Likelihood Models of Somatic Mutation and Codon Substitution in Cancer Genes

Ziheng Yang,* Simon Ro[†] and Bruce Rannala[†,1]

[†]*Department of Medical Genetics, University of Alberta, Edmonton, Alberta T6G 2H7, Canada and*
**Department of Biology, Galton Laboratory, University College London, London WC1E 6BT, England*

Manuscript received March 27, 2002
Accepted for publication May 13, 2003

ABSTRACT

The role of somatic mutation in cancer is well established and several genes have been identified that are frequent targets. This has enabled large-scale screening studies of the spectrum of somatic mutations in cancers of particular organs. Cancer gene mutation databases compile the results of many studies and can provide insight into the importance of specific amino acid sequences and functional domains in cancer, as well as elucidate aspects of the mutation process. Past studies of the spectrum of cancer mutations (in particular genes) have examined overall frequencies of mutation (at specific nucleotides) and of missense, nonsense, and silent substitution (at specific codons) both in the sequence as a whole and in a specific functional domain. Existing methods ignore features of the genetic code that allow some codons to mutate to missense, or stop, codons more readily than others (*i.e.*, by one nucleotide change, *vs.* two or three). A new codon-based method to estimate the relative rate of substitution (fixation of a somatic mutation in a cancer cell lineage) of nonsense *vs.* missense mutations in different functional domains and in different tumor tissues is presented. Models that account for several potential influences on rates of somatic mutation and substitution in cancer progenitor cells and allow biases of mutation rates for particular dinucleotide sequences (CGs and dipyrimidines), transition *vs.* transversion bias, and variable rates of silent substitution across functional domains (useful in detecting investigator sampling bias) are considered. Likelihood-ratio tests are used to choose among models, using cancer gene mutation data. The method is applied to analyze published data on the spectrum of p53 mutations in cancers. A novel finding is that the ratio of the probability of nonsense to missense substitution is much lower in the DNA-binding and transactivation domains (ratios near 1) than in structural domains such as the linker, tetramerization (oligomerization), and proline-rich domains (ratios exceeding 100 in some tissues), implying that the specific amino acid sequence may be less critical in structural domains (*e.g.*, amino acid changes less often lead to cancer). The transition *vs.* transversion bias and effect of CpG dinucleotides on mutation rates in p53 varied greatly across cancers of different organs, likely reflecting effects of different endogenous and exogenous factors influencing mutation in specific organs.

IN the last two decades, many genes that display a tendency to undergo somatic mutation in various cancers have been identified. As a result, the connection between somatic gene mutation and cancer initiation, and progression, is now much better understood (reviewed in Hanahan and Weinberg 2000; Ponder 2001). Somatic mutations in such cancer-associated genes are known to cause various abnormal cell characteristics, such as an enhanced rate of cell division, vascularization, and other properties that facilitate cancer development. The genes found to undergo genetic alterations in cancer have been placed into two categories: oncogenes and tumor suppressor genes (Bishop 1991; Ozoren and El-Deiry 2000). Oncogenes are mutated forms of normal cellular genes, called proto-oncogenes, whose products are elements of the cellular growth-signaling network (*i.e.*, growth factors, cytoplasmic protein kinases, transcriptional factors, cell cycle regulators, etc.). In normal cells, proto-oncogenes promote cell growth only in the presence of a relevant growth signal. However, oncogenic conversion from a proto-oncogene causes constitutively active cellular growth signaling. Conversion of a proto-oncogene into an oncogene can be mediated by various kinds of genetic modifications such as point mutations (RAS oncogene), chromosomal translocations (ABL oncogene), gene amplifications (MYC oncogene), etc. An oncogenic mutation is dominant, and thus genetic modification of only one copy is sufficient for an allele to gain a new function despite the presence of its normal counterpart.

