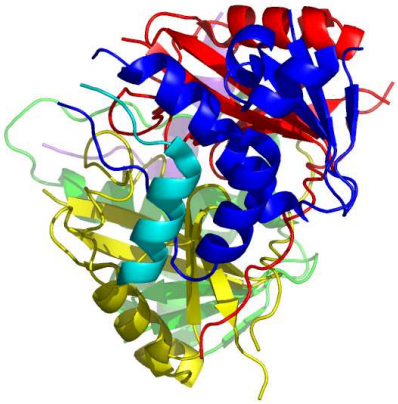
3dcg

Evolutionary trace report by **report_maker**

April 20, 2010



CONTENTS

1 Introduction 1

2 Chain 3dcgA 1
- 2.1 Q15370 overview 1
- 2.2 Multiple sequence alignment for 3dcgA 1
- 2.3 Residue ranking in 3dcgA 2
- 2.4 Top ranking residues in 3dcgA and their position on the structure 2
  - 2.4.1 Clustering of residues at 25% coverage. 2
  - 2.4.2 Overlap with known functional surfaces at 25% coverage. 2
  - 2.4.3 Possible novel functional surfaces at 25% coverage. 4

3 Chain 3dcgD 5
- 3.1 Q502M9 overview 5
- 3.2 Multiple sequence alignment for 3dcgD 5
- 3.3 Residue ranking in 3dcgD 5
- 3.4 Top ranking residues in 3dcgD and their position on the structure 6
  - 3.4.1 Clustering of residues at 24% coverage. 6
  - 3.4.2 Overlap with known functional surfaces at 24% coverage. 6
  - 3.4.3 Possible novel functional surfaces at 24% coverage. 8

4 Notes on using trace results 9
- 4.1 Coverage 9
- 4.2 Known substitutions 9
- 4.3 Surface 9
- 4.4 Number of contacts 9
- 4.5 Annotation 10
- 4.6 Mutation suggestions 10

5 Appendix 10
- 5.1 File formats 10
- 5.2 Color schemes used 10
- 5.3 Credits 10
  - 5.3.1 **Alistat** 10
  - 5.3.2 **CE** 10
  - 5.3.3 **DSSP** 10
  - 5.3.4 **HSSP** 10
  - 5.3.5 **LaTex** 11
  - 5.3.6 **Muscle** 11
  - 5.3.7 **Pymol** 11
- 5.4 Note about ET Viewer 11
- 5.5 Citing this work 11
- 5.6 About report_maker 11
- 5.7 Attachments 11

1 INTRODUCTION

From the original Protein Data Bank entry (PDB id 3dcg):

Title: Crystal structure of the hiv vif bc-box in complex with human elonginb and elonginc

Compound: Mol id: 1; molecule: transcription elongation factor b polypeptide 2; chain: a, c; synonym: rna polymerase ii transcription factor siii subunit b, siii p18, elongin b, elob, elongin 18 kda subunit; engineered: yes; mol id: 2; molecule: transcription elongation factor b polypeptide 1; chain: b, d; fragment: unp residues 17-112; synonym: rna polymerase ii transcription factor siii subunit c, siii p15, elongin-c, eloc, elongin 15 kda subunit; engineered: yes; mol id: 3; molecule: virion infectivity factor; chain: e, f; fragment: unp residues 139-176; synonym: vif, sor protein, virion infectivity factor p17, virion infectivity factor p7; engineered: yes

Organism, scientific name: Human Immunodeficiency Virus Type 1

3dcg contains unique chains 3dcgA (99 residues) and 3dcgD (86 residues) 3dcgC is a homologue of chain 3dcgA. 3dcgB is a homologue of chain 3dcgD. Chains 3dcgE and 3dcgF are too short to permit statistically significant analysis, and were treated as a peptide ligands.

