Dali Tutorial

1 Introduction

Dali is a protein structure comparison server. The server has been running continuously for over 20 years. The server operated first in Heidelberg (Germany), then Hinxton (UK), now Helsinki (Finland). Dali is based on distance matrix comparison (see References for methods). In favourable cases, structure comparison can reveal distant evolutionary relationships not seen by sequence comparison.

The server (http://ekhidna2.biocenter.helsinki.fi/dali/) supports three types of requests:

1. PDB search
2. Pairwise comparison
3. All against all comparison

The server takes the 3D coordinates of protein structures as input and returns a list of similar structures, structural alignments and superimposed structures. The all against all comparison also returns a structural dendrogram and a projection from protein structure space. The results are linked to sequence search and function prediction servers.

This tutorial explains the web interface of the Dali server using live examples.

2 Inputs

The PDB format is based on records with keywords. A sample PDB structure is given in Appendix A. Only ATOM records are required by Dali. The full specification of the format can be found at http://www.wwpdb.org/docs.html.

The following restrictions apply:

- The structure must contain the coordinates of the backbone atoms: N, CA, C and O. If your structure has only the C-alpha coordinates, you can generate a complete backbone using the MaxSprout server at http://www.ebi.ac.uk/maxsprout.
- The structure must contain at least 30 residues. Shorter chains are ignored by Dali.

Publicly available repositories of protein structures are RCSB, PDBe, and PDBj.

PDB entries have a PDB identifier, which is four characters long and consists of a digit followed by three letters or digits, for example, 3ubp. You can find the PDB entries matching a keyword search at RCSB, http://www.rcsb.org/. Each entry can contain one or more chains. The chain identifier is one character. For example, PDB entry 3ubp has three chains A, B and C. In the submission forms, the chain identifier must be concatenated with the PDB identifier, for example, 3ubpC. The PDB search submission form gives hints on possible continuations when you start typing the PDB identifier.

The Dali server does not accept an amino acid sequence as input. If you know only the amino acid sequence of your protein, you can search for a related PDB structure using sequence comparison with servers like SANSparallel (http://ekhidna2.biocenter.helsinki.fi/sans/, search against PDB database). Comparative modeling servers like SwissModel generate a model which only replaces the side chains (according to a sequence alignment) while the backbone stays very close to the template structure. More adventurous servers may generate a model ab initio when the query sequence has no obvious homolog of known structure. For example, PHYRE, I-TASSER and ROBETTA have been some of the top performers in CASP (Critical Assessment of Structure Prediction).

2.1 Submission

The submission forms (Figure 1) for PDB search, pairwise comparison and all against all comparison accept one, two to 11, and three to 64 input structures, respectively. In pairwise and all against comparison, you must click on the +/- buttons to create the required number of input fields. All against all comparison has two alternative submission forms, one for input sets with uploaded structures and an alternative one for input sets composed only of PDB identifiers.

PDB searches and all against all comparisons are time consuming, upwards of 10-15 minutes. A queueing system is in use, so you have to wait even longer if there are many simultaneous requests. If you left your email address, you will receive an email notification when the results are ready. Otherwise you must stay on the result page or bookmark it so that you can return to it later.

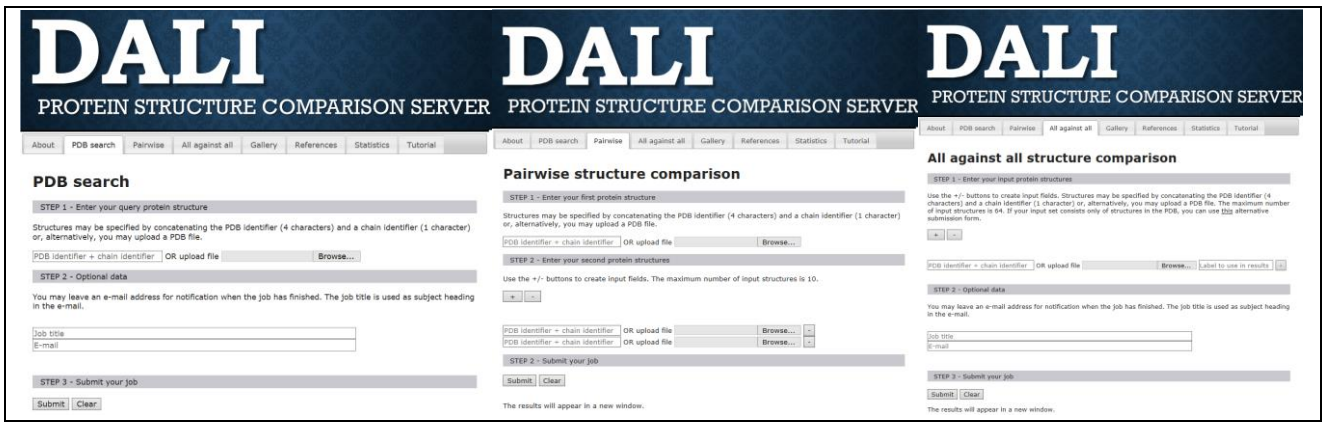


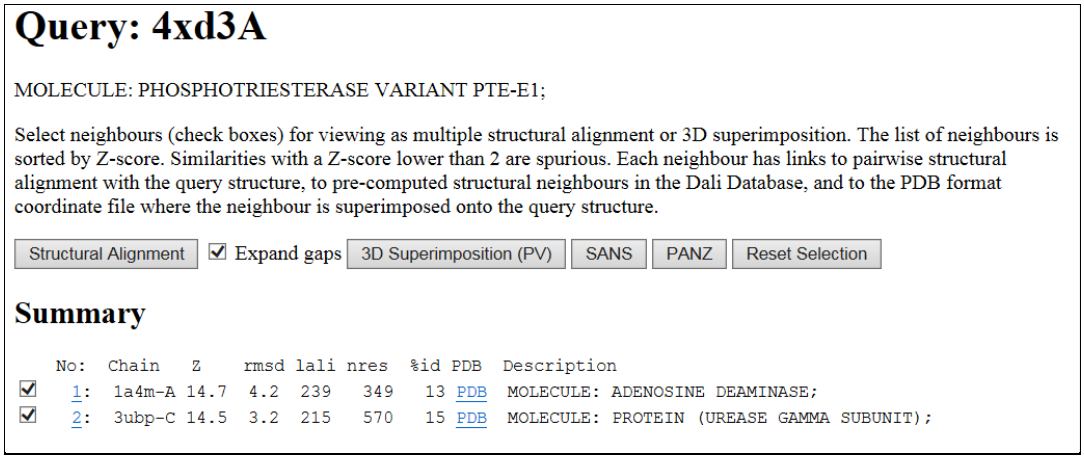
Figure 1. Submission forms.

3 Outputs

In this tutorial, we use the amidohydrolase superfamily as example. The amidohydrolase superfamily was first discovered based on structural similarities between urease, adenosine deaminase and phosphotriesterase. Let's see how they are related structurally.

1. Go to the submission form for pairwise comparison (Figure 1 middle; from the main page click on the "Pairwise" tab).
2. If you do not know the PDB identifiers of the structures you are interested in, make a detour to RCSB's keyword search. For example, from PDB Text we find a phosphotriesterase entry 4xd3 with chains A and G.
3. Type 4xd3A in the box for first protein structure.
4. Press the plus button twice to create two input fields for second structures.
5. Type 3ubpC and 1a4mA in the boxes for second protein structures.
6. Press submit.
7. Wait for the result.

All request types produce match lists in the same format (Figure 2). The matches are sorted by Dali Z-score. The number column has hyperlinks to the pairwise structural alignment between the query and match structure. Each matched structure is also hyperlinked to the PDB entry. The coordinates of the PDB entry are superimposed on the query structure by rigid-body rotation and translation. The checkboxes are used to select a subset of matches for interactive visualization. The buttons above the match list launch structural alignments in 1D ("Structural Alignment") or 3D ("3D Superimposition (PV)"), and provide links to sequence analysis tools ("SANS" sequence database search, "PANZ" function prediction). The following sections demonstrate the interactive visualization options. Use your result from the above exercise or this live example.



Query: 4xd3A

MOLECULE: PHOSPHOTRIESTERASE VARIANT PTE-E1;

Select neighbours (check boxes) for viewing as multiple structural alignment or 3D superimposition. The list of neighbours is sorted by Z-score. Similarities with a Z-score lower than 2 are spurious. Each neighbour has links to pairwise structural alignment with the query structure, to pre-computed structural neighbours in the Dali Database, and to the PDB format coordinate file where the neighbour is superimposed onto the query structure.

Structural Alignment | ☑ Expand gaps | 3D Superimposition (PV) | SANS | PANZ | Reset Selection

Summary

	No:	Chain	Z	rmsd	lali	nres	%id	PDB	Description
☑	1:	1a4m-A	14.7	4.2	239	349	13	PDB	MOLECULE: ADENOSINE DEAMINASE;
☑	2:	3ubp-C	14.5	3.2	215	570	15	PDB	MOLECULE: PROTEIN (UREASE GAMMA SUBUNIT);

Figure 2. Result from pairwise comparison.

3.1 Structural alignment view

The multiple alignment view opens in a new window and displays the alignment of the query structure and the selected matches. The upper block shows the amino acid sequences and the lower block the secondary structure states (H: helix, E: sheet, L: coil). The most frequent symbol in each column is coloured. The alignment view has an option 'expand gaps'. If the option is checked, the complete sequence of all proteins is shown. Residues without a match in the query structure are shown in lowercase. If the option is not checked, all matching sequences are shown stacked on the query sequence, and insertions relative to the query sequence are hidden (Figure 3). We are comparing distantly related proteins but they have a striking signature of invariant amino acids, including the histidines at position 20 and 22 in Figure 3.

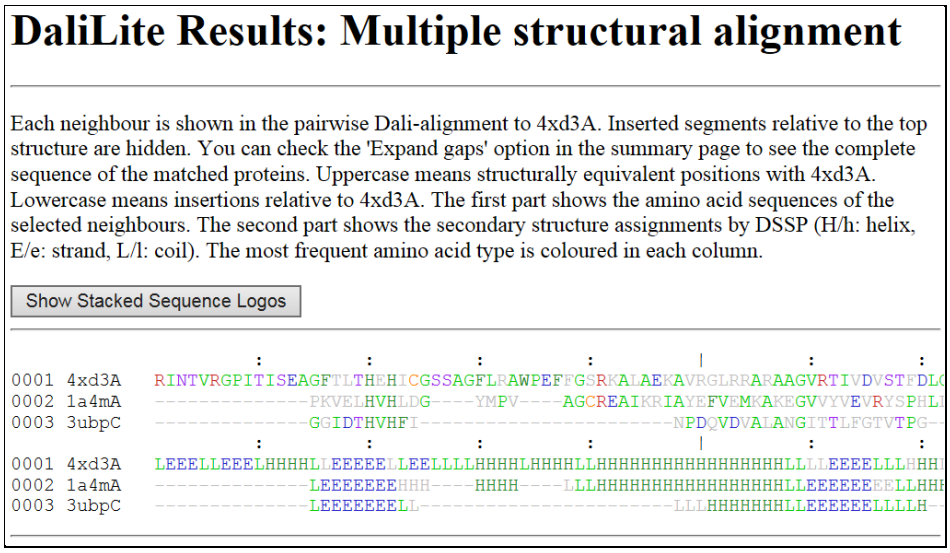


Figure 3. Structural Alignment view (Expand gaps option off).