Genes whose normal role is to prevent damaged cells from escaping regulation and that undergo inactivation of both alleles during tumor development are called tumor suppressor genes. The loss of function of a tumor suppressor gene is typically caused by a nonsense or mis-

[1] *Corresponding author:* Department of Medical Genetics, 8-39 Medical Sciences Bldg., University of Alberta, Edmonton, AB T6G 2H7, Canada. E-mail: brannala@ualberta.ca

sense mutation, a chromosomal deletion, methylation, etc. The role of the normal tumor suppressor gene in noncancerous cells has, in several cases, become clearer following its discovery: p53 arrests the cell cycle at G1 phase, or induces apoptosis, in cells with damaged DNA (Levine 1997), RB blocks cell cycle progression at the G1 phase (Weinberg 1995), and BRCA1 is involved in DNA damage responses and repair pathways (Venkitaraman 2002). The discovery of the important role of somatic gene mutation in cancer has stimulated large-scale screenings for somatic mutations of known human cancer genes in tumor tissues. The best example is the p53 mutation database (IARC TP53 database at http://www.iarc.fr/P53/ and Thierry Soussi's p53 website at http://p53.curie.fr/), which catalogs somatic mutations in the p53 gene from over 15,000 tumors.

The availability of large databases of tumor mutations has enabled cancer biologists to compare frequencies of mutations in different functional domains of a gene and in different tissues. Such studies can potentially clarify the role of these domains in gene function as it relates to cancer. Moreover, the existence of such databases has stimulated a search for mutational hotspots that may be caused by features of the primary sequence (*i.e.*, CpG dinucleotides, etc.) that make a region more susceptible to mutation. Comparative studies of homologous genes across species revealed that highly conserved regions (*i.e.*, regions under strong negative selection) coincide with mutational hotspots in the mutational database (Soussi *et al.* 1990; Walker *et al.* 1999).

Yet another approach for studying cancer mutations examines the frequencies of germline mutations in a population, testing the fit of alternative population genetic models assuming either neutral evolution or positive or negative selection. Slatkin and Rannala (1997) used this approach to study the spectrum of BRCA1 germline mutations. A common feature of published comparative studies, and existing studies of mutational spectra in tumor databases, is that they have focused primarily on overall frequencies of nucleotide changes, rather than on changes in specific codons and their effect on the amino acid sequence. No published analyses (to our knowledge) make explicit use of a codon-based substitution model.

Features of the mutation process should be reasonably well described by a nucleotide-based approach; biases in rates of substitution at particular dinucleotides, for example, can be indicative of exogenous *vs.* endogenous mutagens. One might also expect the mutational spectrum to differ among tumors from different tissues because some organs, such as skin or lung, may be exposed to exogenous (environmental) mutagens (*e.g.*, UV light and tobacco smoke) more heavily than others such as brain (Brash *et al.* 1991; Rodin and Rodin 2000). A number of authors have argued that much of the observed pattern of nucleotide substitutions in cancer genes may be due to fixation of mutations in tumor lineages under the force of natural selection acting at the cellular level so that more aggressive cancer cells (resulting from particular somatic mutations) will dominate in the cellular population dynamics (see, *e.g.*, Vogelstein *et al.* 1988). Substitution probabilities will then be a consequence of both mutational bias in the nucleotide sequence and selective pressure on the amino acid sequence.

Studies that have examined the spectrum of codon substitutions in cancer gene databases, such as the p53 database, have generally used the simple approach of counting the frequencies of missense, nonsense, or silent substitutions observed at a site (*e.g.*, Levine *et al.* 1995; Bennett *et al.* 1999; Hussain and Harris 1999). However, because the codons at different sites in the normal p53 sequence will have different probabilities of undergoing missense, nonsense, or silent changes, this approach will be biased if the codon substitution process is not explicitly modeled. The probability that two mutations occur in a given cell lineage is very small by comparison with the probability of a single mutation. Therefore, a codon that can mutate to a stop codon by a single-nucleotide change will display nonsense substitutions more often than a codon requiring at least two mutations to generate a stop codon. This implies that a codon-based substitution model should be used to study the spectrum of mutations in tumors; the parameters of interest are the probabilities of substitution (*i.e.*, fixation of the mutation in a cell lineage giving rise to a tumor) given that a mutation produces a codon that causes a missense (amino acid substitution), nonsense (nontranslated or truncated protein), or no (silent) amino acid change.

