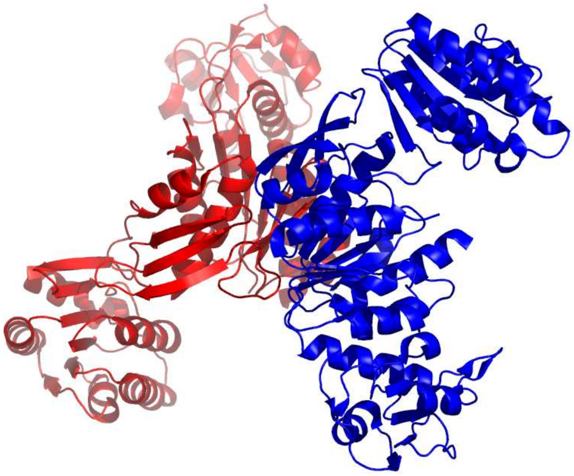
1gg4

Evolutionary trace report by **report_maker**

July 17, 2010



CONTENTS

1 INTRODUCTION

From the original Protein Data Bank entry (PDB id 1gg4):
Title: Crystal structure of escherichia coli udpmurnac-tripeptide d-alanyl-d-alanine-adding enzyme (murf) at 2.3 angstrom resolution
Compound: Mol id: 1; molecule: udp-n-acetylmuramoylalanyl-d-glutamyl-2,6- diaminopimelate-d-alanyl-d-alanyl ligase; chain: a, b; synonym: d-alanyl-d-alanine-adding enzyme, udp-murnac- pentapeptide synthetase; ec: 6.3.2.15; engineered: yes
Organism, scientific name: Escherichia Coli;

1gg4 contains a single unique chain 1gg4A (439 residues long) and its homologue 1gg4B.

2 CHAIN 1GG4A

2.1 P11880 overview

From SwissProt, id P11880, 90% identical to 1gg4A:
Description: UDP-N-acetylmuramoyl-tripeptide–D-alanyl-D-alanine ligase (EC 6.3.2.10) (UDP-MurNAc-pentapeptide synthetase) (D-alanyl-D- alanine-adding enzyme).
Organism, scientific name: Escherichia coli.
Taxonomy: Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia.
Function: Involved in cell wall formation. Catalyzes the final step in the synthesis of UDP-N-acetylmuramoyl-pentapeptide, the precursor of murein.
Catalytic activity: ATP + UDP-N-acetylmuramoyl-L-alanyl-gamma-D- glutamyl-L-lysine + D-alanyl-D-alanine = ADP + phosphate + UDP-N- acetylmuramoyl-L-alanyl-gamma-D-glutamyl-L-lysyl-D-alanyl-D- alanine.
Pathway: Peptidoglycan biosynthesis.
Subunit: Monomer.
Interaction:
Subcellular location: Cytoplasmic.
Similarity: Belongs to the murCDEF 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.

Lichtarge lab 2006

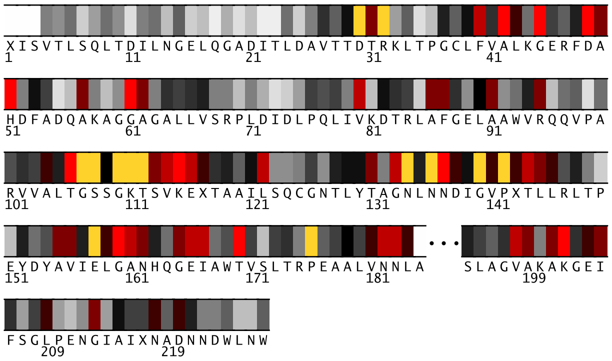


Fig. 1. Residues 1-227 in 1gg4A colored by their relative importance. (See Appendix, Fig.7, for the coloring scheme.)

