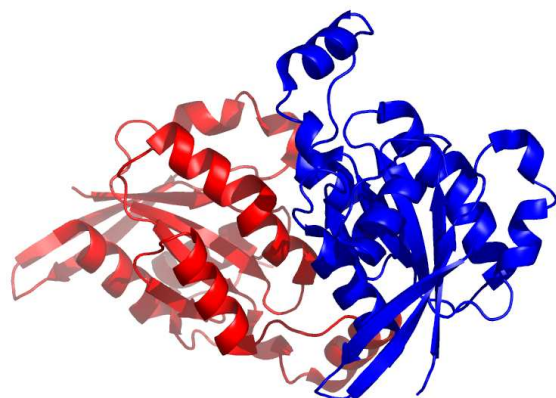


2c2h

Evolutionary trace report by **report_maker**

September 11, 2009



CONTENTS

1 Introduction

2 Chain 2c2hB

- 2.1 Q5U0M8 overview
- 2.2 Multiple sequence alignment for 2c2hB
- 2.3 Residue ranking in 2c2hB
- 2.4 Top ranking residues in 2c2hB 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.
 - 2.4.3 Possible novel 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	10
4.3.2	CE	10
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	11

1 INTRODUCTION

From the original Protein Data Bank entry (PDB id 2c2h):

Title: Crystal structure of the human rac3 in complex with gdp

Compound: Mol id: 1; molecule: ras-related c3 botulinum toxin substrate 3; chain: a, b; synonym: gtpase rac3, p21-rac3; engineered: yes; mutation: yes

Organism, scientific name: Homo Sapiens;

1 2c2h contains a single unique chain 2c2hB (173 residues long) and its homologue 2c2hA.

1

1

1

1

2 CHAIN 2C2HB

1

2

2.1 Q5U0M8 overview

From SwissProt, id Q5U0M8, 97% identical to 2c2hB:

2

2 **Description:** Ras-related C3 botulinum toxin substrate 3 (Rho family, small GTP binding protein Rac3).

7

Organism, scientific name: Homo sapiens (Human).

9

Taxonomy: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;

9

Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates;

9

Catarrhini; Hominidae; Homo.

9

9

9

9

2.2 Multiple sequence alignment for 2c2hB

9

For the chain 2c2hB, the alignment 2c2hB.msf (attached) with 913 sequences was used. The alignment was downloaded from the HSSP

9

9 database, and fragments shorter than 75% of the query as well as duplicate sequences were removed. It can be found in the attachment

10

10 to this report, under the name of 2c2hB.msf. Its statistics, from the *alistat* program are the following:

10

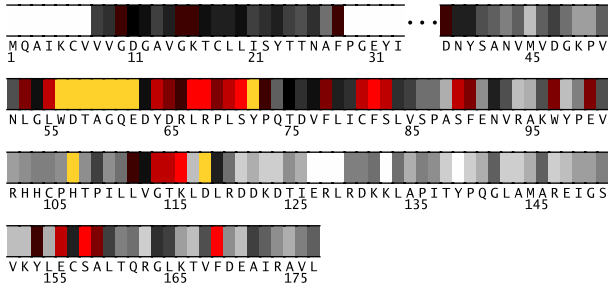


Fig. 1. Residues 1-177 in 2c2hB colored by their relative importance. (See Appendix, Fig.14, for the coloring scheme.)

```

Format:                MSF
Number of sequences:   913
Total number of residues: 140787
Smallest:              65
Largest:              173
Average length:       154.2
Alignment length:     173
Average identity:     41%
Most related pair:    99%
Most unrelated pair:  0%
Most distant seq:    34%

```

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

The alignment consists of 45% eukaryotic (13% vertebrata, 1% arthropoda, 12% fungi, 8% 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 2c2hB.descr.

2.3 Residue ranking in 2c2hB

The 2c2hB sequence is shown in Fig. 1, with each residue colored according to its estimated importance. The full listing of residues in 2c2hB can be found in the file called 2c2hB.ranks.sorted in the attachment.

2.4 Top ranking residues in 2c2hB 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 2c2hB 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.