Modeling rates of somatic codon substitution in tumor development over an individual's lifespan is in many ways similar to modeling rates of germline codon substitution among species over evolutionary time. The problem, in that context, is to estimate relative rates of missense *vs.* silent substitution among sites in a comparative analysis of genes from different species (Yang 2001). In this article, we exploit similarities between these two areas of research to develop some simple codon-based models for studying the spectrum of mutations in cancers. We make an effort to take account of the most important factors influencing mutation and cancer development by studying models of varying complexity, allowing for differences in substitution patterns among tumor tissue types and among p53 functional domains. We use the likelihood-ratio test to compare different models. As more is learned about the process of somatic mutation, and of tumorigenesis, these models can be readily modified using this general framework. We illustrate the utility of the models by applying them to the p53 tumor mutation database.

METHODS

Let $\mathbf{Y} = \{Y_l\}$ be the codon sequence of the normal gene, where Y_l is the codon at site l as determined from the reference sequence and l ranges from 1 to L, where L is the total number of codons in the gene. Let $\mathbf{X} = \{X_{ij}\}$, where X_{ij} is the number of sampled tumor gene sequences with a single-nucleotide substitution replacing normal codon i with mutant codon j (for all $j \neq i$). Note that i, j, and Y_l are each 1 of 64 possible distinct codons with the constraint that one nucleotide difference separates i and j. For example, $j = \{\text{AAG}\}$ and $i = \{\text{ATG}\}$. The codon substitution process acting on a cancer gene in the somatic cells of an individual that will ultimately develop a tumor is modeled as a continuous-time Markov process. The instantaneous substitution rate of this process will depend on many factors such as the rate of mutation to different nucleotides, the selective advantage to tumorigenesis (in promoting cell division, etc.) of cells carrying particular mutant forms of the gene, and so on. In this article, we consider several simple models that incorporate some of these influences. It is shown that the details of the demographic process of cancer cell proliferation can be ignored if we condition on a single-nucleotide substitution having occurred in a given cancer cell lineage. This assumption is satisfied for most of the tumors in the p53 database that we use to illustrate the method.

Constant rates model: To model nucleotide mutation we initially use a model with two parameters to describe the nucleotide mutation process: the average rate of mutation per site, μ, and the ratio of transitions to transversions, κ. To model codon substitution, we use a model with three parameters, $\overline{\beta} = \{\beta_S, \beta_M, \beta_N\}$, where these are the probabilities that a newly arisen synonymous, missense, or nonsense mutation, respectively, ultimately becomes fixed in a tumor lineage. We refer to this as model M_0, or the constant rate (CR) model because it assumes that the same mutation and substitution rates apply across all functional domains of a gene and across all primary tumor tissue types.

Let $\mathbf{Q} = \{q_{ij}\}$, where q_{ij} is the instantaneous rate of substitution from codon i to codon j, $q_i = \Sigma_{j \neq i} q_{ij}$ and $q_{ii} = -q_i$. The off-diagonal elements of $\mathbf{Q}$ are products of the instantaneous nucleotide mutation rates and codon fixation probability. For example, if $i = \{\text{TCG}\}$ and $j = \{\text{TTG}\}$, then $q_{ij} = \mu\kappa/(2 + \kappa) \times \beta_M$. Define $m_j = \Sigma_i \mathbf{I}(Y_i, j)$, where $\mathbf{I}(Y_i, j)$ equals 1 if $Y_i = j$ and 0 otherwise (*i.e.*, the number of codons in the normal sequence that are of type j). It is assumed that each codon undergoes an independent substitution process. A Markov process can be uniquely characterized as a sojourn process (Taylor and Karlin 1984). If the process is initially in state i, the waiting time, t, until an event occurs is exponentially distributed with parameter q_i and the probability density function (pdf) is

$$f(t) = q_i e^{-q_i t}. \tag{1}$$

If a substitution event occurs, and the initial state is i, it is a substitution to state j with probability q_{ij}/q_i. The waiting time, t_i, until the first substitution at any site bearing codon i in the normal sequence is then the smallest order statistic of m_i iid exponential random variables with common parameter q_i. The pdf of t_i is