3.1.1 Stacked sequence logos

The structural alignment view only considers the amino acid sequences of the selected protein structures. The button labelled "Show Stacked Sequence Logos" generates a sequence profile around each structurally known protein by searching for homologs in Uniprot. The sequence profiles are then showed stacked against the query protein (Figure 4, live example). The sequence profiles require optimizing an HMMer model for each structure and this is a slow process, taking about 8 seconds per logo. When you are displaying many logos, zoom out with CTRL-minus so that more rows fit on the screen. Zoom back in with CTRL-plus. The stacked sequence logos show which positions are conserved across the whole set rather than matching by chance in the structural alignment of a small number of structures. The sharp sequence signature motifs often pinpoint the active site of enzyme superfamilies (Figure 5).

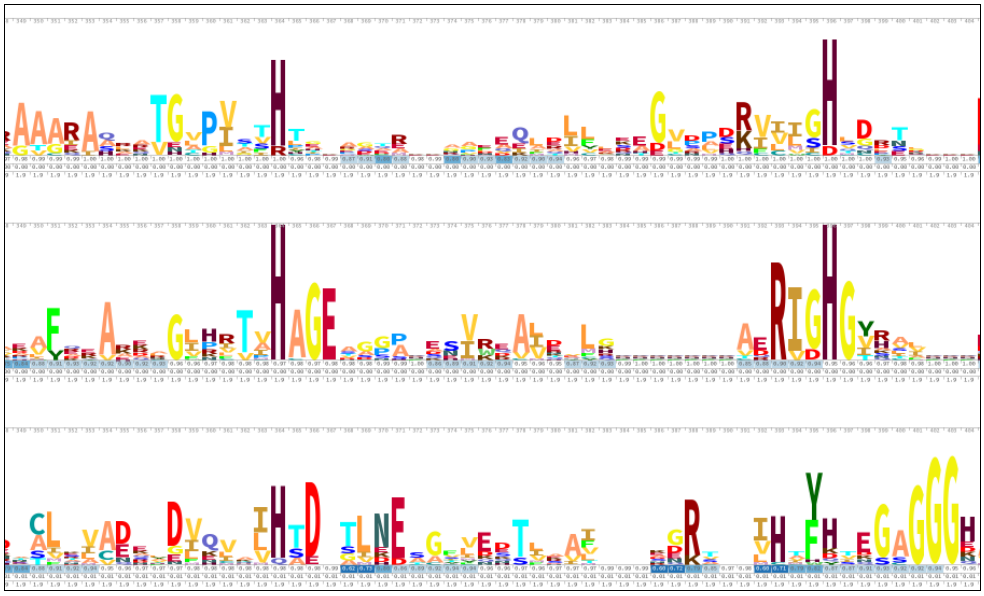


Figure 4. Showing a section of Stacked Sequence Logos view (Expand gaps option on).

3.2 3D superimposition view

The 3D superimposition view opens in a new window and displays the selected structures superimposed on the query structure (Figure 5). This is a rigid body transformation, which may look ugly if the RMSD is high. (Dali does not optimize RMSD, it matches contacts). The web page has simple toggles to select display

styles. In particular, sequence and structure conservation can be mapped to the query structure and shown by colour. The colour map for conservation mapping goes from blue for the highest values through green to red for the lowest values. Sequence conservation is calculated as the relative entropy of a column, SUM p(i) log(p(i)/q(i)), where the sum is over twenty amino acid types I and p(i)=n(i)/N where n(i) is the number of occurrences and N is the number of rows in the alignment, and q(i) are the frequencies of amino acid types in the sequence database. The logarithm is taken in base 2 so the unit of relative entropy is bits. Structure conservation is simply the fraction of selected structures that are structurally aligned to the query structure.

We use PV as structure viewer. PV is a Javascript based viewer which works on modern browsers. PV works as advertised with Chrome and Firefox. With Internet Explorer there are some quirks with asynchronous refreshing of the image, which tends to disappear altogether after the user clicks options but can be restored by moving the cursor or clicking the show/hide options repeatedly.

1. Check 3ubpC and 1a4mA in your summary page and press the "3D Superimposition (PV)" button. Alternatively, you may use this live example.
2. Scale the window to full screen size. This places the viewer area (with light blue passepartout) and option checkboxes side by side.
3. You should see a spaghetti of multiple C-alpha traces and side chains. Click on the radio button labelled "Cartoon". Uncheck "All" and check "Query" in the Show/hide structures options and you should see a green cartoon representation of the query protein.
4. Click on "Structure conservation" for Query colour. Dark blue regions are structurally aligned in all three structures. Hold down the left mouse button and move the cursor in the viewer area to rotate the structure. Hold down the middle button and move the cursor up/down to zoom in/out.
5. Switch to "Sequence conservation" for Query colour and check "Query" in Show/hide side chains options.
6. Clear up the messy picture by removing less conserved side chains. Move sliding ruler under "Query side chains > 0 bits" to the left until the value is 4.15 bits. Now conserved residues at the active site are highlighted. Click on the side chains to see their labels with residue numbers.
7. Check the Show/hide ligands option. Click atoms to see their labels.

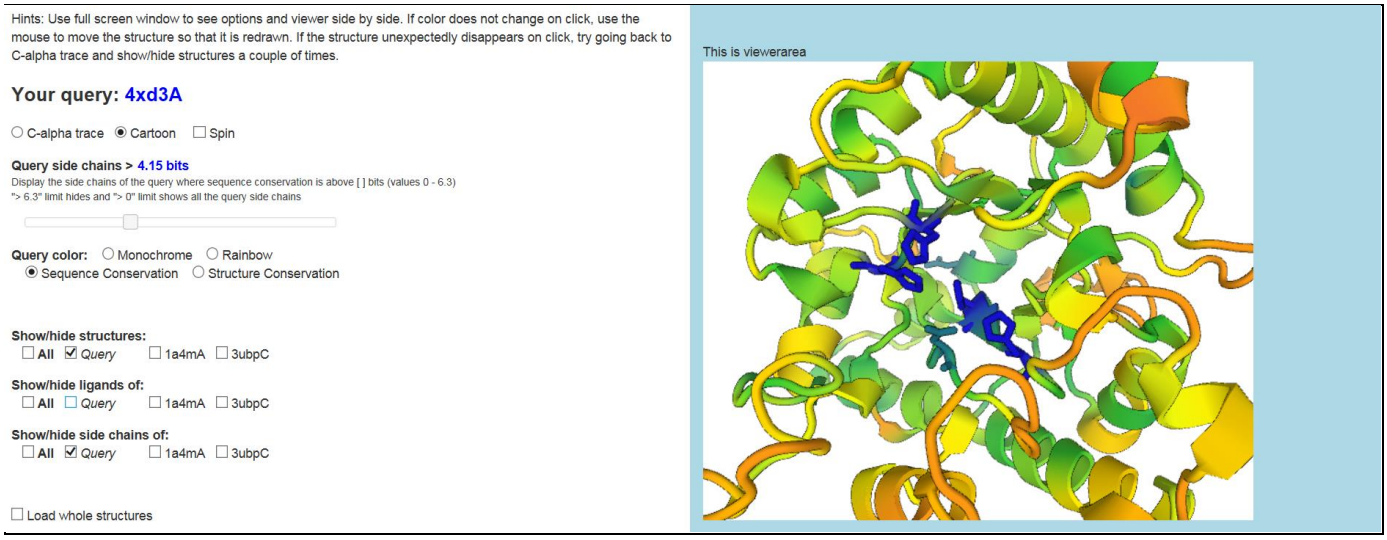


Figure 5. Screenshot of protein viewer (PV).

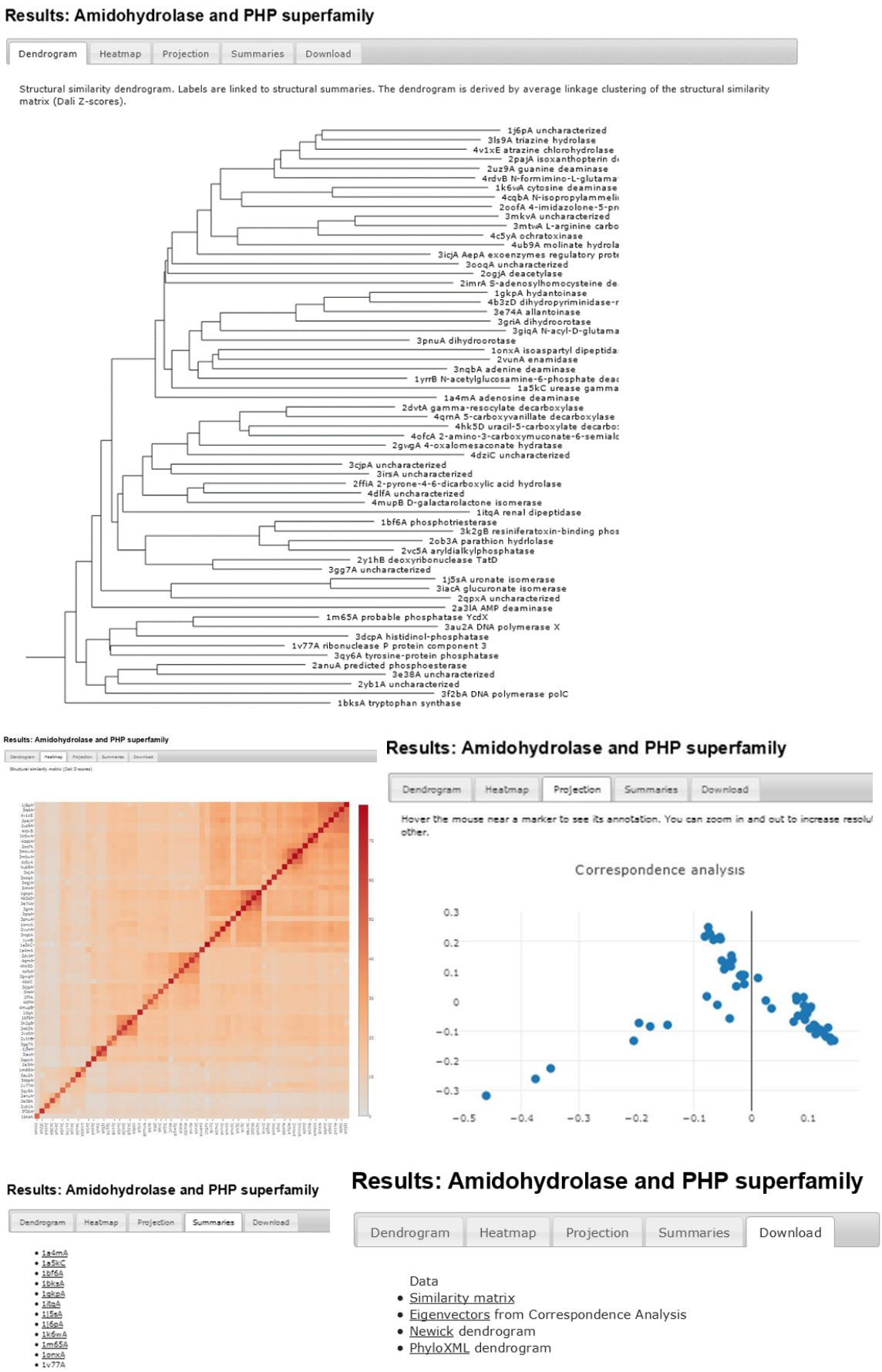


Figure 6. The results of all against all comparison are divided under five tabs.

