Integrating Mutation Data and Structural Analysis of the p53 Tumour-Suppressor Protein

Andrew C.R. Martin[1], Angelo M. Facchiano[2], Alison L. Cuff[1], Tina Hernandez-Boussard[3], Pierre Hainaut[3], Janet M. Thornton[4,5]

[1]School of Animal & Microbial Sciences
University of Reading
Whiteknights, P.O. Box 228, Reading RG6 6AJ, U.K.

[2]CRISCEB — Research Center of Computational
and Biotechnological Sciences
Second University of Naples
via Costantinopoli 16, 80138 Napoli, Italy

[3]International Agency for Research on Cancer,
150 cours Albert Thomas, Lyon 69372, France

[4]Biomolecular Structure & Modelling Unit,
Department of Biochemistry & Molecular Biology,
University College London,
Gower Steet, London WC1E 6BT, U.K.

[5]Department of Crystallography,
Birkbeck College,
Malet Street, London WC1E 7HX, U.K.

Running Title: p53 Mutation Data and Structural Analysis

Key words: p53, relational database, mutations, structural analysis

Abstract **p53 is a nuclear phosphoprotein with cancer-inhibiting properties. When DNA is damaged, it halts the progression of the cell cycle to allow repair enzymes to act or, if damage is too severe, it initiates apoptosis.**

p53 consists of 3 domains: an N-terminal transcription domain, a C-terminal oligomerisation domain and a DNA binding core domain. Mutations in p53 are associated with more than 50% of human cancers and 90% are in the core domain. These mutations affect the structural integrity and/or p53-DNA interactions, leading to the partial or complete loss of the protein's function. In some cases, function can be restored using second-site supressor mutations. Since p53 mediates cell killing in chemo-therapy and radio-therapy, the possibililty of designing drugs that restore functional activity of p53 is of obvious significance in cancer therapy.

Here we attempt to classify mutations in the core domain according to their effects on the structure of p53. A structural analysis was performed on

the p53 crystal structure and the results stored in a relational database. Raw mutation data were collected and imported into the database, which was then used to correlate mutation with structural effect in an automated manner.

The results of this analysis are published on the web (`http://www.bioinf.org.uk/p53/` or `http://www.rubic.rdg.ac.uk/p53/`). In summary, 304 of the 822 distinct muations were explained in structural terms, increasing to 515 when mutations to amino acids 100% conserved between diverse species were included.

In future, classifying p53 mutations into structural groups may provide an explanation for such properties as dominant-negative activity, temperature senstivity and oncogenic potential. The automated method of structural analysis developed here may also be applied to other mutations such as those of dystrophin, BRCA-I and G6PD.

Introduction

From the discovery of p53 in 1979, to the illucidation of its roles in the cell, the interest in this protein has increased continuously[1, 2]. p53 is a nuclear phosphoprotein with cancer-inhibiting properties[3, 4, 5, 6]. It is a multi-functional transcription factor with roles in control of cell cycle progression and apoptosis. Under normal conditions, p53 exists in an inactive state and is maintained at low levels. However, the level increases rapidly in response to DNA damage, hypoxia and nucleotide deprivation[7]. DNA damage increases the ability of p53 to bind DNA and activate a number of genes.

The mechanism of the p53 mediated suppression of cell cycle progression involves arrest within the G1 phase[6, 8] as a consequence of the p53 induced synthesis of p21, an inhibitor of cyclin E/cdk2 and cyclin A/cdk2 kinases. In this way, p53 gives DNA repair mechanisms time to correct damage before the genome is replicated. If damage to the cell is severe, p53 initiates apoptosis by inducing transcription of genes encoding proapoptotic factors[7, 9].

Tumour specific p53 mutations were first identified in 1989[10]; point mutations occur in more than 250 codons and are common in many forms of human cancer. Comparisons of p53 sequences from different species indicate 5 blocks of highly conserved residues which coincide with mutation clusters found in p53 in human cancers. 90% of mutations identified in p53 are in the core domain for which a crystal structure is available (note, however, that this value may be overestimated since most workers have concentrated their research on the core domain). 20% of the mutations are concentrated at 5 'hotspot' codons: 175, 245, 248, 249 and 273.

Endogenous processes, including methylation and deamination of cytosine at CpG residues, free radical damage, and errors that may occur during the synthesis or repair of DNA can result in p53 mutations[11]. Mutations can also occur via DNA damage induced by exogenous, physical or chemical carcinogens. In some cases "mutagen fingerprints" have been identified where certain carcinogens are responsible for specific mutations[12, 13]. For example, cigarette smoke causes G:C to T:A transversions in lung cancers[14] while aflatoxin B1 (AFB1) in the diet, particularly in China and Africa, causes G:C to T:A transversions specifically at the third base pair of codon 249 (AGG→AGT) and is associated with liver cancers. Similarly, UVB exposure is associated with CC:GG to TT:AA dipyrimadine transitions in skin cancers[15].

Inherited p53 mutations are rare. Li *et al.*[16] suggest 0.01% in the normal population and 0.1–1% in various cancer patients while Guinn and Padua[17] state that only 5% of p53 mutations are inherited. Germ-line mutations in the p53 gene have been observed in several families with Li-Fraumeni syndrome[18, 19]. This results in an inherited predisposition to a broad spectrum of cancers including breast cancer, osteosarcomas, soft tissue sarcoma, melanoma, adenocortical carcinomas and leukemias all of which appear at an early age.

More than 50% of all cancers involve the decreased or total loss of function of p53. This is caused, in most cases, by point mututions in one p53 allele. These mutations assert a dominant-negative effect over the remaining wild-type allele, resulting in genetic instability, loss-of-hetrozygosity and a deterimental effect on the function of p53[20]. Some may also exert their own oncogenic activity[8]. Correct functioning of p53 is critical to radiation and chemotherapy since both rely on causing DNA damage which triggers apoptosis *via* p53[20].

Raw mutation data have been collected over a number of years by groups in Germany and France. The databank of mutations, maintained by Hainaut[11], now in Release 4, consists of more than 14000 mutations affecting over 300 residues and linked with more than 60 different tumours. This collection of data is now being expanded with information on the pathology and clinical outcome of different mutations and tumours.

