1a3f
Evolutionary trace report by report_maker
September 18, 2008

CONTENTS

1 Introduction

2 Chain 1a3fA
   2.1 Q65ZF5 overview
   2.2 Multiple sequence alignment for 1a3fA
   2.3 Residue ranking in 1a3fA
   2.4 Top ranking residues in 1a3fA and their position on the structure
      2.4.1 Clustering of residues at 25% coverage.
      2.4.2 Overlap with known functional surfaces at 25% coverage.
      2.4.3 Possible novel functional surfaces at 25% coverage.

3 Notes on using trace results
   3.1 Coverage
   3.2 Known substitutions
   3.3 Surface
   3.4 Number of contacts
   3.5 Annotation
   3.6 Mutation suggestions

4 Appendix
   4.1 File formats
   4.2 Color schemes used
   4.3 Credits

1 INTRODUCTION

From the original Protein Data Bank entry (PDB id 1a3f):
Title: Phospholipase a2 (pla2) from naja naja venom
Compound: Mol id: 1; molecule: phospholipase a2; chain: a, b, c;
ec: 3.1.1.4
Organism, scientific name: Naja Naja;
1a3f contains a single unique chain 1a3fA (119 residues long) and
its homologues 1a3fC and 1a3fB.

2 CHAIN 1A3FA

2.1 Q65ZF5 overview
From SwissProt, id Q65ZF5, 100% identical to 1a3fA:
Description: Phospholipase a2 (EC 3.1.1.4).
Taxonomy: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
Euteleostomi; Lepidosauria; Squamata; Scleroglossa; Serpentes;
Colubroidea; Elapidae; Elapinae; Naja.

2.2 Multiple sequence alignment for 1a3fA

For the chain 1a3fA, the alignment 1a3fA.msf (attached) with 72
sequences was used. The alignment was assembled through combi-
nation of BLAST searching on the UniProt database and alignment
using Muscle program. It can be found in the attachment to this
report, under the name of 1a3fA.msf. Its statistics, from the alistat
program are the following:
Fig. 1. Residues 1-119 in 1a3fA colored by their relative importance. (See Appendix, Fig. 6, for the coloring scheme.)

<table>
<thead>
<tr>
<th>Format:</th>
<th>MSF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sequences:</td>
<td>72</td>
</tr>
<tr>
<td>Total number of residues:</td>
<td>8434</td>
</tr>
<tr>
<td>Smallest:</td>
<td>112</td>
</tr>
<tr>
<td>Largest:</td>
<td>119</td>
</tr>
<tr>
<td>Average length:</td>
<td>117.1</td>
</tr>
<tr>
<td>Alignment length:</td>
<td>119</td>
</tr>
<tr>
<td>Average identity:</td>
<td>54%</td>
</tr>
<tr>
<td>Most related pair:</td>
<td>98%</td>
</tr>
<tr>
<td>Most unrelated pair:</td>
<td>35%</td>
</tr>
<tr>
<td>Most distant seq:</td>
<td>67%</td>
</tr>
</tbody>
</table>

Furthermore, 13% of residues show as conserved in this alignment. The alignment consists of 98% eukaryotic (98% vertebrata) sequences. (Descriptions of some sequences were not readily available.) The file containing the sequence descriptions can be found in the attachment, under the name 1a3fA.descr.

2.3 Residue ranking in 1a3fA

The 1a3fA sequence is shown in Fig. 1, with each residue colored according to its estimated importance. The full listing of residues in 1a3fA can be found in the file called 1a3fA.ranks sorted in the attachment.

2.4 Top ranking residues in 1a3fA and their position on the structure

In the following we consider residues ranking among top 25% of residues in the protein. Figure 2 shows residues in 1a3fA 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 25% coverage. Fig. 3 shows the top 25% 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 1. Clusters of top ranking residues in 1a3fA.