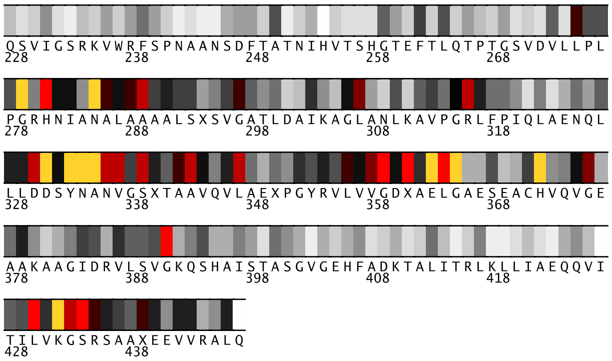


Fig. 2. Residues 228-447 in 1gg4A colored by their relative importance. (See Appendix, Fig.7, for the coloring scheme.)

2.2 Multiple sequence alignment for 1gg4A

For the chain 1gg4A, the alignment 1gg4A.msf (attached) with 413 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 1gg4A.msf. Its statistics, from the *alistat* program are the following:

```
Format:                MSF
Number of sequences: 413
Total number of residues:     177606
Smallest:              384
Largest:               439
Average length:        430.0
Alignment length:      439
Average identity:      35%
Most related pair:     99%
Most unrelated pair: 22%
Most distant seq:      34%
```

Furthermore, 1% of residues show as conserved in this alignment.

The alignment consists of 17% prokaryotic sequences. (Descriptions of some sequences were not readily available.) The file containing the sequence descriptions can be found in the attachment, under the name 1gg4A.descr.

2.3 Residue ranking in 1gg4A

The 1gg4A sequence is shown in Figs. 1–2, with each residue colored according to its estimated importance. The full listing of residues in 1gg4A can be found in the file called 1gg4A.ranks_sorted in the attachment.

2.4 Top ranking residues in 1gg4A and their position on the structure

In the following we consider residues ranking among top 25% of residues in the protein . Figure 3 shows residues in 1gg4A 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.

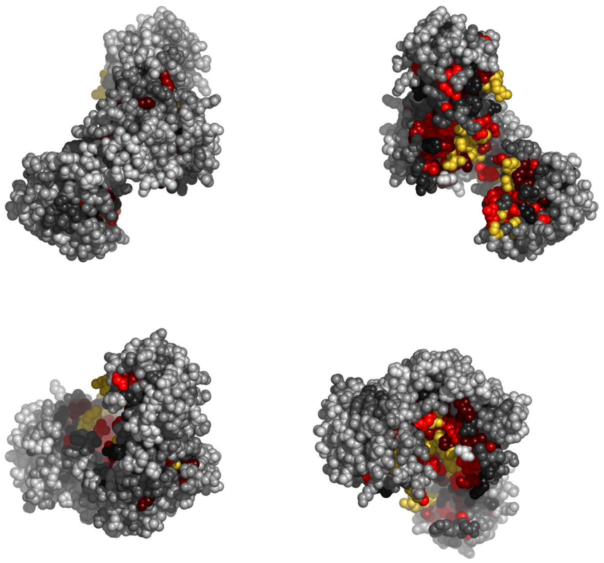


Fig. 3. Residues in 1gg4A, colored by their relative importance. Clockwise: front, back, top and bottom views.

2.4.1 Clustering of residues at 25% coverage. Fig. 4 shows the top 25% of all residues, this time colored according to clusters they belong to. The clusters in Fig.4 are composed of the residues listed in Table 1.

Table 1.		
cluster color	**size**	**member residues**
red	59	80,86,87,91,95,106,107,108 109,110,111,112,113,114,115 116,117,131,133,134,136,137 138,140,141,142,143,144,145 146,155,156,158,159,160,161 162,165,166,167,170,181,182
		continued in next column

