terminal, substrate binding fragment, *2 domains: (1) _-sandwich: 8 strands in 2 sheets; (2) α-helical bundle*

1dkz-D1 Domain 1 fragment, residues Val389-Ala521
1dkz-D2 Domain 2 fragment, residues Asn522-Gln603

The second step I undertook was to demonstrate that a more complex domain structure could be segmented into separate pdb files, which could be used as queries. One structure was selected from the SCOP database with the following characteristics: a single polypeptide structure with two visually distinct domains connected by two loops. The two domains could be separated via two scissions of the polypeptide to produce one domain of a single peptide fragment and a second domain of two peptide fragments. The selected structure, SCOP classifications, and sequence fragments are:

1ad2 ribosomal protein L1 mutant (Ser179Cys); 224 residues; Class: multi-domain proteins (α and _); Fold: Ribosomal protein L1, *2 domains: (1) α+_; (2) α/_ (interrupts domain 1)*

1ad2-D1 Domain 1 fragment, two polypeptides, residues Lys5-Gly67 and Arg160-Ser228
1ad2-D2 Domain 2 fragment, single polypeptide, residues Leu68-Gly159

Finally, I applied the above techniques to the structure for the poliovirus RNA-dependent RNA polymerase, 1rdr. This multi-domain protein has a relatively complex structure. As for all known DNA and RNA polymerase structures, this structure bears resemblance to a right hand, where the active site for phosphodiester bond formation lies in the palm domain. By comparison to other structures, the template:primer complex is believed to bind to the poliovirus polymerase across the base of the thumb domain, with the single-stranded template extending across the fingers domain (Hansen *et al*, 1997). Three structural fragments were created from 1rdr.

1rdr poliovirus RNA-dependent RNA polymerase; 461 residues; Class: multi-domain proteins (α and _); Fold: DNA/RNA polymerases, *"palm" domain has a ferredoxin-like fold, related to that of an adenylyl cyclase domain*

1rdr-RBD Putative RNA-binding motif, residues Lys228-Ala380
1rdr-motifs Canonical RNA polymerase motifs A-E, residues Lys228-Ser240, Gly292-Thr312, Leu321-Tyr334, Ala340-Thr355, Val371-Ala380

Files containing structural information for the domain fragments were created as follows. Pdb files for the parent structures were downloaded from the Protein Data Bank (Berman *et al*, 2000). To determine the sites for cleavage, structures were analyzed in Swiss PDB Viewer (Kaplan and Littlejohn, 2001). After identifying the residues bordering the scission points, the pdb files were opened in Microsoft Word as text files. The titles for the structures were changed to reflect the final fragment. And, atom coordinates for the superfluous parts of the structure were deleted, including all HETAMT and CONECT lines. The final TER line was updated to

reflect the final residue for the fragment. All other data remained unchanged, including line numbers. Thus, when an internal fragment was removed, the line numbers were no longer strictly sequential. Individual pdb files were created for each domain and individual chains (when the parental structural file contained more than one polypeptide chain or bound substrates). Since the DALI help files indicated that only ATOM lines were read by the alignment algorithm, it was believed that no changes were required in the REMARK, SEQRES, etc. lines. To confirm the modified structure files could be read, they were opened and viewed with Swiss PDB Viewer and aligned against self using the LOCK algorithm (Singh and Brutlag, 1997). By these assessments and the results returned from the DALI server, it was concluded that the modified pdb files were decipherable.

Structural alignments in this study were performed using the DALI algorithm Version 2.0 (Holm and Sander, 1993) for a database search of the query structure against the Protein Data Bank (Berman *et al*, 2000). Structural files were submitted to the remote server interactively. Output files were returned by email. Consequently, execution times for the searches are unknown. The DALI algorithm utilizes two-dimensional distances matrices of intra-molecular Cα-Cα distances. Distance matrices for query and target are compared pairwise, using Monte Carlo simulation to align submatrices with the similarity score as a guide (Mount, 2001). The number of hits reported was limited to alignments with Z-score (elastic similarity score) ≥ 2.0. This particular algorithm was selected for several reasons in additional to the fact it is considered the best structural superposition tool currently available (Brutlag, 1999). Since the algorithm parses the distance matrices into hexapeptide fragments, I believed the algorithm would be able to align structures composed of non-contiguous peptide sequences, as is the case for structural domains composed of several peptide fragments. In the DALI help files, it was stated that, for pdb files with multiple chains, the chains were read and aligned individually. As I did not want this to occur, the modified pdb files were created such that an indication of chain termination (designated by the TER line) was given only at the end of the coordinates for atoms in all peptide fragments. Given the results returned, it appears that the DALI algorithm did indeed read the files in their entirety as single structures. Initially, the DALI algorithm was favored because it did not require co-linearity of secondary structural elements in the aligned structures. Unfortunately, the DALI help files indicated that alignments were constrained to be sequential. Ideally, this limitation should be lifted.

Results derived from these searches were analyzed in various ways, as described in the Results and Discussion section. No gold standard list was derived before analyzing the searches. Rather, true positives and false positives were defined as follows. For the four, two domain structures selected from the SCOP database, the annotations listed were used as guides. For example, the SCOP Fold for 1daa was annotated as: *2 domains: (1)* α+_*: _3-α2-_2; (2)* α/_*, a part of its mixed sheet forms barrel: n=6, S=8*. Thus, true positives for Domain 1 included proteins classified as the parent structure (SCOP Class, multi-domain; SCOP Fold, D-amino acid aminotransferase-like PLP-dependent enzymes) or classified in the SCOP class, α+_. While it would have been preferable to use SCOP Class and Fold designations for the domain hits, this was not possible because the domain descriptions did not correlate readily with the SCOP Fold groupings. In some cases, other criteria were considered when defining true positive, as described in the Results and Discussion section.

The method of assigning true positive and false positive for searches with 1rdr, the poliovirus RNA-dependent RNA polymerase, differed. Here, true positive was defined as nucleic acid binding protein. All other proteins were considered false positives. While this definition appears to be based on function, it is indeed a structural classification. As mentioned above, the structures of RNA and DNA polymerases can be likened to that of a right hand, with the active site lying in the palm subdomain, the template:primer complex binding at the base of the thumb, and the single-stranded template extending across the fingers domain (Hansen *et al*, 1997; Kohlstaedt *et al*, 1992). The palm subdomain of RNA polymerases contains a core of five highly conserved motifs, A-E. DNA polymerases lack motif E. The structural arrangement of these canonical motifs is also structurally homologous to the RNA recognition motif (RRM), an RNA binding domain in splicing factors, ribosomal and tRNA binding proteins, and other nucleic acid binding proteins (Lindahl *et al*, 1994; Nagai *et al*, 1990; Goldgur *et al*, 1997). Thus, it is valid to define true positives as proteins that fall into these classes of nucleic acid binding proteins. (This does not imply that all nucleic acid binding proteins bear structural homology to RNA polymerases. There are distinct classes, such as DNA-binding motifs often found in transcription factors.)

It is worth mention that creation of a gold standard list for RNA polymerases was attempted using keyword searches of Protein Data Bank (Berman *et al*, 2000). This was not possible for two main reasons. First, all keyword searches attempt, using queries such as "RNA polymerase" or "nucleotidyltransferase", failed to return a list of only polymerase structures. For example, the list returned for "nucleotidyltransferase" contained many records for capping proteins. And second, not all polymerase structures have been given the same classification designation. For example, 1rdr, the poliovirus RNA-dependent RNA polymerase, is classified as a nucleotidyltransferase. Yet, 1hhs, the RNA-dependent RNA polymerase from dsRNA bacteriophage ϕ6 (a structural homolog of 1rdr) is classified as an RNA polymerase. Furthermore, the classification "nucleotidyltransferase" includes enzymes that transfer nucleotides to protein and small molecule substrates (other than polynucleotides). This lack of consistency in classification of structures in the Protein Data Bank makes it extremely difficult to identify all known structures for a particular type of protein.

