1 INTRODUCTION

From the original Protein Data Bank entry (PDB id 1a8k):

**Title:** Crystallographic analysis of human immunodeficiency virus 1 protease with an analog of the conserved ca-p2 substrate: interactions with frequently occurring glutamic acid residue at p2’ position of substrates

**Compound:** Mol id: 1; molecule: hiv protease; chain: a, b, d, e; ec: 3.4.23.16; engineered: yes; mutation: yes; mol id: 2; molecule: peptide; chain: c, f; engineered: yes; other details: analog of the conserved ca-p2 substrate

**Organism, scientific name:** Human Immunodeficiency Virus 1:

1a8k contains a single unique chain 1a8kA (99 residues long) and its homologues 1a8kD, 1a8kE, and 1a8kB. Chains 1a8kC and 1a8kF are too short to permit statistically significant analysis, and were treated as a peptide ligands.

2 CHAIN 1A8KA

2.1 Q9EKL4 overview

From SwissProt, id Q9EKL4, 96% identical to 1a8kA:

**Description:** Pol protein (Fragment).

**Organism, scientific name:** Human Immunodeficiency Virus 1.

2.2 Multiple sequence alignment for 1a8kA

For the chain 1a8kA, the alignment 1a8kA.msf (attached) with 230 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 1a8kA.msf. Its statistics, from the alstat program are the following:
Furthermore, 64% of residues show as conserved in this alignment. The alignment consists of 62% viral sequences. (Descriptions of some sequences were not readily available.) The file containing the sequence descriptions can be found in the attachment, under the name 1a8kA.descr.

### 2.3 Residue ranking in 1a8kA

The 1a8kA sequence is shown in Fig. 1, with each residue colored according to its estimated importance. The full listing of residues in 1a8kA can be found in the file called 1a8kA.ranks_sorted in the attachment.

### 2.4 Top ranking residues in 1a8kA and their position on the structure

In the following we consider residues ranking among top 65% of residues in the protein (the closest this analysis allows us to get to 25%). Figure 2 shows residues in 1a8kA 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 65% coverage

Fig. 3 shows the top 65% 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</th>
<th>size</th>
<th>member residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>red</td>
<td>59</td>
<td>8, 9, 11, 16, 17, 18, 21, 22, 23, 24 25, 26, 27, 28, 29, 30, 31, 32, 34 38, 39, 40, 42, 43, 44, 45, 46, 48 49, 51, 52, 53, 55, 56, 59, 60, 65 67, 68, 69, 71, 73, 74, 76, 78, 79 80, 81, 83, 86, 88, 91, 92, 94, 95</td>
</tr>
</tbody>
</table>

**Table 1. continued**

<table>
<thead>
<tr>
<th>cluster</th>
<th>size</th>
<th>member residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td>5</td>
<td>96, 97, 98, 99</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1, 2, 4, 5, 6</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 surface accessible might be 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 [LPNQDEM1K], large [W FYHR], hydrophobic [LPV AMWF], polar [GTCY]; positively [KHR], or negatively [DE] charged, aromatic [W FYH], long aliphatic chain [ERQKM], 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 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

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, Light Slate Blue, 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. 4.
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 range and 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_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 10 Å, which is roughly the area needed for one water molecule to come in 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@cmbHg.de November 18, 2002, http://www.cmbi.kun.nl/gv/dssp/descrip.html.


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).


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 Mikhail. The 1D ranking visualization program was written by Ivica Res. report_maker is copyrighted by Lichtarge Lab, Baylor College of Medicine, Houston.

4.7 Attachments

The following files should accompany this report:

- 1a8kA.complex.pdb - coordinates of 1a8kA with all of its interacting partners
- 1a8kA.etvx - ET viewer input file for 1a8kA
- 1a8kA.cluster_report.summary - Cluster report summary for 1a8kA
- 1a8kA.ranks - Ranks file in sequence order for 1a8kA
- 1a8kA.clusters - Cluster descriptions for 1a8kA
- 1a8kA.msf - the multiple sequence alignment used for the chain 1a8kA
- 1a8kA.cluster - Cluster description for 1a8kA
- 1a8kA.descr - description of sequences used in 1a8kA msf
- 1a8kA.ranks_sorted - full listing of residues and their ranking for 1a8kA