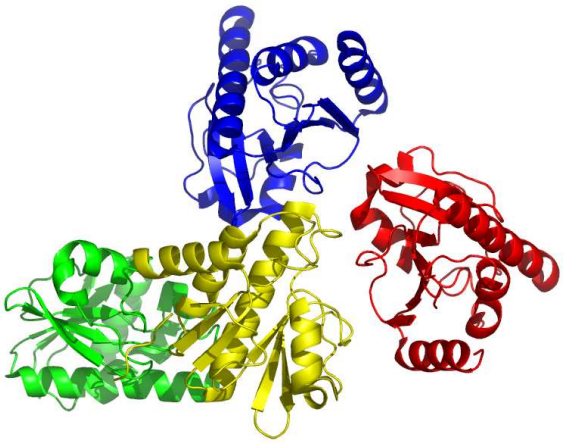
3gpg

Evolutionary trace report by **report_maker**

January 31, 2010



CONTENTS

1 INTRODUCTION

From the original Protein Data Bank entry (PDB id 3gpg):
Title: Crystal structure of macro domain of chikungunya virus
Compound: Mol id: 1; molecule: non-structural protein 3; chain: a, b, c, d; fragment: sequence database residues 1334-1493; synonym: nsp3; engineered: yes
Organism, scientific name: Chikungunya Virus;

3gpg contains a single unique chain 3gpgA (162 residues long) and its homologues 3gpgD, 3gpgC, and 3gpgB.

2 CHAIN 3GPGA

2.1 Q8JUX6 overview

From SwissProt, id Q8JUX6, 100% identical to 3gpgA:
Description: Nonstructural polyprotein.
Organism, scientific name: Chikungunya virus.
Taxonomy: Viruses; ssRNA positive-strand viruses, no DNA stage; Togaviridae; Alphavirus; SFV complex.

2.2 Multiple sequence alignment for 3gpgA

For the chain 3gpgA, the alignment 3gpgA.msf (attached) with 21 sequences was used. The alignment was assembled through combination 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 3gpgA.msf. Its statistics, from the *alistat* program are the following:

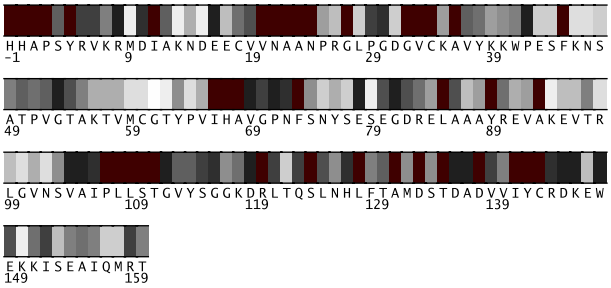


Fig. 1. Residues -1-160 in 3gpgA colored by their relative importance. (See Appendix, Fig.7, for the coloring scheme.)

```
Format:               MSF
Number of sequences: 21
Total number of residues:    3400
Smallest:             161
Largest:              162
Average length:       161.9
Alignment length:     162
Average identity:     60%
Most related pair:    99%
Most unrelated pair: 43%
Most distant seq:     66%
```

Furthermore, 24% of residues show as conserved in this alignment.

The alignment consists of 95% 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 3gpgA.descr.

2.3 Residue ranking in 3gpgA

The 3gpgA sequence is shown in Fig. 1, with each residue colored according to its estimated importance. The full listing of residues in 3gpgA can be found in the file called 3gpgA.ranks_sorted in the attachment.

2.4 Top ranking residues in 3gpgA 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 3gpgA 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 1.

cluster color	size	member residues
red	34	4,11,20,21,22,23,24,27,32,33 34,36,45,66,67,68,73,85,89

continued in next column

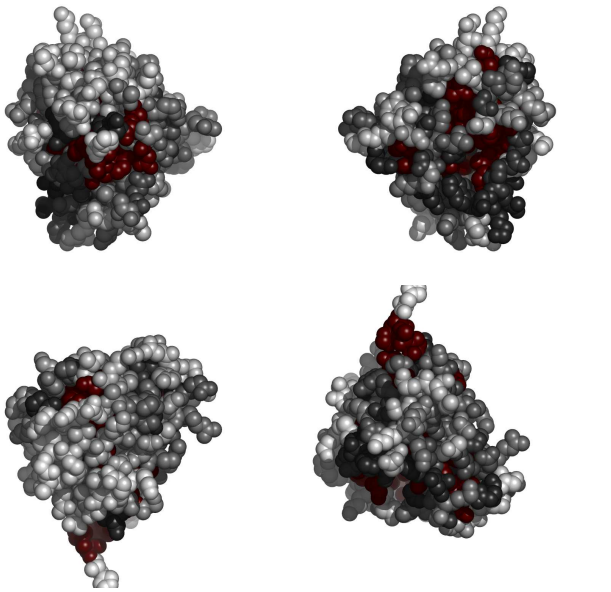


Fig. 2. Residues in 3gpgA, colored by their relative importance. Clockwise: front, back, top and bottom views.