Results and Discussion

The theory behind the approach used here is that it should be possible to decrease the number of false positives in a structural superposition search by limiting a query to the single domain or structural motif of interest from within a more complex structure. In addition to reducing the amount of computation required to accomplish the alignment search, this method is predicted to enrich for hits of biological relevance. For example, a search for proteins with similar structures to the active site of the query could be conducted. This approach would be useful for researches who, having identified a functionally active domain through site-directed site-directed mutagenesis and biochemical assays, wished to find structural homologs for this domain of interest.

To confirm that quality structural alignments could be returned for fragmented structure files, two domain structures were selected from the SCOP database (Murzin *et al*, 1995). As described in the Methods section, four parent structures were selected: in three structures the two domains were connected by a single polypeptide loop. The domains of the fourth structure were connected by two polypeptide loops. This fourth structure was included to confirm that a query

containing coordinates for discontiguos peptide fragments could be accurately read and aligned by the DALI algorithm (Holm and Sander, 1993). Several predications were made before performing the searches. First, the structural superposition algorithm should be able to align the single domains with structures containing these domains. And second, the hits returned for the entire chain should encompass hits returned for the individual domains, with few additions. These few novel hits returned for the whole structure presumably would align across the connection between the two domains, requiring both domains to be present to receive a Z-score ≥ 2.0.

The first structure used in this analysis was 1daa, a D-amino acid aminotransferase from *Bacillus sp.* strain YM-1 (Sugio *et al*, 1995) (Figure 1). Domain 1 (120 residues, α+_) returned 22 hits – 16 true positives (TP) and 6 false positives (FP). Domain 2 (157 residues, α/_) returned 7 hits, all true positives. The five hits with greatest statistical significance (Z-score) were identical for the two domains. No other hits were shared between the individual domains. Clearly, they are the most significant alignments. Indeed, these five proteins are classified in the same SCOP family as the parent structure, 1daa. In all three searches, the Z-score decreased greatly for all hits ranking lower than these five (i.e. any structure aligning to a single domain). The results for chain A and the parent pdb file (containing identical chains A and B) yielded identical results – 26 hits with 19 TP and 7 FP. This was predicted because DALI aligns the chains separately. The two hits of lowest statistical significance (Z = 2.0) were not reported for either Domain 1 or Domain 2 individually. One of these was a TP. All hits for the individual domains were also hits for the entire chain A. The ranking of false positives within the lists for Domain 1 and chain A were equally distributed, as can be seen in the modified Receiver-Operator Curve (number of true positives plotted vs. number of false positive), Figure 1D.

The second parent structure analyzed was 1dqs, the dehydroquinate synthase of *Aspergillus nidulans* (Carpenter *et al*, 1998) (Figure 2). Domain 1 (381 residues, α/_ of a Rossmann-fold topology) returned 459 hits, all of which were included in the hits list for 1dqs:a (chain A). Of the first 86, 81 were TP and 5 FP by defining true positive as α/_. Random hits selected from the remaining 300 structures were all true positives. Given that α/_ is a very general structural classification, it is likely that many of these true positives were indeed false positives. Randomly selecting hits and viewing alignments through LOCK and Chime confirmed this prediction. While a more stringent definition of true positive would have been ideal, it was not possible to derive such on the basis of SCOP classifications because the domain 1 structure did fit well with one and only one SCOP Fold group. Indeed, α/_ TPs fell into numerous Fold groupings. Since it was not realistic (in terms of time) to view all structural alignments pairwise, the assignment of TP and FP was limited to the top 30 hits for all lists longer than 30. Modified ROC curves were generated for these and included in the Figure 2.

1dqs-domain 2 (201 residues, multihelical array) returned 22 hits, with 17 TP and 5 FP when true positive was defined as all α. The 18th hit (1eld, Z = 2.1) was unique in comparison to aligned structures for domain 2 and chain A. The quality of this alignment was assessed using the LOCK algorithm and viewed with Chime. 1eld is a large, complex structure with high packing density. The structural alignment produced by LOCK for 1eld and 1dqs was not good – few secondary structural elements were actually superimposed. It is likely that the significance for this alignment was high enough for DALI to report because the intermolecular Cα atomic distances were sufficiently low, as a direct consequence of the compact nature of 1eld. This suggests that, in general, compact structures are more likely than open structures to be selected as false positives.

Search results for 1dqs:a and the parent pdb file 1dqs (containing coordinates for identical chains A and B) were the same: 487 hits, 10 of which were unique in comparison to the search results for domain 1 or domain 2. The unique hits were not clustered with respect to Z-score. Structural alignments for a few were viewed with LOCK and Chime. The quality of the alignments appeared to vary independently of the Z-score. When assessing TP and FP in the first 30 hits, it was noted that none of the other structures classified by SCOP as the same class, fold, and superfamily were listed. An exhaustive search of all hits confirmed that none were returned as statistically similar structures. Within the first 30 hits returned for 1dqs:a, 28 were true positives and 2 false positives (Figure 2D).

The third structure analyzed was 1dkz, the C-terminal substrate-binding domain from DnaK, a molecular chaperone from *Escherichia coli* (Zhu *et al*, 1996) (Figure 3). Domain 1 (133 residues, _-sandwich) returned 28 hits – 17 TP and 11 FP. Of the false positives, 9 had α+_ structures. This high number of false positives (see ROC curve 1, Figure 3F), was initially attributed to the site at which domain 1 was separated from domain 2. As depicted in Figure 3A, the site for cleavage selected created a domain 1 containing a _-sandwich with a single attached helix. When viewing the parent structure, these two domains appear to be individually folding units. The presence of the helix seemed the likely cause for the α+_ false positives. In light of the true composition of domain 1 (rather than SCOP annotation), these FPs should be considered TP, for a total of 26 TP and 2 FP. To confirm this prediction, the DALI search was repeated using an all _ domain 1 (domain 1beta, Figure 3D), with cleavage after Leu507. Unexpectedly, the results were not improved: 30 hits were reported, with 17 TP and 13 FP. Both new false positives were classified in the SCOP class, small protein. Neither of these alignments was very good, as determined by visual inspection of Chime representations derived from structural alignment using the LOCK algorithm. Although this result refutes the prediction, it demonstrates that the DALI algorithm was not "distracted" by the lone helix. In other words, the reported hits were globally aligned across the more extensive _-sheet and not localized over the helix.

1dkz domain 2 (82 residues, α-helical bundle) returned 315 hits, all included in the results for 1dks:a (chain A) and 1dks (original pdb containing coordinates for a substrate fragment). Analysis of the top 30 hits revealed 25 TP and 5 FP. Nine of the TP were classified in the same SCOP Fold, four-helical up-and-down bundle.

Search results for 1dkz:a (215 residues, entire A chain) listed 374 hits, including 42 unique structures not returned for domains 1 or 2. These unique hits were clustered together and had low Z-scores (mean = 2.26). Of the top 30 hits, 21 were true positives and 9 were FP (Figure 3F). One fewer hits (373 total) were returned when the original 1dkz pdb file. This file contained coordinates for a small peptide substrate in addition to those for chain A. The missing hit was ranked 309, Z = 2.2, in the 1dkz:a list. The fact that addition of a few atoms resulted in loss of significance for this alignment suggests it was not a true positive, supported by the low statistical significance of the observed alignment. True and false positives were not identified from results of these queries.