The open reading frame of human p53 codes for 393 amino acids with a central DNA-binding core domain (from approximately residue 100–300). The three-dimensional structure of this domain, complexed with DNA has been determined[21] and is shown in Figure 1. The N-terminal domain contains a strong transcription activation signal[22] while the C-terminal domain mediates oligomerisation. The core domain consists of a large β-sandwich of two anti-parallel sheets of 4 and 5 strands, respectively. This acts as a scaffold supporting 3 loop-based regions — a loop/β-sheet/α-helix motif (L1), and two large loops (L2 and L3). L2 and L3 are stabilised by zinc coordination and side-chain interactions[21, 23]. DNA is bound by L1 and L3 — the helix and loop for L1 slot into the major groove and L3 binds in the minor groove. The L2 loop stabilises L3 by packing against it. It has been proposed that p53 binds as a tetramer[24] and Pavletich *et al.*[25] stated the interactions occur through the C-terminal domain (residues 325–356)[26].

p53 mutations at or near the core domain are split into two distinct categories. The majority of distinct mutations affect residues essential for the DNA-binding domain's structural integrity (structural mutations). p53 has been shown to be only marginally stable at body temperature[27], so any mutation which further reduces stability is likely to lead to unfolding/misfolding *in vivo*. A smaller class of mutations (functional mutations) affect residues involved in p53–DNA interactions[20, 27], or in interactions with other proteins.

In theory, it should be possible to restore at least some functional activity to tumour-derived p53 mutants by (1) enhancing the stability of the protein in its folded state and/or (2) providing additional DNA contacts[20, 27]. It is possible to rescue some p53 mutations using second-site suppressor mutations. For example, the "hotspot" mutation G245S causes structural changes in L2 and L3, suggestive of distortion of the conformation necessary for DNA binding. Nikolova *et al.*[27] found that the suppressor mutant N239Y restored the stability of G245S and resulted in an improvement in DNA binding. They observed similar results using other second-site suppressors to restore some degree of normal function to other p53 mutations[27]. The marginal stability of p53 suggests that it may be possible to restore wild-type activity through design of drugs which bind the correctly folded form, thus moving the equilibrium through simple mass action[20, 23, 27].

Michalovitz *et al.*[28] suggested a genetic classification of mutations based on the dominance of their activity. Here we take a different approach to classifying mutations. We attempt to explain the effects of mutations in structural terms. Each of the observed mutations is classified in terms of the effect it is likely to have on the three-dimensional structure. We can define three categories of structural effect: (a) those which prevent the protein from folding into the correct conformation, (b) those which destabilise the folded protein (and may be temperature sensitive), (c)

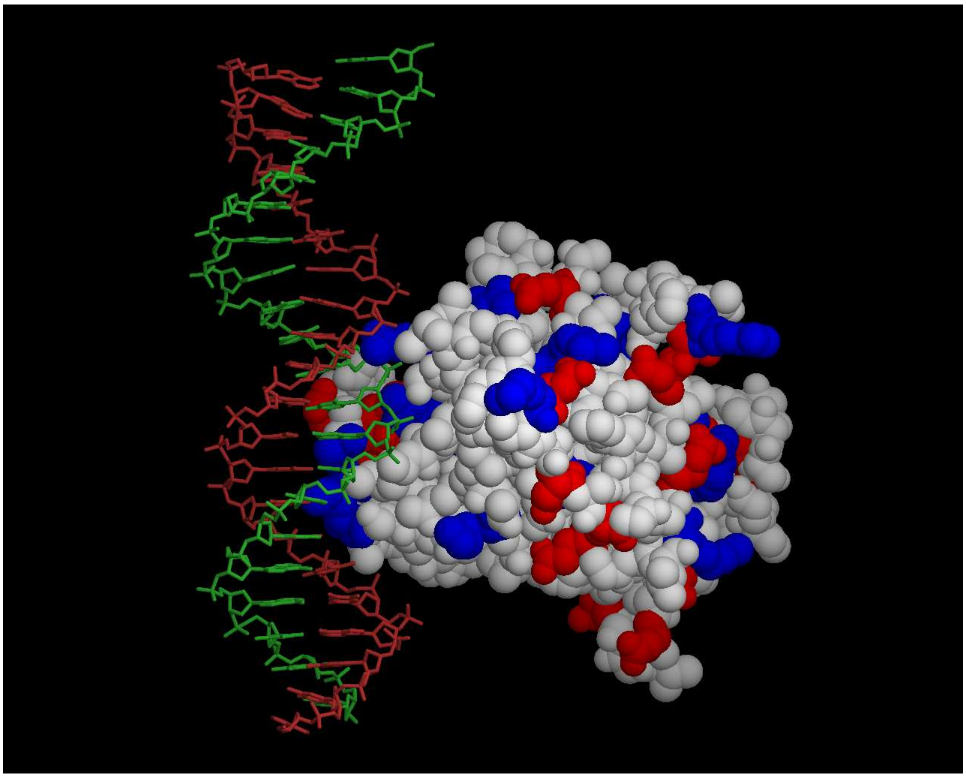

Figure 1: Crystal structure of the core domain of p53 bound to DNA as solved by Cho *et al.*[21].

those which are on the surface of p53 and interfere with the interactions of p53 with DNA or other proteins.

We find we are able to rationalise the effects of 34.4% of distinct mutations on purely structural grounds. If we also consider residues which are 100% conserved across a range of species (and therefore likely to be important for the function of p53), this percentage rises to 58.4%. This actually represents 80.5% of the total observed mutations and those which we cannot explain are thus relatively rare mutation events. Unexplained mutations will fall into one of three classes: (a) those which are not involved in cancer and are non-pathogenic; (b) those which we have genuinely failed to identify, possibly because they have only a slight destabilising effect; (c) those which are on the surface of the p53 core domain and are involved in interactions with the other p53 domains or with other proteins. Mutants in the first category may prove useful as markers to indicate that DNA damage has occurred and this will add to epidemiological information; those in the second category represent a deficiency in the current methodology; those in the third are clearly the most interesting.

We performed a structural analysis of the p53 crystal structure, calculating secondary structure, backbone torsion angles, solvent accessibility and hydrogen bonding parameters and stored these data in a relational database. By also storing mutant data in the database we can correlate structural effects with mutations in a relatively automated fashion.

