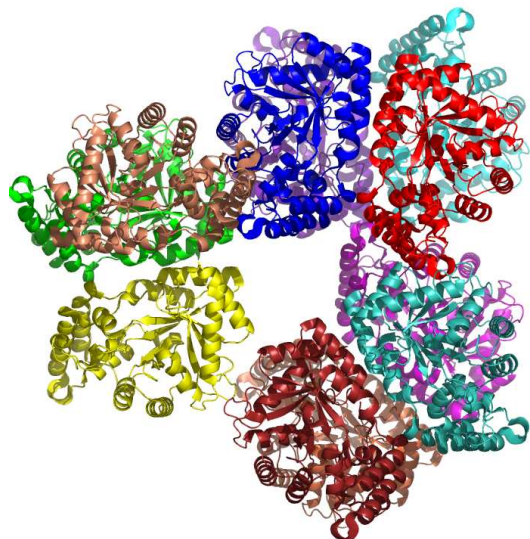


2clp

Evolutionary trace report by **report_maker**

December 4, 2009



4.3.3	DSSP	10
4.3.4	HSSP	10
4.3.5	LaTex	10
4.3.6	Muscle	10
4.3.7	Pymol	10
4.4	Note about ET Viewer	10
4.5	Citing this work	10
4.6	About report_maker	10
4.7	Attachments	10

CONTENTS

1 Introduction

2 Chain 2clpA

- 2.1 O95154 overview
- 2.2 Multiple sequence alignment for 2clpA
- 2.3 Residue ranking in 2clpA
- 2.4 Top ranking residues in 2clpA 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.

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
 - 4.3.1 **Alistat**
 - 4.3.2 **CE**

1 INTRODUCTION

From the original Protein Data Bank entry (PDB id 2clp):

Title: Crystal structure of human aflatoxin b1 aldehyde reductase member 3

Compound: Mol id: 1; molecule: aflatoxin b1 aldehyde reductase member 3; chain: a, b, c, d, e, f, g, h, i, j, k; synonym: afb1-ar 2; engineered: yes

Organism, scientific name: Homo Sapiens;

2clp contains a single unique chain 2clpA (323 residues long) and its homologues 2clpF, 2clpJ, 2clpK, 2clpE, 2clpB, 2clpH, 2clpC, 2clpD, 2clpI, and 2clpG.

1

2 CHAIN 2CLPA

1

2.1 O95154 overview

1

From SwissProt, id O95154, 90% identical to 2clpA:

1

Description: Aflatoxin B1 aldehyde reductase member 3 (EC 1.-.-.) (AFB1-AR 2).

1

Organism, scientific name: Homo sapiens (Human).

2

Taxonomy: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini; Hominidae; Homo.

2

Function: Can reduce the dialdehyde protein-binding form of aflatoxin B1 (AFB1) to the nonbinding AFB1 dialcohol. May be involved in protection of liver against the toxic and carcinogenic effects of AFB1, a potent hepatocarcinogen (By similarity).

9

Cofactor: NADH or NADPH (By similarity).

9

Subunit: Probably act as a homodimer (By similarity).

9

Subcellular location: Cytoplasmic.

9

Tissue specificity: Expressed in colon, kidney, pancreas, adenocarcinoma and endometrium.

9

Similarity: Belongs to the aldo/keto reductase 2 family.

9

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.

9

9

9

10

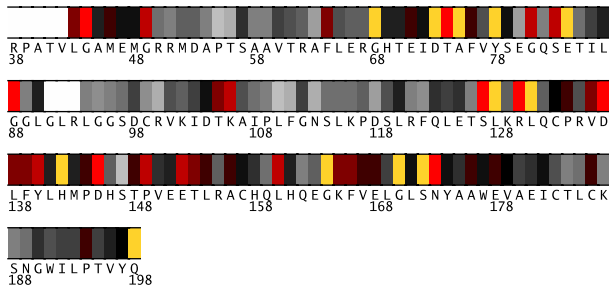


Fig. 1. Residues 38-198 in 2clpA colored by their relative importance. (See Appendix, Fig.11, for the coloring scheme.)