The final set of queries for this stage of the analysis was derived from 1ad2, a mutant form of the ribosomal protein L1 from *Thermus thermophilus* (Unge *et al*, 1997) (Figure 4). As discussed above, two cuts of the polypeptide backbone were required to generate domains 1 and 2. Domain 1 (two chains, 132 residues, α+_) returned 85 hits, fourteen of which were also returned for domain 2. This was the largest number of coincident hits among the four sets of searches, and the time that domain 1 – domain 2 coincident hits were spread evenly down the ranking. In all previous cases, 5 or fewer such hits were returned, all clustered at the top of the Z-

score ranking. Of the top 30 hits for domain 1, there were 24 TP and 6 FP. Domain 2 (single chain, 92 residues, α/_) returned 265 hits, one of which was unique from domain 1 and chain A results. This unique hit was not significant, as it was a false positive with Z = 2.0 and rank # 262. Of the top 30 hits for domain 2, 25 were TP and 5 were FP. Two false positives (1i3c and 1iow) were labeled as such because corresponding entries were not found in the SCOP database. Chain A (224 residues) returned 339 hits, 4 of which were unique. All four unique hits were false positives (when true positive is defined as the same SCOP class and fold as the query) with low Z-scores (Z < 2.2). Of the top 30 hits, there were 29 TP and 1 FP, when TP was defined to include TP of domains 1 and 2. While scanning the hits, it was noted that many were structures for DNA or RNA polymerases and other nucleic acid binding proteins. By the defining true positive as the same SCOP class and fold as the query, these would be classified as FP. However (see discussion below), previous analyses of these structures revealed a conserved core structure (Hansen *et al*, 1997, and references therein). Thus, these false positives can be considered true positives using different criteria. This example demonstrates how prior knowledge about both the query structure and hits can be a useful guide in distinguishing true positive from false positive. Also, it reveals a weakness in the SCOP classification and calls into question the utility of this database in distinguishing true positives from false positives.

Several general conclusions can be drawn from the first stage of searches with these relatively simple structures. First, DALI is capable of accurately aligning domains composed of more than a single polypeptide. While the example used here, 1ad2, had only two chains, there is no obvious reason why a domain composed of several chains couldn't be used. Thus, it should be possible to select the active site domain from an enzyme, and use this subdomain as a query against the PDB in a search for structural neighbors. Such as search was successfully attempted, as discussed below. Second, in all four cases above, the few hits that were returned for both individual domains and the entire chain had large Z-scores, and thus are readily accepted as true structural neighbors. For all other hits, however, visual inspection of pairwise superpositions is required to confirm the quality of the alignment. Third, in no case were all true positives returned as hits: no search reported all the structures classified by SCOP as the same class, fold, and superfamily as the parent structure. For 1daa:a, the results included one structure from each family. However, for the remaining three structures, no additional family members were returned. There are two possible explanations for this occurrence: 1) DALI reports only non-redundant hits; or 2) despite their SCOP classification, the other structures vary significantly. For example, two structures grouped into the same family by SCOP could vary in size. Thus, superimposing one structure upon the other would not yield a good alignment, despite the similarity in secondary structure arrangement. Finally, more stringent (and meaningful) definitions of true positive are needed to improve the analyses.

The second stage of this was to apply the above methods to a more complex protein structure, the RNA-dependent RNA polymerase of poliovirus (Figure 5). An initial search for structural neighbors of 1rdr (the poliovirus RNA-dependent RNA polymerase) using the CE algorithm (Shindyalov and Bourne, 2001) yielded few biologically relevant hits and many obvious false positives (41 hits; 18 TP, 23FP). Yet, there were also a number of intriguing hits – proteins with dissimilar biological functions that also utilized RNA or nucleotides as substrates. Such structural neighbors are of interest because the RNA binding domain of 1rdr has yet to be defined experimentally. It was predicted that these alignments reflected similarities

among the RNA binding domains of the different structure. To test this prediction, the superimposed structures for 1rdr and 1aud (U1 small nuclear ribonucleoprotein A, part of the U1 snRNP of the spliceosomal complex that contains a classic RNA-binding motif (Allain *et al*, 1997)) were viewed using Protein Explorer, a link provided from the CE website. It was clear in this alignment that the RNA-binding domain of U1A had been aligned with a portion of the poliovirus polymerase. Using the sequence alignment based on the pairwise structural alignment provided by CE through a link on the results page, the 1rdr sequence fragment corresponding to superimposed secondary structures was identified. A structural file for the putative RNA-binding domain fragment of 1rdr created (1rdr-RBD) and used as a query for structural superposition against the PDB using the DALI algorithm (Holm and Sander 1993). Fewer hits were returned for 1rdr-RBD than for the complete 1rdr structure: 30 vs. 41, respectively. Twenty of the hits were common between the two searches, including all RNA-dependent RNA polymerases and some false positives. (The number of false positives varies according to how stringently true positive is defined. See below.) And, 10 of the 1rdr-RBD hits were distinct from those for 1RDR. Only one of these 10 was a false positive, for a total of 25 TP and 5 FP (see ROC, Figure 5C). Importantly, one of the new hits was for the HIV-1 reverse transcriptase, an RNA-dependent RNA polymerase previously recognized as structurally homologous to the poliovirus 3D polymerase (Hansen *et al*, 1997).

It should be noted that the new hits returned for the putative RNA-binding domain of 1rdr are not necessarily new in the sense that they were not recognized as structurally similar in the query with 1RDR. Rather, it is likely that removing the superfluous portions of 1rdr increased the alignment score given to these new hits. The DALI server only returns hits with a Z-score greater than 2. Thus, when only a small portion of two structures align, the overall score for the alignment will be lower than the $Z<2.0$ threshold. By limiting your query to a single domain of interest, the overall score for this alignment would increase above the threshold. Thus, limiting the query enriches the results for true positives.

The definition of true positive used to analyze these results was based on the previous observation that the RNA recognition motif (RRM) bears significant structural similarity to the canonical RNA and DNA polymerase motifs (Kohlstaedt *et al*, 1992; Lindahl *et al*, 1994; Nagai *et al*, 1990; Goldgur *et al*, 1997). Thus, true positive was defined as all polymerases and other proteins bearing an RRM (splicing factors, ribosomal proteins, and tRNA binding proteins). This definition is broader than using the SCOP classification for 1rdr (class multi-domain proteins, fold DNA/RNA polymerases).

Further analysis of the putative 1rdr RNA binding domain (highlighted in Figure 5B) revealed that it contained the canonical RNA polymerase motifs A-E, depicted in Figure 5A. Comparison of sequences for the canonical polymerase motifs (Hansen *et al*, 1997) to the 1rdr-RBD sequence fragment revealed that the motifs formed a smaller domain, excluding some helices and loops found in 1rdr-RBD. To assess whether this smaller domain could further reduce the number of false positives, a second structural file was created, 1rdr-motifs, which included coordinates only for atoms within the canonical RNA polymerase motifs. A search with DALI returned 23 hits, with 19 TP and 4 FP. While this is one few false positive than the results for 1rdr-RBD, there are also 6 fewer true positives. By analysis of the ROCs for each of the 1rdr searches (Figure 5C), it can be seen that 1rdr-RBD and 1rdr-motifs yield hits from the DALI algorithm with equal efficiency and specificity – the rise of these two curves is quite similar. Both are clearly superior to the entire polymerase structure 1rdr is use a structural superposition

query. Given the significant loss of true positives with 1rdr-motifs, these results suggest a moderately limited structure gives the best results from the DALI structural alignment algorithm. It should be noted that the sequence segments composing 1rdr-motifs were highly conservative with respect to size – only those regions bearing high homology to at least three other polymerases were selected. It would be interesting to assess the cause for the loss of true positives by creating additional 1rdr motifs structures with motif segments of sizes ranging in size from 1rdr-RBD to the 1rdr-motifs used in this study.