Materials and Methods

Mutation data

The raw mutation data available from `ftp://ftp.ebi.ac.uk/pub/databases/p53/` were imported into a PostgreSQL relational database (`http://www.PostgreSQL.org/`) using a script written in Perl to make small changes to the format. The raw data[11] contain p53 mutations associated with human cancers identified by sequencing and published in the literature. These data include mutations found in normal, pre-neoplastic and neoplastic tissues, including metastases, as well as cell lines derived from such tissues. The data file contains 34 columns and includes data on cell-line, codon, DNA base and amino acid substitution, International Classification of Diseases for Oncology (ICD-O) tumour-site, tumour morphology and histology, tumour grade or stage, and risk factors (sex, country of origin, smoking status and alcohol consumption).

We considered both in-frame and out-of-frame insertions in the same manner; in both cases it is clear that the function of p53 could be disrupted. We also flagged silent point mutations. Earlier versions of the p53 data required considerable clean-up during this procedure; the current dataset required minimal clean-up (some frameshift mutants classified as 'point' rather than 'del' or 'ins', minor changes to the page numbering format of references, etc.). For completeness, the citation data were also imported into a second database table.

Structural data

Our structural analysis was based on PDB file 1tsr[21]. The parameters calculated were: secondary structure using DSSP[29], hydrogen bonding using HBPlus[30], backbone torsion angles and solvent accessibility[31] using NAccess (Simon Hubbard, unpublished). These data were imported into a third database table keyed by residue (codon) number.

Sequence variability

Five regions of conserved residues were defined using the PRINTS procedure of Attwood *et al.*[32, 33]. PRINTS defines 'fingerprint' regions which contain no insertions or deletions. Sequences used in this analysis came from human, cat, golden hamster, bovine, sheep, mouse, rainbow trout, rat, chicken, North European squid, dog, green monkey (*Cercopithecus aethiops*), *Macaca mulatta*, *Xenopus laevis* and *Spermophilus beecheyi.*

We considered sequence variability on the basis that residues which are 100% conserved across such a diverse selection of species must be conserved for functional reasons. Thus we may not have direct structural explanations of why mutations to these residues might affect the function of p53, but we know that these residues are critical to the function of p53 and this is likely to be as a result of interactions with other proteins.

At each residue position in the fingerprint regions, the sequence variability was assessed using a score based on the PET91 mutation matrix[34] normalised such that all scores on the diagonal are maximal and equal. The score is calculated as the average pairwise sum of the matrix scores normalised by the maximum score in the matrix:

$$S_n = \left(\sum_{i=1}^{N} \sum_{j=i+1}^{N} s_{ij} / {}_N C_2 \right) / s_{\max}$$

where n is the position in the sequence, N is the number of sequences, s is a score from the mutation matrix and ${}_N C_2$ is the number of combinations of two elements from the set of N elements (${}_n C_r = n!/((n-r)!r!)$). In this scheme, complete conservation scores 1.0; lower levels of conservation score values down to 0.0, depending not only on the raw variability (as is the case with statistical entropy based scores[35]), but also on the nature of the mutation. The sequence variability scores are stored in the structural data table and are illustrated in Figure 2.

Assessing sidechain replacements

For the present study, very simple assessments of the effects of changes in the structural properties were used. For example, if a residue was involved in donating a sidechain hydrogen bond and is replaced by a residue without hydrogen-bond donor potential we claim to have explained the structural effect of the mutation. If the replacement sidechain is also able to donate a hydrogen-bond, the geometry of the new hydrogen-bond is not tested, it is assumed that small changes in the structure can be accommodated. We thus take a cautious approach and do not classify such mutants as explained even though they may, in fact, be explained in this way.

Each unique sidechain replacement is also assessed on the basis of steric acceptability. The current procedure is again very simple; we adopt a minimum perturbation protocol (MPP)[36] to model the new sidechain into the 3D crystal structure of the p53 core domain and then count any bad clashes with the substituted sidechain. MPP proceeds as follows:

1. Perform a maximum overlap protocol (MOP)[36] replacement of the sidechain where torsion angles are inherited from the parent sidechain where possible.

2. Build a near neighbours list using a cutoff of 8Å (this is greater than the longest sidechain, tryptophan).

3. Spin the sidechain about $\chi 1$ and $\chi 2$ torsion angles in 30° steps flagging each position as either making bad contacts or not.

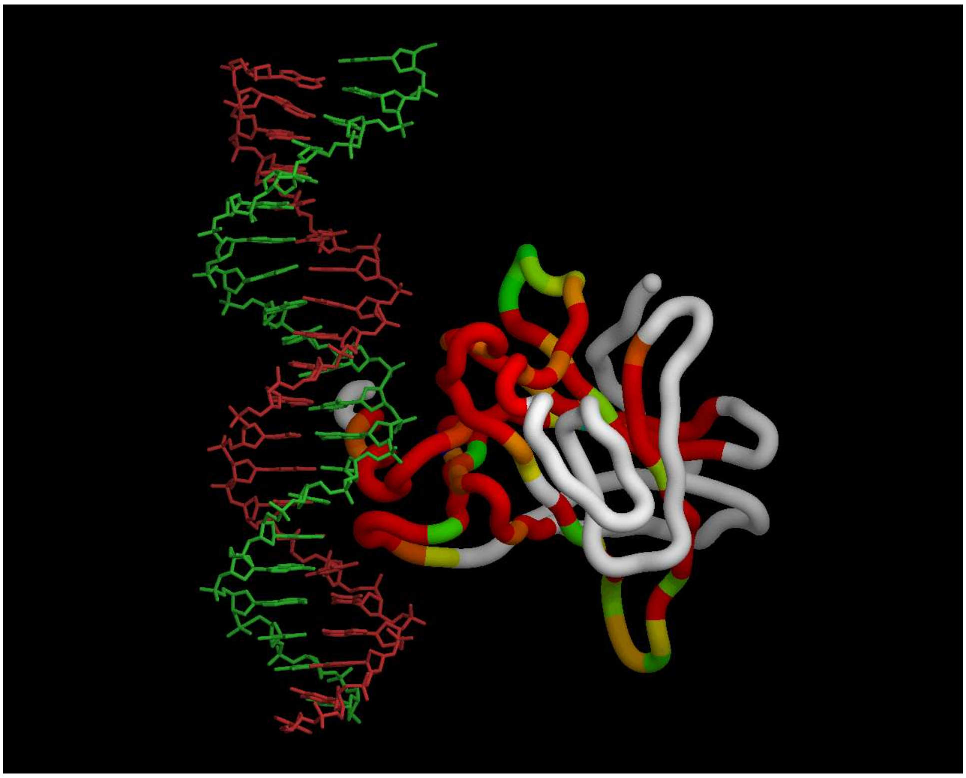

Figure 2: Structure of p53 showing conservation of the fingerprint regions. Non-fingerprint regions are shown in grey with the fingerprint regions coloured from blue (low conservation) to red (high conservation) based on the sequence variability score (see text).