2 CHAIN 3DCGA

2.1 Q15370 overview

From SwissProt, id Q15370, 99% identical to 3dcgA:
Description: Transcription elongation factor B polypeptide 2 (RNA polymerase II transcription factor SIII subunit B) (SIII p18) (Elongin B) (EloB) (Elongin 18 kDa subunit).
Organism, scientific name: Homo sapiens (Human).
Taxonomy: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini; Hominidae; Homo.
Function: SIII, also known as elongin, is a general transcription elongation factor that increases the RNA polymerase II transcription elongation past template-encoded arresting sites. Subunit A is transcriptionally active and its transcription activity is strongly enhanced by binding to the dimeric complex of the SIII regulatory subunits B and C (elongin BC complex).
Function: The elongin BC complex seems to be involved as an adapter protein in the proteasomal degradation of target proteins via different E3 ubiquitin ligase complexes, including the von Hippel-Lindau ubiquitination complex CBC(VHL). By binding to BC- box motifs it seems to link target recruitment subunits, like VHL and members of the SOCS box family, to Cullin/RBX1 modules that activate E2 ubiquitination enzymes.
Pathway: Ubiquitin conjugation; third step.
Subunit: Heterotrimer of an A (A1, A2 or A3), B and C subunit. The elongin BC complex interacts with SOCS1. The elongin BC complex is part of a complex with VHL and hydroxylated HIF1A. The elongin BC complex is part of multisubunit CBC(VHL) E3 ubiquitin ligase complexes with VHL, CUL2 or CUL5 and RBX1; and elongin A/TCEB3 or SOCS1 or WSB1, CUL5 and RBX1 (By similarity).
Subcellular location: Nuclear (Probable).
Similarity: Contains 1 ubiquitin-like domain.
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.

2.2 Multiple sequence alignment for 3dcgA

For the chain 3dcgA, the alignment 3dcgA.msf (attached) with 32 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 3dcgA.msf. Its statistics, from the *alistat* program are the following:

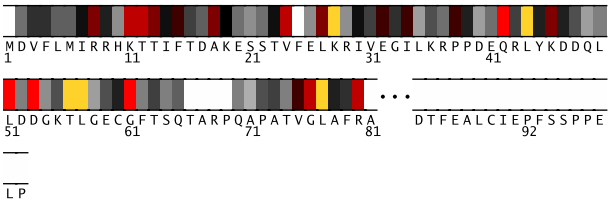


Fig. 1. Residues 1-100 in 3dcgA colored by their relative importance. (See Appendix, Fig.14, for the coloring scheme.)

```
Format:              MSF
Number of sequences: 32
Total number of residues:    2564
Smallest:            59
Largest:             99
Average length:      80.1
Alignment length:    99
Average identity:    44%
Most related pair:   98%
Most unrelated pair: 17%
Most distant seq:    42%
```

Furthermore, 1% of residues show as conserved in this alignment.
The alignment consists of 59% eukaryotic (25% vertebrata, 6% arthropoda, 3% fungi, 6% plantae) sequences. (Descriptions of some sequences were not readily available.) The file containing the sequence descriptions can be found in the attachment, under the name 3dcgA.descr.

2.3 Residue ranking in 3dcgA

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

2.4 Top ranking residues in 3dcgA 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 3dcgA 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.

cluster color	size	member residues
red	24	8,11,12,13,15,18,19,24,27,28 32,34,38,42,44,46,51,53,56 57,61,75,76,77

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

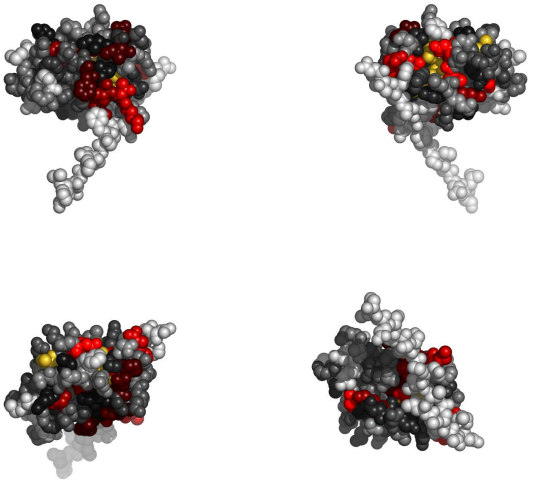


Fig. 2. Residues in 3dcgA, colored by their relative importance. Clockwise: front, back, top and bottom views.