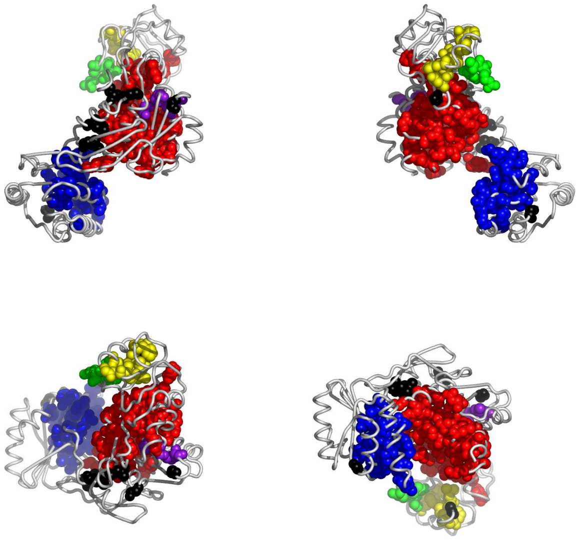


Fig. 4. Residues in 1gg4A, colored according to the cluster they belong to: red, followed by blue and yellow are the largest clusters (see Appendix for the coloring scheme). Clockwise: front, back, top and bottom views. The corresponding Pymol script is attached.

Table 1. *continued*

cluster color	size	member residues
		183,184,198,199,201,202,204,205,218,220,279,281,285,286,288,289
blue	28	315,316,330,331,333,334,335,336,337,339,342,343,347,356,358,359,361,363,364,365,372,376,430,432,433,434,435,439
yellow	9	40,42,43,49,50,51,57,61,62
green	3	30,31,32
purple	3	104,176,209

Table 1. Clusters of top ranking residues in 1gg4A.

2.4.2 Possible novel functional surfaces at 25% coverage. One group of residues is conserved on the 1gg4A surface, away from (or susbtantially larger than) other functional sites and interfaces recognizable in PDB entry 1gg4. It is shown in Fig. 5. 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).

Table 2.

res	type	substitutions(%)	cvg
110	G	G(100)	0.02
111	K	K(100)	0.02
112	T	T(100)	0.02
136	N	N(100)	0.02
30	D	D(98)N.	0.03
279	G	G(98)SAN	0.03
32	R	R(98)N.K	0.04
108	S	S(98)TN	0.05
134	N	N(78)S(21)	0.05
285	N	N(96)T(1)D(1)APS	0.05
45	G	G(97)A(1)NDPT	0.06
137	N	N(95)S(3)GTA	0.06
202	K	K(87)N(4)H(5)E(2)	0.06
115	K	K(92)T(2)R(4)	0.07
170	T	T(27)L(69)M(2)ANS	0.07
281	H	H(96)RY(1)CFQAGP	0.07
61	G	G(95)TN(1)QAERPK	0.08
160	G	G(92)A(3)S(3)	0.08
49	D	D(92)N(7)CT	0.09
51	H	H(92)N(5)EWQLGR	0.09
40	F	F(92)Y(7)A	0.10
182	N	N(35)T(49)VCL(12)SA.	0.11
183	N	N(83)T(8)HAL(1)S(1)K.CRQMG	0.11
116	E	E(69)D(20)Q(2)N(2)A(2)TSGR	0.12
161	A	A(43)M(47)I(3)T(3)VSG	0.12
144	T	T(80)S(7)M(8)VA(2)LN	0.13
166	E	E(83)H(5)N(1)Q(4)D(5)	0.13

continued in next column

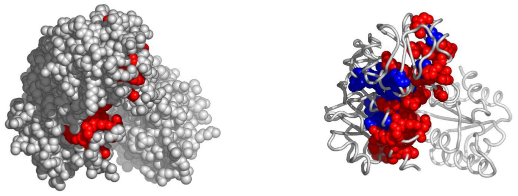


Fig. 5. A possible active surface on the chain 1gg4A. The larger cluster it belongs to is shown in blue.