	Total Observed	Distinct
Total number of mutations	14050	1729
Complex mutations	69	60
Deletions	1250	253
Insertions	357	152
Tandem mutations	200	67
Silent mutations	5	1
Point mutations	12138	1363
Of these:		
Tandem/Point mutations resulting in an amino acid substitution	10204	1083
Tandem/Point mutations resulting in an amino acid substitution in core domain	9812	882

Table I: Summary of p53 mutation data

4. If the parent conformation (resulting from MOP) makes zero or 1 bad contacts then that conformation is accepted.

5. If that fails, then for all the conformations with zero or 1 bad contacts, a choice is made from allowed rotamers.

6. If that fails, the first conformation with a minimal number of clashes is selected.

A bad contact is defined as two atoms whose centres are closer than 2.5Å — this is a simple good/bad assessment; no degree of bad contact is calculated. We take 3 clashes as being indicative of a sidechain replacement which cannot be accommodated. Again this is a conservative decision; it appears that 2 clashes are sufficient to disrupt the structure in many cases.

By using the ability of PostgreSQL to allow user-defined functions, the clash assessment can be performed on-the-fly. In practice, for speed reasons, it is useful to cache the results of all unique sidechain replacements into a column in another database table. This can be achieved by performing a single SQL query on the database. These data were stored in a fourth table keyed by residue (codon) number and replacement residue type.

Analysing the data

Analysis of the data was performed using a set of Perl routines which query the database and extract and format the data. The procedure has been completely automated such that it can be repeated on new datasets as these become available. The results of this analysis are available on the Web (`http://www.bioinf.org.uk/p53/` or `http://www.rubic.rdg.ac.uk/p53/`).

Results

Summary of data

Table I summarizes the mutation data from Release 4 of the p53 mutation databank. For the purposes of this investigation, we have concentrated on analyzing the

	Total mutations involving H-bonding residues		H-bonding potential not conserved	
	Obs	Distinct	Obs	Distinct
Donor	3856	205	1422	104
Acceptor	1479	155	881	85

Table II: Mutations to residues involved in hydrogen-bonding. The first pair of columns shows the numbers of mutations and the second pair of columns shows the numbers mutations where hydrogen-bonding potential is lost.

distinct mutations which result in a simple amino acid substitution in the core domain for which a crystal structure is available. As the table shows, there are 882 of these. This is approximately 51% of the total number of the distinct mutations; the remaining 49% are either more complex mutations, insertions, deletions, or occur outside the core domain. These simple substitution mutations in the core domain represent 69.8% of the total number of observed mutations.

Mutations affecting hydrogen bonding

Hydrogen bonds stabilise the structure of a protein and hydrogen bonding ability must be satisfied throughout. If a residue involved in a hydrogen bond is substituted by another residue unable to form the hydrogen-bond the protein will be destabilised.

As described by Baker and Hubbard[37] the following residues are classified as able to donate a hydrogen bond: H,K,N,Q,R,S,T,W,Y while the following residues can accept a hydrogen bond: D,E,H,N,Q,S,T,Y. There is a total of 4703 substitution mutations (309 distinct mutations) involving hydrogen bonding residues. Using our conservative assessment of explaining hydrogen bonding mutations (described in the Methods) where we do not consider the precise geometry and assume that a small local rearrangement can be accomodated, we find that we can explain 43.2% of observed mutations to hydrogen bonding residues (52.5% of distinct mutations). See Table II.

Mutations to proline

Owing to the cyclic sidechain of proline, the backbone is more restricted in the conformations which it can adopt. Thus mutations from other residues to proline may result in distortion of the structure if the parent amino acid did not adopt a backbone conformation permitted for proline. In addition, proline residues will break an α-helix and all but edge β-strands since cyclisation of the sidechain prevents the regular backbone H-bond formation. Thus mutations to proline in these circumstances will lead to an incorrectly folded protein. A total of 332 point/tandem mutations (58 distinct mutations) result in a mutation to proline (in addition there are 77 silent mutations at 20 distinct sites involving proline, of which 62 occur in the core at 12 distinct sites).

Of the 332 mutations resulting in a substitution by proline, 320 occur in the core at 50 distinct sites. Table III shows these core domain substitutions together with the backbone torsion angles of the parent structure. Those combinations which are disallowed regions for proline are indicated. We define the allowed regions for Proline as $-70° \leq \phi \leq -50°$ and ($-70° \leq \psi \leq -50°$ or $110° \leq \psi \leq 130°$). 47 of the 50 mutations (94%) are disallowed and will thus result in disruption of the

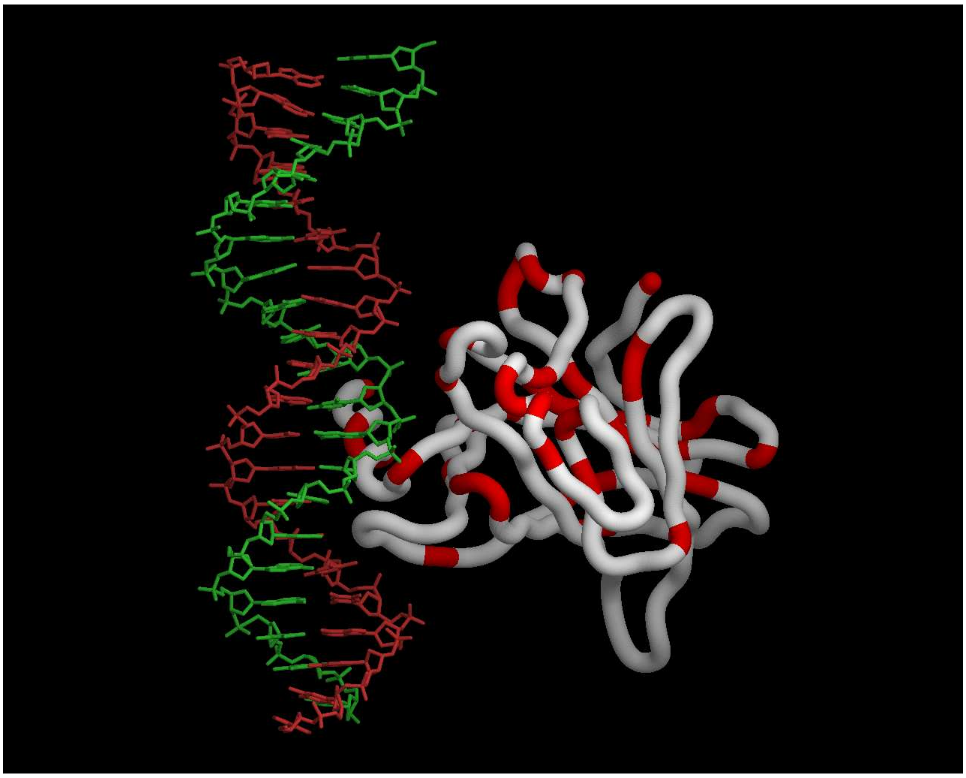

Figure 3: The 47 sites where disallowed proline substitutions are observed are indicated in red.