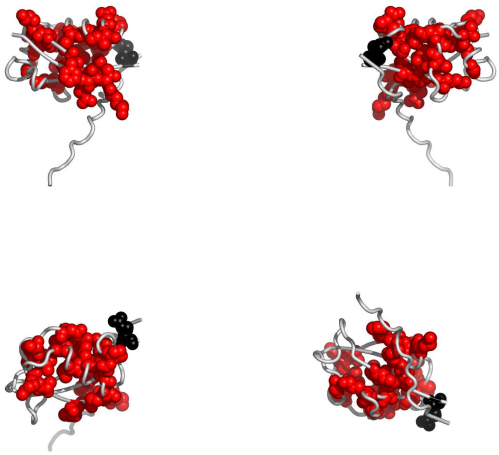


Fig. 3. Residues in 3dcgA, colored according to the cluster they belong to: red, followed by blue and yellow are the largest clusters (see Appendix for the coloring scheme). Clockwise: front, back, top and bottom views. The corresponding Pymol script is attached.

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 3dcgB. Table 2 lists the top 25% of residues at the interface with 3dcgB. The following table (Table 3) suggests possible disruptive replacements for these residues (see Section 4.6).

Table 2.					
res	**type**	**subst's (%)**	**cvg**	**noc/ bb**	**dist (Å)**
12	T	T(87) S(3) L(6) K(3)	0.11	31/14	2.98
11	K	K(84) R(6) N(6) G(3)	0.14	72/33	2.87
8	R	P(3) L(21) R(59) T(6) K(6) .(3)	0.15	4/0	3.68
13	T	T(62) I(18) H(9) E(6) S(3)	0.16	45/32	2.68
15	F	F(65) L(21) Y(6) H(3) Q(3)	0.20	73/19	2.79
34	I	I(71) K(12) Q(3) R(3) L(6) E(3)	0.24	18/0	3.89

Table 2. The top 25% of residues in 3dcgA at the interface with 3dcgB. (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**
12	T	(R)(FKWH)(QM)(E)
11	K	(Y)(FW)(T)(SVA)
8	R	(T)(YD)(CG)(SE)
13	T	(R)(K)(FQWH)(M)
15	F	(E)(K)(T)(D)
34	I	(Y)(T)(H)(R)

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

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

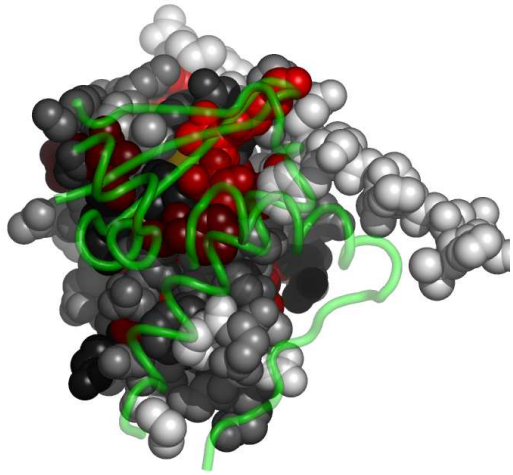


Fig. 4. Residues in 3dcgA, at the interface with 3dcgB, colored by their relative importance. 3dcgB is shown in backbone representation (See Appendix for the coloring scheme for the protein chain 3dcgA.)

Interface with 3dcgC. Table 4 lists the top 25% of residues at the interface with 3dcgC. The following table (Table 5) suggests possible disruptive replacements for these residues (see Section 4.6).

Table 4.					
res	**type**	**subst's (%)**	**cvg**	**noc/ bb**	**dist (Å)**
61	G	.(3) N(21) G(68) K(6)	0.09	10/10	3.02
80	R	.(18) R(78) K(3)	0.10	11/0	2.86
8	R	P(3) L(21) R(59) T(6) K(6) .(3)	0.15	12/0	2.69
46	K	R(3) F(21) K(56) N(6) L(12)	0.18	3/0	4.32

Table 4. The top 25% of residues in 3dcgA at the interface with 3dcgC. (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**
61	G	(FEW)(HR)(YD)(KM)
80	R	(T)(D)(Y)(VCAG)
8	R	(T)(YD)(CG)(SE)
46	K	(Y)(T)(S)(FCWG)

Table 5. List of disruptive mutations for the top 25% of residues in 3dcgA, that are at the interface with 3dcgC.

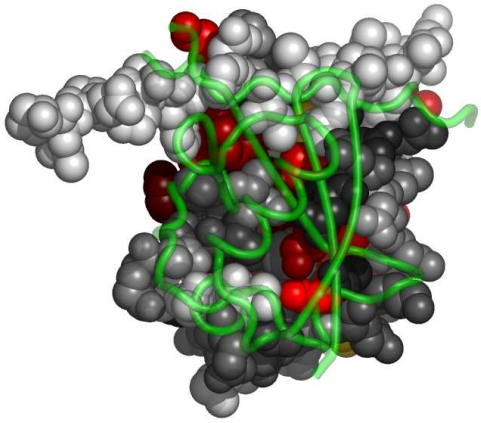


Fig. 5. Residues in 3dcgA, at the interface with 3dcgC, colored by their relative importance. 3dcgC is shown in backbone representation (See Appendix for the coloring scheme for the protein chain 3dcgA.)