3.3 Integrated sequence search tools

Sometimes there are uncharacterized proteins in the summary list. From the interactive summary (Figure 2) you can send the amino acid sequences of selected subsets to search Uniprot by SANSparallel (SANS button) or predict function by PANNZER2 (PANZ button). These tools do not use structure information in any way and they are provided for convenience. The mapping of structures to Uniprot brings the great advantage of crosslinks to literature and protein family classification resources (e.g. PFAM).

3.4 Visualization of protein structure space

All against all comparison generates an overview of protein structure space for a set of input structures. The Example section below will take you through the discovery process for structural classification with Dali. At this point, we just show the outputs. The output of all against all comparison has a different layout from PDB search and pairwise comparison. In addition to the summary lists, there are tabs for plots generated from the all against all similarity matrix (Figure 6 and live example).

The dendrogram is clickable. The leaves are linked to the summary list of that structure, which shows the structural alignments of the other structures of the input set aligned against it. The similarity matrix and scatterplot are interactive, responding to hovering the mouse over a data point. A toolbar appears at the upper right corner of the scatterplot. Click on the single arrow to see the label attached to the nearest data point. Structural alignment summaries are linked both to the leaves of the dendrogram and the list under the Summaries tab. The similarity matrix, eigenvector analysis results and pseudo-phylogenetic trees in Newick and PhyloXML format are available under the Download tab. Appendix B shows an example where the eigenvectors from correspondence analysis were downloaded and grouped according to branches in the dendrogram for plotting in Excel.

4 Example

In this example, we revisit the amidohydrolase superfamily 20 years after it was first discovered. We do PDB searches to find a representative set of current members. We shed light on the relationship between amidohydrolases and the PHP superfamily, which has intriguing structural and sequence similarities to amidohydrolases. We do all against all comparison to get an overview of the position of these two superfamilies in protein structure space. The final result can be seen in this live example.

4.1 Selecting the input set

We already know about three structures in the amidohydrolase superfamily from section 3. Using these structures as queries in PDB searches, we can collect more members from the current PDB. To decide whether a protein is a member or not, we look at characteristic features of the superfamily such as: are the nearest structural neighbours all amidohydrolases? is the structural core conserved? is the sequence signature conserved? Dali's search algorithm uses heuristics and is not guaranteed to deliver the optimal alignment. Therefore we performed a number of searches from diverse seeds and combined the results.

Figure 7 shows the result from one PDB search. All the structures listed in Figure 7 are legitimate members of the amidohydrolase superfamily, despite having diverse molecular functions. PDB search results are reported for PDB90 and PDB100. PDB100 contains all structures in the PDB. PDB90 is a non-redundant subset of PDB structures with less than 90 % sequence identity to each other. PDB and PDB90 are updated weekly. PDB90 representatives are not necessarily stable from week to week.

Merging the structural neighbour lists of a few seed structures and removal of redundant structures (with more than 30-40 % sequence identity) resulted in the set shown in Figure 8. The set includes members of the amidohydrolase superfamily and of the PHP superfamily. The PHP superfamily has structural similarities to amidohydrolases and partially overlapping sequence motifs. Tryptophan synthase was added as an outgroup which is not thought to be related to either of the aforementioned superfamilies.

```
# Query: 4xd3A
# No:  Chain   Z    rmsd lali nres  %id PDB  Description
   1:  1qw7-A 59.4  0.5  327   336   96   MOLECULE: PARATHION HYDROLASE;
   2:  3a3w-A 58.3  0.7  326   329   87   MOLECULE: PHOSPHOTRIESTERASE;
   3:  4if2-A 43.2  1.7  308   324   36   MOLECULE: PHOSPHOTRIESTERASE HOMOLOGY PROTEIN;
   4:  2vc5-A 42.6  1.8  305   314   34   MOLECULE: ARYLDIALKYLPHOSPHATASE;
   5:  4rdz-B 42.1  1.8  305   316   30   MOLECULE: PARATHION HYDROLASE;
   6:  2zc1-A 41.3  2.1  303   333   33   MOLECULE: PHOSPHOTRIESTERASE;
   7:  5ch9-A 41.1  2.2  307   328   31   MOLECULE: PHOSPHOTRIESTERASE;
   8:  3k2g-B 39.2  2.2  295   358   29   MOLECULE: RESINIFERATOXIN-BINDING, PHOSPHOTRIESTERASE-
   9:  3rhg-A 38.9  2.2  296   363   27   MOLECULE: PUTATIVE PHOPHOTRIESTERASE;
  10:  3pnz-A 38.2  2.5  294   329   28   MOLECULE: PHOSPHOTRIESTERASE FAMILY PROTEIN;
  11:  3msr-A 37.1  2.1  288   353   25   MOLECULE: AMIDOHYDROLASES;
  12:  1bf6-A 36.7  2.1  276   291   30   MOLECULE: PHOSPHOTRIESTERASE HOMOLOGY PROTEIN;
  13:  3guw-A 19.8  2.5  213   233   17   MOLECULE: UNCHARACTERIZED PROTEIN AF_1765;
  14:  1j6o-A 19.8  3.1  237   260   14   MOLECULE: TATD-RELATED DEOXYRIBONUCLEASE;
  15:  1zzm-A 19.8  3.4  236   259   18   MOLECULE: PUTATIVE DEOXYRIBONUCLEASE YJJV;
  16:  4p5u-A 19.8  3.0  232   262   15   MOLECULE: TAT-LINKED QUALITY CONTROL PROTEIN TATD;
  17:  3rcm-A 19.7  3.1  236   279   16   MOLECULE: TATD FAMILY HYDROLASE;
  18:  1yix-A 19.6  3.2  234   265   18   MOLECULE: DEOXYRIBONUCLEASE YCFH;
  19:  2gzx-A 19.4  3.2  232   253   16   MOLECULE: PUTATIVE TATD RELATED DNASE;
  20:  3ipw-A 19.3  3.3  243   301   12   MOLECULE: HYDROLASE TATD FAMILY PROTEIN;
  21:  2y1h-B 19.2  3.2  230   265   16   MOLECULE: PUTATIVE DEOXYRIBONUCLEASE TATDN3;
  22:  1onx-A 19.1  3.0  235   390   17   MOLECULE: ISOASPARTYL DIPEPTIDASE;
  23:  2xio-A 19.0  3.1  238   293   11   MOLECULE: PUTATIVE DEOXYRIBONUCLEASE TATDN1;
  24:  3be7-A 18.7  3.6  243   399   15   MOLECULE: ZN-DEPENDENT ARGININE CARBOXYPEPTIDASE;
  25:  3mkv-A 18.7  3.7  246   414   15   MOLECULE: PUTATIVE AMIDOHYDROLASE;
  26:  2qs8-A 18.6  3.7  252   407   17   MOLECULE: XAA-PRO DIPEPTIDASE;
  27:  3mtw-A 18.1  3.8  244   404   14   MOLECULE: L-ARGININE CARBOXYPEPTIDASE CC2672;
  28:  2ftw-A 18.0  3.1  245   484   13   MOLECULE: DIHYDROPYRIMIDINE AMIDOHYDROLASE;
  29:  2vr2-A 18.0  2.9  240   478   13   MOLECULE: DIHYDROPYRIMIDINASE;
  30:  2vm8-A 18.0  3.2  246   477   10   MOLECULE: DIHYDROPYRIMIDINASE-RELATED PROTEIN 2;
  31:  3e2v-A 17.9  3.8  244   363   10   MOLECULE: 3'-5'-EXONUCLEASE;
  32:  4b3z-D 17.9  3.4  250   477   12   MOLECULE: DIHYDROPYRIMIDINASE-RELATED PROTEIN 1;
  33:  4cnu-A 17.9  3.4  250   488   12   MOLECULE: DIHYDROPYRIMIDINASE-LIKE 3;
  34:  4gz7-A 17.8  3.3  247   492   11   MOLECULE: DIHYDROPYRIMIDINASE;
  35:  3dc8-A 17.6  3.3  248   483   15   MOLECULE: DIHYDROPYRIMIDINASE;
  36:  4l9x-B 17.6  3.9  243   458   17   MOLECULE: TRIAZINE HYDROLASE;
  37:  1gkr-A 17.5  3.6  240   451   13   MOLECULE: NON-ATP DEPENDENT L-SELECTIVE HYDANTOINASE;
  38:  3giq-A 17.5  3.1  243   475   17   MOLECULE: N-ACYL-D-GLUTAMATE DEACYLASE;
  39:  4b90-A 17.5  3.2  242   485   14   MOLECULE: DIHYDROPYRIMIDINASE-RELATED PROTEIN 5;
  40:  4c5y-A 17.5  3.6  237   436   15   MOLECULE: OCHRATOXINASE;
  41:  1nfg-A 17.4  3.4  245   457   16   MOLECULE: D-HYDANTOINASE;
  42:  3cjp-A 17.3  2.6  212   262   13   MOLECULE: PREDICTED AMIDOHYDROLASE, DIHYDROOROTASE FAMILY;
  43:  1m7j-A 17.2  3.2  245   474   16   MOLECULE: D-AMINOACYLASE;
  44:  2qpx-A 17.2  3.1  235   376   10   MOLECULE: PREDICTED METAL-DEPENDENT HYDROLASE OF THE TIM-BA
  45:  2p9b-A 17.1  3.8  239   407   16   MOLECULE: POSSIBLE PROLIDASE;
  46:  2paj-A 17.1  3.4  230   421   14   MOLECULE: PUTATIVE CYTOSINE/GUANINE DEAMINASE;
  47:  4i6k-A 17.1  3.4  233   267   11   MOLECULE: AMIDOHYDROLASE FAMILY PROTEIN;
  48:  4tqt-D 17.0  3.4  249   481   13   MOLECULE: D-HYDANTOINASE;
  49:  3gg7-A 16.8  3.4  220   243   18   MOLECULE: UNCHARACTERIZED METALLOPROTEIN;
  50:  3d6n-A 16.8  3.7  245   422   17   MOLECULE: DIHYDROOROTASE;
  51:  2ffi-A 16.7  3.4  236   273   13   MOLECULE: 2-PYRONE-4,6-DICARBOXYLIC ACID HYDROLASE, PUTATIV
  52:  3s2j-A 16.5  3.4  241   393   12   MOLECULE: DIPEPTIDASE;
  53:  2z00-A 16.4  3.5  239   426   16   MOLECULE: DIHYDROOROTASE;
  54:  2vun-A 16.4  3.1  218   385   14   MOLECULE: ENAMIDASE;
  55:  2oof-A 16.4  3.9  234   403   19   MOLECULE: 4-IMIDAZOLONE-5-PROPANOATE AMIDOHYDROLASE;
  56:  2g3f-A 16.4  4.0  236   414   19   MOLECULE: IMIDAZOLONEPROPIONASE;
  57:  3irs-A 16.4  3.5  226   281   11   MOLECULE: UNCHARACTERIZED PROTEIN BB4693;
  58:  2i5g-A 16.2  3.3  236   325   13   MOLECULE: AMIDOHYDROLASE;
  59:  2gok-A 16.2  4.1  233   404   19   MOLECULE: IMIDAZOLONEPROPIONASE;
  60:  3b40-A 16.1  3.3  242   400   10   MOLECULE: PROBABLE DIPEPTIDASE;
  61:  3lu2-A 16.1  3.2  229   309   12   MOLECULE: LMO2462 PROTEIN;
  62:  4v1x-E 16.1  3.4  233   474   15   MOLECULE: ATRAZINE CHLOROHYDROLASE;
  63:  2fty-A 15.8  3.5  245   532   13   MOLECULE: DIHYDROPYRIMIDINASE;
  64:  3hm7-B 15.7  3.0  219   437   13   MOLECULE: ALLANTOINASE;
  65:  3gri-B 15.6  3.4  231   423   11   MOLECULE: DIHYDROOROTASE;
  66:  2z26-A 15.6  3.7  242   344   11   MOLECULE: DIHYDROOROTASE;
  67:  3nqb-A 15.6  3.1  214   587   18   MOLECULE: ADENINE DEAMINASE 2;
  68:  1itq-A 15.6  3.4  235   369    9   MOLECULE: RENAL DIPEPTIDASE;
  69:  4lfy-B 15.6  3.7  241   353   11   MOLECULE: DIHYDROOROTASE;
  70:  3e74-B 15.5  2.9  220   433   13   MOLECULE: ALLANTOINASE;
  71:  3ly0-B 15.4  3.3  237   352   11   MOLECULE: DIPEPTIDASE AC. METALLO PEPTIDASE. MEROPS FAMILY
  72:  2amx-A 15.4  3.5  233   364   12   MOLECULE: ADENOSINE DEAMINASE;
  73:  2wjd-A 15.2  3.2  206   244   13   MOLECULE: TYROSINE-PROTEIN PHOSPHATASE CPSB;
  74:  3lnp-A 15.2  3.9  229   441   16   MOLECULE: AMIDOHYDROLASE FAMILY PROTEIN OLEI01672_1_465;
  75:  2dvt-A 15.0  3.5  221   325   16   MOLECULE: THERMOPHILIC REVERSIBLE GAMMA-RESORCYLATE DECARBO
  76:  2ics-A 14.9  3.4  220   368   12   MOLECULE: ADENINE DEAMINASE;
  77:  3ewd-A 14.9  4.1  235   364   11   MOLECULE: ADENOSINE DEAMINASE;
  78:  2gwg-A 14.9  3.2  223   329   10   MOLECULE: 4-OXALOMESACONATE HYDRATASE;
  79:  4dyk-A 14.9  3.7  227   437   12   MOLECULE: AMIDOHYDROLASE;
  80:  1j6p-A 14.8  3.9  229   407   12   MOLECULE: METAL-DEPENDENT HYDROLASE OF
  81:  3rys-A 14.8  4.1  233   335   13   MOLECULE: ADENOSINE DEAMINASE 1;
  82:  3pnu-A 14.7  3.7  235   338   10   MOLECULE: DIHYDROOROTASE;
  83:  1ie7-C 14.7  3.2  216   570   17   MOLECULE: UREASE GAMMA SUBUNIT;
  84:  1a5k-C 14.7  3.2  217   566   17   MOLECULE: UREASE (GAMMA SUBUNIT);
  85:  4icm-A 14.7  3.6  221   335   14   MOLECULE: 5-CARBOXYVANILLATE DECARBOXYLASE;
```