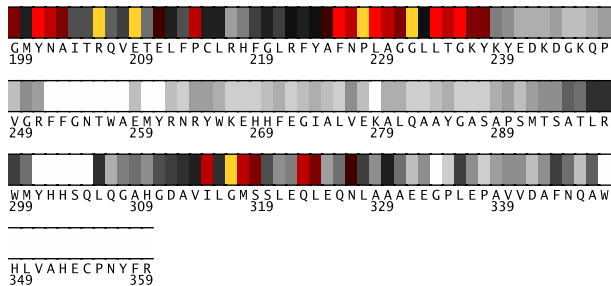


Fig. 2. Residues 199-360 in 2clpA colored by their relative importance. (See Appendix, Fig.11, for the coloring scheme.)

2.2 Multiple sequence alignment for 2clpA

For the chain 2clpA, the alignment 2clpA.msf (attached) with 533 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 2clpA.msf. Its statistics, from the *alistat* program are the following:

```

Format:                MSF
Number of sequences:  533
Total number of residues: 154958
Smallest:              149
Largest:               323
Average length:        290.7
Alignment length:      323
Average identity:       31%
Most related pair:     99%
Most unrelated pair:   3%
Most distant seq:      34%
  
```

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

The alignment consists of 2% eukaryotic (1% vertebrata, 1% fungi), 9% prokaryotic, and <1% archaean sequences. (Descriptions of some sequences were not readily available.) The file containing the sequence descriptions can be found in the attachment, under the name 2clpA.descr.

2.3 Residue ranking in 2clpA

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

2.4 Top ranking residues in 2clpA 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 2clpA 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. 4 shows the top 25% of all residues, this time colored according to clusters they

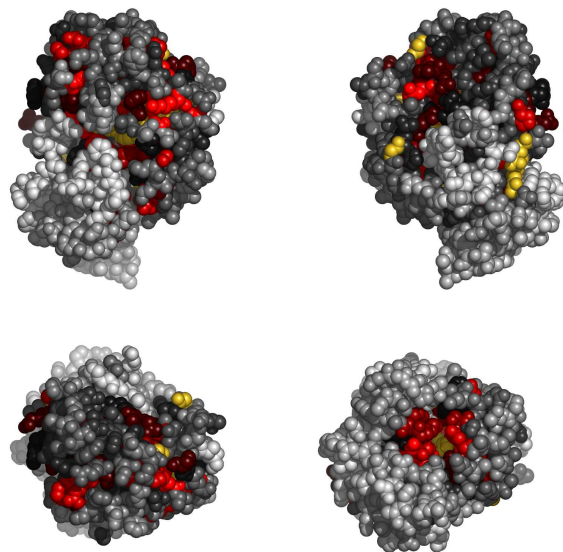


Fig. 3. Residues in 2clpA, colored by their relative importance. Clockwise: front, back, top and bottom views.

belong to. The clusters in Fig.4 are composed of the residues listed in Table 1.

Table 1.		
cluster color	size	member residues
red	77	43, 44, 46, 64, 71, 73, 74, 75, 76 78, 81, 83, 84, 88, 105, 106, 127 128, 130, 131, 133, 134, 136, 137 138, 139, 140, 142, 144, 145, 148 149, 152, 153, 154, 156, 160, 164 165, 166, 167, 168, 170, 172, 173 176, 178, 197, 198, 199, 201, 202 203, 206, 209, 211, 214, 225, 226 227, 228, 229, 230, 231, 232, 234 235, 236, 237, 238, 315, 317, 318