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

2.4.3 Possible novel functional surfaces at 25% coverage. One group of residues is conserved on the 3dcgA surface, away from (or susbtantially larger than) other functional sites and interfaces recognizable in PDB entry 3dcg. 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 4.6).

Table 6.			
res	**type**	**substitutions(%)**	**cvg**
28	K	K(100)	0.01
77	L	.(9)L(90)	0.04
42	Q	S(3)Q(84)M(12)	0.06
53	D	R(3)D(87)E(6)	0.07
continued in next column			

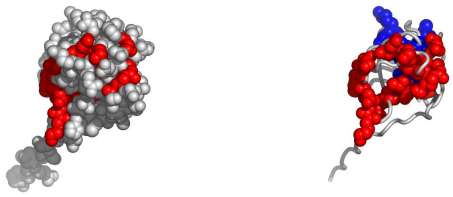


Fig. 6. A possible active surface on the chain 3dcgA. The larger cluster it belongs to is shown in blue.

Table 6. *continued*			
res	**type**	**substitutions(%)**	**cvg**
		N(3)	
51	L	L(78)M(18)I(3)	0.08
61	G	.(3)N(21)G(68) K(6)	0.09
80	R	.(18)R(78)K(3)	0.10
12	T	T(87)S(3)L(6) K(3)	0.11
76	G	.(9)H(18)G(59) A(9)Q(3)	0.13
11	K	K(84)R(6)N(6) G(3)	0.14
8	R	P(3)L(21)R(59) T(6)K(6).(3)	0.15
13	T	T(62)I(18)H(9) E(6)S(3)	0.16
46	K	R(3)F(21)K(56) N(6)L(12)	0.18
15	F	F(65)L(21)Y(6) H(3)Q(3)	0.20
32	E	E(68)Q(15)A(12) H(3)	0.22
38	P	P(78)Q(6)D(6) K(3)T(3)A(3)	0.23
34	I	I(71)K(12)Q(3) R(3)L(6)E(3)	0.24

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

Table 7.		
res	**type**	**disruptive mutations**
28	K	(Y)(FTW)(SVCAG)(HD)
77	L	(YR)(TH)(SCG)(KE)
42	Q	(Y)(H)(FW)(T)
53	D	(FW)(HR)(Y)(VCAG)
51	L	(Y)(R)(TH)(CG)
61	G	(FEW)(HR)(YD)(KM)
80	R	(T)(D)(Y)(VCAG)
continued in next column		

Table 7. *continued*		
res	**type**	**disruptive mutations**
12	T	(R)(FKWH)(QM)(E)
76	G	(E)(KR)(D)(QM)
11	K	(Y)(FW)(T)(SVA)
8	R	(T)(YD)(CG)(SE)
13	T	(R)(K)(FQWH)(M)
46	K	(Y)(T)(S)(FCWG)
15	F	(E)(K)(T)(D)
32	E	(FYWH)(CRG)(TVA)(SK)
38	P	(Y)(R)(H)(T)
34	I	(Y)(T)(H)(R)

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

3 CHAIN 3DCGD

3.1 Q502M9 overview

From SwissProt, id Q502M9, 89% identical to 3dcgD:
Description: Zgc:101879 protein.
Organism, scientific name: Brachydanio rerio (Zebrafish) (Danio rerio).
Taxonomy: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Actinopterygii; Neopterygii; Teleostei; Ostariophysi; Cypriniformes; Cyprinidae; Danio.

3.2 Multiple sequence alignment for 3dcgD

For the chain 3dcgD, the alignment 3dcgD.msf (attached) with 48 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 3dcgD.msf. Its statistics, from the *alistat* program are the following:

```
Format:              MSF
Number of sequences: 48
Total number of residues:    3573
Smallest:            48
Largest:             86
Average length:      74.4
Alignment length:    86
Average identity:    47%
Most related pair:   99%
Most unrelated pair: 24%
Most distant seq:    51%
```

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

The alignment consists of 25% eukaryotic (2% vertebrata, 12% fungi, 4% plantae) sequences. (Descriptions of some sequences were not readily available.) The file containing the sequence descriptions can be found in the attachment, under the name 3dcgD.descr.