Figure 7. Top part of PDB90 summary list for PDB search.

```
1a4mA   adenosine deaminase
1a5kC   urease gamma subunit
1bf6A   phosphotriesterase
1bksA   tryptophan synthase
1gkpA   hydantoinase
1itqA   renal dipeptidase
1j5sA   uronate isomerase
1j6pA   uncharacterized
1k6wA   cytosine deaminase
1m65A   probable phosphatase YcdX
1onxA   isoaspartyl dipeptidase
1v77A   ribonuclease P protein component 3
1yrrB   N-acetylglucosamine-6-phosphate deacetylase
2a3lA   AMP deaminase
2anuA   predicted phosphoesterase
2dvtA   gamma-resocylate decarboxylase
2ffiA   2-pyrone-4-6-dicarboxylic acid hydrolase
2gwgA   4-oxalomesaconate hydratase
2imrA   S-adenosylhomocysteine deaminase
2ob3A   parathion hydrlolase
2ogjA   deacetylase
2oofA   4-imidazolone-5-propanoate amidohydrolase
2pajA   isoxanthopterin deaminase
2qpxA   uncharacterized
2uz9A   guanine deaminase
2vc5A   aryldialkylphosphatase
2vunA   enamidase
2y1hB   deoxyribonuclease TatD
2yb1A   uncharacterized
3au2A   DNA polymerase X
3cjpA   uncharacterized
3dcpA   histidinol-phosphatase
3e38A   uncharacterized
3e74A   allantoinase
3f2bA   DNA polymerase polC
3gg7A   uncharacterized
3giqA   N-acyl-D-glutamate deacylase
3griA   dihydroorotase
3iacA   glucuronate isomerase
3icjA   AepA exoenzymes regulatory protein
3irsA   uncharacterized
3k2gB   resiniferatoxin-binding phosphotriesterase
3ls9A   triazine hydrolase
3mkvA   uncharacterized
3mtwA   L-arginine carboxypeptidase
3nqbA   adenine deaminase
3ooqA   uncharacterized
3pnuA   dihydroorotase
3qy6A   tyrosine-protein phosphatase
4b3zD   dihydropyriminidase-related
4c5yA   ochratoxinase
4cqbA   N-isopropylammelide isopropyl amidohydrolase
4dlfA   uncharacterized
4dziC   uncharacterized
4hk5D   uracil-5-carboxylate decarboxylase
4mupB   D-galactarolactone isomerase
4ofcA   2-amino-3-carboxymuconate-6-semialdehyde decarboxylase
4qrnA   5-carboxyvanillate decarboxylase
4rdvB   N-formimino-L-glutamate iminiohydrolase
```

Figure 8. Input set used in all against all comparison example.

4.2 Position of amidohydrolases and PHP superfamily in structure space

The set of structures from Figure 8 was submitted to all against all comparison, yielding the results shown in Figure 6. You can use this live example to reproduce the figures. The dendrogram (Figure 9) and correspondence analysis plot (Figure 10) agree quite well in grouping the most strongly similar structures. However, branching order nearer the root becomes more or less arbitrary. For example, adenosine deaminase and AMP deaminase (Group A) are far apart in the dendrogram but adjacent in the correspondence analysis plot. Although members of the PHP superfamily occasionally appear in the structural neighbour lists of amidohydrolases, the two superfamilies appear as structurally distinct in our analyses. In the correspondence analysis plot, the first eigenvector (horizontal) separates PHP domain

proteins from amidohydrolases. PHP domains have a 7-stranded beta barrel, while amidohydrolases have an 8-stranded beta barrel. The second eigenvector (vertical) separates amidohydrolases with the catalytic and small domain from those with only the catalytic domain. The outgroup is near the origin, indicating that it has no special affinity towards any of the other groups.

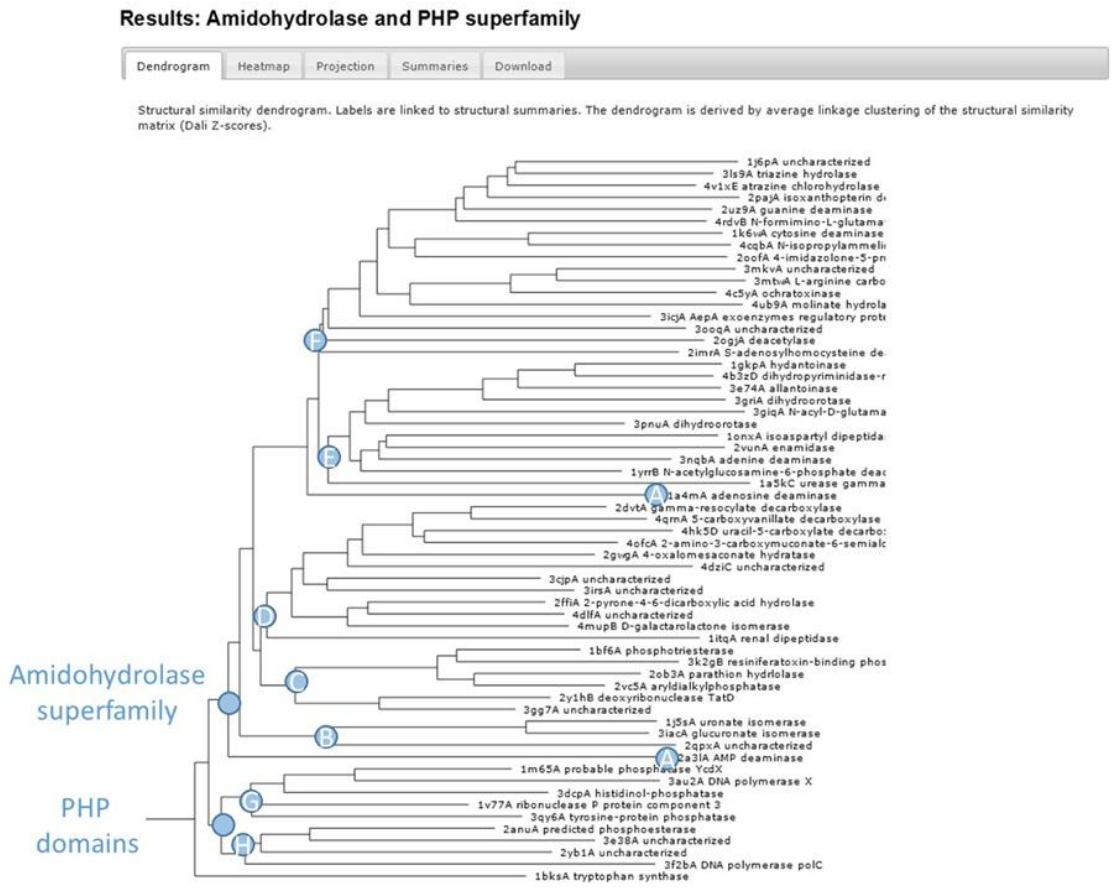


Figure 9: Dendrogram representation of protein structure space. Node labels were added manually.

