Oncogene (1999) 19, 211–218
© 1999 Stockton Press All rights reserved 0950–9232/99 $12.00
http://www.stockton-press.co.uk/onc

Evolutionary conservation and somatic mutation hotspot maps of *p53*: correlation with p53 protein structural and functional features

D Roland Walker[1,5], Jeffrey P Bond[2], Robert E Tarone[3], Curtis C Harris[4], Wojciech Makalowski[1], Mark S Boguski[1] and Marc S Greenblatt*[,4]

[1]National Center for Biotechnology Information, National Library of Medicine, Bld. 38A, National Cancer Institute, NIH, Bethesda, Maryland 20894; [2]Department of Microbiology and Molecular Genetics. University of Vermont, Burlington, Vermont 05401; [3]Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, Maryland 20892; [4]Laboratory of Human Carcinogenesis, Bld. 37, Rm. 2C04, National Cancer Institute, NIH, Bethesda, Maryland 20892, USA; [5]Department of Biology, Johns Hopkins University, Baltimore, Maryland 21218, USA

Missense mutations in *p53* frequently occur at 'hotspot' amino acids which are highly conserved and represent regions of structural or functional importance. Using the *p53* mutation database and the *p53* DNA sequences for 11 species, we more precisely defined the relationships among conservation, mutation frequency and protein structure. We aligned the *p53* sequences codon-by-codon and determined the degree of substitution among them. As a whole, *p53* is evolving at an average rate for a mammalian protein-coding gene. As expected, the DNA binding domain is evolving more slowly than the carboxy and amino termini. A detailed map of evolutionary conservation shows that within the DNA binding domain there are repeating peaks and valleys of higher and lower evolutionary constraint. Mutation hotspots were identified by comparing the observed distribution of mutations to the pattern expected from a random multinomial distribution. Seventy-three hotspots were identified; these 19% of codons account for 88% of all reported *p53* mutations. Both high evolutionary constraint and mutation hotspots are noted at amino acids close to the protein-DNA interface and at others more distant from DNA, often buried within the core of the folded protein but sometimes on its surface. The results indicate that targeting highly conserved regions for mutational and functional analysis may be efficient strategies for the study of cancer-related genes.

Keywords: molecular evolution; missense mutation; crystal structure; DNA binding; domain

Introduction

Somatic mutations in the *p53* tumor suppressor gene are critical events in most types of human cancer (Greenblatt *et al*., 1994). The gene is highly conserved through vertebrate evolution; the central DNA binding domain is more highly conserved than the amino and carboxy termini. The first identification of evolution-arily conserved regions was qualitative (Soussi *et al*., 1990), describing five conserved 'domains' by visual analysis of aligned sequences. Subsequent analysis using the Multiple Alignment Construction and Analysis Workbench (MACAW, Schuler *et al*., 1991) assessed the statistical significance of conserved regions, and found that four of these domains were encompassed within a large highly conserved block between codons 97 and 292 (Greenblatt *et al*., 1994). Variability in degree of conservation was noted within this region, although MACAW did not identify a statistically significant difference for any domain. The region near the amino terminus originally denoted 'Domain I' contained sequences which are absent in chicken *p53*; a weakly significant conserved sequence was defined that overlapped the originally described domain. An additional, moderately well conserved region was found in the carboxyl terminus at the site of the *p53* oligomerization domain. Recently a human gene was described on chromosome 1, *p73*, which has considerable sequence similarity to p53, can function as a transcription factor for *p53* target genes and can induce apoptosis (Kaghad *et al*., 1997; Jost *et al*., 1997).

Databases of somatic *p53* mutations (Hollstein *et al*., 1994) are the largest compendia of disease-related human somatic mutations. At the time of this analysis, the available database contained in excess of 3500 mutations (missense, nonsense, frameshift, splicing and silent; Hollstein *et al*., 1996) and it continues to be updated regularly (administered by IARC and available on the World Wide Web, http://www.iarc.fr/p53/homepage.htm). Analysis of these data can generate and test many hypotheses, including the relationship of mutagenesis to evolutionary gene conservation, protein function, and the contributions of different mutagenic processes to carcinogenesis. Most *p53* missense mutations found in human cancers are located in the highly conserved central DNA binding domain, suggesting a correlation between degree of evolutionary constraint and structural or functional importance of individual amino acid residues (Greenblatt *et al*., 1994).

'Hotspot' somatic mutations in cancers represent protein alterations which provide a selective growth advantage to the cell. The location of a hotspot is thought to denote a functionally critical amino acid residue. Other factors potentially contributing to hotspot formation include high susceptibility of a codon to mutation and slow DNA repair rates (Tornaletti and Pfeifer, 1994). Missense mutations at six codons account for a quarter of all *p53* somatic point mutations (Greenblatt *et al*., 1994). Five of these six hotspots are at CpG dinucleotides; mutations at the

*Correspondence: MS Greenblatt, Hematology/Oncology Unit, Department of Medicine, University of Vermont College of Medicine, Patrick 534, MCHV Campus, Burlington, VT 05401, USA
Received 3 November 1997; revised 20 July 1998; accepted 21 July 1998

other site (codon 249) are associated with exposure to aflatoxin B1 in hepatocellular carcinomas (Hsu *et al*., 1991). All of these amino acids are adjacent to or within the DNA binding interface of the p53 protein (Cho *et al*., 1994). Other putative hotspots exist at many codons within the central domain; both CpG and non-CpG sites frequently show missense mutations. It has been unclear whether all codons at which missense mutations occur denote 'hotspots'. More precise information of what constitutes a 'hotspot' can help identify other potential relationships between mutation, protein structure and function, and carcinogenesis.

Using the *p53* mutation database and the reported *p53* DNA sequences for 11 species, we have refined the earlier estimates of evolutionary constraints and mutational hotspots. Our analysis more precisely defines the relationship between these features, demonstrating a correlation between the variable degree of evolutionary constraint within the large conserved central DNA binding domain and the location of mutational hotspots. We also show that the regions within the three dimensional structure of the p53 protein which are most likely to be conserved and mutated include many residues which contact DNA, many of which form the core of the folded protein, and a few surface residues on the opposite side from DNA.