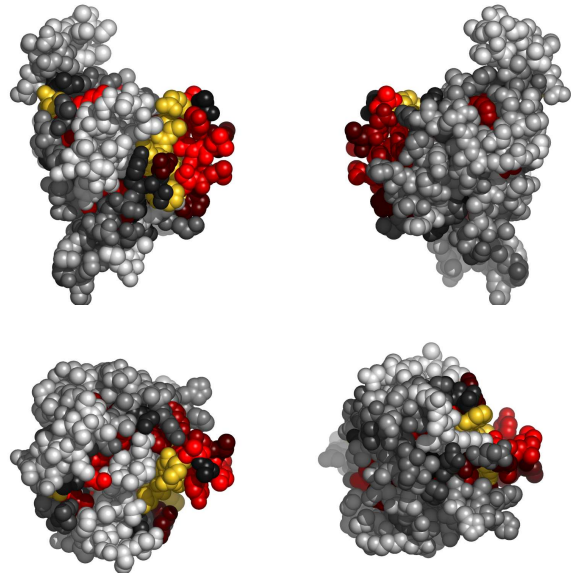


Fig. 2. Residues in 2c2hB, colored by their relative importance. Clockwise: front, back, top and bottom views.

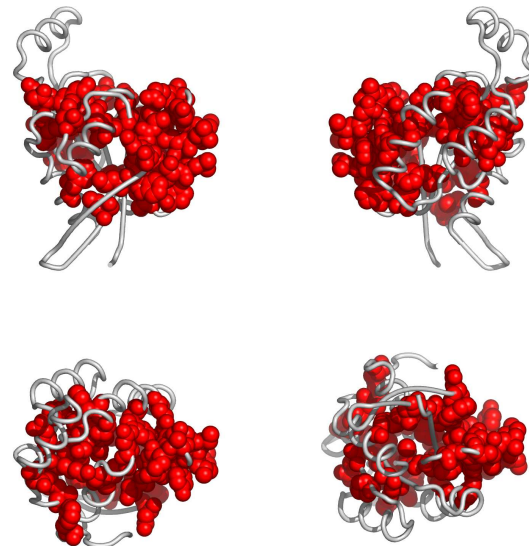


Fig. 3. Residues in 2c2hB, colored according to the cluster they belong to: red, followed by blue and yellow are the largest clusters (see Appendix for the coloring scheme). Clockwise: front, back, top and bottom views. The corresponding Pymol script is attached.

Table 1.		
cluster color	size	member residues
<i>continued in next column</i>		

cluster color	size	member residues
red	44	10, 11, 15, 16, 28, 38, 53, 55, 56 57, 58, 59, 60, 61, 62, 64, 65, 66 67, 68, 69, 70, 71, 72, 73, 75, 78 81, 82, 83, 89, 90, 97, 100, 112 114, 115, 116, 118, 154, 156, 158 159, 169

Table 1. Clusters of top ranking residues in 2c2hB.

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.

Calcium ion binding site. Table 2 lists the top 25% of residues at the interface with 2c2hBCA1182 (calcium ion). The following table (Table 3) suggests possible disruptive replacements for these residues (see Section 3.6).

res	type	subst's (%)	cvg	noc/ bb	dist (Å)	antn
107	H	103.(1) 44	0.01	5/0	2.05	site

Table 2. The top 25% of residues in 2c2hB at the interface with calcium ion. (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.)

res	type	disruptive mutations
107	H	(E)(TQMD)(K)(N)

Table 3. List of disruptive mutations for the top 25% of residues in 2c2hB, that are at the interface with calcium ion.

Figure 4 shows residues in 2c2hB colored by their importance, at the interface with 2c2hBCA1182.

GDP binding site. Table 4 lists the top 25% of residues at the interface with 2c2hBGDP1180 (gdp). The following table (Table 5) suggests possible disruptive replacements for these residues (see Section 3.6).

res	type	subst's (%)	cvg	noc/ bb	dist (Å)
73	P	R(30) P(44) N(1) Q(3)	0.23	1/0	4.96