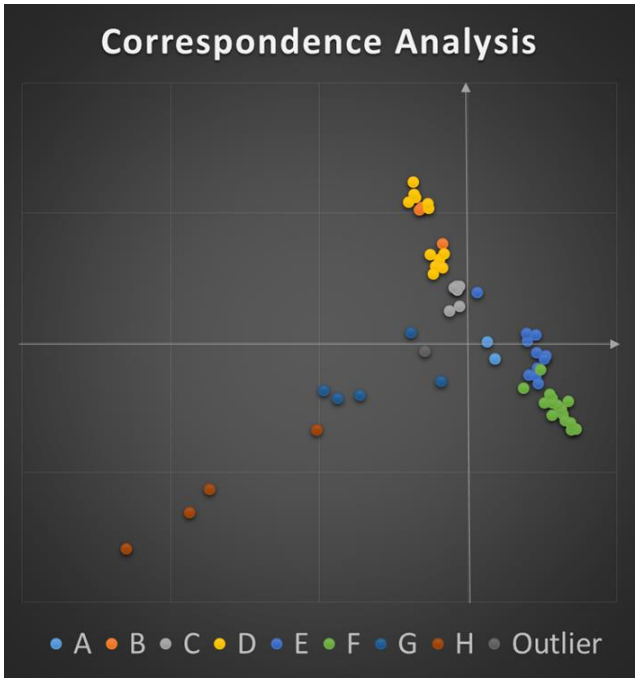


Figure 10. Correspondence analysis plot. Groups were manually defined and correspond to those in the dendrogram (Figure 9).

4.3 Overlapping sequence motifs of amidohydrolases and PHP superfamily

Both PHP domains and amidohydrolases bind metal ions in their active site. Figure 11 shows a stacked sequence logo comparison of two amidohydrolases and two PHP domains. The representatives were chosen from the extremes of the PHP cloud and amidohydrolase cloud in the correspondence analysis plot. The idea is to bring out invariant aminodhydrolase features (conserved in both amidohydrolases) and invariant PHP domain features (conserved in both PHP domains). We see that although some residues are commonly conserved in both superfamilies, each has unique features missing from the other. It is interesting that these distinct sequence motifs coincide with the structural distinction of PHP domains from amidohydrolases (Figure 10).

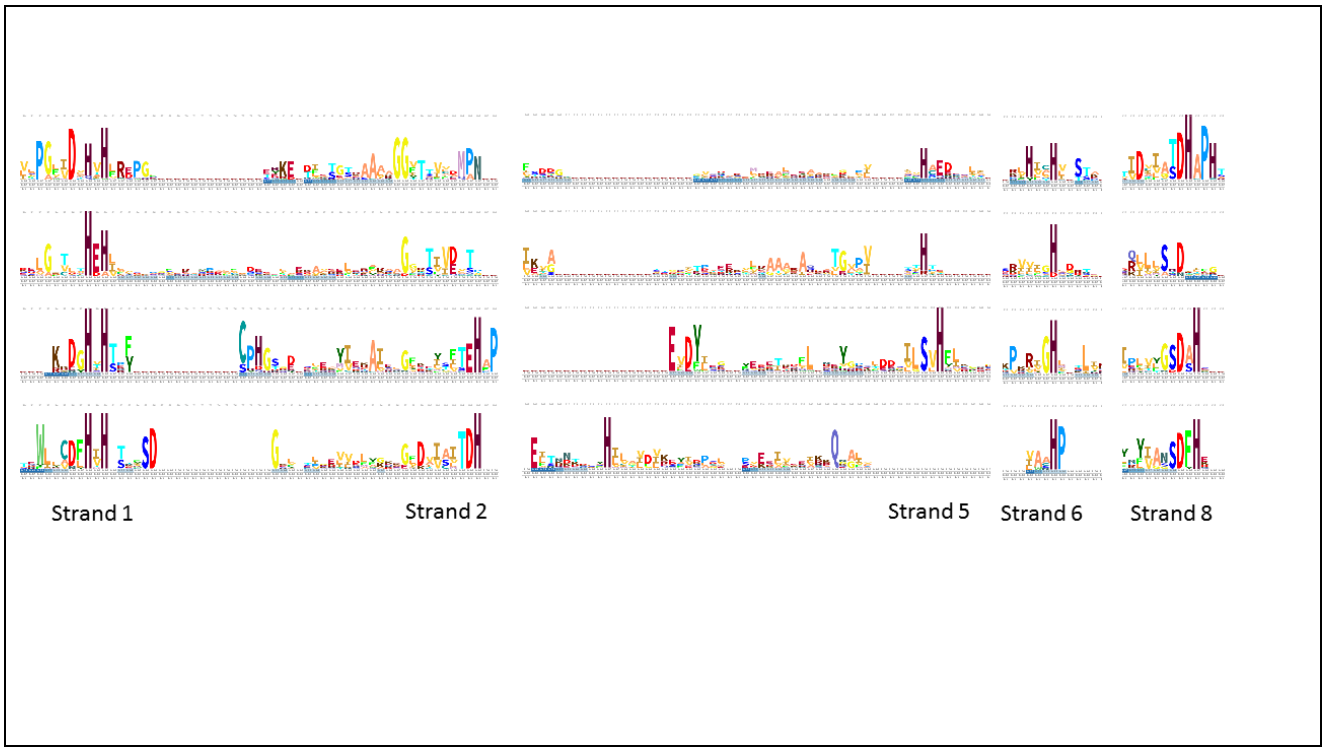


Figure 11. Sequence logos for two amidohydrolases (top) and two PHP proteins (bottom).

4.4 Looking into PHP superfamily

The dendrogram (Figure 9) shows a deep divide within the PHP superfamily between phosphatases on the one hand and phosphoesterases on the other. In the correspondence analysis plot, they are separated by the first eigenvector (Figure 10). The structural similarity between the two subclasses dips almost as low as the outlier. Without the unifying sequence motif, we would even question whether they are evolutionarily related. Inspection of 3D superimpositions and multiple structural alignments revealed that while one subsclass has a parallel alpha/beta barrel like amidohydrolases, one of the beta strands has reversed direction in the other subclass (Figure 12). Because Dali reports only sequential alignments, this explains why the alignment score is lower between the subclasses.

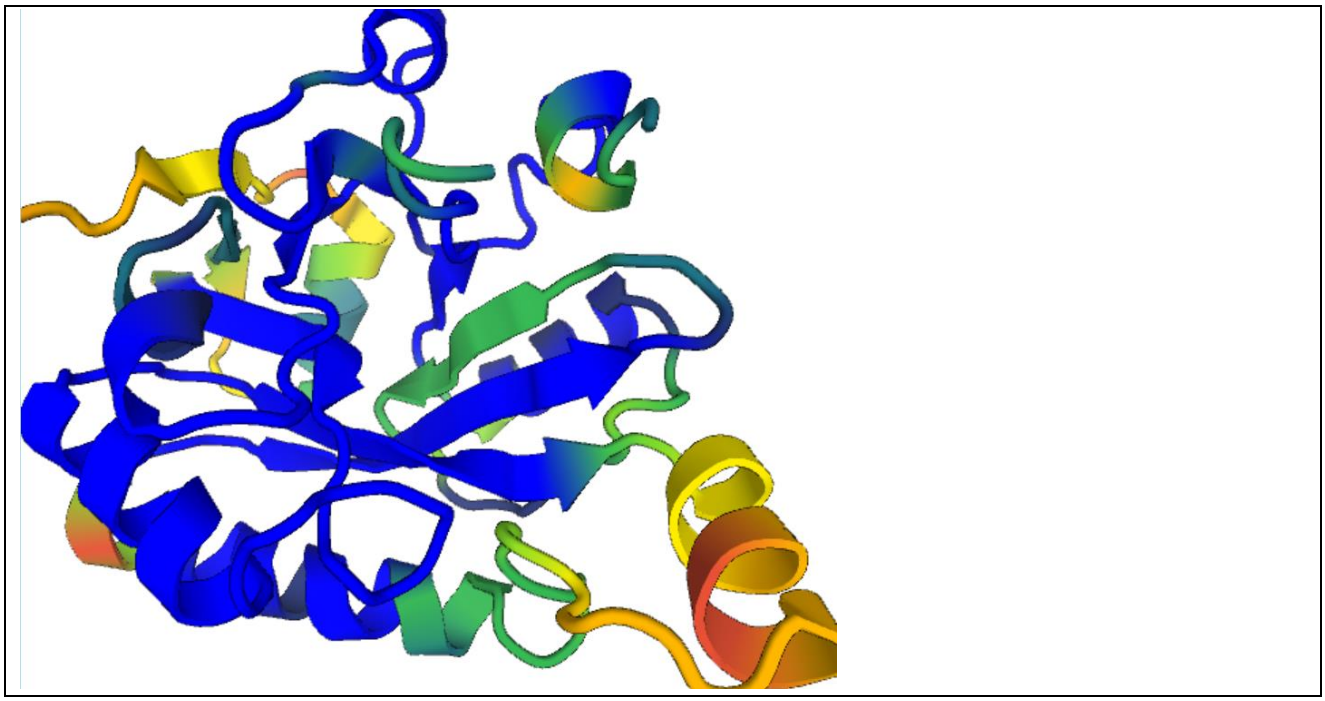


Figure 12. Structural conservation of the eight PHP domains in our data set. Blue regions are aligned in all eight PHP domain structures, the green strand of the barrel has reversed direction in some.

4.5 Diversity of molecular functions in amidohydrolase superfamily

The structural dendrogram is labelled with the proteins' functional descriptions. We can observe that, in general, structural groupings coincide with functional categories. This observation fits well with the idea of evolutionary continuity of structure and function. As a corollary, functions in incongruent positions in the dendrogram should alert to possible misclassification. For example, deacetylase 2ogjA is incorrectly annotated as dihydroorotase in PDB. Though corrected to deacetylase for this protein in Uniprot, the incorrect annotation has spread to sequence neighbours which remain annotated as dihydroorotases in Uniprot (as you will see in a SANS search). Our input set contains a number of uncharacterized proteins. Functionally characterized neighbours in structure space can direct the formulation of testable hypotheses of their molecular function, at least regarding the class of enzyme function if not precise substrate specificity. For example, ochratoxinase and molinate hydrolase are recently evolved new enzymes.

4.6 Conclusion

The Dali server is a useful aid in structural classification. It generates overviews of selected portions of structure space and sequence space with just a few mouse clicks. Similar narratives as above can be developed for any structurally characterized superfamily.

5 Downloads

The DaliLite software is available for academic use from
http://ekhidna.biocenter.helsinki.fi/dali/downloads/download.html.