continued in next column

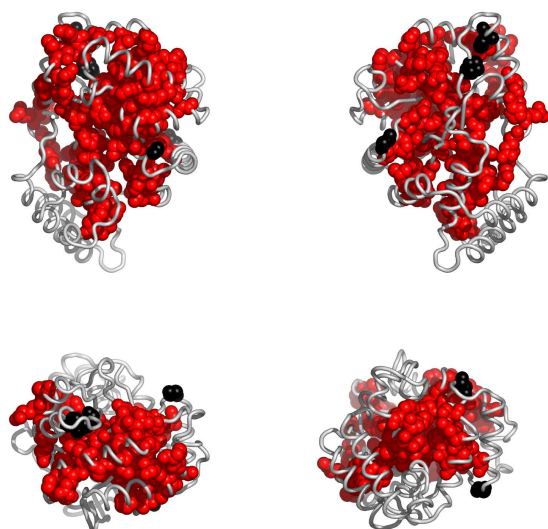


Fig. 4. Residues in 2clpA, colored according to the cluster they belong to: red, followed by blue and yellow are the largest clusters (see Appendix for the coloring scheme). Clockwise: front, back, top and bottom views. The corresponding Pymol script is attached.

Table 1. continued		
cluster color	size	member residues
		319, 323, 324, 327

Table 1. Clusters of top ranking residues in 2clpA.

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 2clpE. By analogy with 2clpD – 2clpE interface. Table 2 lists the top 25% of residues at the interface with 2clpE. 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)
165	K	R(2) Q(1) W(1) K(83)F L(4)V.G H(2)ITY DSA(1)E	0.15	2/0	4.08

Table 2. The top 25% of residues in 2clpA at the interface with 2clpE. (Field names: res: residue number in the PDB entry; type: amino acid type; subst: 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 3.		
res	type	disruptive mutations
165	K	(Y) (T) (FW) (CG)

Table 3. List of disruptive mutations for the top 25% of residues in 2clpA, that are at the interface with 2clpE.

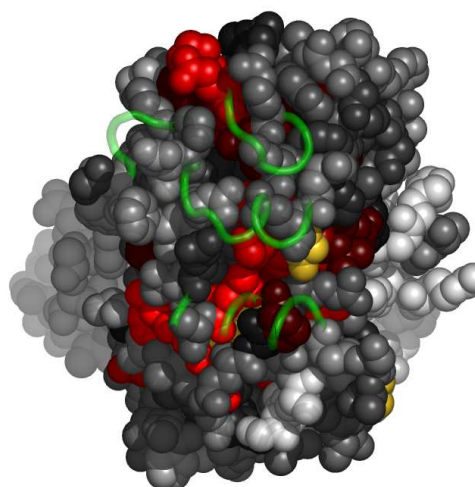


Fig. 5. Residues in 2clpA, at the interface with 2clpE, colored by their relative importance. 2clpE is shown in backbone representation (See Appendix for the coloring scheme for the protein chain 2clpA.)

Figure 5 shows residues in 2clpA colored by their importance, at the interface with 2clpE.

Interface with 2clpH. By analogy with 2clpI – 2clpH interface. Table 4 lists the top 25% of residues at the interface with 2clpH. The following table (Table 5) suggests possible disruptive replacements for these residues (see Section 3.6).

Table 4.					
res	type	subst's (%)	cvg	noc/ bb	dist (Å)
130	R	R(87) Q(2) K(2)N A(2) E(3).DS L	0.08	10/0	3.87

continued in next column

Table 4. continued					
res	type	subst's (%)	cvg	noc/ bb	dist (Å)
81	G	G(88)S Y(2) E(1)H F(2)N . (1)KWP TD	0.10	1/1	4.65
76	F	D(45)W E(10) N(22) A(1) H(1) R(6)GL Q(1) V(1).T F(2)P Y(2)IKM	0.22	17/1	3.94