<table>
<thead>
<tr>
<th>cluster color</th>
<th>size</th>
<th>member residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>29</td>
<td>5, 8, 9, 21, 24, 25, 26, 28, 36, 38, 40, 41, 43, 44, 47, 50, 51, 60, 62, 67, 85, 90, 92, 93, 96, 97, 99, 100, 105</td>
</tr>
</tbody>
</table>

2.4.2 Overlap with known functional surfaces at 25% coverage. The name of the ligand is composed of the source PDB identifier and the heteroatom name used in that file.

Interface with 1a3fC. Table 2 lists the top 25% of residues at the interface with 1a3fC. The following table (Table 3) suggests possible disruptive replacements for these residues (see Section 3.6).
Table 2. The top 25% of residues in 1a3fA at the interface with 1a3fC. (Field names: res: residue number in the PDB entry; type: amino acid type; subst’s: substitutions seen in the alignment, with the percentage of each type in the bracket; noc/bb: number of contacts with the ligand, with the number of contacts realized through backbone atoms given in the bracket; dist: distance of closest approach to the ligand.)

<table>
<thead>
<tr>
<th>res</th>
<th>type</th>
<th>subst’s</th>
<th>cvg (%)</th>
<th>noc/bb</th>
<th>dist (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>Y</td>
<td>Y(100)</td>
<td>0.13</td>
<td>2/0</td>
<td>4.56</td>
</tr>
<tr>
<td>62</td>
<td>P</td>
<td>P(100)</td>
<td>0.13</td>
<td>24/12</td>
<td>3.42</td>
</tr>
</tbody>
</table>

Table 3. List of disruptive mutations for the top 25% of residues in 1a3fA, that are at the interface with 1a3fC.

<table>
<thead>
<tr>
<th>res</th>
<th>type</th>
<th>disruptive mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>Y</td>
<td>(K) (QM) (NVLAP) (D)</td>
</tr>
<tr>
<td>62</td>
<td>P</td>
<td>(Y) (TH) (SKECG) (FQWD)</td>
</tr>
</tbody>
</table>

Figure 4 shows residues in 1a3fA colored by their importance, at the interface with 1a3fC.

2.4.3 Possible novel functional surfaces at 25% coverage. One group of residues is conserved on the 1a3fA surface, away from (or substantially larger than) other functional sites and interfaces recognizable in PDB entry 1a3f. It is shown in Fig. 5. The right panel shows (in blue) the rest of the larger cluster this surface belongs to.
Fig. 5. A possible active surface on the chain 1a3fA. The larger cluster it belongs to is shown in blue.

The residues belonging to this surface "patch" are listed in Table 4, while Table 5 suggests possible disruptive replacements for these residues (see Section 3.6).

### Table 4.

<table>
<thead>
<tr>
<th>res</th>
<th>type</th>
<th>substitutions(%)</th>
<th>cvg</th>
<th>antn</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>C</td>
<td>C(100)</td>
<td>0.13</td>
<td>S-S</td>
</tr>
<tr>
<td>43</td>
<td>C</td>
<td>C(100)</td>
<td>0.13</td>
<td>S-S</td>
</tr>
<tr>
<td>44</td>
<td>C</td>
<td>C(100)</td>
<td>0.13</td>
<td>S-S</td>
</tr>
<tr>
<td>50</td>
<td>C</td>
<td>C(100)</td>
<td>0.13</td>
<td>S-S</td>
</tr>
<tr>
<td>51</td>
<td>Y</td>
<td>Y(100)</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>C</td>
<td>C(100)</td>
<td>0.13</td>
<td>S-S</td>
</tr>
<tr>
<td>62</td>
<td>P</td>
<td>P(100)</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>99</td>
<td>C</td>
<td>C(100)</td>
<td>0.13</td>
<td>S-S</td>
</tr>
<tr>
<td>100</td>
<td>F</td>
<td>F(100)</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>I</td>
<td>I(98)V(1)</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>67</td>
<td>Y</td>
<td>Y(98)C(1)</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>105</td>
<td>Y</td>
<td>Y(95)W(2)F(1)</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>96</td>
<td>A</td>
<td>A(98)T(1)</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td>F(98)Y(1)</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>H</td>
<td>H(98)N(1)</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>85</td>
<td>C</td>
<td>C(98)S(1)</td>
<td>0.23</td>
<td>S-S</td>
</tr>
<tr>
<td>92</td>
<td>C</td>
<td>C(98)Y(1)</td>
<td>0.23</td>
<td>S-S</td>
</tr>
<tr>
<td>21</td>
<td>F</td>
<td>Y(79)F(20)</td>
<td>0.25</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Residues forming surface "patch" in 1a3fA.