Appendix A: Sample PDB entry

```
HEADER    PANCREATIC HORMONE                      16-JAN-81   1PPT
TITLE     X-RAY ANALYSIS (1.4-ANGSTROMS RESOLUTION) OF AVIAN PANCREATIC
TITLE    2 POLYPEPTIDE. SMALL GLOBULAR PROTEIN HORMONE
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: AVIAN PANCREATIC POLYPEPTIDE;
COMPND   3 CHAIN: A;
COMPND   4 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: MELEAGRIS GALLOPAVO;
SOURCE   3 ORGANISM_COMMON: TURKEY;
SOURCE   4 ORGANISM_TAXID: 9103
AUTHOR    T.L.BLUNDELL,J.E.PITTS,I.J.TICKLE,S.P.WOOD
JRNL        AUTH   T.L.BLUNDELL,J.E.PITTS,I.J.TICKLE,S.P.WOOD,C.W.WU
JRNL        TITL   X-RAY ANALYSIS (1. 4-A RESOLUTION) OF AVIAN PANCREATIC
JRNL        TITL 2 POLYPEPTIDE: SMALL GLOBULAR PROTEIN HORMONE.
JRNL        REF    PROC.NATL.ACAD.SCI.USA        V.  78  4175 1981
JRNL        REFN                   ISSN 0027-8424
JRNL        PMID   16593056
JRNL        DOI    10.1073/PNAS.78.7.4175
ATOM      1  N   GLY A   1       2.296  -9.636  18.253  1.00  0.00           N
ATOM      2  CA  GLY A   1       1.470  -9.017  17.255  1.00  0.00           C
ATOM      3  C   GLY A   1       0.448  -9.983  16.703  1.00  0.00           C
ATOM      4  O   GLY A   1       0.208 -11.066  17.345  1.00  0.00           O
ATOM      5  N   PRO A   2      -0.170  -9.672  15.624  1.00  0.00           N
ATOM      6  CA  PRO A   2      -1.135 -10.606  14.958  1.00  0.00           C
ATOM      7  C   PRO A   2      -0.376 -11.824  14.490  1.00  0.00           C
ATOM      8  O   PRO A   2       0.776 -11.860  14.075  1.00  0.00           O
ATOM      9  CB  PRO A   2      -1.717  -9.829  13.776  1.00  0.00           C
ATOM     10  CG  PRO A   2      -0.817  -8.685  13.546  1.00  0.00           C
ATOM     11  CD  PRO A   2       0.108  -8.454  14.780  1.00  0.00           C
ATOM     12  N   SER A   3      -1.184 -12.918  14.566  1.00  0.00           N
ATOM     13  CA  SER A   3      -0.626 -14.187  14.053  1.00  0.00           C
ATOM     14  C   SER A   3      -0.642 -14.190  12.493  1.00  0.00           C
ATOM     15  O   SER A   3      -1.149 -13.332  11.830  1.00  0.00           O
ATOM     16  CB  SER A   3      -1.360 -15.359  14.573  1.00  0.00           C
ATOM     17  OG  SER A   3      -2.655 -15.234  14.212  1.00  0.00           O
ATOM     18  N   GLN A   4       0.243 -14.995  11.964  1.00  0.00           N
ATOM     19  CA  GLN A   4       0.489 -14.940  10.481  1.00  0.00           C
ATOM     20  C   GLN A   4      -0.766 -15.384   9.734  1.00  0.00           C
ATOM     21  O   GLN A   4      -1.330 -16.452  10.019  1.00  0.00           O
ATOM     22  CB  GLN A   4       1.639 -15.895  10.114  1.00  0.00           C
ATOM     23  CG  GLN A   4       2.182 -15.697   8.704  1.00  0.00           C
ATOM     24  CD  GLN A   4       3.315 -16.670   8.366  1.00  0.00           C
ATOM     25  OE1 GLN A   4       3.718 -16.761   7.207  1.00  0.00           O
ATOM     26  NE2 GLN A   4       3.864 -17.403   9.317  1.00  0.00           N
ATOM     27  N   PRO A   5      -1.196 -14.647   8.750  1.00  0.00           N
ATOM     28  CA  PRO A   5      -2.414 -14.970   8.087  1.00  0.00           C
ATOM     29  C   PRO A   5      -2.264 -16.297   7.258  1.00  0.00           C
ATOM     30  O   PRO A   5      -1.184 -16.595   6.819  1.00  0.00           O
ATOM     31  CB  PRO A   5      -2.798 -13.854   7.153  1.00  0.00           C
ATOM     32  CG  PRO A   5      -1.809 -12.748   7.438  1.00  0.00           C
ATOM     33  CD  PRO A   5      -0.768 -13.190   8.408  1.00  0.00           C
ATOM     34  N   THR A   6      -3.381 -16.917   7.174  1.00  0.00           N
ATOM     35  CA  THR A   6      -3.548 -18.158   6.308  1.00  0.00           C
ATOM     36  C   THR A   6      -3.745 -17.747   4.861  1.00  0.00           C
ATOM     37  O   THR A   6      -4.693 -17.045   4.518  1.00  0.00           O
ATOM     38  CB  THR A   6      -4.752 -18.911   6.884  1.00  0.00           C
ATOM     39  OG1 THR A   6      -4.040 -19.502   8.074  1.00  0.00           O
ATOM     40  CG2 THR A   6      -4.799 -20.260   6.058  1.00  0.00           C
ATOM     41  N   TYR A   7      -2.893 -18.207   3.953  1.00  0.00           N
ATOM     42  CA  TYR A   7      -3.065 -18.017   2.495  1.00  0.00           C
ATOM     43  C   TYR A   7      -4.327 -18.738   2.010  1.00  0.00           C
ATOM     44  O   TYR A   7      -4.536 -19.927   2.291  1.00  0.00           O
ATOM     45  CB  TYR A   7      -1.828 -18.587   1.791  1.00  0.00           C
ATOM     46  CG  TYR A   7      -1.913 -18.407   0.265  1.00  0.00           C
ATOM     47  CD1 TYR A   7      -2.029 -17.122  -0.283  1.00  0.00           C
ATOM     48  CD2 TYR A   7      -1.884 -19.519  -0.588  1.00  0.00           C
ATOM     49  CE1 TYR A   7      -2.090 -16.948  -1.671  1.00  0.00           C
ATOM     50  CE2 TYR A   7      -1.943 -19.344  -1.978  1.00  0.00           C
```

```
ATOM     51  CZ  TYR A   7      -2.039 -18.057  -2.521  1.00  0.00           C
ATOM     52  OH  TYR A   7      -2.067 -17.876  -3.868  1.00  0.00           O
ATOM     53  N   PRO A   8      -5.261 -18.068   1.439  1.00  0.00           N
ATOM     54  CA  PRO A   8      -6.566 -18.626   0.996  1.00  0.00           C
ATOM     55  C   PRO A   8      -6.492 -19.530  -0.193  1.00  0.00           C
ATOM     56  O   PRO A   8      -7.584 -20.240  -0.510  1.00  0.00           O
ATOM     57  CB  PRO A   8      -7.488 -17.397   0.798  1.00  0.00           C
ATOM     58  CG  PRO A   8      -6.583 -16.313   0.428  1.00  0.00           C
ATOM     59  CD  PRO A   8      -5.230 -16.608   1.173  1.00  0.00           C
ATOM     60  N   GLY A   9      -5.375 -19.740  -0.857  1.00  0.00           N
ATOM     61  CA  GLY A   9      -5.423 -20.730  -1.983  1.00  0.00           C
ATOM     62  C   GLY A   9      -5.256 -19.861  -3.214  1.00  0.00           C
ATOM     63  O   GLY A   9      -5.790 -18.783  -3.338  1.00  0.00           O
ATOM     64  N   ASP A  10      -4.626 -20.463  -4.248  1.00  0.00           N
ATOM     65  CA  ASP A  10      -4.521 -19.865  -5.611  1.00  0.00           C
ATOM     66  C   ASP A  10      -5.891 -19.644  -6.232  1.00  0.00           C
ATOM     67  O   ASP A  10      -6.079 -18.696  -7.006  1.00  0.00           O
ATOM     68  CB  ASP A  10      -3.697 -20.772  -6.523  1.00  0.00           C
ATOM     69  CG  ASP A  10      -2.225 -20.895  -6.117  1.00  0.00           C
ATOM     70  OD1 ASP A  10      -1.521 -21.886  -6.544  1.00  0.00           O
ATOM     71  OD2 ASP A  10      -1.682 -20.014  -5.347  1.00  0.00           O
ATOM     72  N   ASP A  11      -6.810 -20.490  -5.917  1.00  0.00           N
ATOM     73  CA  ASP A  11      -8.106 -20.411  -6.537  1.00  0.00           C
ATOM     74  C   ASP A  11      -9.141 -19.681  -5.693  1.00  0.00           C
ATOM     75  O   ASP A  11     -10.273 -19.451  -6.151  1.00  0.00           O
ATOM     76  CB  ASP A  11      -8.681 -21.809  -6.852  1.00  0.00           C
ATOM     77  CG  ASP A  11      -7.791 -22.570  -7.829  1.00  0.00           C
ATOM     78  OD1 ASP A  11      -7.396 -21.995  -8.913  1.00  0.00           O
ATOM     79  OD2 ASP A  11      -7.431 -23.778  -7.563  1.00  0.00           O
ATOM     80  N   ALA A  12      -8.612 -18.887  -4.775  1.00  0.00           N
ATOM     81  CA  ALA A  12      -9.622 -18.132  -4.017  1.00  0.00           C
ATOM     82  C   ALA A  12     -10.101 -16.933  -4.820  1.00  0.00           C
ATOM     83  O   ALA A  12      -9.482 -16.473  -5.779  1.00  0.00           O
ATOM     84  CB  ALA A  12      -8.829 -17.575  -2.793  1.00  0.00           C
ATOM     85  N   PRO A  13     -11.366 -16.547  -4.687  1.00  0.00           N
ATOM     86  CA  PRO A  13     -11.981 -15.406  -5.466  1.00  0.00           C
ATOM     87  C   PRO A  13     -11.187 -14.121  -5.215  1.00  0.00           C
ATOM     88  O   PRO A  13     -10.522 -13.958  -4.032  1.00  0.00           O
ATOM     89  CB  PRO A  13     -13.424 -15.243  -4.908  1.00  0.00           C
ATOM     90  CG  PRO A  13     -13.500 -16.009  -3.659  1.00  0.00           C
ATOM     91  CD  PRO A  13     -12.236 -16.882  -3.480  1.00  0.00           C
ATOM     92  N   VAL A  14     -11.180 -13.190  -6.126  1.00  0.00           N
ATOM     93  CA  VAL A  14     -10.341 -11.949  -5.914  1.00  0.00           C
ATOM     94  C   VAL A  14     -10.673 -11.235  -4.610  1.00  0.00           C
ATOM     95  O   VAL A  14      -9.729 -10.720  -3.902  1.00  0.00           O
ATOM     96  CB  VAL A  14     -10.477 -11.110  -7.162  1.00  0.00           C
ATOM     97  CG1 VAL A  14      -9.809  -9.750  -7.062  1.00  0.00           C
ATOM     98  CG2 VAL A  14     -10.013 -11.873  -8.431  1.00  0.00           C
ATOM     99  N   GLU A  15     -11.842 -11.320  -4.113  1.00  0.00           N
ATOM    100  CA  GLU A  15     -12.142 -10.553  -2.855  1.00  0.00           C
ATOM    101  C   GLU A  15     -11.451 -11.174  -1.704  1.00  0.00           C
ATOM    102  O   GLU A  15     -11.170 -10.380  -0.760  1.00  0.00           O
ATOM    103  CB  GLU A  15     -13.711 -10.553  -2.589  1.00  0.00           C
ATOM    104  CG  GLU A  15     -14.210 -11.854  -1.955  1.00  0.00           C
ATOM    105  CD  GLU A  15     -15.561 -11.698  -1.254  1.00  0.00           C
ATOM    106  OE1 GLU A  15     -16.152 -12.731  -0.757  1.00  0.00           O
ATOM    107  OE2 GLU A  15     -16.108 -10.534  -1.161  1.00  0.00           O
ATOM    108  N   ASP A  16     -11.319 -12.496  -1.702  1.00  0.00           N
ATOM    109  CA  ASP A  16     -10.488 -13.190  -0.722  1.00  0.00           C
ATOM    110  C   ASP A  16      -9.000 -12.982  -0.958  1.00  0.00           C
ATOM    111  O   ASP A  16      -8.238 -12.840   0.013  1.00  0.00           O
ATOM    112  CB  ASP A  16     -10.706 -14.670  -0.638  1.00  0.00           C
ATOM    113  CG  ASP A  16     -12.106 -15.022  -0.152  1.00  0.00           C
ATOM    114  OD1 ASP A  16     -12.571 -16.208  -0.343  1.00  0.00           O
ATOM    115  OD2 ASP A  16     -12.821 -14.131   0.443  1.00  0.00           O
ATOM    116  N   LEU A  17      -8.476 -12.788  -2.105  1.00  0.00           N
ATOM    117  CA  LEU A  17      -7.028 -12.438  -2.126  1.00  0.00           C
ATOM    118  C   LEU A  17      -6.810 -10.983  -1.717  1.00  0.00           C
ATOM    119  O   LEU A  17      -5.812 -10.718  -1.159  1.00  0.00           O
ATOM    120  CB  LEU A  17      -6.647 -12.470  -3.630  1.00  0.00           C
ATOM    121  CG  LEU A  17      -6.525 -13.931  -4.064  1.00  0.00           C
```

```
ATOM    122  CD1 LEU A  17      -6.250 -13.916  -5.582  1.00  0.00           C
ATOM    123  CD2 LEU A  17      -5.372 -14.656  -3.263  1.00  0.00           C
ATOM    124  N   ILE A  18      -7.786 -10.075  -1.877  1.00  0.00           N
ATOM    125  CA  ILE A  18      -7.676  -8.682  -1.354  1.00  0.00           C
ATOM    126  C   ILE A  18      -7.789  -8.754   0.184  1.00  0.00           C
ATOM    127  O   ILE A  18      -6.937  -8.096   0.818  1.00  0.00           O
ATOM    128  CB  ILE A  18      -8.822  -7.866  -1.960  1.00  0.00           C
ATOM    129  CG1 ILE A  18      -8.514  -7.564  -3.452  1.00  0.00           C
ATOM    130  CG2 ILE A  18      -8.973  -6.595  -1.116  1.00  0.00           C
ATOM    131  CD1 ILE A  18      -9.640  -6.945  -4.117  1.00  0.00           C
ATOM    132  N   ARG A  19      -8.698  -9.540   0.771  1.00  0.00           N
ATOM    133  CA  ARG A  19      -8.636  -9.644   2.284  1.00  0.00           C
ATOM    134  C   ARG A  19      -7.271 -10.180   2.719  1.00  0.00           C
ATOM    135  O   ARG A  19      -6.742  -9.797   3.773  1.00  0.00           O
ATOM    136  CB  ARG A  19      -9.736 -10.584   2.604  1.00  0.00           C
ATOM    137  CG  ARG A  19     -11.010  -9.860   3.115  1.00  0.00           C
ATOM    138  CD  ARG A  19     -12.209 -10.742   3.604  1.00  0.00           C
ATOM    139  NE  ARG A  19     -13.244 -10.589   2.620  1.00  0.00           N
ATOM    140  CZ  ARG A  19     -14.562 -10.256   2.523  1.00  0.00           C
ATOM    141  NH1 ARG A  19     -15.649  -9.931   3.450  1.00  0.00           N
ATOM    142  NH2 ARG A  19     -14.877 -10.167   1.331  1.00  0.00           N
ATOM    143  N   PHE A  20      -6.704 -11.222   2.134  1.00  0.00           N
ATOM    144  CA  PHE A  20      -5.441 -11.780   2.548  1.00  0.00           C
ATOM    145  C   PHE A  20      -4.374 -10.694   2.394  1.00  0.00           C