$$f_i(t_i) = q_i m_i e^{-(q_i m_i) t_i}. \tag{2}$$

The density function is the same for sites bearing any other codon $l \neq i$ in the normal sequence provided that q_i and m_i are replaced by q_l and m_l. The first substitution occurs at a site with codon i in the normal sequence if $t_i < t_j$ for all $j \neq i$. The joint density of t_i and $t_i < t_j$ for all $j \neq i$ is

$$f(t_i, t_i < t_j, \forall j \neq i) = q_i m_i e^{-(q_i m_i) t_i} e^{-\Sigma_{j \neq i} q_j m_j t_i}. \tag{3}$$

The marginal probability that $t_i < t_j$, averaged over all possible values of t_i, is

$$\Pr(t_i < t_j) = \int_{t_i=0}^{t_i=\infty} q_i m_i e^{-(q_i m_i t_i)} e^{-\Sigma_{j \neq i}(q_j m_j t_i)} dt_i = \frac{m_i q_i}{\Sigma_j m_j q_j}. \tag{4}$$

If it is assumed that no more than one codon substitution has occurred in a gene in the development of a particular tumor lineage, then the probability of a change from codon i to j is the probability that a substitution occurs at a site with codon i in the normal sequence (given by Equation 4 above) multiplied by the probability of a transition from i to j, given that a substitution has occurred, which is q_{ij}/q_i as noted above. Thus, the probability that one substitution occurs from i to j is

$$\phi_{ij} = \left(\frac{m_i q_i}{\Sigma_j m_j q_j}\right) \frac{q_{ij}}{q_i} = \frac{m_i q_{ij}}{\Sigma_j m_j q_j}. \tag{5}$$

Because both q_{ij} and q_j are linear functions of μ, the mutation rate cancels out. The remaining parameters $\overline{\beta}$ and κ can be estimated from the data using maximum likelihood. The substitution probabilities always occur in ratios in Equation 5, so one of these parameters is not identifiable. We instead estimate the three identifiable parameters $\alpha_N = \beta_N/\beta_S$, $\alpha_M = \beta_M/\beta_S$, and κ. The likelihood function is

$$\mathrm{L}(\mathbf{X}|\mathbf{Y}, \alpha_N, \alpha_M, \kappa) = C \prod_i \prod_{i \neq j} \phi_{ij}^{X_{ij}}. \tag{6}$$

We used numerical methods to maximize the log-likelihood function ($\log L$) with respect to these parameters, where $\log L$ is

$$\log L(\mathbf{X}|\mathbf{Y}, \alpha_M, \alpha_N, \kappa) = \sum_i \sum_{i \neq j} X_{ij} \log \left(\frac{m_i q_{ij}}{\Sigma_j m_j q_j}\right). \tag{7}$$

The CR model assumes that rates of nucleotide mutation and codon substitution are identical across nucleotides over the entire coding region of the gene.

Variable rates models: The CR model M_0 presented above can be readily extended to develop a hierarchy of variable rates models; here we present several models that allow rates of substitution to vary across known functional domains of a tumor suppressor gene (or oncogene) and/or across tumors of different tissues. Moreover, models that allow mutation rates to be influenced by the primary nucleotide sequence, for example, to account for the well-known influence of CpG dinucleo-