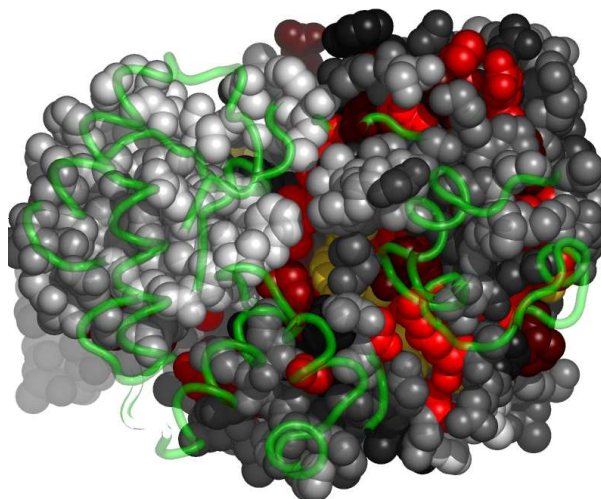


Table 4. The top 25% of residues in 2clpA at the interface with 2clpH. (Field names: res: residue number in the PDB entry; type: amino acid type; substs: substitutions seen in the alignment; with the percentage of each type in the bracket; noc/bb: number of contacts with the ligand, with the number of contacts realized through backbone atoms given in the bracket; dist: distance of closest approach to the ligand.)

Fig. 6. Residues in 2clpA, at the interface with 2clpH, colored by their relative importance. 2clpH is shown in backbone representation (See Appendix for the coloring scheme for the protein chain 2clpA.)

Table 5.		
res	type	disruptive mutations
130	R	(TY)(CDG)(FW)(VA)
81	G	(R)(K)(E)(Q)
76	F	(E)(K)(T)(D)

Table 5. List of disruptive mutations for the top 25% of residues in 2clpA, that are at the interface with 2clpH.

Figure 6 shows residues in 2clpA colored by their importance, at the interface with 2clpH.

Interface with 2clpI. By analogy with 2clpJ – 2clpI interface. Table 6 lists the top 25% of residues at the interface with 2clpI. The following table (Table 7) suggests possible disruptive replacements for these residues (see Section 3.6).

Table 6.					
res	type	subst's (%)	cvg	noc/ bb	dist (Å)
214	P	P(80) R(2) E(4) D(3) A(4)SKI THQYGLN V	0.14	5/3	4.55
211	E	E(72)	0.21	65/7	3.05

continued in next column

Table 6. continued					
res	type	subst's (%)	cvg	noc/ bb	dist (Å)
176	A	D(12) G(5)S T(1) A(3)HL N(1)K P(1)VQ G(17) A(62)V P(6) S(5) L(2) . (2)RKI FE	0.24	19/13	3.31

Table 6. The top 25% of residues in 2clpA at the interface with 2clpI. (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 approach to the ligand.)

Table 7.		
res	type	disruptive mutations
214	P	(Y)(R)(H)(T)
211	E	(H)(FW)(Y)(R)

continued in next column

Table 7. continued		
res	type	disruptive mutations
176	A	(Y) (R) (E) (K)

Table 7. List of disruptive mutations for the top 25% of residues in 2clpA, that are at the interface with 2clpI.

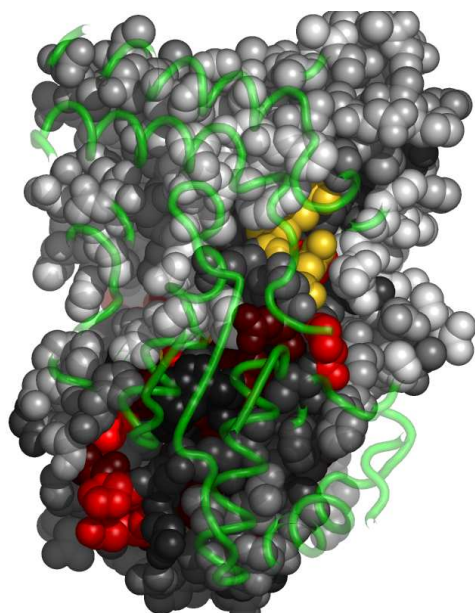


Fig. 7. Residues in 2clpA, at the interface with 2clpI, colored by their relative importance. 2clpI is shown in backbone representation (See Appendix for the coloring scheme for the protein chain 2clpA.)

Figure 7 shows residues in 2clpA colored by their importance, at the interface with 2clpI.

Interface with 2clpJ. By analogy with 2clpE – 2clpJ interface. Table 8 lists the top 25% of residues at the interface with 2clpJ. The following table (Table 9) suggests possible disruptive replacements for these residues (see Section 3.6).