Conclusions

The results reported here suggest a new structural superposition tool could be developed to assist in limited structural searches. As a first step, the query structure would be analyzed for presence of distinct structural domains and motifs. These parts would then be queried individually against a structure database (e.g. PDB) for structural homologs, using the DALI algorithm (Holm and Sander, 1993). The output data would include a summary list of the domains and motifs recognized, followed by lists of hits found for each part. An option the query using the entire structure should be included. Also included would be links to view structural alignments in sequence format (i.e. what portions of the hit and query amino acid sequences were aligned in the structural superposition).

Despite the favorable outcomes described above, the results obtained from these searches confirm the general inability of structural alignment algorithms to correctly identify all known true positives. More than a hundred structures are available for RNA and DNA polymerases. Yet even for the best limited query, only 24 true positives were reported – 11 polymerase structures and 13 nucleic acid binding proteins, whose structures have previously been identified as highly similar to the conserved polymerase structure (Lindahl *et al*, 1994; Nagai *et al*, 1990; Goldgur *et al*, 1997). One possible explanation for the low number of returned true positives in the searches described here is that DALI reports only non-redundant hits, including only one representative structure for each protein (i.e. only one of many structures for the Hepatitis C Virus RNA-dependent RNA polymerase was returned). Unfortunately, this rationalization is not sufficient. On the one hand, there was no statement in the DALI help files to indicate that searches were non-redundant. If they were, it would be advantageous to have the option of a redundant search, for example allowing for direct comparison of conformational re-arrangements observed under different crystallization conditions. On the other hand, numerous non-redundant true positives were not returned in this search for polymerase structural neighbors. While it would be comforting to presume that this weakness will eventually be ameliorated by tweaking and re-writing the algorithms, this is highly unlikely because the process of alignment is in opposition to the biological reality of protein structure. The ability to align two sequences or structures depends on their static nature. In contrast, proteins are highly flexible structures, whose function depends upon that flexibility. A structure file, whether derived from crystallography or NMR studies, gives a single three-dimensional representation of this structure. Thus, it is possible that a structural superposition algorithm could fail to align two structures of the same protein when those structures represent extremes in protein conformation. For instance, image a protein was accurately represented by a right hand. To the human eye, the hand is a hand whether fingers are extended or curled in a fist. To a structural superposition algorithm, however, these two

conformations represent distinct structures. This illustrates the limitations of utilizing structural similarity to imply functional similarity. However, the limitations go further. While a child's hand may be smaller than his mother's, they are both hands. Yet, even when held open in the same conformation a structural superposition algorithm would fail to recognize the similarity due to the difference in size – analogous points between the two would not be superimposable. This limitation seems the most obvious explanation for the low number of true positive hits for the poliovirus RNA-dependent RNA polymerase. For many years, it has been recognized that, despite the low level of sequence homolog, polymerases have highly conserved structures, especially within the palm subdomain that contains the active site for phosphodiester bond formation (Hansen *et al*, 1997 and references therein). However, the relative sizes of polymerase molecules differ significantly. Furthermore, the placement of the canonical motifs within the overall right-hand architecture can vary from directly between the fingers and thumb domains, as in the poliovirus polymerase, to shifted slightly towards the fingers domain, as in HIV-1 reverse transcriptase (Hansen *et al*, 1997). Although the algorithms lack the flexibility of the human mind, structural superposition should not be abandoned altogether. Rather, analysis of search results must be conducted with an eye of skepticism, keeping in mind that many true positives will be missing and than many of the intriguing hits will be mere false positives. The results from a structural alignment are a starting place for future thoughtful research, not an end in and of themselves.

References

Allain, F.H., Howe, P.W., Neuhaus, D., and Varani, G. (1997) Structural basis of the RNA-binding specificity of human U1A protein. *EMBO J.* **16**, 5764-72.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyaloy, I.N., and Bourne, P.E. (2000) The Protein Databank. *Nucleic Acids Research* **28**, 235-242.

Brutlag, D.L. (1999) Protein folds and protein structure superposition. Lecture for Computational Molecular Biology, Stanford University School of Medicine, November 24, 1999.

Carpenter, E.P., Hawkins, A.R., Frost, J.W., and Brown, K.A. (1998) Structure of dehydroquinate synthase reveals an active site capable of multistep catalysis. *Nature* **394**, 299-302.

Goldgur, Y *et al* and Safro, M. (1997) The crystal structure of phenylalanyl-tRNA synthetase from *Thermus thermophilus* complexed with cognate $tRNA^{Phe}$. *Structure*, **5**, 59-68.

Hansen, J.L., Long, A.M., and Schultz, S.C. (1997) Structure of the RNA-dependent RNA polymerase of poliovirus. *Structure* **5**, 1109-1122.

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

Kaplan, W., and Littlejohn, T.G. (2001) Swiss-PDB Viewer (Deep View). *Briefings in Bioinformatics* **2**, 195-197.

Kohlstaedt, L.A., Wang, J., Friedman, J.M., Rice, P.A., and Steitz, T.A. (1992) Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science* **256**, 1783-1790.

Lindahl, M. *et al*, and Amons, R. (1994). Crystal structure of the ribosomal protein S6 from *Thermus thermophilus. EMBO J.* **13,** 1249-1254.

Lyle, J.M. (2002) Kindly provided POV-RAY image of PV 3Dpolymerase and assistance in using POV-RAY.

Mount, D.W. (2001) Chapter 9: Protein Classification and Structure Prediction. In *BioInformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press. Pp. 381-478.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.

Nagai, K., Oubridge, C., Jessen, T.H., Li, J. and Evans, P.R. (1990) Crystal structure of the RNA-binding domain of the U1 small nuclear ribonucleoprotein A. *Nature*, **348**, 515-520.

Shindyalov, I.N., and Bourne, P.E. (2001) A database and tools for 3D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. *Nucleic Acids Research* **29**, 228-229.

Singh, A.P. and Brutlag, D.L. (unknown) Protein structure alignment: a comparison of methods.