3.3 Residue ranking in 3dcgD

The 3dcgD sequence is shown in Fig. 7, with each residue colored according to its estimated importance. The full listing of residues in 3dcgD can be found in the file called 3dcgD.ranks_sorted in the attachment.

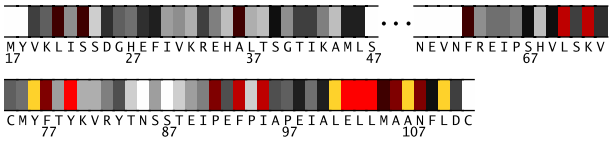


Fig. 7. Residues 17-112 in 3dcgD colored by their relative importance. (See Appendix, Fig.14, for the coloring scheme.)

3.4 Top ranking residues in 3dcgD 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 8 shows residues in 3dcgD 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.

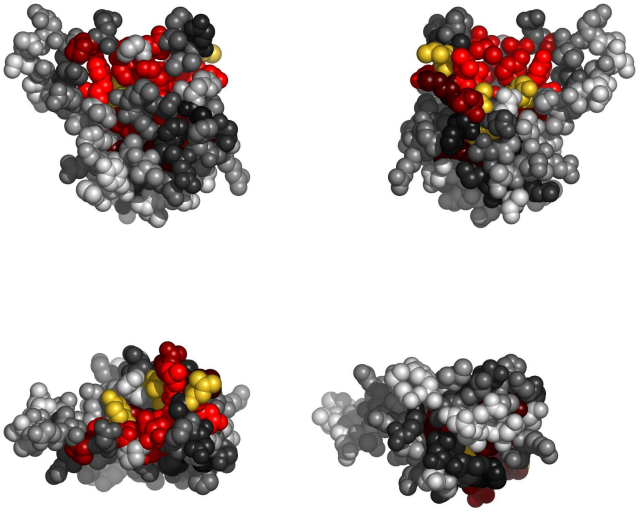


Fig. 8. Residues in 3dcgD, colored by their relative importance. Clockwise: front, back, top and bottom views.

3.4.1 Clustering of residues at 24% coverage. Fig. 9 shows the top 24% of all residues, this time colored according to clusters they belong to. The clusters in Fig.9 are composed of the residues listed in Table 8.

Table 8.		
cluster color	**size**	**member residues**
red	21	21,23,36,62,70,72,76,77,79
continued in next column		

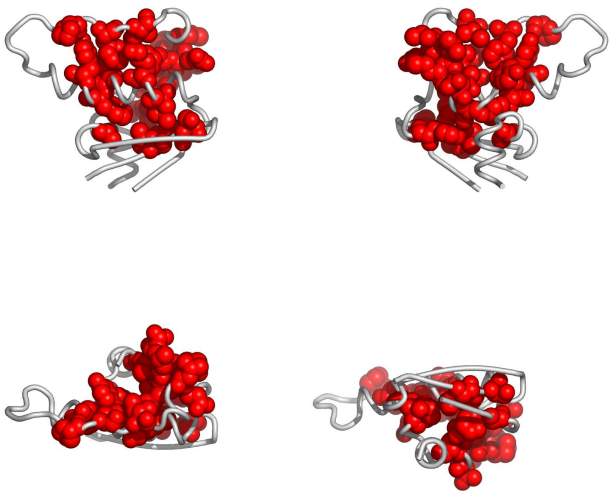


Fig. 9. Residues in 3dcgD, colored according to the cluster they belong to: red, followed by blue and yellow are the largest clusters (see Appendix for the coloring scheme). Clockwise: front, back, top and bottom views. The corresponding Pymol script is attached.

Table 8. *continued*

cluster color	size	member residues
		91,93,95,101,102,103,104,105 106,107,108,110

Table 8. Clusters of top ranking residues in 3dcgD.

3.4.2 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 the peptide 3dcgE. Table 9 lists the top 24% of residues at the interface with 3dcgE. The following table (Table 10) suggests possible disruptive replacements for these residues (see Section 4.6).