Table 8.					
res	type	subst's (%)	cvg	noc/ bb	dist (Å)
152	E	E (85) D (9) Q (1) . A (1) TVP H	0.12	3/0	4.42

Table 8. The top 25% of residues in 2clpA at the interface with 2clpJ. (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 9.		
res	type	disruptive mutations
152	E	(H) (FW) (R) (Y)

Table 9. List of disruptive mutations for the top 25% of residues in 2clpA, that are at the interface with 2clpJ.

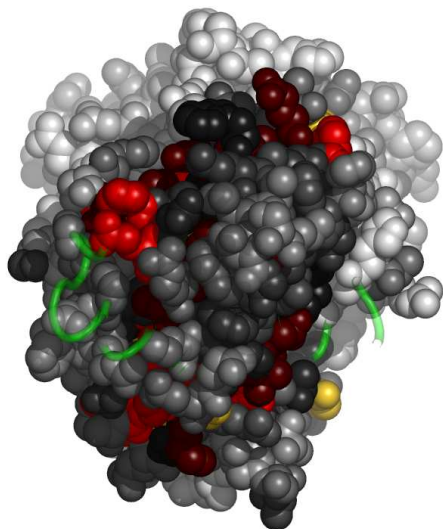


Fig. 8. Residues in 2clpA, at the interface with 2clpJ, colored by their relative importance. 2clpJ is shown in backbone representation (See Appendix for the coloring scheme for the protein chain 2clpA.)

Figure 8 shows residues in 2clpA colored by their importance, at the interface with 2clpJ.

Interface with 2clpD. By analogy with 2clpE – 2clpD interface. Table 10 lists the top 25% of residues at the interface with 2clpD. The following table (Table 11) suggests possible disruptive replacements for these residues (see Section 3.6).

Table 10.					
res	type	subst's (%)	cvg	noc/ bb	dist (Å)
88	G	G (92) A (4) R (1) S . K EQ	0.07	2/2	4.35

Table 10. The top 25% of residues in 2clpA at the interface with 2clpD. (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 11.		
res	type	disruptive mutations
88	G	(FW) (EHR) (KY) (D)

Table 11. List of disruptive mutations for the top 25% of residues in 2clpA, that are at the interface with 2clpD.

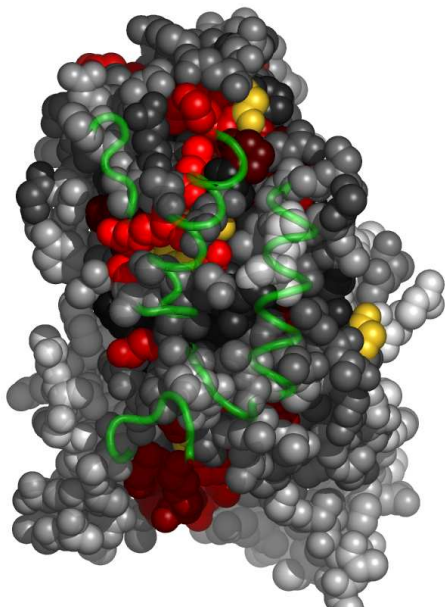


Fig. 9. Residues in 2clpA, at the interface with 2clpD, colored by their relative importance. 2clpD is shown in backbone representation (See Appendix for the coloring scheme for the protein chain 2clpA.)

Figure 9 shows residues in 2clpA colored by their importance, at the interface with 2clpD.

NDP binding site. Table 12 lists the top 25% of residues at the interface with 2clpANDP1361 (ndp). The following table (Table 13) suggests possible disruptive replacements for these residues (see Section 3.6).