continued in next column

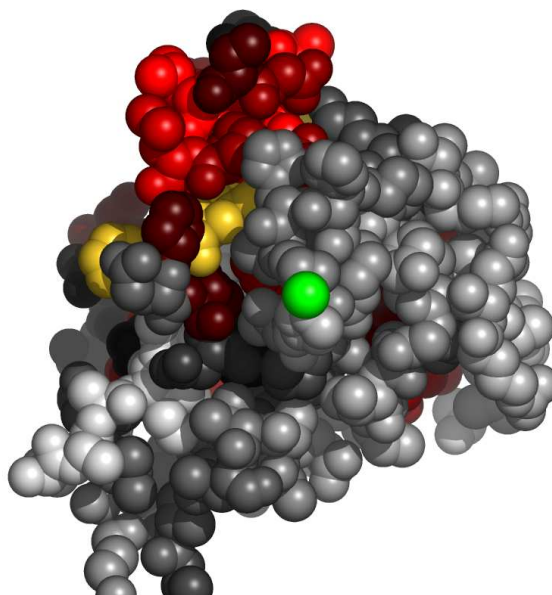


Fig. 4. Residues in 2c2hB, at the interface with calcium ion, colored by their relative importance. The ligand (calcium ion) is colored green. Atoms further than 30Å away from the geometric center of the ligand, as well as on the line of sight to the ligand were removed. (See Appendix for the coloring scheme for the protein chain 2c2hB.)

res	type	subst's (%)	cvg	noc/ bb	dist (Å)
		K(2) E(1) T(1) S(6) D(1) A(2) G(1).IY LHF			

Table 4. The top 25% of residues in 2c2hB at the interface with GDP. (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.)

res	type	disruptive mutations
73	P	(Y)(R)(H)(T)

Table 5. List of disruptive mutations for the top 25% of residues in 2c2hB, that are at the interface with GDP.

Figure 5 shows residues in 2c2hB colored by their importance, at the interface with 2c2hBGDP1180.

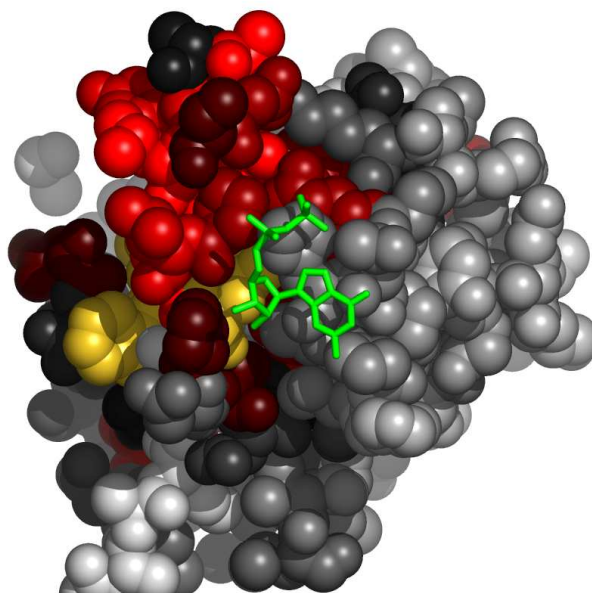


Fig. 5. Residues in 2c2hB, at the interface with GDP, colored by their relative importance. The ligand (GDP) is colored green. Atoms further than 30Å away from the geometric center of the ligand, as well as on the line of sight to the ligand were removed. (See Appendix for the coloring scheme for the protein chain 2c2hB.)

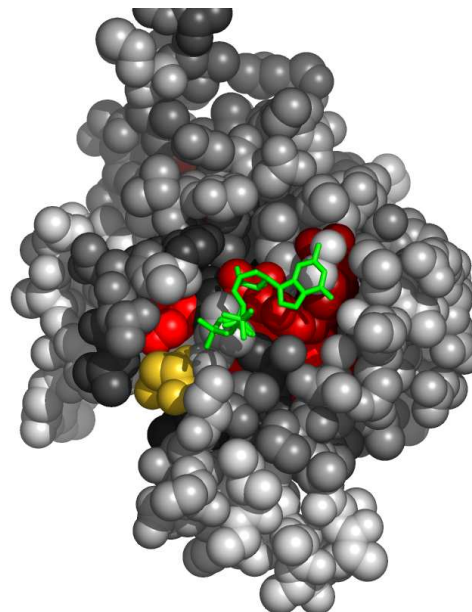


Fig. 6. Residues in 2c2hB, at the interface with GDP, colored by their relative importance. The ligand (GDP) is colored green. Atoms further than 30Å away from the geometric center of the ligand, as well as on the line of sight to the ligand were removed. (See Appendix for the coloring scheme for the protein chain 2c2hB.)

GDP binding site. By analogy with 2c2hA – 2c2hAGDP1182 interface. Table 6 lists the top 25% of residues at the interface with 2c2hAGDP1182 (gdp). 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)
156	E	E(92)L . (1)N Q(1)SAK RDTYGVI FP	0.13	24/5	2.67

Table 6. The top 25% of residues in 2c2hB at the interface with GDP.(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
156	E	(H) (FW) (Y) (R)

Table 7. List of disruptive mutations for the top 25% of residues in 2c2hB, that are at the interface with GDP.