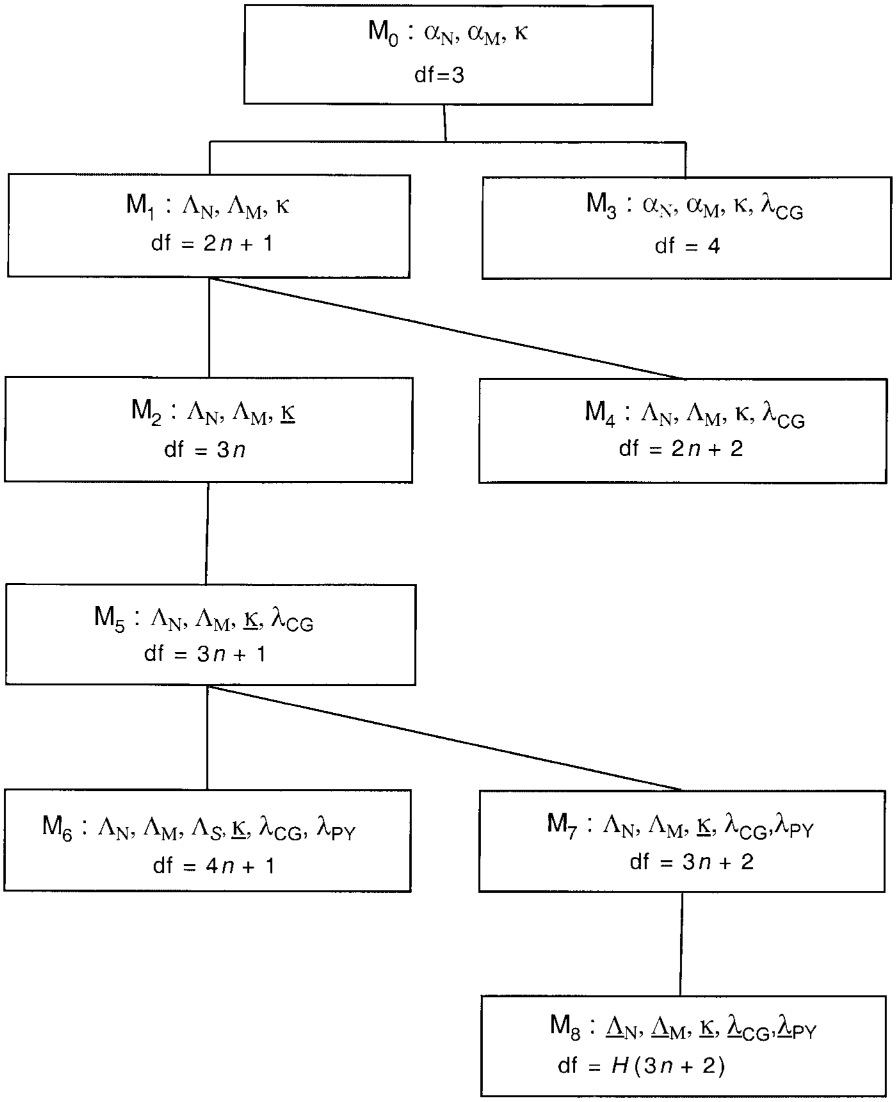
$M_0 : \alpha_N, \alpha_M, \kappa$
df=3
$M_1 : \Lambda_N, \Lambda_M, \kappa$
df = $2n + 1$
$M_3 : \alpha_N, \alpha_M, \kappa, \lambda_{CG}$
df = 4
$M_2 : \Lambda_N, \Lambda_M, \underline{\kappa}$
df = $3n$
$M_4 : \Lambda_N, \Lambda_M, \kappa, \lambda_{CG}$
df = $2n + 2$
$M_5 : \Lambda_N, \Lambda_M, \underline{\kappa}, \lambda_{CG}$
df = $3n + 1$
$M_6 : \Lambda_N, \Lambda_M, \Lambda_S, \underline{\kappa}, \lambda_{CG}, \lambda_{PY}$
df = $4n + 1$
$M_7 : \Lambda_N, \Lambda_M, \underline{\kappa}, \lambda_{CG}, \lambda_{PY}$
df = $3n + 2$
$M_8 : \underline{\Lambda}_N, \underline{\Lambda}_M, \underline{\kappa}, \underline{\lambda}_{CG}, \underline{\lambda}_{PY}$
df = $H(3n + 2)$