Results

Evolutionary constraint

Substitution rates were determined as described in the Materials and methods section. In mammals, rates of nonsynonymous substitution (resulting in an amino acid change) range from $0.01-0.13\times10^{-9}$ base substitutions/year for highly conserved genes such as β-actin and insulin to $2.21-2.79\times10^{-9}$ for less conserved genes such as the interferons (Li and Grauer, 1991).

As a whole, *p53* is evolving at an average rate for a mammalian protein-coding gene. The nonsynonymous substitution distance, K_a, comparing the coding sequence of human with mouse *p53* is 0.146 base substitutions/site (rate $=0.91\times10^{-9}$ base substitutions/year). Estimates of the time of divergence of humans and rodents vary from 75–125 million years; we used a value of 80 million years, the time of the great mammalian radiation (Li, 1997). The average K_a for coding sequences among 1138 genes shared by humans and mice is 0.088 substitutions/site (rate $=0.550\times10^{-9}$ base substitutions/year). Among 1212 genes shared by humans and rats the values are 0.078 and 0.49×10^{-9} (Makalowski and Boguski, 1998). The values for human-mouse and human-rat *p53*, corrected for evolutionary distance, reflect the mean and median values for pairwise comparisons of all species in this study (data not shown). The synonymous substitution distance, K_s, was 0.519 (rate $=3.2\times10^{-9}$/year), which approximates the average K_s among human-mouse gene pairs ($K_s=0.46$, rate $=2.91\times10^{-9}$/year).

As expected, the rates of evolution differ among the amino, central DNA binding and carboxyl regions of *p53*. K_a comparing human and mouse *p53* sequences in these regions were 0.293 (rate $=1.83\times10^{-9}$/year), 0.084 (0.525×10^{-9}/year) and 0.142 (0.089×10^{-9}/year), respectively (Table 1). The DNA binding region is thus evolving at a rate slightly below average, but about half as rapidly as the gene as a whole. The N and C termini are both evolving more rapidly than average, with the N terminus evolving at twice the rate of the C terminus. This supports experimental data which suggest that the transactivation domain provides a more generic function, whereas the oligomerization domain provides an important structure with specific biological function (Wang *et al*., 1996, reviewed in Greenblatt *et al*., 1994). K_s values were comparable among the three regions of the gene (Table 1). It has been proposed that there are four conserved 'domains' within the central region of the protein (Soussi *et al*., 1990). We calculated codon-by-codon evolutionary distances using a sliding window of seven codons across the entire gene. Within the DNA binding region there is an oscillating pattern of peaks and valleys of higher and lower evolutionary sequence variation (Figure 1), with seven major conserved stretches, including absolute conservation between codons 236–251 and codons 275–282.

Hotspot analysis

We used a version of the *p53* mutation database in which 2870 missense mutations are found, representing 737 different base substitutions at 229 different codons. We define a codon as a hotspot if the number of

Table 1 Rates of base substitutions between human and mouse p53

	Nonsynonymous substitutions		*Synonymous substitutions*	
Sequence	*Evolutionary distance (K_a, muts/site)*	*Evolutionary rate ($\times10^{-9}$ year)*[a]	*Evolutionary distance (K_S, muts/site)*	*Evolutionary rate ($\times10^{-9}$ year)*[a]
Entire p53 gene	0.146	0.88	0.519	3.2
N-Terminus (104 Codons)	0.293	1.83	0.578	3.61
DNA binding (191 Codons)	0.084	0.525	0.503	3.14
C-Terminus (98 Codons)	0.142	0.888	0.366	2.29
Mean for 1138 human-mouse gene pairs[b]	0.088	0.550	0.466	2.91

[a]Rate = K/2T, T = Time of divergence between species, estimated at 80 million years. [b]From Makalowski and Boguski, 1998