Figure 6 shows residues in 2c2hB colored by their importance, at the interface with 2c2hAGDP1182.

GDP binding site. Table 8 lists the top 25% of residues at the interface with 2c2hBGDP1179 (gdp). 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)	antn
100	E	E(81) D(3) I(7) R(1)A L(1)VQG WK.PNHS YF	0.18	18/12	3.30	site

Table 8. The top 25% of residues in 2c2hB at the interface with GDP.(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 9.		
res	type	disruptive mutations
100	E	(H) (FW) (Y) (R)

Table 9. List of disruptive mutations for the top 25% of residues in 2c2hB, that are at the interface with GDP.

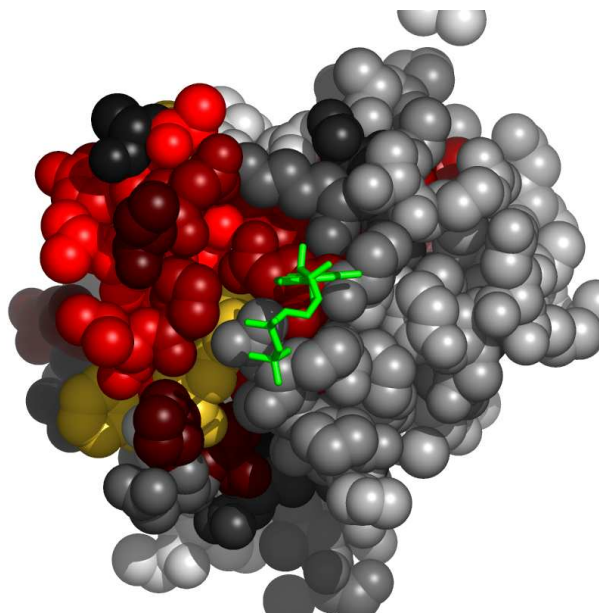


Fig. 7. Residues in 2c2hB, at the interface with GDP, colored by their relative importance. The ligand (GDP) is colored green. Atoms further than 30Å away from the geometric center of the ligand, as well as on the line of sight to the ligand were removed. (See Appendix for the coloring scheme for the protein chain 2c2hB.)

Figure 7 shows residues in 2c2hB colored by their importance, at the interface with 2c2hBGDP1179.

Magnesium ion binding site. Table 10 lists the top 25% of residues at the interface with 2c2hBMG1189 (magnesium ion). 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)	antn
57	D	D(96) . (1)GCE NSI	0.01	3/0	3.97	
58	T	T(96) . (1) S(1)VIA NGPC	0.02	1/1	4.61	
16	K	K(70)	0.21	4/1	4.75	site

continued in next column

Table 10. continued						
res	type	subst's (%)	cvg	noc/bb	dist (Å)	antn
		. (29)E				

Table 10. The top 25% of residues in 2c2hB at the interface with magnesium ion.(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 11.		
res	type	disruptive mutations
57	D	(R) (H) (FW) (Y)
58	T	(R) (K) (H) (Q)
16	K	(Y) (FW) (T) (VCAG)

Table 11. List of disruptive mutations for the top 25% of residues in 2c2hB, that are at the interface with magnesium ion.

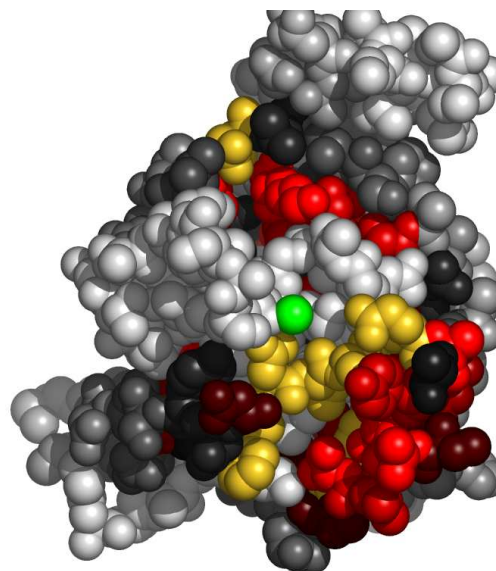


Fig. 8. Residues in 2c2hB, at the interface with magnesium ion, colored by their relative importance. The ligand (magnesium ion) is colored green. Atoms further than 30Å away from the geometric center of the ligand, as well as on the line of sight to the ligand were removed. (See Appendix for the coloring scheme for the protein chain 2c2hB.)