structure. Some of these, however, are borderline and may be accomodated by a very small rearrangement (e.g. Leu137→Pro). The 47 disallowed Proline mutations sites are illustrated in Figure 3.

Mutations from glycine

A total of 809 mutations (70 distinct mutations) are observed from a native glycine to another residue (in addition there are 71 silent mutations of glycine at 14 distinct sites). 771 of these (53 distinct) occur in the core region.

Because it has no sidechain, glycine is able to adopt conformations which are sterically hindered for other amino acids. Substitution of any native glycine residues which adopt one of these conformations will thus result in disruption of the structure resulting in an incorrectly folded protein.

The allowed regions of the Ramachandran plot for non-glycine/non-proline residues are, for this purpose, defined as: $(-180.0 \leq \phi \leq -30.0/60.0 \leq \psi \leq 180.0)$ or $(-155.0 \leq \phi \leq -15.0/-90.0 \leq \psi \leq 60.0)$ or $(-180.0 \leq \phi \leq -45.0/-180.0 \leq \psi \leq -120.0)$ or $(30.0 \leq \phi \leq 90.0/-20.0 \leq \psi \leq 105.0)$. All non-glycine residues in the p53 crystal structure fall within these limits.

With the exception of the glycine residues at codons 117, 154, 187, 244, 245 and 262, all the others fall in regions allowed for other amino acids. Therefore, only mutations to these 6 glycines will result in disruption of the structure. These sites are illustrated in Figure 4. Table IV shows the substitutions of glycine residues by other amino acids and it can be seen that 32 of 53 core region distinct mutations (60.4%) are disallowed.

Codon	Amino acid	Number of mutations	Secondary structure	ϕ	ψ	Disallowed?
96	ser	1	–	–	-157.73	$\surd$
110	arg	4	E	-144.137	151.752	$\surd$
111	leu	2	E	-113.239	147.026	$\surd$
116	ser	1	–	-76.39	-39.217	$\surd$
127	ser	3	E	-109.949	100.816	$\surd$
136	gln	2	–	-100.488	152.672	$\surd$
137	leu	3	T	-56.519	131.157	$\surd$
138	ala	14	T	71.046	-6.909	$\surd$
140	thr	2	–	-66.207	125.858	
144	gln	5	E	-92.668	136.71	$\surd$
145	leu	8	E	-100.898	118.692	$\surd$
149	ser	3	S	-146.477	133.016	$\surd$
155	thr	11	–	-61.553	129.245	
156	arg	22	E	-119.194	161.537	$\surd$
158	arg	13	E	-127.57	139.746	$\surd$
159	ala	21	E	-119.554	146.868	$\surd$
161	ala	1	E	-106.536	157.296	$\surd$
165	gln	3	S	-71.981	138.63	$\surd$
166	ser	1	T	-40.607	-64.889	$\surd$
168	his	3	G	-66.524	-13.923	$\surd$
170	thr	1	G	-88.444	-18.605	$\surd$
175	arg	4	–	-65.109	147.266	$\surd$
178	his	7	H	-56.791	-61.784	$\surd$
179	his	3	H	-57.75	-31.935	$\surd$
181	arg	11	H	-62.976	-43.551	$\surd$
183	ser	2	–	-62.278	154.117	$\surd$
189	ala	3	–	-63.187	129.726	
193	his	7	–	-70.794	133.586	$\surd$
194	leu	9	S	-81.987	-44.986	$\surd$
196	arg	11	E	-122.688	161.578	$\surd$
202	arg	3	T	-115.442	7.853	$\surd$
213	arg	3	–	-37.348	132.738	$\surd$
214	his	1	E	-102.983	146.49	$\surd$
241	ser	6	T	-92.341	12.954	$\surd$
247	asn	1	T	44.311	28.416	$\surd$
248	arg	11	T	86.753	7.922	$\surd$
252	leu	8	E	-118.307	137.257	$\surd$
253	thr	2	E	-103.296	127.718	$\surd$
257	leu	10	E	-83.297	136.58	$\surd$
260	ser	1	T	-47.916	-18.882	$\surd$
264	leu	1	E	-63.609	126.162	$\surd$
265	leu	10	E	-114.011	-21.378	$\surd$
267	arg	8	E	-158.106	140.954	$\surd$
271	glu	1	E	-89.422	163.063	$\surd$
273	arg	20	E	-126.608	110.838	$\surd$
276	ala	15	S	-90.697	143.622	$\surd$
282	arg	10	H	-80.193	-38.59	$\surd$
283	arg	21	H	-50.147	-57.613	$\surd$
284	thr	4	H	-52.739	-57.402	$\surd$
289	leu	3	–	90.022	—	$\surd$

Table III: Mutations to proline occurring in the DNA binding domain of p53. Secondary structure as assigned by DSSP in the parent is indicated (E: β-strand, H: α-helix, T: turn, S: bend, G: 3_{10}-helix). Mutations where the parent amino acid was in a region disallowed for proline are flagged.