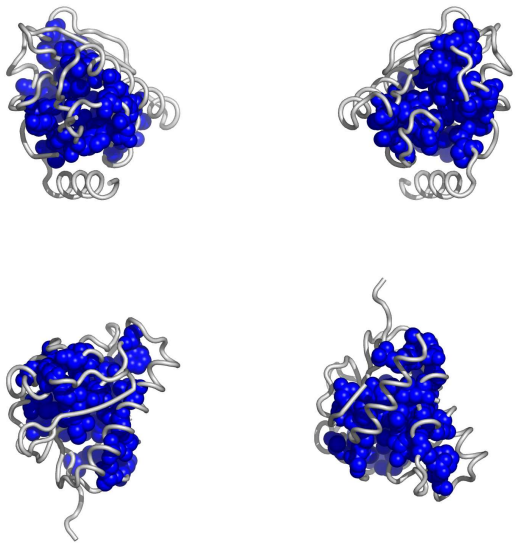


Fig. 3. Residues in 3gpgA, colored according to the cluster they belong to: red, followed by blue and yellow are the largest clusters (see Appendix for the coloring scheme). Clockwise: front, back, top and bottom views. The corresponding Pymol script is attached.

Table 1. *continued*

cluster color	size	member residues
		93,107,108,109,110,111,120 124,128,131,133,135,141,142 143

continued in next column

Table 1. *continued*		
cluster color	**size**	**member residues**
blue	5	-1,0,1,2,138

Table 1. Clusters of top ranking residues in 3gpgA.

2.4.2 Overlap with known functional surfaces at 24% coverage. The name of the ligand is composed of the source PDB identifier and the heteroatom name used in that file.

Interface with 3gpgD. By analogy with 3gpgC – 3gpgD interface. Table 2 lists the top 24% of residues at the interface with 3gpgD. The following table (Table 3) suggests possible disruptive replacements for these residues (see Section 3.6).

Table 2.					
res	**type**	**subst's (%)**	**cvg**	**noc/ bb**	**dist (Å)**
-1	H	H(100)	0.24	35/10	3.13
0	H	H(100)	0.24	10/10	4.06
1	A	A(100)	0.24	4/3	4.36
2	P	P(100)	0.24	35/9	3.54
4	Y	Y(100)	0.24	24/1	3.90
93	A	A(100)	0.24	17/13	3.53
131	A	A(100)	0.24	2/2	4.45
133	D	D(100)	0.24	34/25	3.01
135	T	T(100)	0.24	31/24	2.83

Table 2. The top 24% of residues in 3gpgA at the interface with 3gpgD. (Field names: res: residue number in the PDB entry; type: amino acid type; substs: 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 apporach to the ligand.)

Table 3.		
res	**type**	**disruptive mutations**
-1	H	(E)(TQMD)(SNKVCLAPIG)(YR)
0	H	(E)(TQMD)(SNKVCLAPIG)(YR)
1	A	(KYER)(QHD)(N)(FTMW)
2	P	(YR)(TH)(SKECG)(FQWD)
4	Y	(K)(QM)(NEVLAPIR)(D)
93	A	(KYER)(QHD)(N)(FTMW)
131	A	(KYER)(QHD)(N)(FTMW)
133	D	(R)(FWH)(KYVCAG)(TQM)
135	T	(KR)(FQMWH)(NELPI)(D)

Table 3. List of disruptive mutations for the top 24% of residues in 3gpgA, that are at the interface with 3gpgD.

Figure 4 shows residues in 3gpgA colored by their importance, at the interface with 3gpgD.

Interface with 3gpgC. By analogy with 3gpgD – 3gpgC interface. Table 4 lists the top 24% of residues at the interface with 3gpgC. The following table (Table 5) suggests possible disruptive replacements for these residues (see Section 3.6).