FIGURE 1.—Relationships among the models developed for analyzing the spectrum of somatic cancer mutations. Each box defines a particular model. The parameters included in each model are listed in each box, as well as the total number of free parameters (df, degrees of freedom). The parameters are defined as follows: α_N and α_M are the relative substitution rates of nonsense *vs.* silent substitutions and missense *vs.* silent substitutions, respectively; κ is the ratio of rates of transition *vs.* transversion; λ_{CG} is the relative rate of substitution at CG dinucleotides; λ_{PY} is the relative rate of substitution at dipyrimidines; Λ_N and Λ_M are vectors $\{\alpha_N(i)\}$ and $\{\alpha_M(i)\}$, respectively, where $\alpha_N(i)$ is the relative rate of nonsense substitutions in the ith functional domain, etc.; $\underline{\Lambda}_N$ and $\underline{\Lambda}_M$ are matrices $\{\alpha_N(i, j)\}$ and $\{\alpha_M(i, j)\}$, respectively, where $\alpha_N(i, j)$ is the relative rate of nonsense substitutions in the ith functional domain of the jth tumor type, etc.; $\underline{\lambda}_{CG}$ is a vector $\{\lambda_{CG}(j)\}$, where $\lambda_{CG}(i)$ is the relative rate of substitution at CG dinucleotides in the jth tumor type; $\underline{\lambda}_{PY}$ is a vector $\{\lambda_{PY}(j)\}$, where $\lambda_{PY}(j)$ is the relative rate of substitution at dipyrimidines in the jth tumor type; $\underline{\kappa}$ is a vector $\{\kappa(i)\}$, where $\kappa(i)$ is the transition-transversion bias for the ith functional domain (for models M_2, M_5, M_6, and M_7), and is a matrix $\kappa(i, j)$, where $\kappa(i, j)$ is the transition-transversion bias for the ith functional domain in the jth tumor type (for model M_8).

tides on mutation rates, are considered (see COOPER and YOUSSOUFIAN 1988; LAIRD and JAENISCH 1996). The models are summarized in Figure 1.

Model M_1 allows the relative rates of missense and nonsense substitution to vary across functional domains. We define $\Lambda_N = \{\alpha_N(i)\}$ and $\Lambda_M = \{\alpha_M(i)\}$, where $\alpha_N(i)$ is the ratio of nonsense to silent substitutions in the ith functional domain, etc. Model M_1 retains a common transition/transversion bias, κ, and a common mutation rate, μ, across functional domains. If a gene has n functional domains, there are $2n + 1$ parameters under this model because μ cannot be estimated from the data if we condition on a single substitution having occurred in each sampled tumor. Model M_2 is similar, but allows the transition/transversion bias parameter, $\kappa(i)$, to also vary across regions. We define $\underline{\kappa} = \{\kappa(i)\}$, where $\kappa(i)$ is the transition/transversion ratio for the ith functional domain. If a gene has n functional domains, there are $3n$ parameters under this model.

Model M_3 assumes constant rates across functional domains but adds an additional parameter λ_{CG} that is the relative rate of substitution at CG dinucleotides *vs.* non-CG sites. The dinucleotide model considers the substitution rate of a "quintet," which includes the nucleotide before the first codon position, the codon itself, and the nucleotide after the third codon position. If a mutation changes a quintet with no CpG into a quintet with CpG (for example, "T TCT A" changing into "T TCG A"), the substitution rate is divided by λ_{CG}. If a mutation changes a quintet with a CpG into a quintet without, the substitution rate is multiplied by λ_{CG}. If the source and target quintets either both lack or both contain CpG doublets, the rate is not changed. Model M_4 allows $\alpha_M(i)$ and $\alpha_N(i)$ to vary across functional domains (as in M_1) but adds an additional parameter λ_{CG} that is assumed to be constant across domains. Model M_5 extends model M_4 by allowing transition/transversion ratios to vary across functional domains; model M_5 is identical to M_2 apart from the additional parameter λ_{CG}. Model M_6 adds n additional parameters to model M_5, allowing the relative

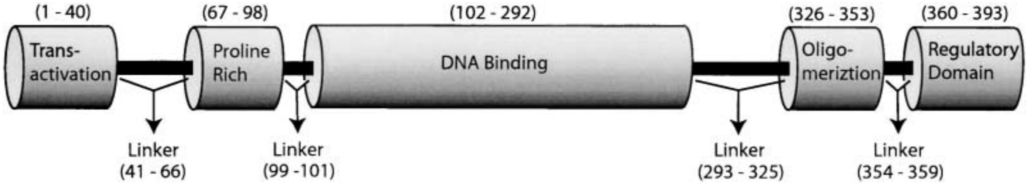

Figure 2.—Diagram showing the functional domains of the p53 protein. The four linker regions are combined and treated as a single functional domain in our analysis. The inferred roles of the domains in p53 protein function are described in the text.

