1ab1
Evolutionary trace report by report_maker
April 30, 2010

CONTENTS
1 Introduction 1
2 Chain 1ab1A 1
   2.1 P01542 overview 1
   2.2 Multiple sequence alignment for 1ab1A 1
   2.3 Residue ranking in 1ab1A 1
   2.4 Top ranking residues in 1ab1A and their position on the structure 1
      2.4.1 Clustering of residues at 24% coverage. 2
      2.4.2 Possible novel functional surfaces at 24% coverage. 3
3 Notes on using trace results 3
   3.1 Coverage 3
   3.2 Known substitutions 3
   3.3 Surface 3
   3.4 Number of contacts 3
   3.5 Annotation 3
   3.6 Mutation suggestions 4
4 Appendix 4
   4.1 File formats 4
   4.2 Color schemes used 4
   4.3 Credits 4
      4.3.1 Alistat 4
      4.3.2 CE 4
   4.4 Note about ET Viewer 5
   4.5 Citing this work 5
   4.6 About report_maker 5
   4.7 Attachments 5

1 INTRODUCTION
From the original Protein Data Bank entry (PDB id 1ab1):
Title: Si form crambin
Compound: Mol id: 1; molecule: crambin (ser22/ile25); chain: a
Organism, scientific name: Crambe Abyssinica;
1ab1 contains a single unique chain 1ab1A (46 residues long).

2 CHAIN 1AB1A
2.1 P01542 overview
From SwissProt, id P01542, 95% identical to 1ab1A:
Description: Crambin.
Organism, scientific name: Crambe abyssinica (Abyssinian crambe).
Taxonomy: Eukaryota; Viridiplantae; Streptophyta; Embryophyta;
Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core
eudicotyledons; rosids; eumirläppige; Brassicales; Brassicaceae;
Crambe.
Function: The function of this hydrophobic plant seed protein is not
known.
Subcellular location: Secreted.
Miscellaneous: Two isoforms exists, a major form PL (shown here)
and a minor form SI.
Similarity: Belongs to the plant thionin (TC 1.C.44) family.
About: This Swiss-Prot entry is copyright. It is produced through a
collaboration between the Swiss Institute of Bioinformatics and the
EMBL outstation - the European Bioinformatics Institute. There are
no restrictions on its use as long as its content is in no way modified
and this statement is not removed.

2.2 Multiple sequence alignment for 1ab1A
For the chain 1ab1A, the alignment 1ab1A.msf (attached) with 41
sequences was used. The alignment was downloaded from the HSSP
database, and fragments shorter than 75% of the query as well as
duplicate sequences were removed. It can be found in the attachment
to this report, under the name of 1ab1A.msf. Its statistics, from the
alistat program are the following:
### 2.3 Residue ranking in 1ab1A

The 1ab1A sequence is shown in Fig. 1, with each residue colored according to its estimated importance. The full listing of residues in 1ab1A can be found in the file called 1ab1A.ranks, sorted in the attachment.

### 2.4 Top ranking residues in 1ab1A and their position on the structure

In the following we consider residues ranking among top 24% of residues in the protein (the closest this analysis allows us to get to 25%). Figure 2 shows residues in 1ab1A colored by their importance: bright red and yellow indicate more conserved/important residues (see Appendix for the coloring scheme). A Pymol script for producing this figure can be found in the attachment.

#### 2.4.1 Clustering of residues at 24% coverage.

Fig. 3 shows the top 24% of all residues, this time colored according to clusters they belong to. The clusters in Fig. 3 are composed of the residues listed in Table 1.

<table>
<thead>
<tr>
<th>cluster color</th>
<th>size</th>
<th>member residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>10</td>
<td>3, 4, 10, 13, 16, 26, 27, 32, 34, 40</td>
</tr>
</tbody>
</table>

Table 1. Clusters of top ranking residues in 1ab1A.

#### 2.4.2 Possible novel functional surfaces at 24% coverage.

One group of residues is conserved on the 1ab1A surface, away from (or substantially larger than) other functional sites and interfaces recognizable in PDB entry 1ab1. It is shown in Fig. 4. The right panel shows (in blue) the rest of the larger cluster this surface belongs to. The residues belonging to this surface "patch" are listed in Table 2, while Table 3 suggests possible disruptive replacements for these residues (see Section 3.6).
Fig. 4. A possible active surface on the chain 1ab1A. The larger cluster it belongs to is shown in blue.

Table 2. Residues forming surface "patch" in 1ab1A.

<table>
<thead>
<tr>
<th>res</th>
<th>type</th>
<th>substitutions(%)</th>
<th>cvg</th>
<th>antn</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>C</td>
<td>C(100)</td>
<td>0.07</td>
<td>S-S</td>
</tr>
<tr>
<td>20</td>
<td>G</td>
<td>G(95) V(2) L(2)</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>R</td>
<td>R(95) K(2) I(2)</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>A</td>
<td>A(92) I(4) C(2)</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>F</td>
<td>Y(87) F(12)</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>I</td>
<td>I (87) D(2) H(2)</td>
<td>0.22</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Disruptive mutations for the surface patch in 1ab1A.