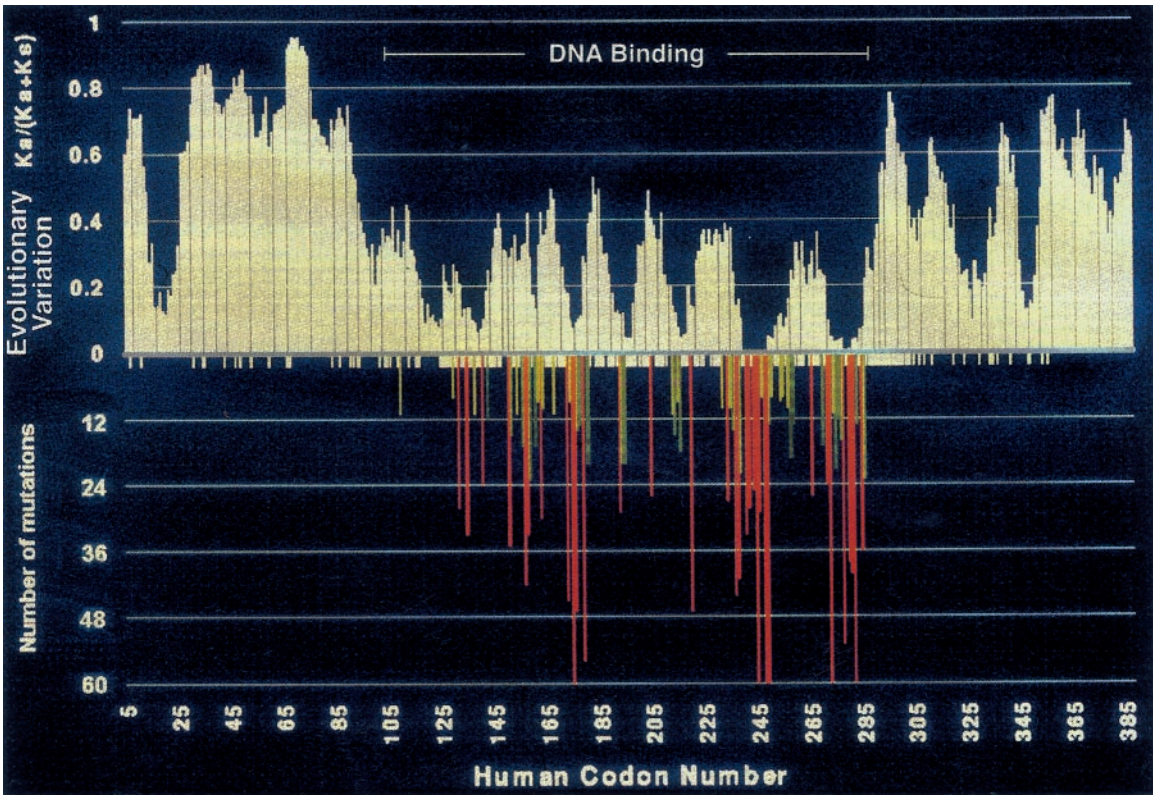

Figure 1 Schematic line chart of p53 gene, depicting evolutionary variation above the X axis and numbers of missense mutations (from the database of p53 mutations from human tumors and cell lines, Hollstein *et al*., 1996) below the X axis. Evolutionary variation is noted as the average ratio $K_a/(K_a+K_s)$ for all pairwise comparisons of p53 interspecies homologs, where K_a = nonsynonymous and K_s = synonymous substitution distances. A low value for $K_a/(K_a+K_s)$ is indicative of low sequence variation (high evolutionary constraint). The mutation values have been rendered on a linear scale with two modifications: (1) all hotspot sites of eight or more mutations are indicated in proportion to number of mutations, except that for the six codons with more than 100 mutations the lines have been truncated at 60; and (2) sites with 1–7 reported mutations are equalized and indicated by a short tick mark. Red lines denote hotspots of $P<0.001$ and $P<0.01$, green denotes $P<0.014$, yellow denotes $P<0.025$ (see text for Materials and methods)

mutations significantly exceeds the number expected from a multinomial distribution with equal probability of mutation at all codons. Hotspots were determined by a four-step procedure described in the Materials and methods section.

This four-step method defined a hotspot ($P<0.05$) as eight or more mutations at a codon, and identified 73 total hotspots accounting for 2519 mutations (88% of all missense mutations and 67% of all recorded *p53* mutations, Table 2). The first step, which defines the hotspots of highest statistical significance, identified 32 hotspots ($P<0.001$) of more than 24 mutations each, accounting for 1958 mutations. The second step identified 23 hotspots ($0.001<P<0.01$) of 12–23 mutations, accounting for 387 mutations; the third step identified 14 hotspots ($0.01<P<0.014$) of 9–11 mutations accounting for 142 mutations, and the last step identified four hotspots ($0.014<P<0.025$) of eight mutations accounting for 32 mutations.

Correlation among p53 evolutionary constraint, structure and somatic mutation hotspots

Comparison of the degree of evolutionary constraint with the hotspot map confirms the expectation that mutations are more likely to be found in areas of great evolutionary constraint. Within the DNA binding region, there appears to be a threshold level of evolutionary constraint for the appearance of hotspots. Clusters of codons which are more variable correspond to gaps in the map of hotspot codons (Figure 1). The longest sequences of absolute evolutionary conservation among the 11 species are codons 236–251 and codons 275–282. Mutation hotspots occur at most codons between 234–259 and 270–286. Other clusters of hotspots which correspond to conserved clusters include codons 153–164, 173–181, 193–195 and 213–216. Not every highly constrained codon within these sequences is a hotspot. Tyrosines at codons 205 and 220, however, are invariant and are hotspots; the codons adjacent to tyr 205 are neither highly constrained nor hotspots and no codons adjacent to tyr 220 are hotspots although val 218, pro 219 and glu 221 are invariant. Codons 130–143 are highly conserved and contain three stronger and three weaker hotspots, whereas codons 115–124 are conserved but are not hotspots for missense mutation.

The solution of the crystal structure of the p53 DNA binding region revealed that the hotspot codons with the most mutations encoded amino acids close to the p53-DNA interface, a region comprised of a loop–sheet–helix motif (L1, S2-S2′, H2) and a second loop (L3) which contact DNA, another large loop (L2) which interacts with and provides structural support for the residues contacting DNA, and a large beta sheet (strands S10, S9, S4, S7 and S6) which provides a core 'barrel'-shaped hydrophobic domain (Cho *et al*., 1994). The residues most often mutated either directly