Figure 8 shows residues in 2c2hB colored by their importance, at the interface with 2c2hBMG1189.

GDP binding site. Table 12 lists the top 25% of residues at the interface with 2c2hBGDP1178 (gdp). 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
57	D	D(96) . (1)GCE NSI	0.01	2/0	4.68	
61	Q	Q(88) S(2)P T(1)L D(1)A N(1)HGM F. (1)RK E	0.03	1/0	4.75	
118	D	D(97)ES .HTPYN	0.04	17/0	2.76	
158	S	S(96)T . (1)CRY FLNDEA	0.06	9/5	3.45	site
116	K	K(82) Q(14)TR .MPNYCH F	0.07	53/10	3.20	site
159	A	S(18) A(75) V(1) . (1)GLI MDHTPF	0.16	12/8	2.96	site
15	G	G(70) . (29)E	0.20	38/38	3.01	site
16	K	K(70) . (29)E	0.21	32/17	2.70	site
28	F	F(65) . (29) Y(2)TIN CL	0.24	27/0	3.66	site
11	D	D(68) . (29) A(1)NGS E	0.25	4/4	3.84	

Table 12. The top 25% of residues in 2c2hB at the interface with GDP.(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 13.		
res	type	disruptive mutations
57	D	(R)(H)(FW)(Y)
61	Q	(Y)(H)(FTW)(CG)
118	D	(R)(FWH)(K)(VA)
158	S	(R)(K)(H)(FQW)
116	K	(Y)(T)(FW)(S)

continued in next column

res	type	disruptive mutations
159	A	(R) (K) (Y) (E)
15	G	(R) (K) (FWH) (M)
16	K	(Y) (FW) (T) (VCAG)
28	F	(K) (E) (QR) (D)
11	D	(R) (H) (FW) (Y)

Table 13. List of disruptive mutations for the top 25% of residues in 2c2hB, that are at the interface with GDP.

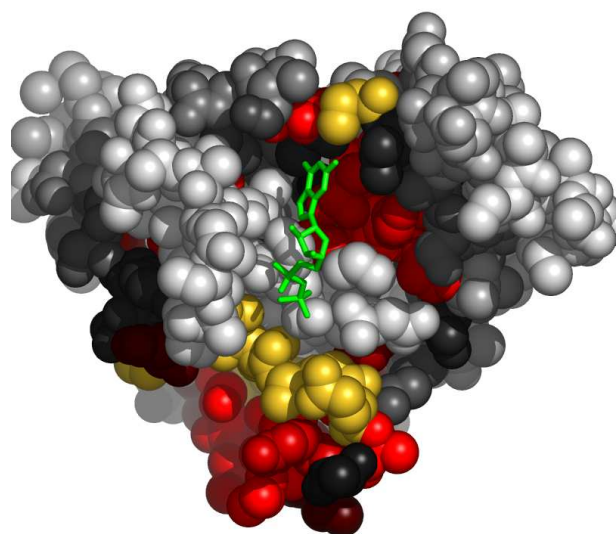


Fig. 9. Residues in 2c2hB, at the interface with GDP, colored by their relative importance. The ligand (GDP) is colored green. Atoms further than 30Å away from the geometric center of the ligand, as well as on the line of sight to the ligand were removed. (See Appendix for the coloring scheme for the protein chain 2c2hB.)

Figure 9 shows residues in 2c2hB colored by their importance, at the interface with 2c2hBGDP1178.

GDP binding site. Table 14 lists the top 25% of residues at the interface with 2c2hBGDP1181 (gdp). The following table (Table 15) suggests possible disruptive replacements for these residues (see Section 3.6).

res	type	subst's (%)	cvg	noc/ bb	dist (Å)	antn
107	H	103.(1) 44	0.01	71/4	3.03	site

Table 14. The top 25% of residues in 2c2hB at the interface with GDP.(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
107	H	(E) (TQMD) (K) (N)

Table 15. List of disruptive mutations for the top 25% of residues in 2c2hB, that are at the interface with GDP.