Codon	Substitution	Number of mutations	Secondary Structure	Phi	Psi	Disallowed?
105	arg	4	–	-101.432	-162.392	
105	val					
108	asp	1	–	73.438	59.243	
112	asp	2	E	-122.894	159.618	
112	ser					
117	arg	4	–	107.595	165.374	√
117	glu					√
154	asp	48	T	93.49	-18.147	√
154	ser					√
154	val					√
187	arg	12	S	96.541	-9.697	√
187	asp					√
187	cys					√
187	ser					√
187	val					√
199	ala	19	S	54.791	42.745	
199	arg					
199	glu					
199	val					
226	ala	4	T	79.884	2.235	
226	asp					
226	ser					
244	ala	116	T	97.695	-23.373	√
244	arg					√
244	asp					√
244	cys					√
244	gln					√
244	glu					√
244	ser					√
244	val					√
245	ala	412	T	-117.227	-115.488	√
245	arg					√
245	asp					√
245	cys					√
245	glu					√
245	his					√
245	leu					√
245	ser					√
245	thr					√
245	val					√
262	asp	9	–	106.316	8.547	√
262	cys					√
262	ser					√
262	val					√
266	ala	108	E	-167.723	162.972	
266	arg					
266	gln					
266	glu					
266	val					
279	arg	32	H	-68.848	-54.811	
279	glu					
279	leu					
279	trp					

Table IV: Mutations from glycine occurring in the DNA binding domain of p53. Secondary structure as assigned by DSSP in the parent is indicated (E: β-strand, H: α-helix, T: turn, S: bend, G: 3_{10}-helix). Mutations where the parent glycine is in a region disallowed for other amino acids are flagged.

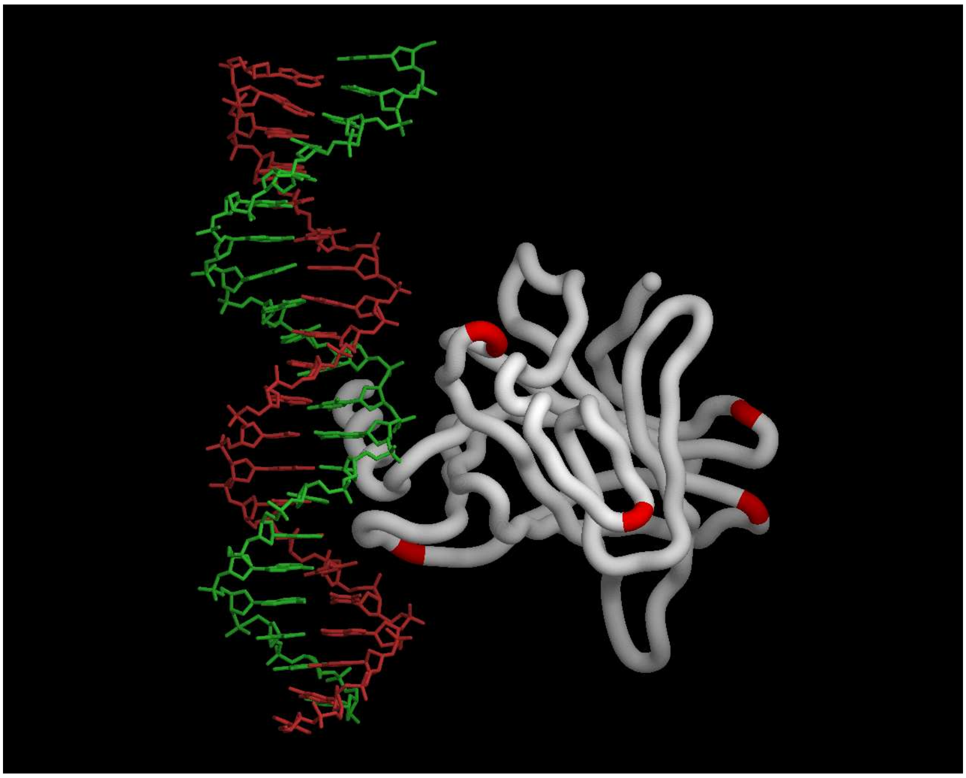

Figure 4: The 6 sites at which glycines adopt backbone conformations disallowed for other residues and where substitutions occur are indicated in red.

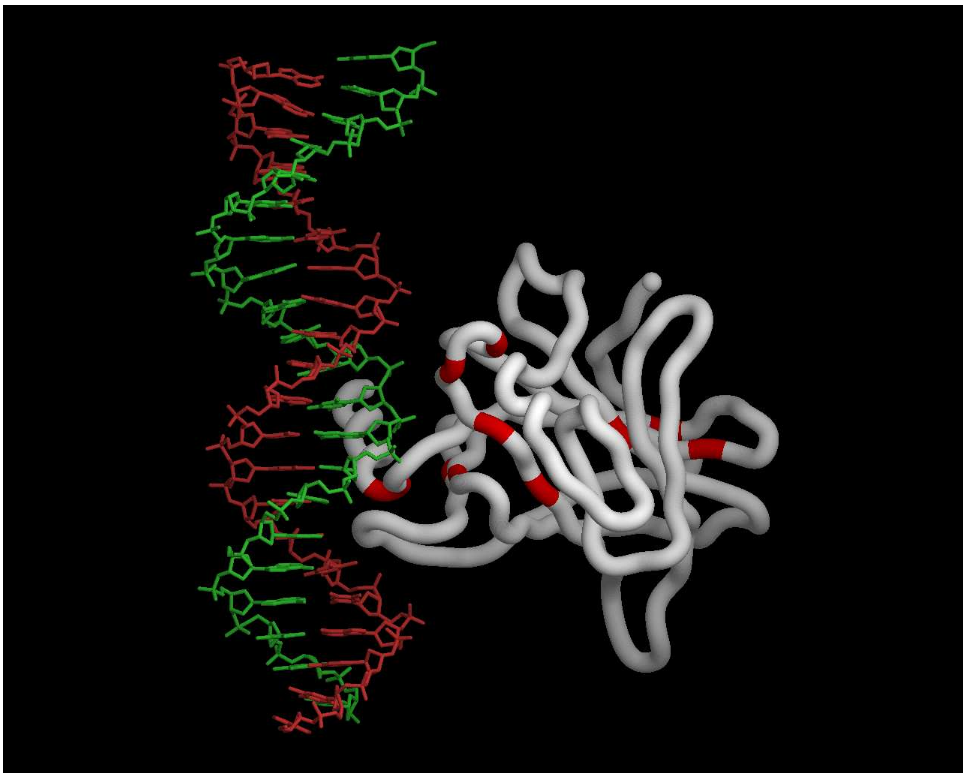

Figure 5: The 11 sites at which substitutions result in bad clashes are indicated in red.

Residue clashes

If a substituted residue is too large for the available space it will lead to distortion of the structure and may result in the protein folding incorrectly. Of the 882 distinct substitution mutations in the core, 24 (2.7%) result in a bad clash (3 or more bad contacts with surrounding atoms in the best sidechain orientation). If we consider that *any* number of bad contacts will disrupt the structure, we can include a further 44 distinct mutations, resulting in a total of 68 (7.7%) mutations resulting in bad clashes.