Table 2 Hotspot codons in p53, structural features

Codon	*# Muts*	P *value*	*AA*[a]	*Secondary structure*[b]	*Amino acid contact*[c]	*Codon*	*# Muts*	P *value*	*AA*[a]	*Secondary structure*[b]	*Amino acid contact*[c]
110	11	$P<0.014$	R	S1	PE	234	27	$P<0.001$	Y	S8	B
130	8	$P<0.025$	L		PE	235	10	$P<0.014$	N	S8	B
132	28	$P<0.001$	K	S2′	B	236	14	$P<0.01$	Y	S8	B
135	33	$P<0.001$	C	S2′	B	237	44	$P<0.001$	M	L3	B
138	11	$P<0.014$	A		PE	238	41	$P<0.001$	C	L3	Zn
141	24	$P<0.001$	C	S3	B	239	22	$P<0.01$	N	L3	DNA
143	12	$P<0.01$	V	S3	B	241	33	$P<0.001$	S	L3	DNA
151	35	$P<0.001$	P		B	242	28	$P<0.001$	C	L3	Zn
152	15	$P<0.01$	P		PE	244	25	$P<0.001$	G	L3	Exp
154	11	$P<0.014$	G		PE	245	126	$P<0.001$	G	L3	B
156	17	$P<0.01$	R	S4	PE	246	29	$P<0.001$	M	L3	B
157	42	$P<0.001$	V	S4	B	247	8	$P<0.025$	N	L3	DNA
158	33	$P<0.001$	R	S4	B	248	238	$P<0.001$	R	L3	DNA
159	23	$P<0.01$	A	S4	B	249	144	$P<0.001$	R	L3	B
161	17	$P<0.01$	A	S4	B	250	13	$P<0.01$	P	L3	PE
162	9	$P<0.014$	I	S4	B	251	8	$P<0.025$	I	S9	B
163	30	$P<0.001$	Y	S4	B	254	9	$P<0.014$	I	S9	B
164	10	$P<0.014$	K	L2	B	255	8	$P<0.025$	I	S9	B
168	11	$P<0.014$	H	L2	PE	257	10	$P<0.014$	L	S9	B
173	45	$P<0.001$	V	L2	B	258	19	$P<0.01$	E	S9	B
174	9	$P<0.014$	R	L2	PE	259	13	$P<0.01$	D		B
175	171	$P<0.001$	R	L2	B	266	26	$P<0.001$	G	S10	B
176	47	$P<0.001$	C	L2	Zn	270	17	$P<0.01$	F	S10	B
177	14	$P<0.01$	P	H1	Exp	272	24	$P<0.001$	V	S10	B
178	13	$P<0.01$	H	H1	Exp	273	255	$P<0.001$	R	S10	DNA
179	56	$P<0.001$	H	H1	Zn	274	11	$P<0.014$	V	S10	B
181	20	$P<0.01$	R	H1	Exp	275	21	$P<0.01$	C		DNA
193	29	$P<0.001$	H		B	276	12	$P<0.01$	A		DNA
194	20	$P<0.01$	L		B	277	16	$P<0.01$	C		DNA
195	20	$P<0.01$	I	S5	B	278	53	$P<0.001$	P	H2	B
205	26	$P<0.001$	Y	S6	B	280	38	$P<0.001$	R	H2	DNA
213	11	$P<0.014$	R		B	281	40	$P<0.001$	D	H2	DNA
214	15	$P<0.01$	H	S7	B	282	105	$P<0.001$	R	H2	B
215	9	$P<0.014$	S	S7	B	283	13	$P<0.01$	R	H2	DNA
216	18	$P<0.01$	V	S7	B	285	36	$P<0.001$	E	H2	PE
220	47	$P<0.001$	Y		B	286	23	$P<0.01$	E	H2	PE
232	10	$P<0.014$	I	S8	B						

[a]AA = Amino acid, one letter code. [b]Structural features, from Cho et al. (1994). H = Helix; L = Loop; S = Beta sheet. [c]Contact features of AA with DNA, Zinc, or other AAs (see Materials and methods). B = Buried, <15% surface area exposed; DNA = Contacts DNA; Exp = Exposed, >50% surface area exposed; PE = Partially Exposed, 15 < <50% surface area exposed; Zn = Binds Zinc

contact DNA (those in the L3 loop, S10 sheet and H2 helix) or closely contact those that do so (codon 175 and others in the L2 loop). Mutations of the buried amino acids destabilize the folded protein (Cho *et al*., 1994). Fewer mutations were noted in the region of the molecule farther from DNA, a region comprised of a second antiparallel beta sheet (strands S5, S8, S3 and S1) and two long loops.

To study the topography of the hotspots and conserved amino acids in more detail, we colored renderings of the crystal structure according to our expanded hotspot map (Figure 2) and the map of the degree of evolutionary constraint of each codon (Figure 3). Table 2 indicates which mutations contact DNA or zinc and which are buried or exposed and within which secondary structures they lie. As expected, most hotspots of highest statistical significance and high evolutionary conservation are near the DNA-protein interface, at amino acids either contacting DNA or supporting those that do. However, some other specific structural features were observed. Specific regions of the protein which are thought to have functional importance show the following characteristics: (1) All four of the residues which bind zinc and stabilize loops L2 and L3 (cysteines 176, 238 and 242 and histidine 179) are hotspots of highest statistical significance (Table 2). Replacement of cysteines 176, 238 and 242 by serine is known to completely block transcriptional activation (Rainwater *et al*., 1995). The correlation of transcriptional function, conservation, and hotspots is variable among other cysteines in p53. Notably, two cysteines (codons 135 and 141) are among the six conserved hotspots between codons 130–143; (2) There are notable hotspots in regions of the protein further from DNA. Mutation of many of these amino acids can be expected to destabilize the protein structure. Tyrosines 205 and 220 are both hotspots of highest significance and absolute evolutionary conservation which have not previously been identified as important functional residues. The amino acids adjacent to tyrosine 205 are nonconserved; amino acids 218, 219 and 221 are conserved but are not hotspots. The side chains of tyrosines 205 and 220 make numerous contacts with side chain and backbone atoms in the protein interior, suggesting that mutation of these residues would destabilize the folded state of the protein and that they are not available for phosphorylation. Other conserved and hotspot amino acids opposite to the DNA contact face include proline 151, valine 157, arginine 158, isoleucine 232, glutamate 258, aspartate 259 and glycine 266 (Table 2, Figures 2 and 3). Mutation of proline 151 and glycine 266 will

likely alter backbone structure and/or side chain interactions, and mutation of valine 157 and isoleucine 232 would likely alter packing of the hydrophobic interface between the beta-sheets. However, the consequences of mutating arg 158, glu 258 and aspartate 259 are less clear. Structural rendering indicates that the side chains of arginine 156, arginine 158, serine 215 and glutamate 258 (all hotspots) interact with each other in a surface groove remote from the DNA binding face of the protein (data not shown). (3) Only four hotspot amino acids have more than 50% of their surface area exposed (pro 177, his 178, arg 181, gly 244). The four are very close to each other and to DNA (three in helix 1 and one in the adjacent loop L3), and appear ideally located for participating in the protein-protein interactions between p53 dimers predicted by Cho *et al*. (1994) to occur in helix 1; (4) Oligomerization domains in the carboxyl terminal (Clore *et al*., 1994; Jeffrey *et al*., 1995) show a high degree of conservation within the beta sheet at residues 326–334 and the alpha helix from residues 335–355. The remainder of the carboxyl terminus is less well conserved. The basic region between amino acids 361 and the end of the protein shows a degree of conservation at or lower than the least conserved residues in the DNA binding region; (5) Repeats of PXXP, thought to contribute to p53 apoptosis function by acting as SH3 ligands (Walker and Levine, 1996), are present between amino acids 61–93, but this region is not highly conserved, and no hotspots occur here. A possible explanation is that changing one amino acid in this linking region has minimal effects on protein function; (6) At CpG sites, two potential mutations occur from deamination. At