ATOM    146  O   PHE A  20      -3.489 -10.541   3.246  1.00  0.00           O
ATOM    147  CB  PHE A  20      -5.101 -12.968   1.646  1.00  0.00           C
ATOM    148  CG  PHE A  20      -3.710 -13.511   1.946  1.00  0.00           C
ATOM    149  CD1 PHE A  20      -2.666 -13.334   1.030  1.00  0.00           C
ATOM    150  CD2 PHE A  20      -3.489 -14.185   3.148  1.00  0.00           C
ATOM    151  CE1 PHE A  20      -1.392 -13.833   1.327  1.00  0.00           C
ATOM    152  CE2 PHE A  20      -2.218 -14.680   3.443  1.00  0.00           C
ATOM    153  CZ  PHE A  20      -1.168 -14.502   2.535  1.00  0.00           C
ATOM    154  N   TYR A  21      -4.347  -9.918   1.280  1.00  0.00           N
ATOM    155  CA  TYR A  21      -3.402  -8.885   1.007  1.00  0.00           C
ATOM    156  C   TYR A  21      -3.411  -7.902   2.181  1.00  0.00           C
ATOM    157  O   TYR A  21      -2.361  -7.533   2.718  1.00  0.00           O
ATOM    158  CB  TYR A  21      -3.821  -8.142  -0.271  1.00  0.00           C
ATOM    159  CG  TYR A  21      -3.056  -6.870  -0.555  1.00  0.00           C
ATOM    160  CD1 TYR A  21      -1.730  -6.925  -0.993  1.00  0.00           C
ATOM    161  CD2 TYR A  21      -3.708  -5.654  -0.376  1.00  0.00           C
ATOM    162  CE1 TYR A  21      -1.042  -5.733  -1.246  1.00  0.00           C
ATOM    163  CE2 TYR A  21      -3.022  -4.467  -0.625  1.00  0.00           C
ATOM    164  CZ  TYR A  21      -1.692  -4.505  -1.058  1.00  0.00           C
ATOM    165  OH  TYR A  21      -1.035  -3.335  -1.287  1.00  0.00           O
ATOM    166  N   ASP A  22      -4.642  -7.469   2.585  1.00  0.00           N
ATOM    167  CA  ASP A  22      -4.728  -6.473   3.688  1.00  0.00           C
ATOM    168  C   ASP A  22      -4.207  -7.089   4.978  1.00  0.00           C
ATOM    169  O   ASP A  22      -3.568  -6.404   5.780  1.00  0.00           O
ATOM    170  CB  ASP A  22      -6.194  -6.063   3.854  1.00  0.00           C
ATOM    171  CG  ASP A  22      -6.707  -5.158   2.746  1.00  0.00           C
ATOM    172  OD1 ASP A  22      -7.967  -5.120   2.494  1.00  0.00           O
ATOM    173  OD2 ASP A  22      -5.890  -4.431   2.071  1.00  0.00           O
ATOM    174  N   ASN A  23      -4.505  -8.385   5.271  1.00  0.00           N
ATOM    175  CA  ASN A  23      -4.038  -8.947   6.542  1.00  0.00           C
ATOM    176  C   ASN A  23      -2.506  -9.142   6.402  1.00  0.00           C
ATOM    177  O   ASN A  23      -1.837  -8.942   7.451  1.00  0.00           O
ATOM    178  CB  ASN A  23      -4.716 -10.254   6.855  1.00  0.00           C
ATOM    179  CG  ASN A  23      -6.186 -10.108   7.257  1.00  0.00           C
ATOM    180  OD1 ASN A  23      -6.841 -11.277   7.242  1.00  0.00           O
ATOM    181  ND2 ASN A  23      -6.691  -9.038   7.581  1.00  0.00           N
ATOM    182  N   LEU A  24      -1.987  -9.505   5.267  1.00  0.00           N
ATOM    183  CA  LEU A  24      -0.451  -9.679   5.213  1.00  0.00           C
ATOM    184  C   LEU A  24       0.173  -8.275   5.430  1.00  0.00           C
ATOM    185  O   LEU A  24       1.220  -8.264   5.981  1.00  0.00           O
ATOM    186  CB  LEU A  24      -0.202 -10.148   3.805  1.00  0.00           C
ATOM    187  CG  LEU A  24       1.277 -10.481   3.548  1.00  0.00           C
ATOM    188  CD1 LEU A  24       1.764 -11.470   4.539  1.00  0.00           C
ATOM    189  CD2 LEU A  24       1.499 -11.012   2.076  1.00  0.00           C
ATOM    190  N   GLN A  25      -0.410  -7.204   4.911  1.00  0.00           N
ATOM    191  CA  GLN A  25       0.085  -5.877   5.096  1.00  0.00           C
ATOM    192  C   GLN A  25       0.191  -5.541   6.582  1.00  0.00           C
```

```
ATOM    193  O   GLN A  25       1.265  -5.150   7.065  1.00  0.00           O
ATOM    194  CB  GLN A  25      -0.806  -4.832   4.429  1.00  0.00           C
ATOM    195  CG  GLN A  25      -0.281  -3.402   4.489  1.00  0.00           C
ATOM    196  CD  GLN A  25      -0.921  -2.504   3.422  1.00  0.00           C
ATOM    197  OE1 GLN A  25      -0.397  -1.428   3.134  1.00  0.00           O
ATOM    198  NE2 GLN A  25      -2.028  -2.888   2.811  1.00  0.00           N
ATOM    199  N   GLN A  26      -0.856  -5.749   7.310  1.00  0.00           N
ATOM    200  CA  GLN A  26      -0.856  -5.563   8.757  1.00  0.00           C
ATOM    201  C   GLN A  26       0.308  -6.289   9.398  1.00  0.00           C
ATOM    202  O   GLN A  26       1.009  -5.921  10.266  1.00  0.00           O
ATOM    203  CB  GLN A  26      -2.266  -5.866   9.301  1.00  0.00           C
ATOM    204  CG  GLN A  26      -2.357  -5.747  10.824  1.00  0.00           C
ATOM    205  CD  GLN A  26      -2.333  -4.297  11.313  1.00  0.00           C
ATOM    206  OE1 GLN A  26      -2.414  -4.053  12.516  1.00  0.00           O
ATOM    207  NE2 GLN A  26      -2.225  -3.309  10.444  1.00  0.00           N
ATOM    208  N   TYR A  27       0.366  -7.626   9.157  1.00  0.00           N
ATOM    209  CA  TYR A  27       1.356  -8.520   9.698  1.00  0.00           C
ATOM    210  C   TYR A  27       2.759  -8.036   9.359  1.00  0.00           C
ATOM    211  O   TYR A  27       3.622  -7.942  10.245  1.00  0.00           O
ATOM    212  CB  TYR A  27       1.111  -9.949   9.116  1.00  0.00           C
ATOM    213  CG  TYR A  27       2.122 -10.986   9.594  1.00  0.00           C
ATOM    214  CD1 TYR A  27       3.124 -11.439   8.729  1.00  0.00           C
ATOM    215  CD2 TYR A  27       2.044 -11.485  10.899  1.00  0.00           C
ATOM    216  CE1 TYR A  27       4.061 -12.377   9.173  1.00  0.00           C
ATOM    217  CE2 TYR A  27       2.984 -12.421  11.346  1.00  0.00           C
ATOM    218  CZ  TYR A  27       3.998 -12.862  10.486  1.00  0.00           C
ATOM    219  OH  TYR A  27       4.932 -13.751  10.922  1.00  0.00           O
ATOM    220  N   LEU A  28       3.051  -7.735   8.171  1.00  0.00           N
ATOM    221  CA  LEU A  28       4.342  -7.279   7.784  1.00  0.00           C
ATOM    222  C   LEU A  28       4.675  -5.907   8.478  1.00  0.00           C
ATOM    223  O   LEU A  28       5.887  -5.811   8.927  1.00  0.00           O
ATOM    224  CB  LEU A  28       4.614  -7.138   6.312  1.00  0.00           C
ATOM    225  CG  LEU A  28       4.490  -8.512   5.562  1.00  0.00           C
ATOM    226  CD1 LEU A  28       4.429  -8.253   3.969  1.00  0.00           C
ATOM    227  CD2 LEU A  28       5.761  -9.362   5.838  1.00  0.00           C
ATOM    228  N   ASN A  29       3.737  -4.993   8.530  1.00  0.00           N
ATOM    229  CA  ASN A  29       4.064  -3.735   9.328  1.00  0.00           C
ATOM    230  C   ASN A  29       4.502  -4.080  10.751  1.00  0.00           C
ATOM    231  O   ASN A  29       5.252  -3.321  11.381  1.00  0.00           O
ATOM    232  CB  ASN A  29       2.748  -2.894   9.387  1.00  0.00           C
ATOM    233  CG  ASN A  29       2.581  -2.132   8.083  1.00  0.00           C
ATOM    234  OD1 ASN A  29       1.565  -1.465   7.896  1.00  0.00           O
ATOM    235  ND2 ASN A  29       3.539  -2.200   7.175  1.00  0.00           N
ATOM    236  N   VAL A  30       3.812  -5.044  11.456  1.00  0.00           N
ATOM    237  CA  VAL A  30       4.069  -5.352  12.824  1.00  0.00           C
ATOM    238  C   VAL A  30       5.379  -6.029  12.954  1.00  0.00           C
ATOM    239  O   VAL A  30       6.270  -5.680  13.834  1.00  0.00           O
ATOM    240  CB  VAL A  30       2.961  -6.250  13.460  1.00  0.00           C
ATOM    241  CG1 VAL A  30       3.526  -6.593  14.954  1.00  0.00           C
ATOM    242  CG2 VAL A  30       1.683  -5.435  13.674  1.00  0.00           C
ATOM    243  N   VAL A  31       5.698  -6.965  12.038  1.00  0.00           N
ATOM    244  CA  VAL A  31       7.001  -7.699  12.162  1.00  0.00           C
ATOM    245  C   VAL A  31       8.157  -6.782  11.957  1.00  0.00           C
ATOM    246  O   VAL A  31       9.302  -7.004  12.397  1.00  0.00           O
ATOM    247  CB  VAL A  31       6.845  -8.810  10.979  1.00  0.00           C
ATOM    248  CG1 VAL A  31       8.217  -9.440  10.757  1.00  0.00           C
ATOM    249  CG2 VAL A  31       5.842  -9.913  11.342  1.00  0.00           C
ATOM    250  N   THR A  32       8.037  -5.850  11.015  1.00  0.00           N
ATOM    251  CA  THR A  32       9.068  -4.861  10.736  1.00  0.00           C
ATOM    252  C   THR A  32       8.975  -3.622  11.693  1.00  0.00           C
ATOM    253  O   THR A  32       9.882  -2.752  11.522  1.00  0.00           O
ATOM    254  CB  THR A  32       8.833  -4.336   9.342  1.00  0.00           C
ATOM    255  OG1 THR A  32       7.762  -3.572   9.058  1.00  0.00           O
ATOM    256  CG2 THR A  32       9.614  -4.848   8.439  1.00  0.00           C
ATOM    257  N   ARG A  33       8.154  -3.603  12.666  1.00  0.00           N
ATOM    258  CA  ARG A  33       7.995  -2.499  13.658  1.00  0.00           C
ATOM    259  C   ARG A  33       7.787  -1.166  12.943  1.00  0.00           C
ATOM    260  O   ARG A  33       8.043  -0.093  13.507  1.00  0.00           O
ATOM    261  CB  ARG A  33       9.235  -2.379  14.561  1.00  0.00           C
ATOM    262  CG  ARG A  33       9.711  -3.734  15.065  1.00  0.00           C
ATOM    263  CD  ARG A  33      10.875  -3.613  16.039  1.00  0.00           C
```

```
ATOM    264  NE  ARG A  33      10.549  -2.811  17.223  1.00  0.00           N
ATOM    265  CZ  ARG A  33       9.938  -3.306  18.303  1.00  0.00           C
ATOM    266  NH1 ARG A  33       9.575  -4.596  18.352  1.00  0.00           N
ATOM    267  NH2 ARG A  33       9.649  -2.584  19.395  1.00  0.00           N
ATOM    268  N   HIS A  34       7.031  -1.228  11.897  1.00  0.00           N
ATOM    269  CA  HIS A  34       6.779   0.039  11.099  1.00  0.00           C
ATOM    270  C   HIS A  34       5.289   0.163  10.798  1.00  0.00           C
ATOM    271  O   HIS A  34       4.835  -0.137   9.689  1.00  0.00           O
ATOM    272  CB  HIS A  34       7.587  -0.011   9.878  1.00  0.00           C
ATOM    273  CG  HIS A  34       7.608   1.293   9.098  1.00  0.00           C
ATOM    274  ND1 HIS A  34       6.953   1.430   7.879  1.00  0.00           N
ATOM    275  CD2 HIS A  34       8.195   2.486   9.363  1.00  0.00           C
ATOM    276  CE1 HIS A  34       7.144   2.668   7.454  1.00  0.00           C
ATOM    277  NE2 HIS A  34       7.884   3.310   8.330  1.00  0.00           N
ATOM    278  N   ARG A  35       4.524   0.544  11.879  1.00  0.00           N
ATOM    279  CA  ARG A  35       3.108   0.616  11.852  1.00  0.00           C
ATOM    280  C   ARG A  35       2.637   1.882  11.134  1.00  0.00           C
ATOM    281  O   ARG A  35       1.446   2.229  11.173  1.00  0.00           O
ATOM    282  CB  ARG A  35       2.550   0.636  13.314  1.00  0.00           C
ATOM    283  CG  ARG A  35       2.848  -0.712  13.994  1.00  0.00           C
ATOM    284  CD  ARG A  35       2.475  -0.788  15.476  1.00  0.00           C
ATOM    285  NE  ARG A  35       3.312  -1.745  16.223  1.00  0.00           N
ATOM    286  CZ  ARG A  35       2.837  -2.773  16.945  1.00  0.00           C
ATOM    287  NH1 ARG A  35       1.521  -3.007  17.037  1.00  0.00           N
ATOM    288  NH2 ARG A  35       3.612  -3.634  17.621  1.00  0.00           N
ATOM    289  N   TYR A  36       3.365   2.609  10.443  1.00  0.00           N
ATOM    290  CA  TYR A  36       2.765   3.446   9.296  1.00  0.00           C
ATOM    291  C   TYR A  36       2.332   2.479   8.197  1.00  0.00           C
ATOM    292  O   TYR A  36       3.166   1.720   7.671  1.00  0.00           O
ATOM    293  CB  TYR A  36       4.021   4.330   8.787  1.00  0.00           C
ATOM    294  CG  TYR A  36       4.734   4.795  10.058  1.00  0.00           C
ATOM    295  CD1 TYR A  36       5.675   3.963  10.681  1.00  0.00           C
ATOM    296  CD2 TYR A  36       4.424   6.040  10.616  1.00  0.00           C
ATOM    297  CE1 TYR A  36       6.332   4.393  11.840  1.00  0.00           C
ATOM    298  CE2 TYR A  36       5.083   6.471  11.773  1.00  0.00           C
ATOM    299  CZ  TYR A  36       6.043   5.652  12.379  1.00  0.00           C
ATOM    300  OH  TYR A  36       6.704   6.088  13.485  1.00  0.00           O
ATOM    301  OXT TYR A  36       1.276   2.139   7.885  1.00  0.00           O
HETATM  303 ZN    ZN A  37       1.119 -11.175  19.270  1.00  0.00          ZN
```