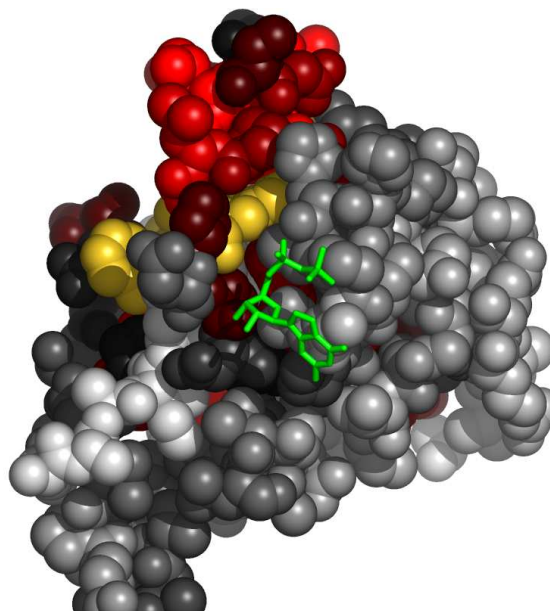


Fig. 10. Residues in 2c2hB, at the interface with GDP, colored by their relative importance. The ligand (GDP) is colored green. Atoms further than 30Å away from the geometric center of the ligand, as well as on the line of sight to the ligand were removed. (See Appendix for the coloring scheme for the protein chain 2c2hB.)

Figure 10 shows residues in 2c2hB colored by their importance, at the interface with 2c2hBGDP1181.

Sulfate ion binding site. Table 16 lists the top 25% of residues at the interface with 2c2hBSO41187 (sulfate ion). The following table (Table 17) suggests possible disruptive replacements for these residues (see Section 3.6).

res	type	subst's (%)	cvg	noc/ bb	dist (Å)	antn
107	H	103.(1) 44	0.01	5/0	3.37	site

Table 16. The top 25% of residues in 2c2hB at the interface with sulfate ion.(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 17.		
res	type	disruptive mutations
107	H	(E)(TQMD)(K)(N)

Table 17. List of disruptive mutations for the top 25% of residues in 2c2hB, that are at the interface with sulfate ion.

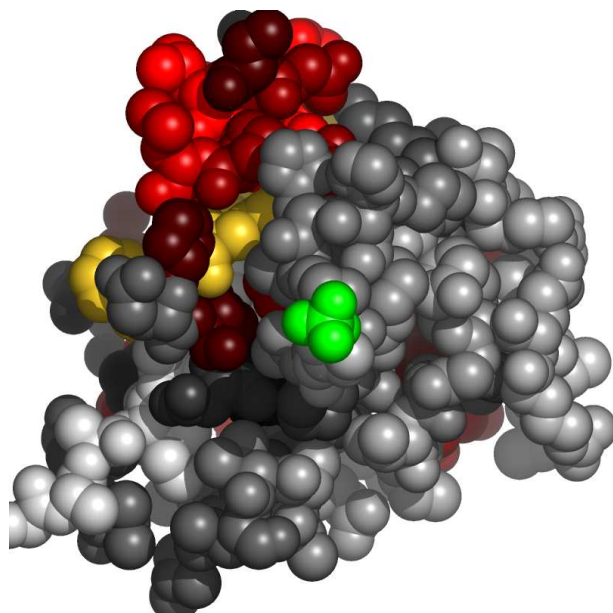


Fig. 11. Residues in 2c2hB, at the interface with sulfate ion, colored by their relative importance. The ligand (sulfate ion) is colored green. Atoms further than 30Å away from the geometric center of the ligand, as well as on the line of sight to the ligand were removed. (See Appendix for the coloring scheme for the protein chain 2c2hB.)

Figure 11 shows residues in 2c2hB colored by their importance, at the interface with 2c2hBSO41187.

Sulfate ion binding site. Table 18 lists the top 25% of residues at the interface with 2c2hBSO41188 (sulfate ion). The following table (Table 19) suggests possible disruptive replacements for these residues (see Section 3.6).

Table 18.						
res	type	subst's (%)	cvg	noc/bb	dist (Å)	antn
107	H	103.(1) 44	0.01	5/0	3.19	site

Table 18. The top 25% of residues in 2c2hB at the interface with sulfate ion.(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 19.		
res	type	disruptive mutations
107	H	(E)(TQMD)(K)(N)

Table 19. List of disruptive mutations for the top 25% of residues in 2c2hB, that are at the interface with sulfate ion.