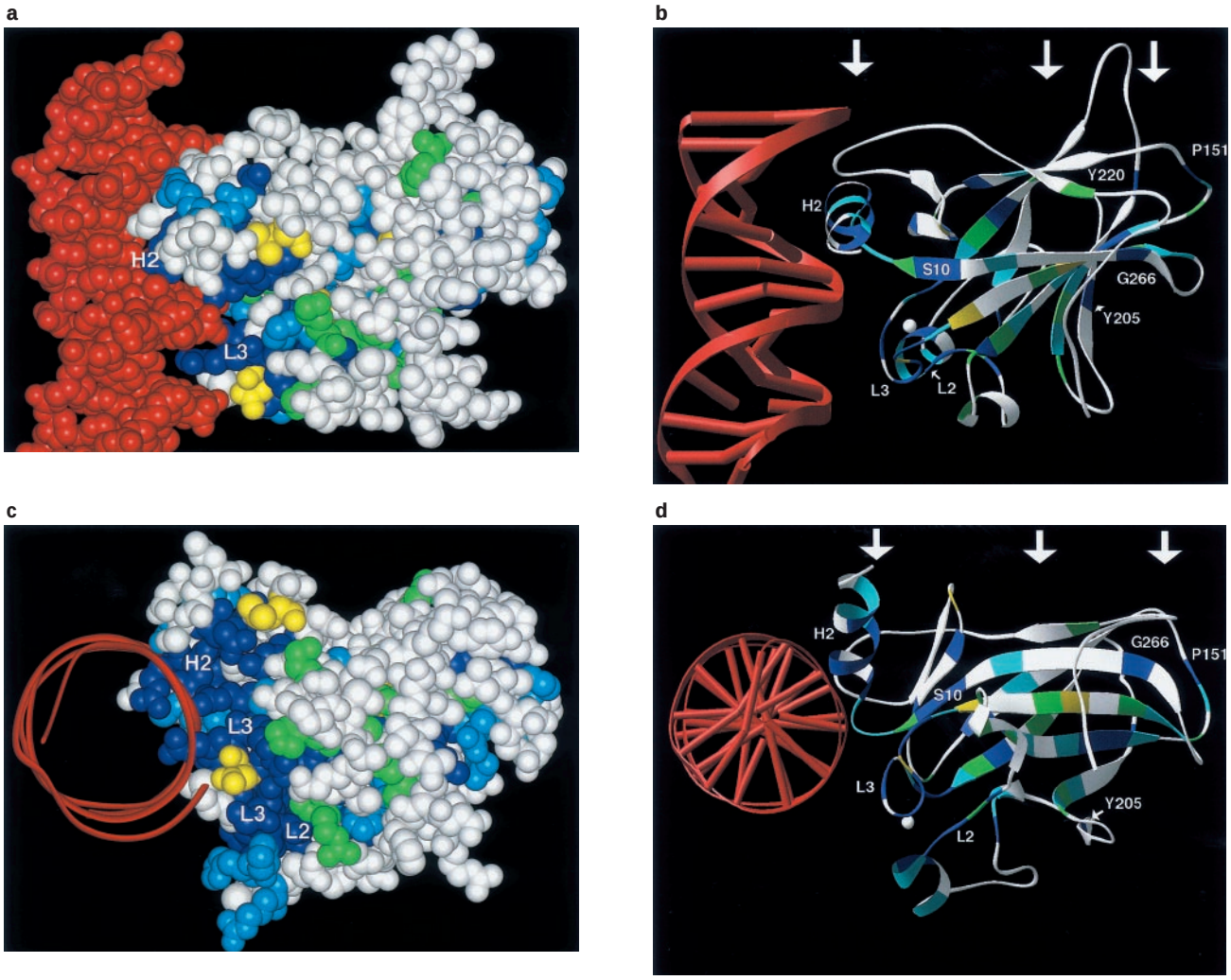

Figure 2 Three dimensional structure of p53 protein: Topography of p53 mutational hotspot codons. White amino acid residues indicate codons which are not hotspots, colored residues indicate codons which are mutational hotspots of varying significance: dark blue = $P < 0.001$; light blue = $P < 0.01$; green = $P < 0.014$; yellow = $P < 0.025$ (see text for explanation of P values). Red denotes DNA helix. Transverse views of DNA: (**a**) Space filling model of p53 protein and DNA showing mutation hotspots at DNA interface and buried within the folded protein. (**b**) Ribbon model of p53 protein backbone and stick model of DNA, showing secondary structural features of mutation hotspot amino acids through folded protein. Arrows indicate three regions of protein: the region adjacent to and contacting DNA, the tightly packed beta sheet 'barrel', and the loops facing away from DNA. White circle represents Zn atom. Helical axis views of DNA; (**c**) space filling and (**d**) ribbon representations of p53 protein and backbone of DNA helix. Hotspot mutations are concentrated in the loops, sheets, and helices close to the DNA interface (including inside the major and minor grooves), but many hotspot amino acids form a core in the central part of the DNA binding region, and some hotspots (e.g., proline 151, glycine 266, and tyrosines 205 and 220, labeled P, G, and Y, respectively) occur in the loops of the molecule which face away from DNA. See text for details