<table>
<thead>
<tr>
<th>res</th>
<th>type</th>
<th>disruptive mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>C</td>
<td>(KER) (FQMWH) (NYLPI) (SVA)</td>
</tr>
<tr>
<td>20</td>
<td>G</td>
<td>(R) (KE) (H) (Y)</td>
</tr>
<tr>
<td>10</td>
<td>R</td>
<td>(T) (Y) (D) (S)</td>
</tr>
<tr>
<td>27</td>
<td>A</td>
<td>(R) (Y) (KE) (H)</td>
</tr>
<tr>
<td>13</td>
<td>F</td>
<td>(K) (E) (Q) (D)</td>
</tr>
<tr>
<td>34</td>
<td>I</td>
<td>(R) (Y) (H) (K)</td>
</tr>
</tbody>
</table>

3 NOTES ON USING TRACE RESULTS

3.1 Coverage

Trace results are commonly expressed in terms of coverage: the residue is important if its "coverage" is small - that is if it belongs to some small top percentage of residues [100% is all of the residues in a chain], according to trace. The ET results are presented in the form of a table, usually limited to top 25% percent of residues (or to some nearby percentage), sorted by the strength of the presumed evolutionary pressure. (i.e., the smaller the coverage, the stronger the pressure on the residue.) Starting from the top of that list, mutating a couple of residues should affect the protein somehow, with the exact effects to be determined experimentally.

3.2 Known substitutions

One of the table columns is "substitutions" - other amino acid types seen at the same position in the alignment. These amino acid types may be interchangeable at that position in the protein, so if one wants to affect the protein by a point mutation, they should be avoided. For example if the substitutions are "RVK" and the original protein has an R at that position, it is advisable to try anything, but RVK. Conversely, when looking for substitutions which will not affect the protein, one may try replacing, R with K, or (perhaps more surprisingly), with V. The percentage of times the substitution appears in the alignment is given in the immediately following bracket. No percentage is given in the cases when it is smaller than 1%. This is meant to be a rough guide - due to rounding errors these percentages often do not add up to 100%.

3.3 Surface

To detect candidates for novel functional interfaces, first we look for residues that are solvent accessible (according to DSSP program) by at least 10 Å², which is roughly the area needed for one water molecule to come in the contact with the residue. Furthermore, we require that these residues form a “cluster” of residues which have neighbor within 5 Å from any of their heavy atoms.

Note, however, that, if our picture of protein evolution is correct, the neighboring residues which are not surface accessible might be equally important in maintaining the interaction specificity - they should not be automatically dropped from consideration when choosing the set for mutagenesis. (Especially if they form a cluster with the surface residues.)

3.4 Number of contacts

Another column worth noting is denoted “noc/bb”; it tells the number of contacts heavy atoms of the residue in question make across the interface, as well as how many of them are realized through the backbone atoms (if all or most contacts are through the backbone, mutation presumably won’t have strong impact). Two heavy atoms are considered to be “in contact” if their centers are closer than 5 Å.

3.5 Annotation

If the residue annotation is available (either from the pdb file or from other sources), another column, with the header “annotation” appears. Annotations carried over from PDB are the following: site (indicating existence of related site record in PDB), S-S (disulfide bond forming residue), hb (hydrogen bond forming residue, jb (james bond forming residue), and sb (for salt bridge forming residue).

3.6 Mutation suggestions

Mutation suggestions are completely heuristic and based on complementarity with the substitutions found in the alignment. Note that they are meant to be disruptive to the interaction of the protein with its ligand. The attempt is made to complement the following properties: small [AVGSTC], medium [LPNQDEMIK], large [WFYHR], hydrophobic [LPVAMWFI], polar [GTCY]; positively [KHR], or negatively [DE] charged, aromatic [WFYH]; long aliphatic chain [EKQRMF], OH-group possession [SDETY], and NH2 group possession [NQRK]. The suggestions are listed according to how different they appear to be from the original amino acid, and they are grouped in round brackets if they appear equally disruptive. From left to right, each bracketed group of amino acid types resembles more strongly the original (i.e. is, presumably, less disruptive) These suggestions are tentative - they might prove disruptive to the fold rather than to the interaction. Many researcher will
choose, however, the straightforward alanine mutations, especially in the beginning stages of their investigation.