Table 2. *continued*

res	type	substitutions(%)	cvg
133	G	G(79)R(1)A(13)K(2)ESQLD	0.14
198	V	V(46)I(51)LT(1).	0.14
50	A	G(67)A(32)	0.15
87	F	M(8)L(81)Y(5)RF(3)IV	0.15
113	S	S(33)T(63)G(3)	0.15
165	G	G(87)D(1)N(5)RASQKH(1)E	0.16
199	A	A(75)L(7)V(7)IC(1)RFK(2)MSQGE.	0.16
62	A	A(88)V(4)C(3)IPSTGN	0.17
162	N	N(59)WR(15)S(19)TYG(2)QD	0.18
31	T	S(58)T(40)G.	0.19
201	A	A(71)E(18)T(6)G(1)SID	0.19
146	L	L(70)A(10)F(12)MSVC(1)TGI(1)D	0.20
204	E	E(77)A(11)S(5)RQTHD(1)KCPVW	0.20
117	X	M(64)V(2)L(22)A(6)F(1)I(1)WYXGT	0.21
43	L	L(76)I(15)KF(5)RVM	0.22
138	D	E(33)H(19)CD(35)L(3)SQ(2)FAN(1)YMR	0.22
184	L	V(37)I(31)L(6)A(2).(22)FC	0.23
218	N	N(72)P(17)ILD(1)MVC(1)FERYGH(1).S	0.23
220	D	D(80)E(8)G(2)FA(3)QN(2)RISY	0.23
143	X	L(76)A(1)K(4)HQ(1)W(1)M(6)FY(1)RVI(2)XE	0.24
109	S	N(34)V(22)S(22)A(7)T(3)C(6)IGQF	0.25

Table 2. Residues forming surface ”patch” in 1gg4A.

Table 3.

res	type	disruptive mutations
110	G	(KER)(FQMWHD)(NYLPI)(SVA)
111	K	(Y)(FTW)(SVCAG)(HD)
112	T	(KR)(FQMWH)(NELPI)(D)
136	N	(Y)(FTWH)(SEVCARG)(MD)
30	D	(R)(FWH)(YVCAG)(T)
279	G	(R)(KE)(H)(FW)
32	R	(T)(YD)(SVCAG)(FELWPI)
108	S	(R)(K)(FWH)(M)
134	N	(Y)(FWH)(R)(TE)
285	N	(Y)(H)(R)(FW)
45	G	(R)(K)(EH)(FW)
137	N	(Y)(H)(R)(FW)
202	K	(Y)(T)(FW)(VCAG)
115	K	(Y)(FW)(T)(VA)
170	T	(R)(K)(H)(FW)
281	H	(E)(D)(Q)(TM)
61	G	(R)(E)(FWH)(Y)
160	G	(KR)(E)(QH)(FMW)
49	D	(R)(FWH)(K)(Y)
51	H	(E)(T)(D)(M)
40	F	(K)(E)(Q)(D)
182	N	(Y)(H)(R)(FW)
183	N	(Y)(FWH)(T)(E)
116	E	(FWH)(Y)(R)(VA)
161	A	(R)(K)(Y)(E)
144	T	(R)(K)(H)(FW)
166	E	(FW)(H)(Y)(VCAG)
133	G	(R)(FWH)(Y)(KE)
198	V	(R)(K)(Y)(E)
50	A	(KER)(Y)(QHD)(N)
87	F	(E)(K)(T)(D)
113	S	(KR)(FQMWH)(E)(NLPI)
165	G	(R)(FW)(E)(H)
199	A	(Y)(R)(E)(K)
62	A	(R)(K)(YE)(H)
162	N	(Y)(H)(FW)(R)
31	T	(KR)(FQMWH)(NELPI)(D)
201	A	(R)(K)(Y)(H)
146	L	(R)(Y)(H)(K)
204	E	(H)(FW)(Y)(R)
117	X	(KR)(E)(Y)(Q)
43	L	(Y)(T)(R)(H)
138	D	(R)(H)(FW)(Y)
184	L	(R)(Y)(H)(KE)
218	N	(Y)(H)(T)(R)
220	D	(R)(H)(FW)(Y)
143	X	(Y)(E)(R)(K)
109	S	(R)(K)(H)(Q)

Table 3. Disruptive mutations for the surface patch in 1gg4A.

Another group of surface residues 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 4, while Table 5 suggests possible disruptive replacements for these residues (see Section 3.6).