### Table 5.

<table>
<thead>
<tr>
<th>res</th>
<th>type</th>
<th>disruptive mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>C</td>
<td>(KER) (FQMWH) (NYLPI) (SVA)</td>
</tr>
<tr>
<td>43</td>
<td>C</td>
<td>(KER) (FQMWH) (NYLPI) (SVA)</td>
</tr>
<tr>
<td>44</td>
<td>C</td>
<td>(KER) (FQMWH) (NYLPI) (SVA)</td>
</tr>
<tr>
<td>50</td>
<td>C</td>
<td>(KER) (FQMWH) (NYLPI) (SVA)</td>
</tr>
<tr>
<td>51</td>
<td>Y</td>
<td>(K) (QM) (NEVLAP) (R) (D)</td>
</tr>
<tr>
<td>60</td>
<td>C</td>
<td>(KER) (FQMWH) (NYLPI) (SVA)</td>
</tr>
<tr>
<td>62</td>
<td>P</td>
<td>(YR) (TH) (SKECG) (FQWD)</td>
</tr>
<tr>
<td>99</td>
<td>C</td>
<td>(KER) (FQMWH) (NYLPI) (SVA)</td>
</tr>
<tr>
<td>100</td>
<td>F</td>
<td>(KE) (TQD) (SNCRG) (M)</td>
</tr>
<tr>
<td>9</td>
<td>I</td>
<td>(YR) (H) (TRE) (SQCDCG)</td>
</tr>
</tbody>
</table>

Table 5. Disruptive mutations for the surface patch in 1a3fA.

### 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 5Å.

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 (disulde
bond forming residue), hb (hydrogen bond forming residue), jh (james
bond forming residue), and sb (for salt bridge forming residue).

3.6 Mutation suggestions

Mutation suggestions are completely heuristic and based on comple-
tmentarity 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 [LPNQDEM], large
[WFYHR], hydrophobic [LPVAMWF], polar [GTCY]; positively
[KHR], or negatively [DE] charged, aromatic [WFYH],
long aliphatic chain [EKRQM], OH-group possession [SDETY],
and NH2 group possession [NQKR]. 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 disrupt-
tive 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 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 the 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

Fig. 6. Coloring scheme used to color residues by their relative importance.

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, tur-
quoise, brown, coral, magenta, LightSalmon, SkyBlue, violet, gold,
biqus, 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. 6.

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 stati-
atics 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 de-
finied 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
maker uses the CE program: 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.

http://swift.cmbi.kun.nl/swift/hssp/

4.3.5 LaTeX  The text for this report was processed using L\LaTeX; Leslie Lamport, "LaTeX: A Document Preparation System Addison-Wesley," Reading, Mass. (1986).


http://www.drive5.com/muscle/

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/. (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 Řeš. report_maker is copyrighted by Lichtarge Lab, Baylor College of Medicine, Houston.

4.7 Attachments

The following files should accompany this report:

- 1a3fA.complex.pdb - coordinates of 1a3fA with all of its interacting partners
- 1a3fA.etvx - ET viewer input file for 1a3fA
- 1a3fA.cluster_report.summary - Cluster report summary for 1a3fA
- 1a3fA.ranks - Ranks file in sequence order for 1a3fA
- 1a3fA.clusters - Cluster descriptions for 1a3fA
- 1a3fA.msf - the multiple sequence alignment used for the chain 1a3fA
- 1a3fA.descr - description of sequences used in 1a3fA msf
- 1a3fA.ranks_sorted - full listing of residues and their ranking for 1a3fA
- 1a3fA.1a3fC.if.pml - Pymol script for Figure 4
- 1a3fA.cbcvg - used by other 1a3fA – related pymol scripts