4 APPENDIX
4.1 File formats

Files with extension “ranks_sorted” are the actual trace results. The fields in the table in this file:

- **alignment#** number of the position in the alignment
- **residue#** residue number in the PDB file
- **type** amino acid type
- **rank** rank of the position according to older version of ET
- **variability** has two subfields:
  1. number of different amino acids appearing in this column of the alignment
  2. their type
- **rho** ET score - the smaller this value, the lesser variability of this position across the branches of the tree (and, presumably, the greater importance for the protein)
- **cvg** coverage - percentage of the residues on the structure which have this rho or smaller
- **gaps** percentage of gaps in this column

4.2 Color schemes used

The following color scheme is used in figures with residues colored by cluster size: black is a single-residue cluster; clusters composed of more than one residue colored according to this hierarchy (ordered by descending size): red, blue, yellow, green, purple, azure, turquoise, brown, coral, magenta, LightSalmon, SkyBlue, violet, gold, bisque, LightSlateBlue, orchid, RosyBrown, MediumAquamarine, DarkOliveGreen, CornflowerBlue, grey55, burlywood, LimeGreen, tan, DarkOrange, DeepPink, maroon, BlanchedAlmond.

The colors used to distinguish the residues by the estimated evolutionary pressure they experience can be seen in Fig. 5.

4.3 Credits

4.3.1 **Alistat** **alistat** reads a multiple sequence alignment from the file and shows a number of simple statistics about it. These statistics include the format, the number of sequences, the total number of residues, the average and range of the sequence lengths, and the alignment length (e.g. including gap characters). Also shown are some percent identities. A percent pairwise alignment identity is defined as (idents / MIN(len1, len2)) where idents is the number of exact identities and len1, len2 are the unaligned lengths of the two sequences. The “average percent identity”, "most related pair", and "most unrelated pair" of the alignment are the average, maximum, and minimum of all (N)(N-1)/2 pairs, respectively. The “most distant seq” is calculated by finding the maximum pairwise identity (best relative) for all N sequences, then finding the minimum of these N numbers (hence, the most outlying sequence). **alistat** is copyrighted by HHMI/Washington University School of Medicine, 1992-2001, and freely distributed under the GNU General Public License.

4.3.2 **CE** To map ligand binding sites from different source structures, **report_make** uses the CE program: [http://cl.sdsc.edu/](http://cl.sdsc.edu/) Shindyalov IN, Bourne PE (1998) “Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.” Protein Engineering 11(9) 739-747.

4.3.3 **DSSP** In this work a residue is considered solvent accessible if the DSSP program finds it exposed to water by at least 10Å², which is roughly the area needed for one water molecule to come in the contact with the residue. **DSSP** is copyrighted by W. Kabsch, C. Sander and MPI-MF, 1983, 1985, 1988, 1994 1995, CMBI version by Elmar.Krieger@cmbi.kun.nl November 18, 2002, [http://www.cmbi.kun.nl/gv/dssp/descrip.html](http://www.cmbi.kun.nl/gv/dssp/descrip.html).

4.3.4 **HSSP** Whenever available, **report_make** uses HSSP alignment as a starting point for the analysis (sequences shorter than 75% of the query are taken out, however); R. Schneider, A. de Daruvar, and C. Sander. “The HSSP database of protein structure-sequence alignments.” Nucleic Acids Res., 25:226–230, 1997. [http://swift.cmbi.kun.nl/swift/hssp/](http://swift.cmbi.kun.nl/swift/hssp/).

4.3.5 **LaTeX** The text for this report was processed using **LATEX**: Leslie Lamport, “**LaTeX: A Document Preparation System Addison-Wesley,” Reading, Mass. (1986).”


4.3.7 **Pymol** The figures in this report were produced using **Pymol**. The scripts can be found in the attachment. **Pymol** is an open-source application copyrighted by DeLano Scientific LLC (2005). For more information about **Pymol** see [http://pymol.sourceforge.net/](http://pymol.sourceforge.net/). (Note for Windows users: the attached package needs to be unzipped for Pymol to read the scripts and launch the viewer.)
4.4 Note about ET Viewer

Dan Morgan from the Lichtarge lab has developed a visualization tool specifically for viewing trace results. If you are interested, please visit:

http://mammoth.bcm.tmc.edu/traceview/

The viewer is self-unpacking and self-installing. Input files to be used with ETV (extension .etvx) can be found in the attachment to the main report.

4.5 Citing this work


4.6 About report maker

report maker was written in 2006 by Ivana Mihalek. The 1D ranking visualization program was written by Ivica Reš. report maker is copyrighted by Lichtarge Lab, Baylor College of Medicine, Houston.

4.7 Attachments

The following files should accompany this report:

- lab1A.complex.pdb - coordinates of lab1A with all of its interacting partners
- lab1A.etvx - ET viewer input file for lab1A
- lab1A.cluster_report.summary - Cluster report summary for lab1A
- lab1A.ranks - Ranks file in sequence order for lab1A
- lab1A.clusters - Cluster descriptions for lab1A
- lab1A.msf - the multiple sequence alignment used for the chain lab1A
- lab1A.descr - description of sequences used in lab1A msf
- lab1A.ranks_sorted - full listing of residues and their ranking for lab1A