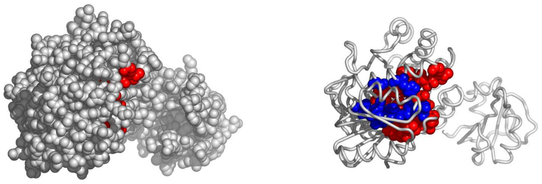


Fig. 6. Another possible active surface on the chain 1gg4A. The larger cluster it belongs to is shown in blue.

Table 4.

res	type	substitutions(%)	cvg
334	N	N(100)	0.02
363	E	E(99)A	0.02
365	G	G(99)A	0.02
432	K	K(99).	0.03
331	D	D(86)E(13)	0.04
333	Y	Y(98)F(1)	0.04
361	X	L(4)M(88)I(5)FV(1)XS	0.07
364	L	L(86)M(2)V(10)I	0.08
316	R	R(92)H(2)KW(3)SGN	0.10
339	S	S(90)A(9)	0.10
343	A	A(89)TG(6)S(3)V	0.11
330	D	D(76)N(20)SLAEHKI	0.12
337	V	P(76)L(3)V(15)S(1)ETDK(1)IA	0.12
336	N	N(75)A(2)G(2)S(19)	0.13
342	A	A(86)V(2)NTLS(6)REKI	0.20
439	X	M(62)L(27)S(1)F(5)V(1)T.WXIA	0.22
435	R	R(73)N(12)L(5)H(4)K(1).FYAQ	0.23
315	G	G(68)L(2)W(1)A(1)K(2)R(5)M(11)N(2)S.QPVDHET	0.25

Table 4. Residues forming surface ”patch” in 1gg4A.

Table 5.

res	type	disruptive mutations
334	N	(Y)(FTWH)(SEVCARG)(MD)
363	E	(H)(FYWR)(CG)(TKVA)
365	G	(KER)(QHD)(FYMW)(N)
432	K	(Y)(FTW)(SVCAG)(HD)
331	D	(R)(FWH)(YVCAG)(K)
333	Y	(K)(Q)(EM)(NR)
361	X	(R)(Y)(K)(E)
364	L	(Y)(R)(H)(T)
316	R	(D)(T)(E)(Y)
339	S	(KR)(QH)(FYEMW)(N)
343	A	(KR)(E)(Y)(QH)
330	D	(R)(H)(FW)(Y)
337	V	(R)(Y)(K)(H)
336	N	(Y)(H)(FWR)(E)
342	A	(Y)(R)(E)(H)
439	X	(R)(K)(E)(Y)
435	R	(T)(D)(E)(Y)
315	G	(R)(E)(K)(H)

Table 5. Disruptive mutations for the surface patch in 1gg4A.

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 $10\AA^2$, 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 $5\AA$ 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

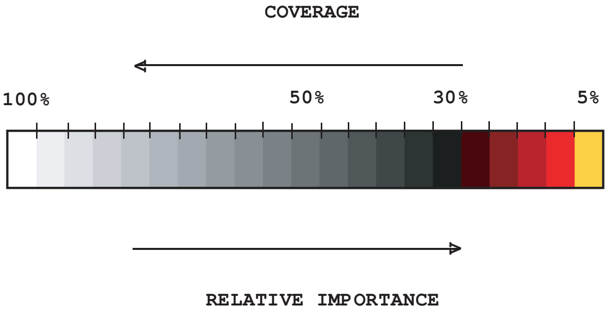


Fig. 7. Coloring scheme used to color residues by their relative importance.

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 10Å^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:

- 1gg4A.complex.pdb - coordinates of 1gg4A with all of its interacting partners
- 1gg4A.etvx - ET viewer input file for 1gg4A
- 1gg4A.cluster_report.summary - Cluster report summary for 1gg4A
- 1gg4A.ranks - Ranks file in sequence order for 1gg4A
- 1gg4A.clusters - Cluster descriptions for 1gg4A
- 1gg4A.msf - the multiple sequence alignment used for the chain 1gg4A
- 1gg4A.descr - description of sequences used in 1gg4A msf
- 1gg4A.ranks_sorted - full listing of residues and their ranking for 1gg4A