Table 9.

res	type	subst's (%)	cvg	noc/ bb	dist (Å)
107	A	A(93) S(6)	0.02	25/13	3.41
76	Y	Y(97) F(2)	0.04	69/0	2.69
101	L	L(89) M(4) V(2) I(4)	0.05	12/6	3.93
104	L	L(89)	0.06	21/4	3.60

continued in next column

Table 9. *continued*

res	type	subst's (%)	cvg	noc/ bb	dist (Å)
103	L	M(10) L(93) V(4) I(2)	0.07	17/5	3.75
79	Y	Y(83) W(12) F(4)	0.09	10/1	3.52
95	I	I(89) T(2) V(6) F(2)	0.10	16/3	3.78
93	F	F(70) M(22) R(2) L(2) D(2)	0.13	8/0	3.86
108	N	N(41) H(6) D(47) G(2) E(2)	0.15	20/14	3.66

Table 9. The top 24% of residues in 3dcgD at the interface with 3dcgE. (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 10.

res	type	disruptive mutations
107	A	(KR)(YE)(QH)(D)
76	Y	(K)(Q)(EM)(NR)
101	L	(Y)(R)(H)(T)
104	L	(Y)(R)(TH)(SCG)
103	L	(YR)(H)(T)(KE)
79	Y	(K)(Q)(E)(M)
95	I	(R)(Y)(K)(EH)
93	F	(T)(KE)(D)(CG)
108	N	(Y)(FW)(H)(R)

Table 10. List of disruptive mutations for the top 24% of residues in 3dcgD, that are at the interface with 3dcgE.

Figure 10 shows residues in 3dcgD colored by their importance, at the interface with 3dcgE.

Interface with 3dcgA. Table 11 lists the top 24% of residues at the interface with 3dcgA. The following table (Table 12) suggests possible disruptive replacements for these residues (see Section 4.6).

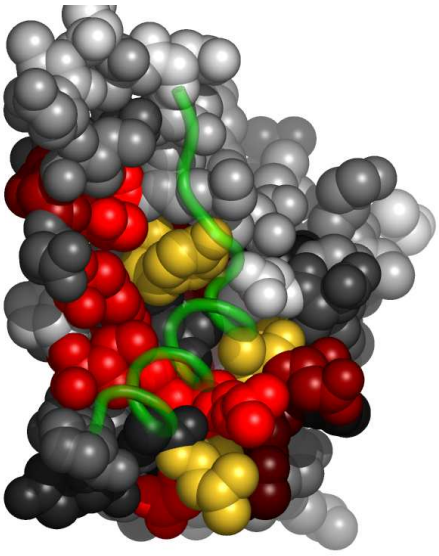


Fig. 10. Residues in 3dcgD, at the interface with 3dcgE, colored by their relative importance. 3dcgE is shown in backbone representation (See Appendix for the coloring scheme for the protein chain 3dcgD.)

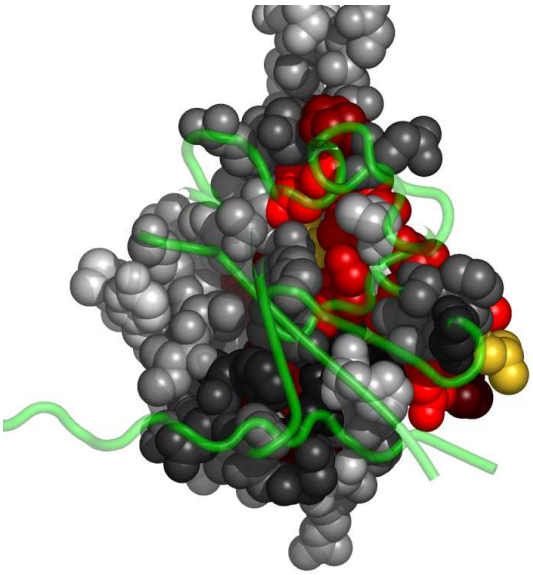


Fig. 11. Residues in 3dcgD, at the interface with 3dcgA, colored by their relative importance. 3dcgA is shown in backbone representation (See Appendix for the coloring scheme for the protein chain 3dcgD.)

Table 11.

res	type	subst's (%)	cvg	noc/ bb	dist (Å)
72	K	K(85) R(8) C(6)	0.12	1/0	4.83

Table 11. The top 24% of residues in 3dcgD at the interface with 3dcgA. (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 12.

res	type	disruptive mutations
72	K	(Y)(FW)(T)(S)

Table 12. List of disruptive mutations for the top 24% of residues in 3dcgD, that are at the interface with 3dcgA.