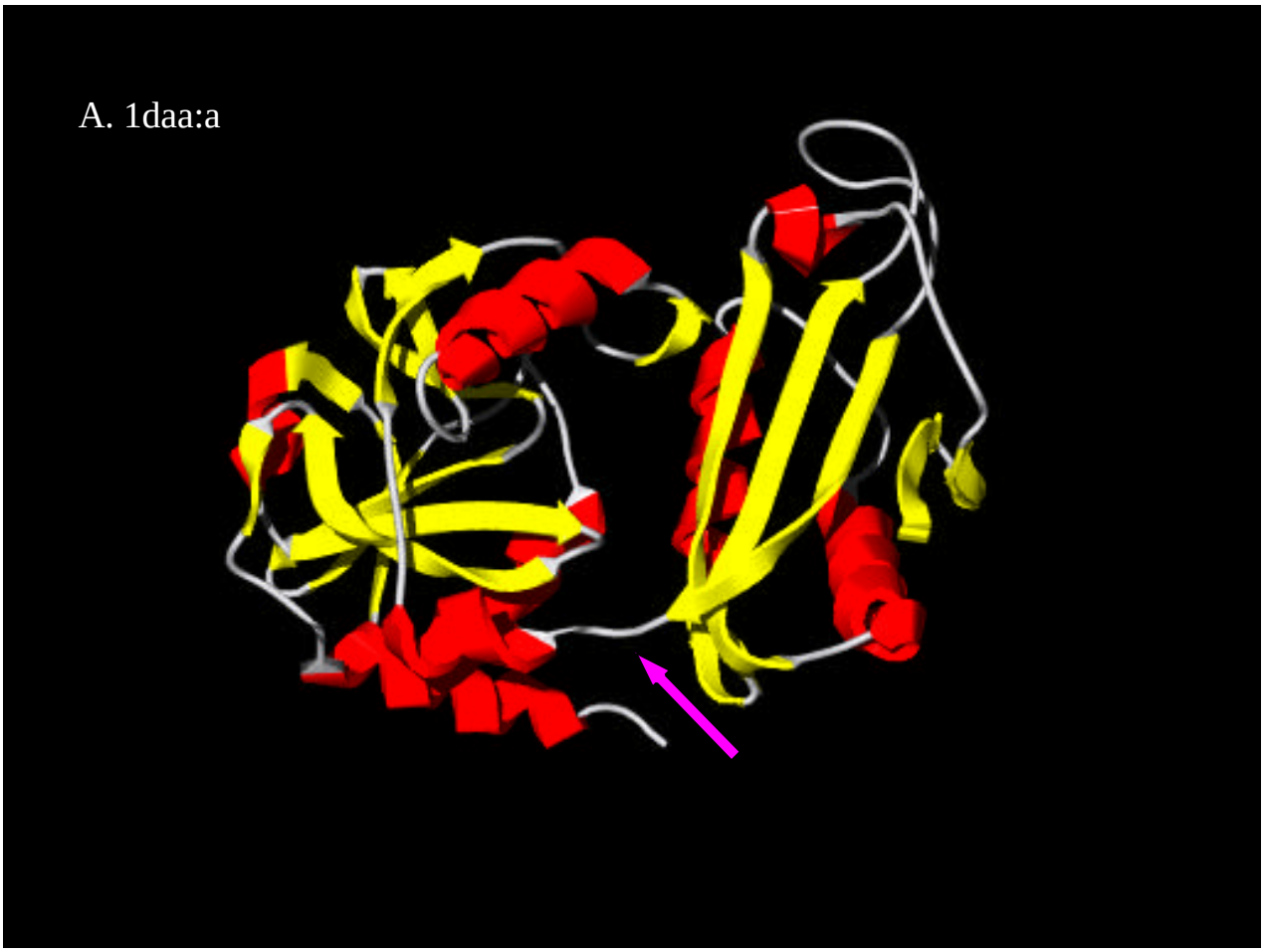
Singh, A.P., and Brutlag, D.L. (1997) Hierarchical Protein Structure Superposition using both Secondary Structure and Atomic Representations. *Proc. Intelligent Systems for Molecular Biology 97.*

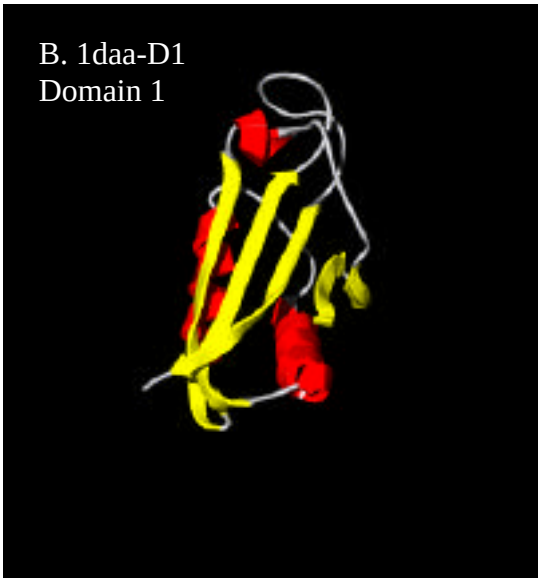
Sugio, S., Petsko, G.A., Manning, J.M., Soda, K., and Ringe, D. (1995) Crystal structure of a D-amino acid aminotransferase: how the protein controls steroselectivity. *Biochemistry* **34**, 9661-9669.

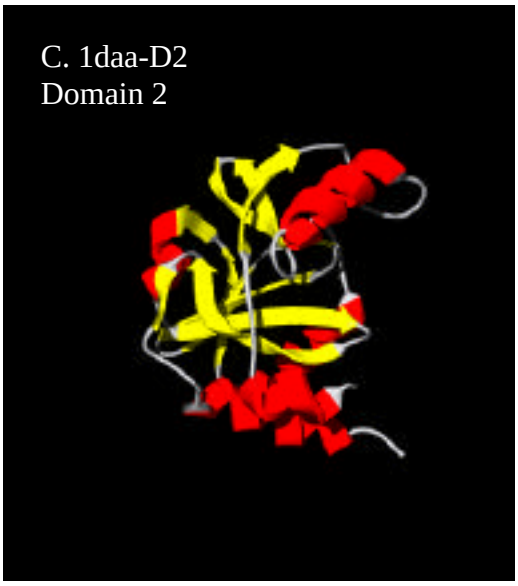
Unge, J., Al-Karadaghi, S., Liljas, A., Konsson, B.H., Eliseikina, I., Ossina, N., Nevskaya, N., Fomenkova, N., Garber, M., and Nikonov, S. (1997) A mutant form of the ribosomal protein L1 reveals conformational flexibility. *FEBS Letters* **411**, 53-9.

Zhu, X., Zhao, X., Burkholder, W.F., Gragerov, A., Ogata, C.M., Gottesman, M.E., and Hendrickson, W.A. (1996) Structural analysis of substrate binding by the molecular chaperone DnaK. *Science* **272**, 1606-1614.

Figure 1: Images of 1daa, D-amino acid aminotransferase. A.1daa:a (chain A) B. 1daa-D1 (domain 1) C. 1daa-D2 (domain 2). Cleavage site used to derive domains 1 and 2 is indicated by arrow in 1daa:a. Modeling was performed using Swiss-PDB Viewer (Kaplan and Littlejohn, 2001) and POV-RAY (POV-Ray, 1999) D. Modified Receiver-Operator-Curves for results from DALI search against the Protein Data Bank.




B. 1daa-D1
Domain 1


C. 1daa-D2
Domain 2

D.