Table 12.						
res	type	subst's (%)	cvg	noc/ bb	dist (Å)	antn
73	D	D(99).N	0.01	14/3	2.87	site
78	Y	Y(99).T	0.01	22/0	3.23	site
142	H	H(98)C. TY	0.01	11/1	2.95	site
232	G	G(99)RA S.	0.01	21/21	3.17	site
172	S	S(95) C(3)T.V G	0.03	5/0	3.43	
198	Q	Q(98)	0.03	16/1	3.21	site

continued in next column

Table 12. continued						
res	type	subst's (%)	cvg	noc/ bb	dist (Å)	antn
228	P	E(1)G.R P(86) S(6) A(5) V(1)GT	0.05	26/12	3.20	site
317	G	G(83) S(14)DI AP.(1)R C	0.05	41/41	3.25	site
44	G	G(94) . (3) P(1) S(1)	0.07	16/16	3.18	site
173	N	N(70) E(10) S(9) D(5) T(1)KV. QA	0.07	7/0	3.06	site
229	L	L(91) M(1) I(4)VFQ GAS	0.08	31/12	2.75	site
226	F	W(39) Y(45) F(13)RG	0.09	66/8	2.75	site
227	N	S(65) Y(7) F(9) N(10) A(1) T(1) Q(1) G(1)HDW PR	0.10	45/24	2.75	
315	I	I(83) V(7) L(6) . (1)WAS MCT	0.12	22/7	3.44	site
323	Q	Q(76) H(12)L Y(1) R(4)NAS G.(1)TE	0.12	74/13	3.03	site
106	K	K(89)QA IV(3) R(1)TNE .GPDS	0.13	8/0	3.78	
230	A	G(24)N A(67) K(2)Q	0.13	7/7	3.06	site

continued in next column

res	type	subst's (%)	cvg	noc/ bb	dist (Å)	antn
318	M	C(1)SRH VELMD. A(63) I(1) P(10) M(2)Q T(8)S G(3) V(5)YE . (1)RN L(1)WC	0.13	23/21	3.55	
235	T	T(63) S(26) G(1)K A(6)RQI L	0.14	17/1	3.02	site
238	Y	Y(61) W(5) L(3)M F(5) I(11) K(1). H(3)ES R(2)P V(1)AD	0.15	7/0	2.71	
319	S	R(37) S(26) T(21)L K(10)V . (1)DNG AIQEF	0.15	25/8	2.90	site
237	K	K(70) N(1) A(7)V R(7) D(2) T(2) P(3) S(1)QGL	0.17	17/0	2.75	site
231	G	R(19) G(38) F(2) N(3) S(14) A(6) W(4) Q(4) K(3) M(1)CEH LVY.	0.18	37/37	2.75	
324	L	L(74) V(10)T	0.18	1/1	4.92	

continued in next column

res	type	subst's (%)	cvg	noc/ bb	dist (Å)	antn
46	M	W(3) F(2) I(4)M A(1)Y . (1)E M(58) . (2) A(3)D N(16) F(1)T G(5) W(1) L(3) H(3)PSY IVR	0.20	32/8	3.47	site
327	N	A(9) S(6) N(59) I(1) L(5) D(1)Q E(1) Y(5) . (2) T(3) V(1)RMF	0.22	19/1	2.99	

Table 12. The top 25% of residues in 2clpA at the interface with NDP.(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 approach to the ligand.)

res	type	disruptive mutations
73	D	(R) (FWH) (YVCAG) (T)
78	Y	(K) (M) (Q) (LPI)
142	H	(E) (Q) (M) (K)
232	G	(E) (K) (R) (D)
172	S	(KR) (Q) (H) (FMW)
198	Q	(Y) (FW) (H) (T)
228	P	(R) (Y) (H) (K)
317	G	(R) (K) (E) (H)
44	G	(R) (K) (E) (H)
173	N	(Y) (H) (FW) (R)
229	L	(YR) (H) (T) (KE)
226	F	(E) (K) (D) (Q)
227	N	(Y) (TEH) (R) (FW)
315	I	(R) (Y) (H) (K)
323	Q	(Y) (FWH) (T) (VA)

continued in next column

Table 13. continued		
res	type	disruptive mutations
106	K	(Y) (FW) (T) (H)
230	A	(Y) (R) (E) (K)
318	M	(Y) (H) (R) (T)
235	T	(R) (H) (K) (FW)
238	Y	(K) (Q) (R) (E)
319	S	(R) (K) (H) (FW)
237	K	(Y) (FW) (T) (H)
231	G	(E) (R) (K) (D)
324	L	(R) (Y) (H) (K)
46	M	(Y) (H) (R) (T)
327	N	(Y) (H) (T) (FW)

Table 13. List of disruptive mutations for the top 25% of residues in 2clpA, that are at the interface with NDP.