Mutations involving DNA binding

The most common mutations observed in p53 are involved in binding DNA. These mutations result in the protein either being unable to bind to p53 or losing specificity of interactions. We define DNA binding residues as those in which the relative accessibility changes by at least 5% between the complexed form of p53 observed in the crystal structure and the same structure of p53 but with the DNA removed. This identifies 14 residues (Ala119, Lys120, Ser121, Asn239, Ser241, Met243, Asn247, Arg248, Arg273, Cys275, Ala276, Cys277, Arg280, Arg283) all of which are seen to have mutations.

At these 14 sites, a total of 2383 mutations is observed, 74 of which are distinct. While mutations at the more peripheral of these sites may, in some circumstances, allow DNA still to bind, the stability of the complex and the specificity of DNA binding is likely to be affected and this will affect the function of the protein.

Mutations involving zinc binding

Zinc binding is essential for the function of p53 — presumably it does not adopt the correct conformation in the absence of zinc binding. Thus mutations to the residues involved in interaction with zinc will result in p53 being non-functional. Examination of the crystal structure shows that Cys176, His 179, Cys238 and Cys242 are all involved in zinc binding. A total of 611 mutations is observed at the 4 sites, 29 of which are distinct. Any mutation to these residues is likely to prevent or destabilise zinc binding, destabilizing the structure and resulting in loss of function.

Mutations to conserved regions

The analysis of fingerprint regions reveals a total of 73 residues in the core domain which are 100% conserved across all species for which p53 sequences were analysed (see methods). These are residues 118, 119, 120, 121, 122, 124, 125, 126, 127, 130, 131, 132, 137, 138, 139, 141, 142, 158, 159, 161, 163, 164, 172, 173, 175, 177, 178, 179, 214, 215, 216, 218, 219, 220, 221, 223, 230, 231, 234, 238, 239, 240, 241, 242, 243, 244, 245, 247, 249, 250, 251, 253, 256, 257, 258, 264, 265, 266, 267, 270, 271, 272, 273, 275, 276, 277, 278, 279, 280, 281, 282, 285 and 286.

While we cannot offer a direct structural explanation for many of these, one can assume that they are conserved throughout evolution for a good reason and, in the case of surface residues, this is likely to be that the amino acid is critical for interactions with other proteins. 6169 mutations resulting in amino acid substitutions occur (395 distinct) to these 73 conserved residues.

Conclusions

This is the first time this type of overview analysis has been performed. Other work on p53 mutations has tended to concentrate on individual mutants of interest rather than attempting to automate the classification of structural effects.

Some mutations can be explained in multiple ways. This is shown in detail on the web site (`http://www.bioinf.org.uk/p53/` or `http://www.rubic.rdg.ac.uk/p53/`). In total, we were able to explain 304 of the 822 distinct mutations resulting in substitutions in the core domain (34.5%) on purely structural grounds. If mutations to 100% conserved amino acids are also considered, then this number rises to 515 of 822 distinct mutations (58.4%).

Of the unexplained mutations, it might be expected that the majority of these will be on the surface. Using a cutoff of 10% accessibility to classify a residue as exposed, we actually find that only 236 of the 367 unexplained distinct mutations (64.31%) are exposed.

Note that our criteria for classifying a mutation as explained are fairly strict. For example we assume that any hydrogen-bonding sidechain substitution will be able to maintain the hydrogen bond if it has donor or acceptor capabilities the same as the parent; in practice, a structural change may be necessary.

Clearly the sidechain replacement assessment could be made much more sophisticated and will be addressed in future work. A minimisation procedure could be incorporated into the sidechain replacement together with a measure of the degree of bad contact rather than a simple yes/no assessment of clashes. In addition we could use X-Site scores to assess the acceptability of sidechain replacements and account for large sidechains being replaced by smaller sidechains thus creating a void in the structure. Similarly rather than simply assessing residues on the basis of ability to donate or accept hydrogen bonds, it would be possible to assess the geometry of replacements which, in principle, are able to maintain the required ability.

Excluding those mutations for which we have genuinely not identified a structural explanation, some mutants may actually have a silent non-pathogenic phentotype. More interesting are those which are on the surface of the p53 core domain and are involved in interactions with the other p53 domains or with other proteins. In future, we intend to apply the patch analysis methodology of Jones and Thornton[38, 39] to identify regions of the protein surface likely to be involved in protein-protein interactions.

In the long term, it is hoped that properties of p53 mutations, such as dominant negative activity, oncogenic potential and temperature-sensitivity may be explained by classification of p53 mutations into structural groups whose molecular basis may then be analysed.

We see this approach not only as a useful tool in examination of p53 mutations, but also as a paradigm for the study of many other diseases caused by point mutations. In the near future, when structural data become available, it will become possible to apply the same forms of analysis to dystrophin, BRCA-I and G6PD — in all cases mutation databanks are available.

References

[1] Matlashewski, G. p53 twenty years on, meeting review. Oncogene 18:7618–7620, 1999.

[2] May, P. and May, E. Twenty years of p53 research: Structural and functional aspects of the p53 protein. Oncogene 18:7621–7636, 1999.

[3] Crawford, L. The 53,000-dalton cellular protein and its role in transformation. Int. Rev. Exp. Pathol. 25:1–50, 1983.

[4] Culotta, E. and Koshland, Jr, D. E. p53 sweeps through cancer research. Science 262:1958–1959, 1993.

[5] Harris, C. C. p53: At the crossroads of molecular carcinogenesis and risk assessment. Science 262:1980–1981, 1993.

[6] Levine, A. J. p53, the cellular gatekeeper for growth and differentiation. Cell 88:323–331, 1997.

[7] Lakin, N. and Jackson, S. Regulation of p53 in response to DNA damage. Oncogene 18:7644–7655, 1999.

[8] Ko, L. and Prives, C. p53: Puzzle and paradigm. Genes Dev. 10:1054–1072, 1996.

[9] Chao, C., Saito, S., Kang, J., Anderson, C., Appella, E., and Xu, Y. p53 transcriptional activity is essential for p53-dependent apoptosis following DNA damage. EMBO J. Sept 15:4967–4975, 2000.

[10] Romano, J. W., Ehrhart, J. C., Duthu, A., Kim, C. M., Appella, E., and May, P. Identification and characterization of a p53 gene in a human osteosarcoma cell line. Oncogene 4:1483–1488, 1989.