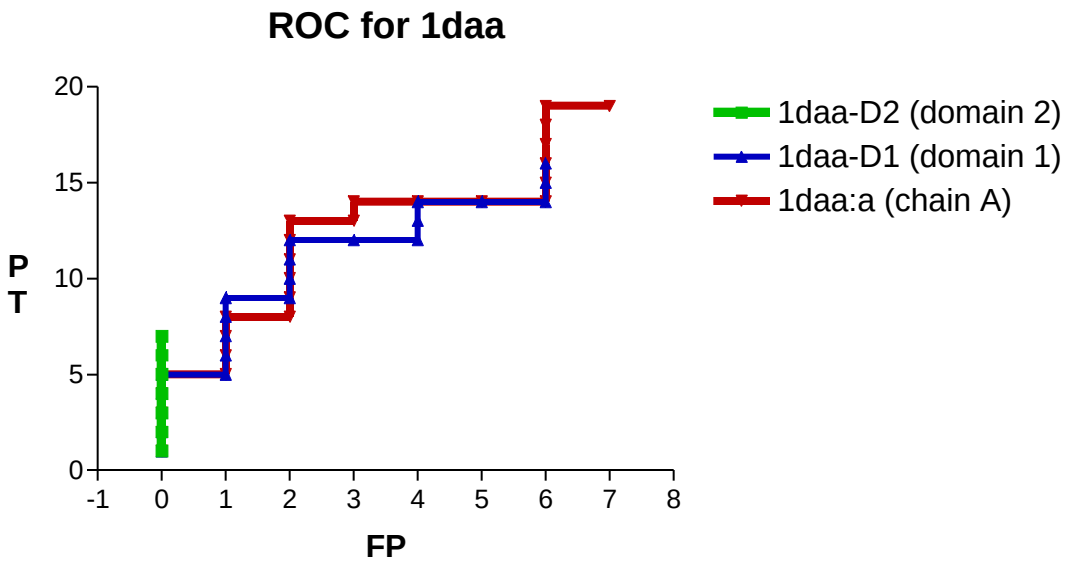
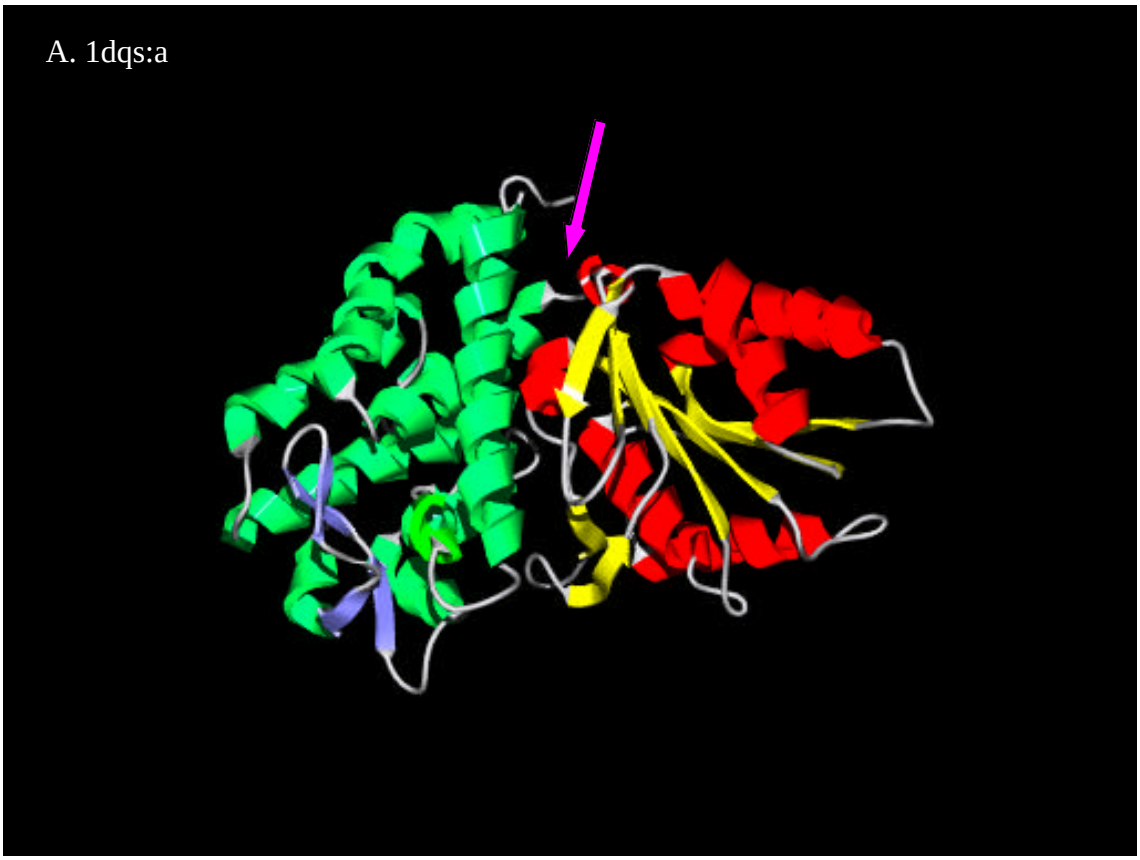
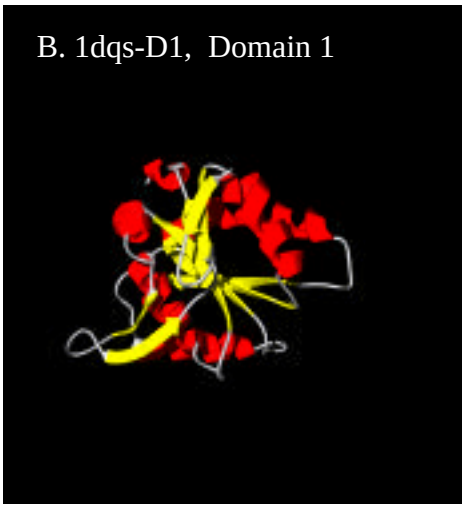
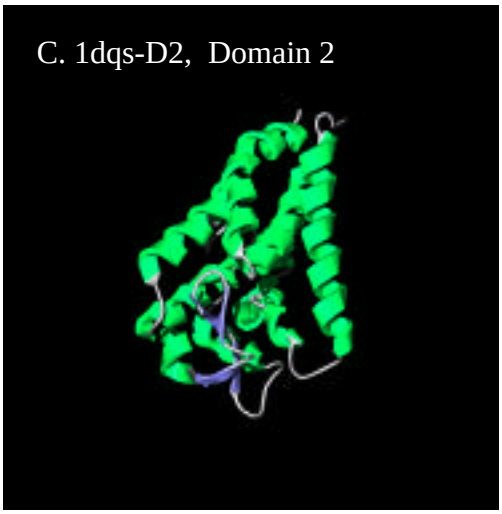

ROC for 1daa
20
15
10
5
0
P
T
-1
0
1
2
3
4
5
6
7
8
FP
1daa-D2 (domain 2)
1daa-D1 (domain 1)
1daa:a (chain A)

Figure 2: Images of 1dqs, dehydroquinate synthase. The two domains are indicated by differential coloring of secondary structures. A. 1dqs:a (chain A) B. 1dqs-D1 (domain 1) C. 1dqs-D2 (domain 2). Cleavage site used to derive domains 1 and 2 is indicated by arrow in 1dqs:a. Modeling was performed using Swiss-PDB Viewer (Kaplan and Littlejohn, 2001) and POV-RAY (POV-Ray, 1999) D. Modified Receiver-Operator-Curves for results from DALI search against the Protein Data Bank.







D.