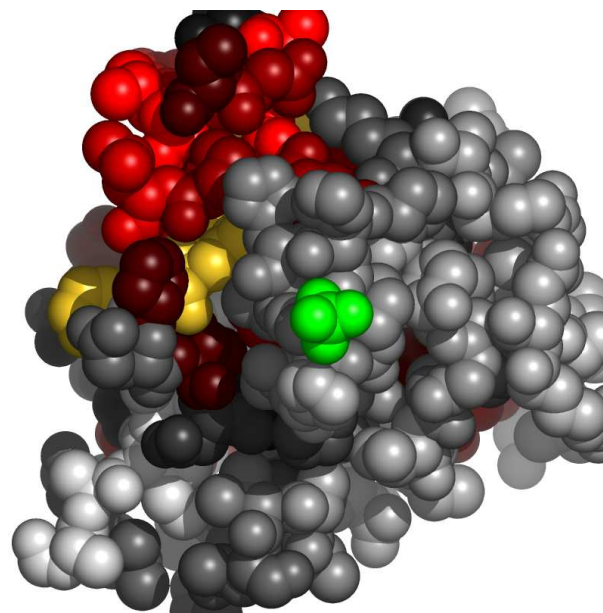


Fig. 12. Residues in 2c2hB, at the interface with sulfate ion, colored by their relative importance. The ligand (sulfate ion) is colored green. Atoms further than 30Å away from the geometric center of the ligand, as well as on the line of sight to the ligand were removed. (See Appendix for the coloring scheme for the protein chain 2c2hB.)

Figure 12 shows residues in 2c2hB colored by their importance, at the interface with 2c2hBSO41188.

2.4.3 Possible novel functional surfaces at 25% coverage. One group of residues is conserved on the 2c2hB surface, away from (or substantially larger than) other functional sites and interfaces recognizable in PDB entry 2c2h. 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 20, while Table 21 suggests possible disruptive replacements for these residues (see Section 3.6).

Table 20.				
res	type	substitutions(%)	cvg	antn
57	D	D(96).(1)GCENSI	0.01	

continued in next column

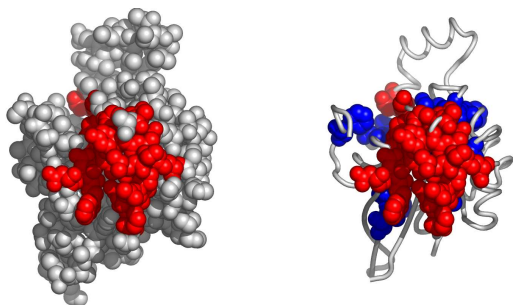


Fig. 13. A possible active surface on the chain 2c2hB. The larger cluster it belongs to is shown in blue.

Table 20. continued				
res	type	substitutions(%)	cvg	antn
60	G	G(96).(1)SADCVQ TNX	0.01	
58	T	T(96).(1)S(1)VI ANGPC	0.02	
59	A	A(91).(1)GCR F(1)S(3)LDTHMPE	0.02	
61	Q	Q(88)S(2)PT(1)L D(1)AN(1)HGMP . (1)RKE	0.03	
72	Y	Y(91)M(2)F(2) I(1)L.PN	0.04	
118	D	D(97)ES.HTPYN	0.04	
56	W	W(77)F(8)L(5) . (1)Y(2)C(3)IHR ANVKM	0.05	
62	E	E(86)SD(5)VH(1) KP(1)A.(1)CGMNT	0.05	
68	R	R(83)T(11)A(1)S IG(1)PK.HNVL	0.06	
67	L	L(75)I(13)M(4) V(3).(2)FNYPT	0.07	
116	K	K(82)Q(14)TR.MP NYCHF	0.07	site
71	S	S(63)Y(19)F(3)G A(5)C(4)WHL.IRT QMEPN	0.08	
64	Y	Y(71)F(22)L(2) K(1)DH.(1)IVNSR PM	0.12	
70	L	L(70)A(6)P(1) Q(3)S(5)N(1) I(1)F(2)M(2) T(1)RY.VGKHE	0.14	
115	T	T(43)N(29)C(12) S(1)L(11)G.AFMI V	0.14	
65	D	N(10)R(13)D(53)	0.17	

continued in next column

Table 20. continued				
res	type	substitutions(%)	cvg	antn
69	P	K(2)S(5)A(3) E(6)THQ(1)MPGV . (1)Y P(65)S(9)T(5) D(4)Q(1)H(1) R(3)A(2)IK(1)NL E(1).VCG	0.18	
100	E	E(81)D(3)I(7) R(1)AL(1)VQGWK. PNHSYF	0.18	site
15	G	G(70).(29)E	0.20	site
16	K	K(70).(29)E	0.21	site
66	R	R(59)A(7)S(9) T(6)N(3)Q(2) G(1)K(3)H.(2)CP EMVIY	0.22	
38	D	D(48).(20)E(29) NQ	0.23	
73	P	R(30)P(44)N(1) Q(3)K(2)E(1) T(1)S(6)D(1) A(2)G(1).IYLHF	0.23	
11	D	D(68).(29)A(1)N GSE	0.25	
75	T	A(31)S(15)T(44) G(5)V(1)NXYC(1) .IE	0.25	