Figure 11 shows residues in 3dcgD colored by their importance, at the interface with 3dcgA.

Interface with 3dcgC. Table 13 lists the top 24% of residues at the interface with 3dcgC. The following table (Table 14) suggests possible disruptive replacements for these residues (see Section 4.6).

Table 13.

res	type	subst's (%)	cvg	noc/ bb	dist (Å)
102	E	E(89) D(10)	0.08	9/0	3.27
79	Y	Y(83) W(12) F(4)	0.09	32/9	3.05
72	K	K(85) R(8) C(6)	0.12	3/0	3.07
93	F	F(70) M(22) R(2) L(2) D(2)	0.13	12/2	3.72
91	P	P(81) S(4) T(8) E(2) F(4)	0.19	6/0	4.18

Table 13. The top 24% of residues in 3dcgD at the interface with 3dcgC. (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 14.		
res	**type**	**disruptive mutations**
102	E	(FWH)(R)(YVCAG)(T)
79	Y	(K)(Q)(E)(M)
72	K	(Y)(FW)(T)(S)
93	F	(T)(KE)(D)(CG)
91	P	(R)(Y)(H)(K)

Table 14. List of disruptive mutations for the top 24% of residues in 3dcgD, that are at the interface with 3dcgC.

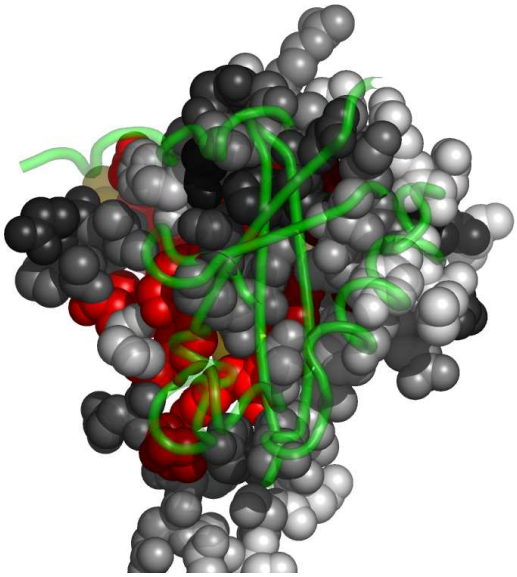


Fig. 12. Residues in 3dcgD, at the interface with 3dcgC, colored by their relative importance. 3dcgC is shown in backbone representation (See Appendix for the coloring scheme for the protein chain 3dcgD.)

Figure 12 shows residues in 3dcgD colored by their importance, at the interface with 3dcgC.

3.4.3 Possible novel functional surfaces at 24% coverage. One group of residues is conserved on the 3dcgD surface, away from (or susbtantially larger than) other functional sites and interfaces recognizable in PDB entry 3dcg. It is shown in Fig. 13. 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 15, while Table 16 suggests possible disruptive replacements for these residues (see Section 4.6).

Table 15.			
res	**type**	**substitutions(%)**	**cvg**
107	A	A(93)S(6)	0.02
76	Y	Y(97)F(2)	0.04
101	L	L(89)M(4)V(2)	0.05
continued in next column			

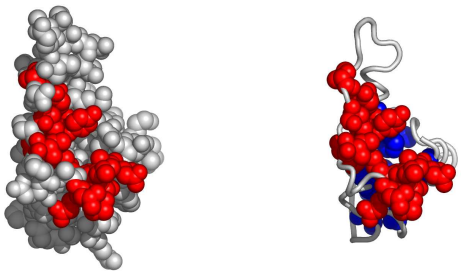


Fig. 13. A possible active surface on the chain 3dcgD. The larger cluster it belongs to is shown in blue.

Table 15. *continued*			
res	**type**	**substitutions(%)**	**cvg**
		I(4)	
104	L	L(89)M(10)	0.06
103	L	L(93)V(4)I(2)	0.07
102	E	E(89)D(10)	0.08
79	Y	Y(83)W(12)F(4)	0.09
95	I	I(89)T(2)V(6) F(2)	0.10
72	K	K(85)R(8)C(6)	0.12
93	F	F(70)M(22)R(2) L(2)D(2)	0.13
108	N	N(41)H(6)D(47) G(2)E(2)	0.15
106	A	A(87)V(2)K(6) T(4)	0.17
91	P	P(81)S(4)T(8) E(2)F(4)	0.19
105	M	M(62)L(16)I(6) V(10)A(2)T(2)	0.22

Table 15. Residues forming surface "patch" in 3dcgD.