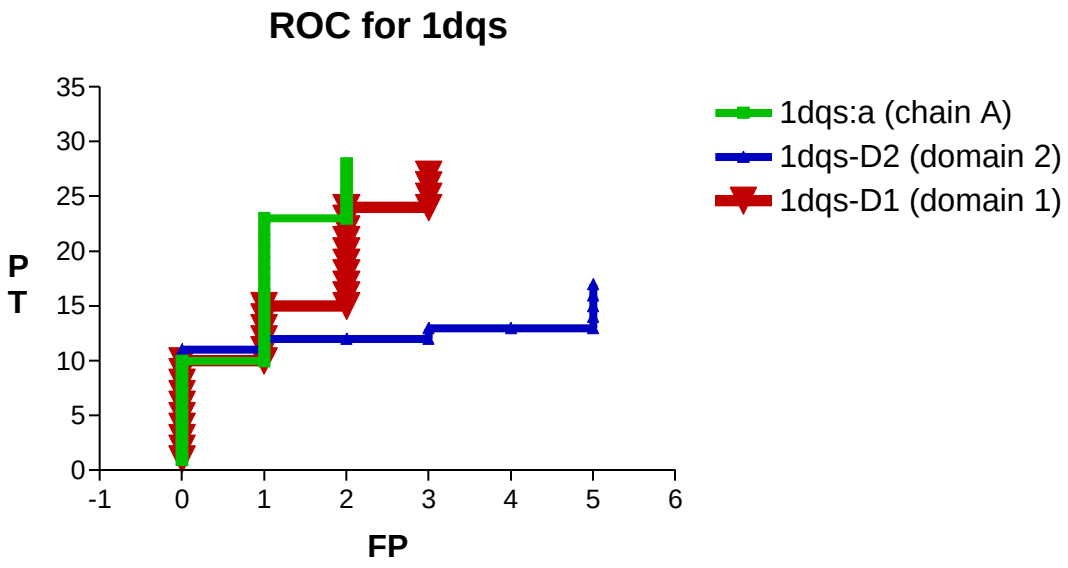
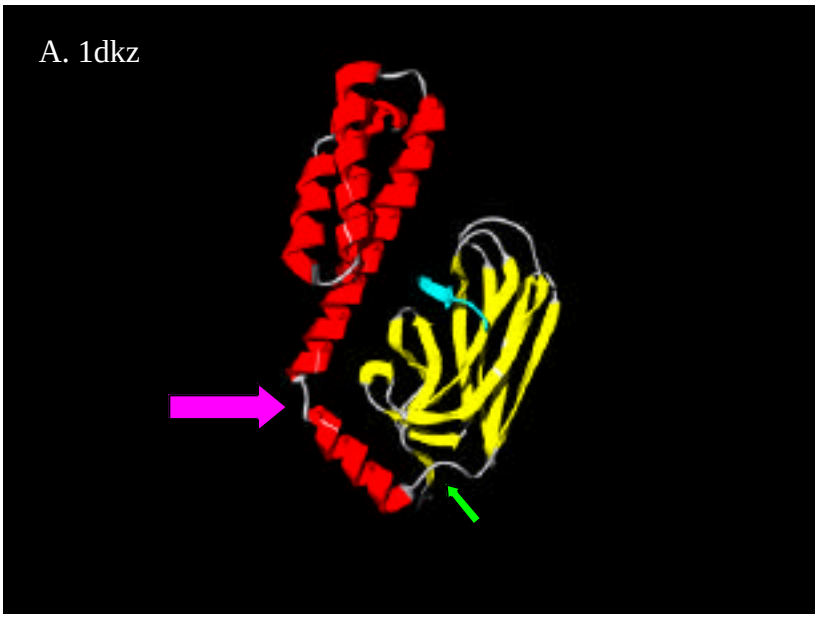
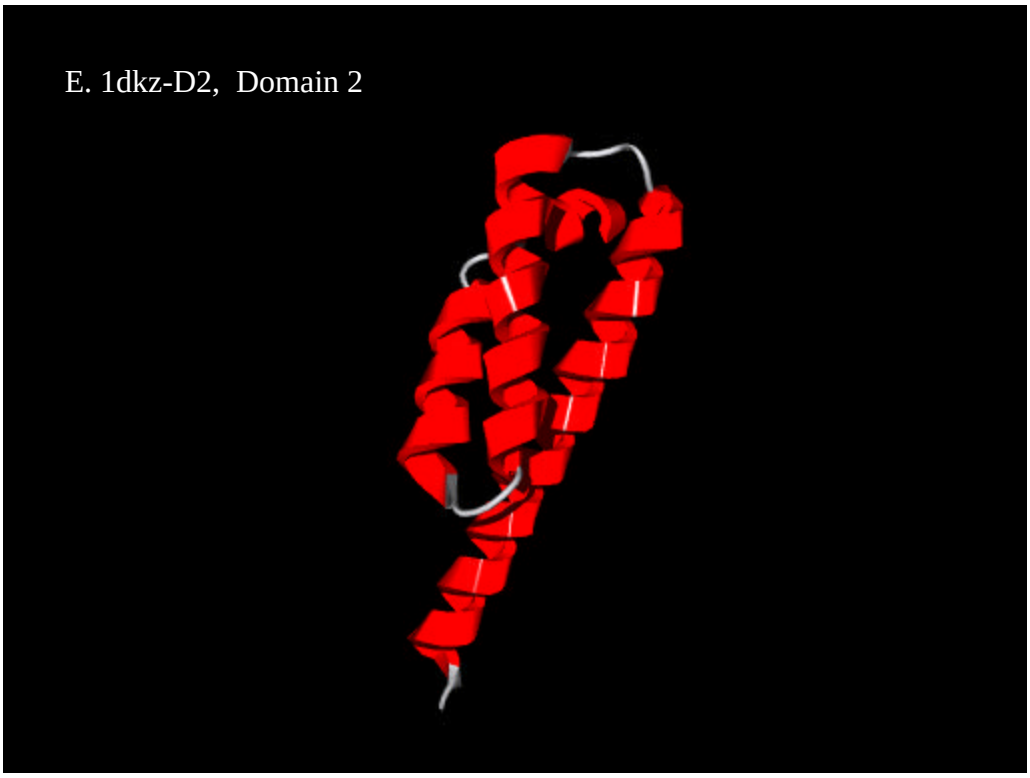

ROC for 1dqs
PT
FP
35
30
25
20
15
10
5
0
-1
0
1
2
3
4
5
6
1dqs:a (chain A)
1dqs-D2 (domain 2)
1dqs-D1 (domain 1)

Figure 3: Images of 1dkz, DnaK. A. 1dkz (chain A with bound peptide substrate) B. 1dkz:a (chain A without substrate) C. 1dkz-D1 (domain 1) D. 1dkz-D1_beta (domain 1Beta) E. 1dkz-D2 (domain 2). Cleavage site used to derive domains 1 and 2 is indicated by the large pink arrow in 1dkz. Cleavage site used to derive domain 1Beta is indicated by the small green arrow in 1dkz. Modeling was performed using Swiss-PDB Viewer (Kaplan and Littlejohn, 2001) and POV-RAY (POV-Ray, 1999) F. Modified Receiver-Operator-Curves for results from DALI search against the Protein Data Bank.



C. 1dkz-D1, Domain 1

D. 1dkz-D1B, Domain 1 beta


E. 1dkz-D2, Domain 2

F.

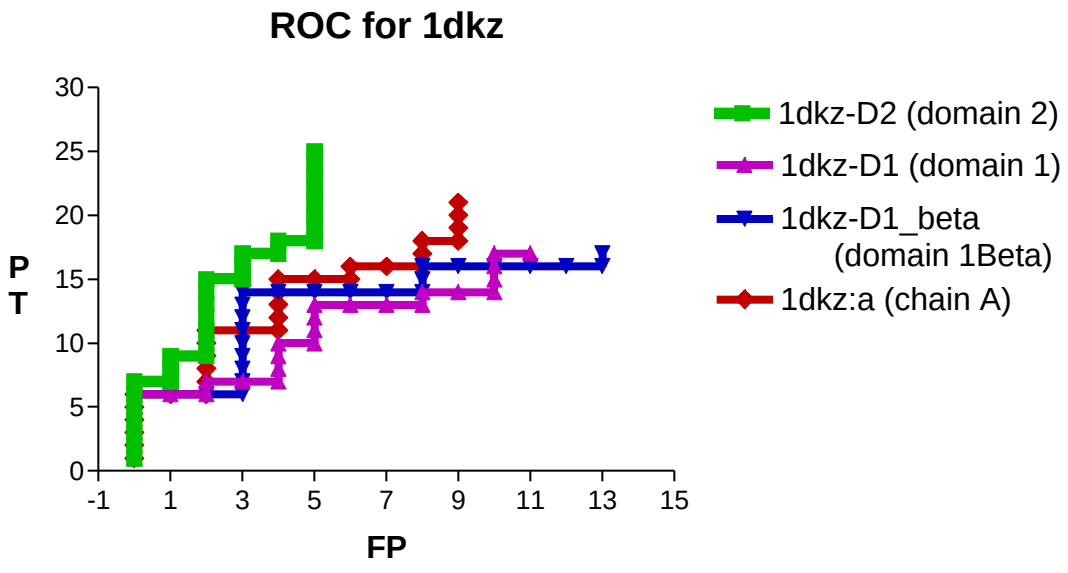
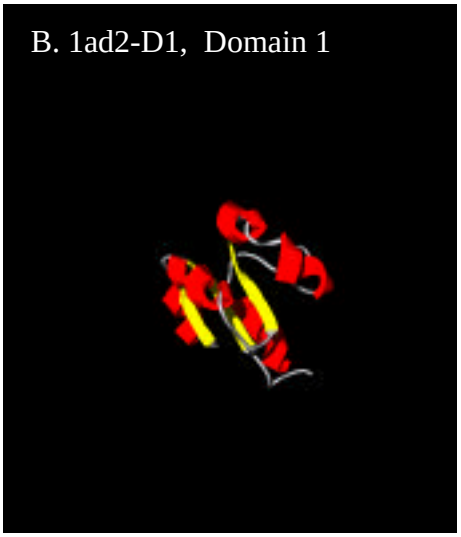

ROC for 1dkz
PT
FP
30
25
20
15
10
5
0
-1
1
3
5
7
9
11
13
15
1dkz-D2 (domain 2)
1dkz-D1 (domain 1)
1dkz-D1_beta
(domain 1Beta)
1dkz:a (chain A)

Figure 4: Images of 1ad2, mutant ribosomal protein L1. A.1ad2. B. 1ad2-D1 (domain 1). C. 1ad2-D2 (domain 2). Cleavage sites used to derive domains 1 and 2 is indicated by the arrows in 1ad2. Modeling was performed using Swiss-PDB Viewer (Kaplan and Littlejohn, 2001) and POV-RAY (POV-Ray, 1999) D. Modified Receiver-Operator-Curves for results from DALI search against the Protein Data Bank.