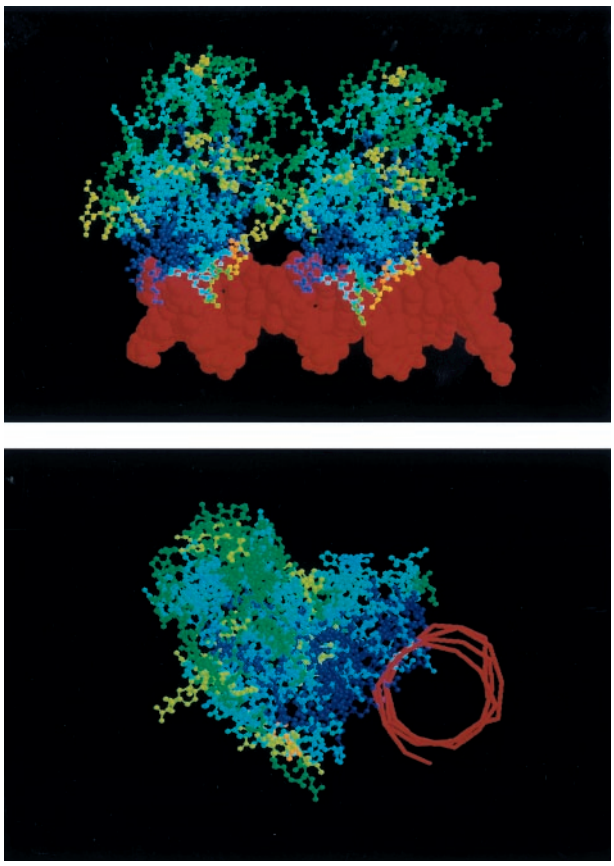

Figure 3 Three dimensional structure of p53 protein: Topography of degree of evolutionary amino acid conservation. Color code in order of highest to lowest conservation = dark blue > light blue > green > yellow > orange. Red denotes DNA helix. (**a** and **b**) both show two adjacent p53 DNA binding domains. (**a**) Transverse view, ball and stick model of p53 protein and space filling model of DNA (**b**). Helical axis view, ball and stick representation of p53 protein and backbone of DNA helix. Most highly conserved residues are concentrated in the loops, sheets, and helices close to the DNA interface (including inside the major and minor grooves), but some involve the core of the folded protein more distant from the DNA interface. These patterns correlate with the hotspot map

many sites both mutations are seen in tumors. However, at codons 175 and 282 one mutation predominates (175 arg > his, 282 arg > trp). Modeling studies suggest that a histidine at codon 175 would interfere sterically with surrounding residues, whereas replacement with cysteine is predicted to have a less dramatic effect on the structure. At codon 282, arg > trp accounts for almost all CpG transitions and arg > gln almost none. Tryptophan at codon 282 would interfere sterically with thr 118, ser 127, and phe 134 (none of which are hotspots), whereas replacement with glutamine appears to cause less structural distortion. (7) Only one hotspot, arginine 110, is more N-terminal than amino acid 130, despite the fact that much of Loop 1 (amino acids 113–124) is highly conserved and adjacent to DNA.

Discussion

Our results refine previous observations of *p53* evolutionary and mutational hotspot data, providing higher resolution of p53 structure/function relationships. These results support the concept that within the conserved DNA binding region there are amino acid residues and sequence motifs which are essential and other sequences which are less critical to p53 structure and function.

The previous description of four 'domains' within the DNA binding region should be refined, as they do not describe true functional or structural protein domains (well defined regions that either perform a specific function or constitute a stable, compact structural unit) (Li, 1997). Rather, they represent conserved sequence motifs within a larger DNA binding domain, whose degree of conservation oscillates as amino acid sequences enter and depart from critical subunits of the domain. Figure 1 shows that within the DNA binding domain there are repeating peaks and valleys of higher and lower evolutionary constraint, corresponding with clusters of mutational hotspots. These regions include portions of beta sheets, helices, loops, and DNA and zinc-contacting residues (Cho *et al*., 1994), demonstrating that hotspots can occur in any structural motif. Many are close to the protein-DNA interface and many are buried within the structure of the folded protein both near and distant to the DNA contact region (Figures 2 and 3). However, several conserved hotspot amino acids far from the DNA contact region are partially exposed in a groove on the p53 surface, suggesting potential protein-protein interactions. These residues (arg 156, arg 158, ser 215 and glu 258) are candidates for site directed mutagenesis studies.

Both functional alterations and mutagenic stresses contribute to hotspot formation. Any hotspot likely denotes a functionally or structurally critical residue. The hotspots of highest statistical significance all occur in the most constrained segments of the gene and can be divided into two groups. The 'hottest' of these occur at CpG dinucleotides, where deamination of cytosine leading to G:C to A:T transition occurs about 10–12 times more frequently than other transitions (Sved and Bird, 1990). The second group of highly significant hotspots ($P < 0.001$) occurs at codons without CpG sites but within highly constrained regions, consistent with strong functional selection of any mutant at these codons. The specific amino acid substitution also may be important; the example of the mutants at codon 175 highlights the value of structural molecular modeling. The C to T arg > his substitution at codon 175 would generally be expected to be less disruptive than the G to A arg > cys (Rodin *et al*., 1998), but the histidine mutation predominates. Modeling the steric interactions of the two substitutions in the context of p53 local structural change suggests the steric interference of histidine but not cysteine at codon 175.

We determined the hotspots using the data for the entire gene. Since many investigators examine only exons 5–8, some ascertainment bias exists favoring detection of mutations in these exons 5–8. However, we feel that the database represents a reasonable approximation of the true mutation patterns in the whole gene. When only data from studies which sequenced all coding exons of p53 are analysed, 87% of all mutations are in exons 5–8. However, only 30% of the mutations outside of these exons were missense (Greenblatt *et al*., 1994). Thus, 95% of all missense mutations occur in exons 5–8, enough to demonstrate

a representative hotspot map. There is only one hotspot outside of exons 5–8 (codon 110, $n=9$), and no other codon approaches the eight mutations required to be designated as a hotspot. Although it is not likely that strong hotspots in this area were missed, it is curious that the Loop 1 is highly conserved and some amino acids are adjacent to DNA, but there are no hotspots.