Table 16.		
res	**type**	**disruptive mutations**
107	A	(KR)(YE)(QH)(D)
76	Y	(K)(Q)(EM)(NR)
101	L	(Y)(R)(H)(T)
104	L	(Y)(R)(TH)(SCG)
103	L	(YR)(H)(T)(KE)
102	E	(FWH)(R)(YVCAG)(T)
79	Y	(K)(Q)(E)(M)
95	I	(R)(Y)(K)(EH)
72	K	(Y)(FW)(T)(S)
93	F	(T)(KE)(D)(CG)
108	N	(Y)(FW)(H)(R)
106	A	(Y)(ER)(K)(H)
91	P	(R)(Y)(H)(K)
continued in next column		

Table 16. *continued*		
res	**type**	**disruptive mutations**
105	M	(Y)(H)(R)(T)

Table 16. Disruptive mutations for the surface patch in 3dcgD.

4 NOTES ON USING TRACE RESULTS

4.1 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.

4.2 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%.

4.3 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.)

4.4 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 $5\mathring{A}$.

4.5 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).

4.6 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.

5 APPENDIX

5.1 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

5.2 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.

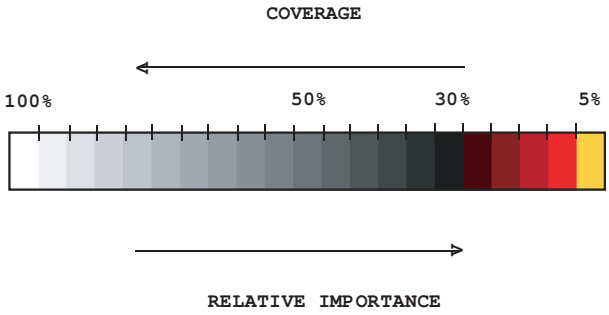


Fig. 14. Coloring scheme used to color residues by their relative importance.

The colors used to distinguish the residues by the estimated evolutionary pressure they experience can be seen in Fig. 14.

5.3 Credits

5.3.1 **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.

5.3.2 **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.

5.3.3 **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.`

5.3.4 **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/`

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

5.3.6 **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/`

5.3.7 **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.)

5.4 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.

5.5 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.

5.6 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.

5.7 Attachments

The following files should accompany this report:

- 3dcgA.complex.pdb - coordinates of 3dcgA with all of its interacting partners
- 3dcgA.etvx - ET viewer input file for 3dcgA
- 3dcgA.cluster_report.summary - Cluster report summary for 3dcgA
- 3dcgA.ranks - Ranks file in sequence order for 3dcgA
- 3dcgA.clusters - Cluster descriptions for 3dcgA

- 3dcgA.msf - the multiple sequence alignment used for the chain 3dcgA
- 3dcgA.descr - description of sequences used in 3dcgA msf
- 3dcgA.ranks_sorted - full listing of residues and their ranking for 3dcgA
- 3dcgA.3dcgB.if.pml - Pymol script for Figure 4
- 3dcgA.cbcvg - used by other 3dcgA – related pymol scripts
- 3dcgA.3dcgC.if.pml - Pymol script for Figure 5
- 3dcgD.complex.pdb - coordinates of 3dcgD with all of its interacting partners
- 3dcgD.etvx - ET viewer input file for 3dcgD
- 3dcgD.cluster_report.summary - Cluster report summary for 3dcgD
- 3dcgD.ranks - Ranks file in sequence order for 3dcgD
- 3dcgD.clusters - Cluster descriptions for 3dcgD
- 3dcgD.msf - the multiple sequence alignment used for the chain 3dcgD
- 3dcgD.descr - description of sequences used in 3dcgD msf
- 3dcgD.ranks_sorted - full listing of residues and their ranking for 3dcgD
- 3dcgD.3dcgE.if.pml - Pymol script for Figure 10
- 3dcgD.cbcvg - used by other 3dcgD – related pymol scripts
- 3dcgD.3dcgA.if.pml - Pymol script for Figure 11
- 3dcgD.3dcgC.if.pml - Pymol script for Figure 12