Appendix B: Input data for plot in Figure 10

PDBID	EV1	A	B	C	D	E	F	G	H	Outlier
1a4mA		0.04	-0.02							
2a3lA		0.03	0.00							
1j5sA		-0.06		0.21						
2qpxA		-0.03		0.15						
3iacA		-0.07		0.21						
1bf6A		-0.02			0.08					
2ob3A		-0.01			0.09					
2vc5A		-0.02			0.09					
2y1hB		-0.01			0.06					
3gg7A		-0.03			0.05					
3k2gB		-0.02			0.09					
1itqA		-0.04				0.13				
2dvtA		-0.07				0.22				
2ffiA		-0.04				0.12				
2gwgA		-0.05				0.21				
3cjpA		-0.04				0.12				
3irsA		-0.03				0.14				
4dlfA		-0.05				0.14				
4dziC		-0.05				0.21				
4hk5D		-0.08				0.22				
4mupB		-0.05				0.11				
4ofcA		-0.07				0.23				
4qrnA		-0.07				0.25				
1a5kC		0.09					0.01			
1gkpA		0.08					0.00			
1onxA		0.09					-0.04			
1yrrB		0.08					-0.05			
2vunA		0.09					-0.05			
3e74A		0.10					-0.02			
3giqA		0.10					-0.02			
3griA		0.09					-0.01			
3nqbA		0.09					-0.06			
3pnuA		0.01					0.08			
4b3zD		0.08					0.02			
1j6pA		0.14						-0.13		

1k6wA	0.10		-0.09			
2imrA	0.07		-0.07			
2ogjA	0.10		-0.04			
2oofA	0.11		-0.11			
2pajA	0.13		-0.11			
2uz9A	0.14		-0.13			
3icjA	0.11		-0.08			
3ls9A	0.13		-0.12			
3mkvA	0.12		-0.10			
3mtwA	0.12		-0.10			
3ooqA	0.13		-0.09			
4c5yA	0.12		-0.10			
4cqbA	0.11		-0.09			
4rdvB	0.14		-0.13			
4ub9A	0.11		-0.08			
4v1xE	0.14		-0.12			
1m65A	-0.18			-0.09		
1v77A	-0.04			-0.06		
3au2A	-0.20			-0.07		
3dcpA	-0.15			-0.08		
3qy6A	-0.08			0.02		
2anuA	-0.35				-0.23	
2yb1A	-0.38				-0.26	
3e38A	-0.46				-0.32	
3f2bA	-0.20				-0.13	
1bksA	-0.06					-0.01