Other locus-specific mutation databases exist; some can be found on the World Wide Web at http://www.cf.ac.uk/uwcm/mg/oth_mut.html. None are as large as the p53 mutation database; i.e., for no other gene does the number of observed missense mutations exceed the number of amino acids. Our method for defining mutation hotspots has not yet been applied to other databases. It is likely that many hotspots will be missed when analysing smaller databases. The next largest database of missense mutations in tumor supressor genes is for the *p16* gene, which is involved in cell cycle regulation in the retinoblastoma pathway. Analyses of evolutionary patterns and missense mutations of *p16* suggest similar correlations (Greenblatt, manuscript in preparation).

The synonymous substitution rate (base substitutions not resulting in an amino acid change) for p53 is about average for mammalian genes (Makalowski and Boguski, 1998). According to the neutral theory of molecular evolution, synonymous mutations should be selectively neutral and should occur randomly throughout the genome. However, K_s is usually higher than the nonsynonymous rate K_a but lower than substitution rates in pseudogenes or other noncoding regions. K_s varies among genes and there is a correlation with K_a (Ticher and Graur, 1989). Thus, although no amino acids change in the protein, some selective pressure does affect synonymous substitutions. Hypotheses to explain the variation in rates include effects of local and regional sequence context and differential tRNA availability resulting in selection at the translational level. A and T nucleotides are positively correlated and G and C nucleotides negatively correlated with synonymous substitutions, and the base composition at both the third and second codon positions affects rates (Ticher and Graur, 1989). Much of the variation in K_s among genes can be ascribed to differences in codon frequency and therefore A/T *vs* G/C frequencies at positions 2 and 3 (Ticher and Graur, 1989).

Areas of high evolutionary conservation encode amino acid residues critical for structure or function. Our analysis suggests that detailed determination of evolutionary conservation is a good surrogate marker for residues which are so critical for tumor suppressor function that they will represent mutational hotspots. The valleys of low sequence variability (high evolutionary constraint) within the DNA binding region correlate closely with mutational hotspots (Figure 1). The results suggest strategies for both molecular epidemiologic and functional analyses of *p53* and other putative cancer-related genes. An efficient strategy to search for missense mutations in newly identified genes may be to target highly conserved intragenic regions for mutational analysis. An efficient approach to determining protein function might be to target for functional analysis carefully defined, highly conserved regions and hotspot mutants. The recently identified *p53* homolog, *p73* (Kaghad *et al*., 1997), will provide further opportunity for study of p53 evolutionary, structural and functional features.

Materials and methods

Statistical methods for determining evolutionary constraint

We previously assessed *p53* evolutionary conservation in large aligned segments of the gene using the Multiple Alignment Construction and Analysis Workbench (MACAW, Schuler *et al*., 1991). Using a multiple alignment of 11 *p53* homologs (human, monkey, sheep, cow, horse, mouse, rat, hamster, chicken, frog, flounder) created in MACAW, we attempted to more precisely estimate the degree of evolutionary constraint along the sequence of *p53*. We used two methods. Synonymous and nonsynonymous mutation distances reported in Table 1 were calculated using method 1 of Ina (1995) that includes a correction for multiple substitutions at single sites based on the two-parameter model of Kimura (1980). Calculations for figures, depicting successive local windows of seven codons, used the program 'subroll'. Subroll determines the degree of synonymous and nonsynonymous substitution (K_s and K_a) between two nucleotide sequences which have been aligned codon-by-codon. The value derived for degree of evolutionary variation describes the number of base substitutions required to produce the observed discrepancy in amino acids. A total of 27 base substitutions between two species are possible at each codon. The term $K_a/(K_a+K_s)$ corrects for the fact that some base substitutions are at degenerate codons and do not change the amino acid. The degree of degeneracy varies among codons (e.g., CCX is threefold degenerate at position 3, whereas TGC is onefold degenerate, and TGG is 0-fold). Subroll uses the assumption that *n*-fold degenerate codons have an n/3 chance of undergoing synonymous substitution. Where codons differ by two or more nucleotide positions, the minimum number of substitutions is assumed and the most favorable path is determined according to the PAM100 matrix. Gaps in the DNA sequence in nonhuman species are included in the analysis, with a codon-to-gap comparison calculated as equivalent to a nonsynonymous substitution. Codons which had no analog in the human *p53* sequence were neglected. Subroll was written in version 5 of perl. Subroll is now available as part of the SEALS package (Walker & Koonin, 1997).

The calculations of substitution must be done in a pairwise fashion (comparing only two sequences at a time), but the results can be combined for multiple sequence alignments by either (1) calculating the substitution rate $r=K/2T$, where T is equal to the time of divergence between each species, or (2) calculating $K_a/(K_a+K_s)$, a heuristic measure of evolutionary constraint which is independent of evolutionary distance. For the human-mouse calculations, the time of divergence was estimated to be 80 million years (Li, 1997).

Statistical methods for hotspot detection in p53

We defined the number of mutations determining a hotspot site using the following calculations. Suppose N total mutations are observed in a gene with M codons (for *p53*, $M=393$). Under the assumption that mutations are equally likely in every codon, the distribution of mutations in the M codons will follow a multinomial distribution with probability 1/M at each codon. A particular codon is defined to be a hotspot if the number of mutations at the codon is too extreme to be consistent with a binomial distribution with probability 1/M. The significance at each codon can be determined using an exact binomial test with a Bonferroni correction (Miller, 1981) for the fact that

there are M codons which are potential hotspots. Thus, using the usual Bonferroni correction with nominal significance level α, a codon with x mutations would be considered a hotspot if the binomial probability of x or more mutations (i.e., the *P*-value) is less than α/M.