D.

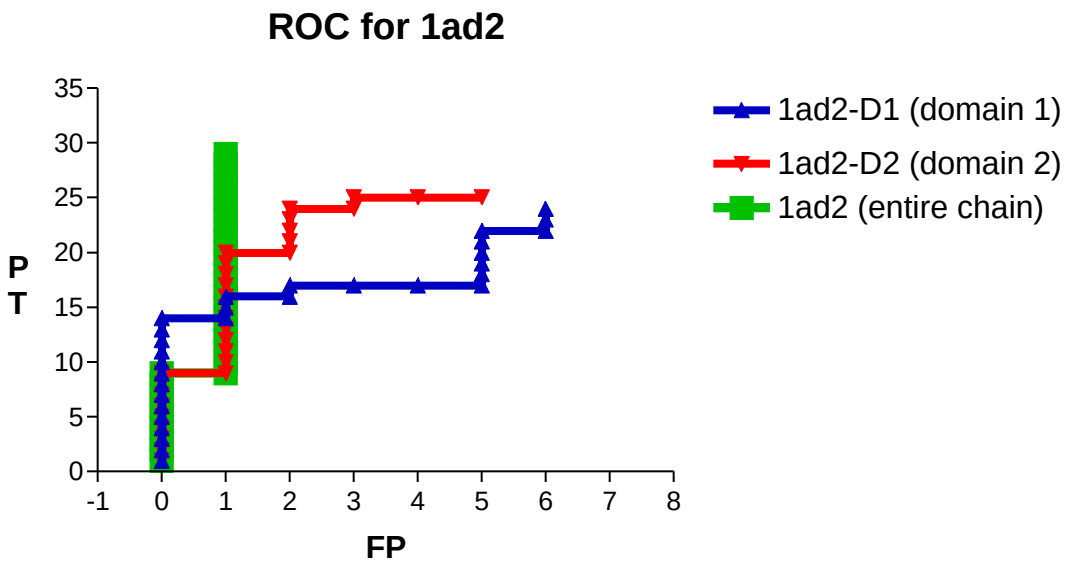
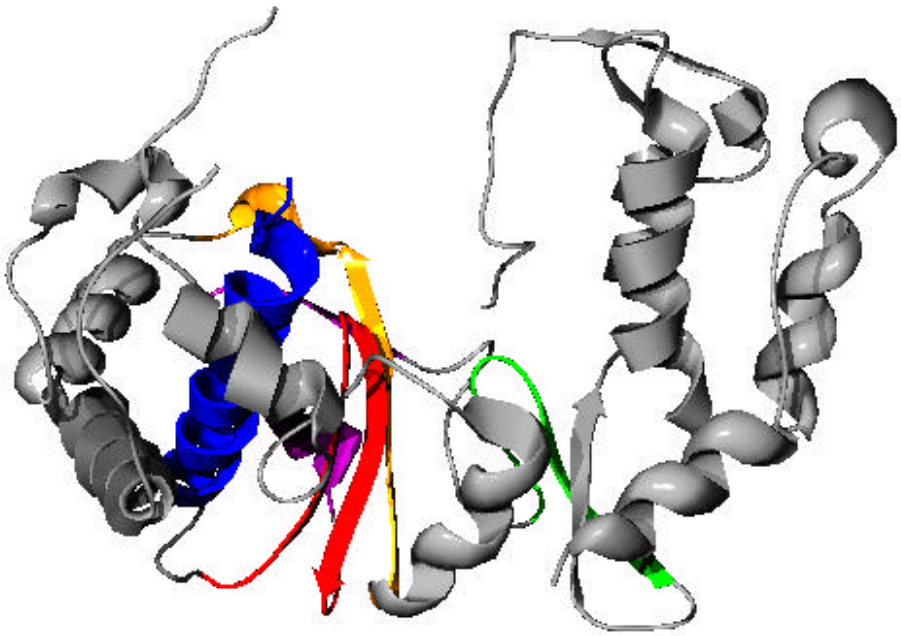

ROC for 1ad2
35
30
25
20
15
10
5
0
P
T
-1
0
1
2
3
4
5
6
7
8
FP
1ad2-D1 (domain 1)
1ad2-D2 (domain 2)
1ad2 (entire chain)

Figure 5: Images of 1rdr, the poliovirus RNA-dependent RNA polymerase. A. 1rdr with canonical RNA polymerase motifs A-E colored in orange, blue, red, purple, and green, respectively (Lyle, 2002). B. 1rdr with putative RNA binding domain highlighted. Modeling was performed using Swiss-PDB Viewer (Kaplan and Littlejohn, 2001) and POV-RAY (POV-Ray, 1999) C. Modified Receiver-Operator-Curves for results from DALI search against the Protein Data Bank.

A.

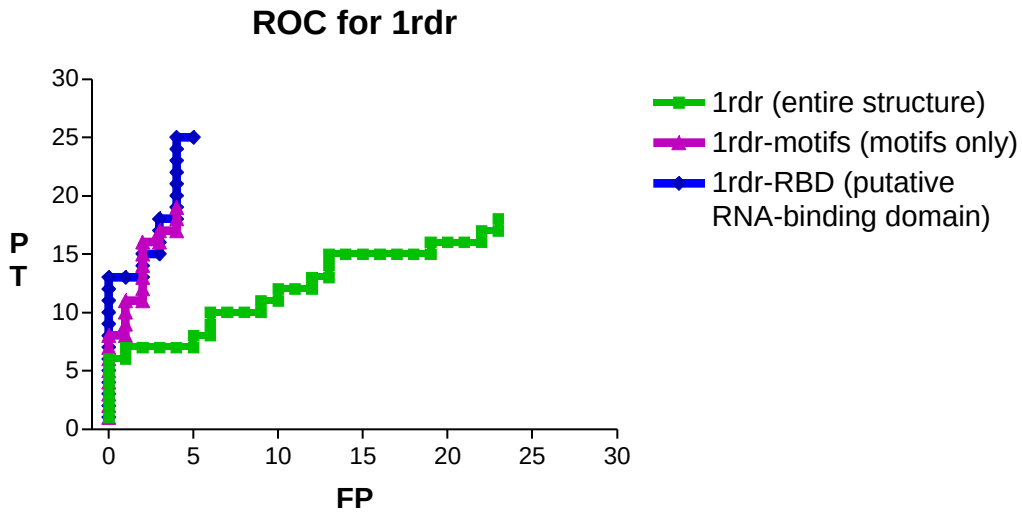
1rdr



B.


1rdr

C.


ROC for 1rdr
PT
FP
0
5
10
15
20
25
30
1rdr (entire structure)
1rdr-motifs (motifs only)
1rdr-RBD (putative RNA-binding domain)