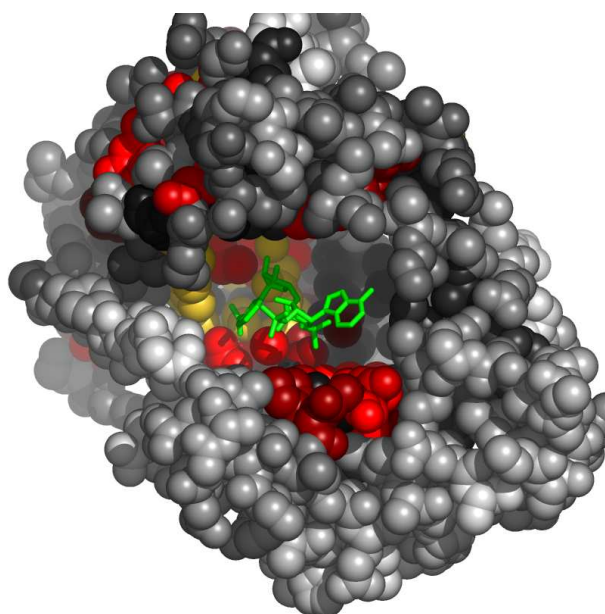


Fig. 10. Residues in 2clpA, at the interface with NDP, colored by their relative importance. The ligand (NDP) is colored green. Atoms further than 30 Å away from the geometric center of the ligand, as well as on the line of sight to the ligand were removed. (See Appendix for the coloring scheme for the protein chain 2clpA.)

Figure 10 shows residues in 2clpA colored by their importance, at the interface with 2clpANDP1361.

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\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\AA .

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],

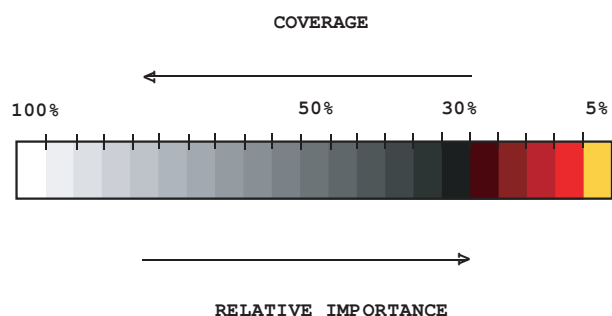


Fig. 11. Coloring scheme used to color residues by their relative importance.

long aliphatic chain [EKRRQM], 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 researchers 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

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. 11.

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 $(\text{idents} / \text{MIN}(\text{len1}, \text{len2}))$ 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\AA^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 L^AT_EX; 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:

- 2clpA.complex.pdb - coordinates of 2clpA with all of its interacting partners
- 2clpA.etvx - ET viewer input file for 2clpA
- 2clpA.cluster_report.summary - Cluster report summary for 2clpA
- 2clpA.ranks - Ranks file in sequence order for 2clpA
- 2clpA.clusters - Cluster descriptions for 2clpA
- 2clpA.msf - the multiple sequence alignment used for the chain 2clpA
- 2clpA.descr - description of sequences used in 2clpA msf
- 2clpA.ranks_sorted - full listing of residues and their ranking for 2clpA
- 2clpA.2clpE.if.pml - Pymol script for Figure 5
- 2clpA.cbcbvg - used by other 2clpA – related pymol scripts
- 2clpA.2clpH.if.pml - Pymol script for Figure 6
- 2clpA.2clpI.if.pml - Pymol script for Figure 7
- 2clpA.2clpJ.if.pml - Pymol script for Figure 8
- 2clpA.2clpD.if.pml - Pymol script for Figure 9
- 2clpA.2clpANDP1361.if.pml - Pymol script for Figure 10