Greater power for detecting hotspots can be obtained by a sequential, four-step application of the Bonferroni method. For the analyses in this study, the overall significance level $\alpha = 0.05$, was subdivided into increasing levels $\alpha_1 = 0.001$, $\alpha_2 = 0.01$, $\alpha_3 = 0.014$ and $\alpha_4 = 0.025$. In the first step the above analysis is performed using nominal level α_1, identifying as hotspots all codons with $P < 0.001/M$, where P is calculated using a binomial distribution with sample size N and probability 1/M. Suppose m_1 hotspots are identified with n_1 total mutations at these m_1 codons. If the distribution of the remaining mutations is random, then these remaining $N\text{-}n_1$ mutations will be distributed among the remaining $M - m_1$ codons according to a multinomial distribution with probability $1/(M - m_1)$ at each codon. Thus at step 2, the above analysis is performed using nominal level α_2, identifying as hotspots all codons with $P < 0.01/(M - m_1)$, where P is calculated using a binomial distribution with sample size $N\text{-}n_1$ and probability $1/(M - m_1)$. Suppose m_2 additional hotspots are identified with n_2 total mutations at these m_2 codons. Step 3 repeats the procedure with significance level α_3, sample size $N\text{-}n_1 - n_2$, and the $M\text{-}m_1 - m_2$ remaining codons as potential hotspots. If this step identifies m_3 additional hotspots with n_3 total mutations, then the last step repeats the procedure using significance level α_4, a sample size of $N\text{-}n_1 - n_2 - n_3$, and the $M\text{-}m_1 - m_2 - m_3$ remaining codons as potential hotspots. The total number of hotspots is then $m_1 + m_2 + m_3 + m_4$, where m_4 is the number of hotspots identified in the last step. Since $\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0.05$, the four-step procedure has an overall nominal significance level of 5%.

Visualization and rendering of the p53 three dimensional structure

The three dimensional structure of p53 (PDB entry 1tsr, Cho *et al*., 1994) was rendered using RIBBONS (Carson, 1991) and InsightII (MSI) for Figures 2 and 3. Residues were assessed as buried or exposed using a combination of calculations based on solvent accessibility (InsightII solvation module) and visual examination. Amino acids in the p53 structure were designated, buried, partially exposed, or exposed (Table 2) depending on whether their solvent accessible surface area was less than 15%, greater than 15% but less than 50%, or greater than 50%, respectively of the accessible surface area in a reference tripeptide (Gly-X-Gly, where X is the amino acid in question, Kyte, 1995).

Acknowledgements
We appreciate the assistance of Gary J Nelson in preparation of the figures.

References

Carson M. (1991). *J. Appl. Cryst.,* **24,** 958 – 961.
Cho Y, Gorina S, Jeffrey P and Pavletich NP. (1994). *Science,* **265,** 346 – 355.
Clore GM, Omichinski JG, Sakaguchi K, Zambrano N, Sakamoto H, Appella E and Gronenborn AN. (1994). *Science,* **265,** 386 – 391.
Greenblatt MS, Hollstein M, Bennett WP and Harris CC. (1994). *Cancer Res.,* **54,** 4855 – 4878.
Hollstein M, Rice K, Greenblatt MS, Soussi T, Fuchs R, Sorlie T, Hovig E, Smith-Sorensen B, Montesano R and Harris CC. (1994). *Nucleic Acids Res.,* **22,** 3551 – 3555.
Hollstein M, Shomer B, Greenblatt M, Soussi T, Hovig E, Montesano R and Harris CC. (1996). *Nucleic Acids Res.,* **24,** 141 – 146.
Hsu I-C, Metcalf RA, Sun T, Welsh J, Wang NJ and Harris CC. (1991). *Nature,* **350,** 427 – 428.
Ina Y. (1995). *J. Mol. Evol.,* **40,** 190 – 226.
Jeffrey PD, Gorina S and Pavletich NP. (1995). *Science,* **267,** 1498 – 1502.
Jost Ca, Marin MC and Kaelin WG. (1997). *Nature,* **389,** 191 – 194.
Kaghad M, Bonnet H, Yang A, Creacer L, Biscan J-C, Valent A, Minty A, Chalon P, Lelias J-M, Dumont X, Ferrara P, McKeon F and Caput D. (1997). *Cell,* **90,** 809 – 819.
Kimura M. (1980). *J. Mol. Evol.,* **16,** 111 – 120.
Kyte J. (1995). *Mechanism in protein chemistry*. Garland Publishers: New York.
Li W-H and Grauer D. (1991). *Fundamentals of Molecular Evolution*. Sunderland MA (ed.). Sinauer Associates, Inc., pp. 69 – 70.
Li W-H. (1997). *Molecular Evolution*. Sunderland MA (ed.). Sinauer Associates, Inc.
Makalowski W and Boguski MS. (1998). *Proc. Natl. Acad. Sci. USA,* in press.
Miller RG. (1981). *Simultaneous Statistical Inference*. Springer-Verlag: New York. pp. 6 – 8.
Rainwater R, Parks D, Anderson ME, Tegtmeyer P and Mann K. (1995). *Molec. Cell. Biol.,* **15,** 3892 – 3903.
Rodin SN, Holmquist GP and Rodin AS. (1998). *Int. J. Molec. Med.,* **1,** 191 – 199.
Schuler GD, Altschul SF and Lipman DJ. (1991). *Proteins,* **9,** 180 – 190.
Soussi T, Caron de Fromental C and May P. (1990). *Oncogene,* **5,** 945 – 952.
Sved J and Bird A. (1990). *Proc. Natl. Acad. Sci. USA,* **87,** 4692 – 4696.
Ticher A and Graur D. (1989). *J. Mol. Evol.,* **28,** 286 – 298.
Tornaletti S and Pfeifer GP. (1994). *Science,* **263,** 1436 – 1439.
Walker DR and Koonin EV. (1997). *ISBM,* **5,** 333 – 339.
Walker KK and Levine AJ. (1996). *Proc. Natl. Acad. Sci. USA,* **93,** 15335 – 15340.
Wang XW, Vermeulen W, Coursen JD, Gibson M, Lupold SE, Forrester K, Xu G, Elmore L, Yeh H, Hoeijmakers JHJ and Harris CC. (1996). *Genes and Dev.,* **10,** 1219 – 1232.