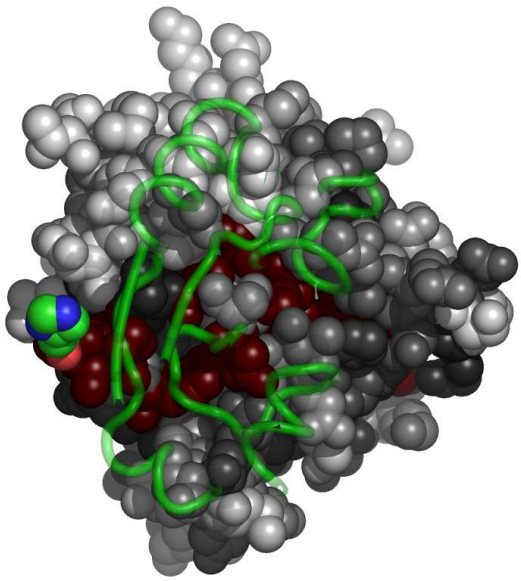


Fig. 4. Residues in 3gpgA, at the interface with 3gpgD, colored by their relative importance. 3gpgD is shown in backbone representation (See Appendix for the coloring scheme for the protein chain 3gpgA.)

Table 4.					
res	type	subst's (%)	cvg	noc/ bb	dist (Å)
24	N	N(100)	0.24	2/0	4.58

Table 4. The top 24% of residues in 3gpgA at the interface with 3gpgC. (Field names: res: residue number in the PDB entry; type: amino acid type; substs: 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 apporach to the ligand.)

Table 5.		
res	type	disruptive mutations
24	N	(Y)(FTWH)(SEVCARG)(MD)

Table 5. List of disruptive mutations for the top 24% of residues in 3gpgA, that are at the interface with 3gpgC.

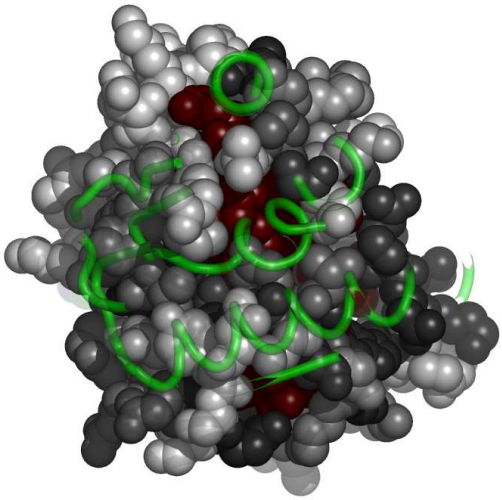


Fig. 5. Residues in 3gpgA, at the interface with 3gpgC, colored by their relative importance. 3gpgC is shown in backbone representation (See Appendix for the coloring scheme for the protein chain 3gpgA.)

Figure 5 shows residues in 3gpgA colored by their importance, at the interface with 3gpgC.

2.4.3 Possible novel functional surfaces at 24% coverage. One group of residues is conserved on the 3gpgA surface, away from (or susbtantially larger than) other functional sites and interfaces recognizable in PDB entry 3gpg. It is shown in Fig. 6. 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 6, while Table 7 suggests possible disruptive replacements for these residues (see Section 3.6).

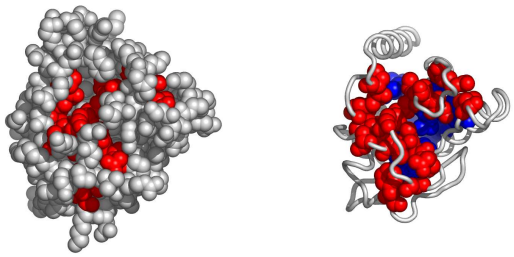


Fig. 6. A possible active surface on the chain 3gpgA. The larger cluster it belongs to is shown in blue.

Table 6.			
res	type	substitutions(%)	cvg
11	I	I(100)	0.24
22	A	A(100)	0.24
23	A	A(100)	0.24
24	N	N(100)	0.24
27	G	G(100)	0.24
32	G	G(100)	0.24
33	V	V(100)	0.24
34	C	C(100)	0.24
36	A	A(100)	0.24
45	F	F(100)	0.24
73	F	F(100)	0.24
107	P	P(100)	0.24
108	L	L(100)	0.24
110	S	S(100)	0.24
111	T	T(100)	0.24
120	R	R(100)	0.24
142	Y	Y(100)	0.24
143	C	C(100)	0.24

Table 6. Residues forming surface "patch" in 3gpgA.