Table 20. Residues forming surface "patch" in 2c2hB.

Table 21.		
res	type	disruptive mutations
57	D	(R)(H)(FW)(Y)
60	G	(R)(K)(E)(H)
58	T	(R)(K)(H)(Q)
59	A	(R)(K)(Y)(E)
61	Q	(Y)(H)(FTW)(CG)
72	Y	(K)(R)(Q)(E)
118	D	(R)(FWH)(K)(VA)
56	W	(E)(K)(TD)(Q)
62	E	(H)(FW)(R)(Y)
68	R	(D)(Y)(T)(E)
67	L	(R)(Y)(H)(T)
116	K	(Y)(T)(FW)(S)
71	S	(R)(K)(H)(Q)
64	Y	(K)(Q)(R)(E)
70	L	(Y)(R)(H)(T)
115	T	(R)(K)(H)(Q)
65	D	(R)(H)(FW)(Y)
69	P	(Y)(R)(H)(T)
100	E	(H)(FW)(Y)(R)

continued in next column

Table 21. <i>continued</i>		
res	type	disruptive mutations
15	G	(R) (K) (FWH) (M)
16	K	(Y) (FW) (T) (VCAG)
66	R	(D) (TY) (E) (CG)
38	D	(FWHR) (Y) (VCAG) (T)
73	P	(Y) (R) (H) (T)
11	D	(R) (H) (FW) (Y)
75	T	(R) (K) (H) (Q)

Table 21. Disruptive mutations for the surface patch in 2c2hB.

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*], 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

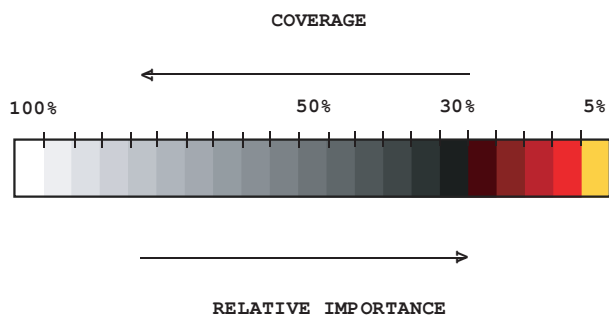


Fig. 14. Coloring scheme used to color residues by their relative importance.

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. 14.

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:

- 2c2hB.complex.pdb - coordinates of 2c2hB with all of its interacting partners
- 2c2hB.etvx - ET viewer input file for 2c2hB
- 2c2hB.cluster_report.summary - Cluster report summary for 2c2hB
- 2c2hB.ranks - Ranks file in sequence order for 2c2hB
- 2c2hB.clusters - Cluster descriptions for 2c2hB
- 2c2hB.msf - the multiple sequence alignment used for the chain 2c2hB
- 2c2hB.descr - description of sequences used in 2c2hB msf
- 2c2hB.ranks_sorted - full listing of residues and their ranking for 2c2hB
- 2c2hB.2c2hBCA1182.if.pml - Pymol script for Figure 4
- 2c2hB.cbcyg - used by other 2c2hB – related pymol scripts
- 2c2hB.2c2hBGDP1180.if.pml - Pymol script for Figure 5
- 2c2hB.2c2hAGDP1182.if.pml - Pymol script for Figure 6
- 2c2hB.2c2hBGDP1179.if.pml - Pymol script for Figure 7
- 2c2hB.2c2hBMG1189.if.pml - Pymol script for Figure 8
- 2c2hB.2c2hBGDP1178.if.pml - Pymol script for Figure 9
- 2c2hB.2c2hBGDP1181.if.pml - Pymol script for Figure 10
- 2c2hB.2c2hBSO41187.if.pml - Pymol script for Figure 11
- 2c2hB.2c2hBSO41188.if.pml - Pymol script for Figure 12