[11] Hainaut, P., Hernandez, T., Robinson, A., Rodriiguez-Tome, P., Flores, T., Hollstein, M., Harris, C. C., and Montesano, R. IARC database of p53 gene mutations in human tumors and cell lines: Updated compilation, revised formats and new visualisation tools. Nuc. Ac. Res. 26:205–213, 1998.

[12] Greenblatt, M. S., Bennett, W. P., Hollstein, M., and Harris, C. C. Mutations in the p53 tumor suppressor gene: Clues to cancer etiology and molecular pathogenesis. Cancer Res. 54:4855–4878, 1994.

[13] Harris, C. C. p53 tumor suppressor gene: From the basic research laboratory to the clinic — an abridged historical perspective. Carcinogenesis 17:1187–1198, 1996.

[14] Chiba, I., Takahashi, T., Nau, M. M., D'Amico, D., Curiel, D. T., Mitsudomi, T., Buchhagen, D. L., Carbone, D., Piantadosi, S., Koga, H., Reissman, P., Slamon, D. J., Holmes, E. C., and Minna, J. D. Mutations in the p53 gene are frequent in primary, resected non-small cell lung cancer. Oncogene 5:1603–1610, 1990.

[15] Brash, D. E., Rudolph, J. A., Simon, J. A., Lin, A., McKenna, G. J., Baden, H. P., Halperin, A. J., and Ponten, J. A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. Proc. Natl. Acad. Sci. USA 88:10124–10128, 1991.

[16] Li, F. P., Garber, J. E., Friend, S. H., Strong, L. C., Patenaude, A. F., Juengst, E. T., Reilly, P. R., Correa, P., and Fraumeni Jr., J. F. Recommendations on predictive testing for germ line p53 mutations among cancer-prone individuals. J. Natl. Cancer Instit. 84:1156–1160, 1992.

[17] Guinn, B. A. and Padua, R. A. p53: A role in the initiation and progression of leukaemia? CANCERJ 8:195–200, 1995.

[18] Malkin, D., Li, F. P., Strong, L. C., Fraumeni, J. F., Nelson, C. E., Kim, D. H., Kassel, J., Gryka, M. A., Bischoff, F. Z., Tainsky, M. A., and Friend, S. H. Germ line p53 mutations in a familial syndrome of breast-cancer, sarcomas, and other neoplasms. Science 250:1233–1238, 1990.

[19] Srivastava, S., Zou, Z. Q., Pirollo, K., Blattner, W., and Chang, E. H. Germline transmission of a mutated p53 gene in a cancer-prone family with Li-Fraumeni syndrome. Nature (London) 348:747–749, 1990.

[20] Brachmann, R. K., Eby, K. Y. Y., Pavletich, N. P., and Boeke, J. D. Genetic selection of intragenic suppressor mutations that reverse the effects of common p53 cancer mutations. EMBO J. 17:1847–1859, 1998.

[21] Cho, Y., Gorina, S., Jeffrey, P. D., and Pavletich, N. P. Crystal structure of a p53 tumor suppressor-DNA complex: Understanding tumorigenic mutations. Science 265:346–355, 1994.

[22] Vogelstein, B. and Kinzler, K. W. X-rays strike p53 again. Nature (London) 370:174–175, 1994.

[23] Wong, K.-B., DeDecker, B. S., Freund, S. M., Proctor, M. R., Bycroft, M., and Fersht, A. R. Hot-spot mutants of p53 core domain evince characteristic local structural changes. Proc. Natl. Acad. Sci. USA 96:8438–8442, 1999.

[24] Vogelstein, B. and Kinzler, K. W. p53 and dysfunction. Cell 70:523–526, 1992.

[25] Pavletich, N. P., Chambers, K. A., and Pabo, C. O. The DNA-binding domain of p53 contains the four conserved regions and the major mutation hot spots. Genes Dev. 7:2556–2564, 1993.

[26] Jeffrey, P. D., Gorina, S., and Pavletich, N. P. Crystal structure of the tetramerization domain of the p53 tumor suppressor at 1.7Ångströms. Science 267:1498–1502, 1995.

[27] Nikolova, P. V., Wong, K.-B., DeDecker, B., Henckel, J., and Fersht, A. R. Mechanism of rescue of common p53 cancer mutations by second-site suppressor mutations. EMBO J. 19:370–378, 2000.

[28] Michalovitz, D., Halevy, O., and Oren, M. p53 mutations — gains or losses. J. Cell. Biochem. 45:22–29, 1991.

[29] Kabsch, W. and Sander, C. Dictionary of protein secondary structure. Biopolymers 22:2577–2637, 1983.

[30] McDonald, I. K. and Thornton, J. M. Satisfying hydrogen-bonding potential in proteins. J. Mol. Biol. 238:777–793, 1994.

[31] Lee, B. K. and Richards, F. M. The interpretation of protein structures: Estimation of static accessibility. J. Mol. Biol. 55:379–400, 1971.

[32] Attwood, T. K. and Beck, M. E. Prints — a protein motif fingerprint database. Protein Eng. 7:841–848, 1994.

[33] Attwood, T. K., Beck, M. E., Bleasby, A. J., and Parry-Smith, D. J. Prints — a database of protein motif fingerprints. Nuc. Ac. Res. 22:3590–3596, 1994.

[34] Jones, D. T., Taylor, W. R., and Thornton, J. M. A mutation data matrix for transmembrane proteins. FEBS Lett. 339:269–275, 1994.

[35] Shenkin, P. S., Erman, B., and Mastrandrea, L. D. Inofmation-theoretical entropy as a measure of sequence variability. Proteins: Struct., Funct., Genet. 11:297–313, 1991.

[36] Snow, M. E. and Amzel, L. M. Calculating three-dimensional changes in protein structure due to amino acid substitutions: The variable domain of immunoglobulins. Proteins: Struct., Funct., Genet. 1:276–279, 1986.

[37] Baker, E. N. and Hubbard, R. E. Hydrogen bonding in globular proteins. Progr. Biophy. Molec. Biol. 44:97–179, 1984.

[38] Jones, S. and Thornton, J. M. Principles of protein-protein interactions. Proc. Natl. Acad. Sci. USA 93:13–20, 1996.

[39] Jones, S. and Thornton, J. M. Prediction of protein-protein interaction sites using patch analysis. J. Mol. Biol. 272:133–143, 1997.