Table 7.		
res	type	disruptive mutations
11	I	(YR)(TH)(SKECG)(FQWD)
22	A	(KYER)(QHD)(N)(FTMW)
23	A	(KYER)(QHD)(N)(FTMW)
24	N	(Y)(FTWH)(SEVCARG)(MD)
27	G	(KER)(FQMWHD)(NYLPI)(SVA)
32	G	(KER)(FQMWHD)(NYLPI)(SVA)
33	V	(KYER)(QHD)(N)(FTMW)
34	C	(KER)(FQMWHD)(NYLPI)(SVA)
36	A	(KYER)(QHD)(N)(FTMW)
45	F	(KE)(TQD)(SNCRG)(M)
73	F	(KE)(TQD)(SNCRG)(M)
		continued in next column

Table 7. *continued*

res	type	disruptive mutations
107	P	(YR)(TH)(SKECG)(FQWD)
108	L	(YR)(TH)(SKECG)(FQWD)
110	S	(KR)(FQMWH)(NYELPI)(D)
111	T	(KR)(FQMWH)(NELPI)(D)
120	R	(TD)(SYEVCLAPIG)(FMW)(N)
142	Y	(K)(QM)(NEVLAPIR)(D)
143	C	(KER)(FQMWHD)(NYLPI)(SVA)

Table 7. Disruptive mutations for the surface patch in 3gpgA.

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 *n*ot 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 $10\mathring{A}^2$, 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 $5\mathring{A}$ 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 $5\mathring{A}$.

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 $[EKRQM]$, 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 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

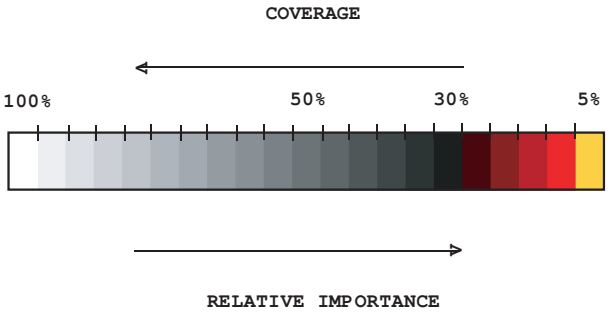


Fig. 7. Coloring scheme used to color residues by their relative importance.

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, 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. 7.

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_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 $10 Å^2$, 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.`

4.3.4 **HSSP** Whenever available, report_maker 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/`

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).

4.3.6 **Muscle** When making alignments "from scratch", report maker uses Muscle alignment program: Edgar, Robert C. (2004), *"MUSCLE: multiple sequence alignment with high accuracy and high throughput."* Nucleic Acids Research 32(5), 1792-97.

`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

The method used to rank residues and make predictions in this report can be found in Mihalek, I., I. Reš, O. Lichtarge. (2004). *"A Family of Evolution-Entropy Hybrid Methods for Ranking of Protein Residues by Importance"* J. Mol. Bio. **336**: 1265-82. For the original version of ET see O. Lichtarge, H.Bourne and F. Cohen (1996). *"An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families"* J. Mol. Bio. **257**: 342-358.

report_maker itself is described in Mihalek I., I. Res and O. Lichtarge (2006). *"Evolutionary Trace Report Maker: a new type of service for comparative analysis of proteins."* Bioinformatics **22**:1656-7.

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:

- 3gpgA.complex.pdb - coordinates of 3gpgA with all of its interacting partners
- 3gpgA.etvx - ET viewer input file for 3gpgA
- 3gpgA.cluster_report.summary - Cluster report summary for 3gpgA
- 3gpgA.ranks - Ranks file in sequence order for 3gpgA
- 3gpgA.clusters - Cluster descriptions for 3gpgA
- 3gpgA.msf - the multiple sequence alignment used for the chain 3gpgA
- 3gpgA.descr - description of sequences used in 3gpgA msf
- 3gpgA.ranks_sorted - full listing of residues and their ranking for 3gpgA
- 3gpgA.3gpgD.if.pml - Pymol script for Figure 4
- 3gpgA.cbcvg - used by other 3gpgA – related pymol scripts
- 3gpgA.3gpgC.if.pml - Pymol script for Figure 5