rate of silent substitution to vary across functional domains; we define $\Lambda_S = \{\alpha_S(i)\}$, where $\alpha_S(i)$ is the silent substitution rate of functional domain i relative to functional domain 1, for all $i \neq 1$. In addition, we add a parameter λ_{PY} that is the relative substitution rate for nucleotides that occur as dipyrimidines *vs.* those that do not (using a quintet codon model of the same form as was used to model λ_{CG}). Model M_7 extends model M_5 by adding the λ_{PY} parameter. Model M_8 is the most parameter-rich model we consider. This model extends M_7 by allowing Λ_M, Λ_N, and λ_{CG} to vary across primary tumors from different tissues, adding $2(H + 2)(n + 2)$ additional parameters, where H is the number of different primary tumor tissues stratified in the database.

ANALYSIS

The p53 tumor suppressor protein was originally identified in several independent studies in 1979 both as a protein that interacts with SV40 virus large T antigen (Lane and Crawford 1979; Linzer and Levine 1979) and as a highly expressed protein in chemically induced tumors (Deleo *et al.* 1979). Initially, it was thought that p53 was an oncogene, but subsequent studies in the late 1980s clearly established that p53 is actually a tumor suppressor gene (May and May 1999). Inactivation of the p53 gene is now known to be the most common alteration in tumors; slightly >50% of human cancers contain mutations in this gene (Hollstein *et al.* 1994). The p53 protein has various functional roles in normal cells. As a transcription factor, p53 upregulates expression of genes involved in cell cycle arrest or apoptosis in response to DNA damage or other kinds of stress such as hypoxia, expression of an oncogene, etc. (Levine 1997). Furthermore, p53 is known to be involved in transcription-independent apoptosis and DNA damage repair (Balint and Vousden 2001).

For our analysis, we used release 5 of the p53 database (http://www.iarc.fr/P53/). This database contained a total of 15,121 tumor entries as of July 1, 2001. The 11 exons of the p53 gene contain 1179 nucleotides coding for 393 amino acids. In total, 222 of the 393 codons have thus far been observed to be targets of mutation in cancer. Mutations include insertions or deletions of nucleotides (most often resulting in a frameshift), as well as point mutations. Because our models condition on a single-point mutation having occurred (in an exon), prior to our analysis we removed sequences from the database that contained insertions, deletions, mutations in introns, or multiple-point mutations. This reduced the total number of tumor entries used in our analysis to 12,759. There are six recognized functional domains in the p53 gene (see Figure 2) but the boundaries of the domains described in the literature often differ by several amino acids (see, *e.g.*, Levine 1997; Roemer 1999). We use the boundaries suggested by Roemer (1999) to define the start and end points for each domain in our analysis.

The transcriptional activation domain (residues 1–40) interacts with the basal transcriptional machinery (*e.g.*, RNA polymerase, other transcription factors, etc.), activating transcription of its target genes; the proline-rich domain (residues 67–98) is involved in the binding of p53 to the nuclear matrix and may play a role in stimulating apoptosis in cells with irreversible DNA damage (Jiang *et al.* 2001); the DNA-binding domain (residues 102–292) interacts with DNA and binds to specific promoters that are a target for p53 in its role as a transcription factor (Levine 1997); the oligomerization domain (residues 326–353), also called the tetramerization domain (TD), is involved in the assembly of p53 molecules into their characteristic tetrameric structure and also plays a role in DNA binding, protein-protein interactions, and post-translational interactions (Chéne 2001); the regulatory domain (residues 360–393) regulates sequence-specific DNA binding (Levine 1997); the four "linker" regions join these five domains and were collectively treated as a sixth distinct functional region (linkers) in our analyses (see Figure 2).

Model M_8 partitions the parameter estimates according to primary tumor tissue type as well as functional domain. To carry out this analysis, we partitioned the data according to the source of the primary tumor as documented in the database. We combined mutations from samples obtained from both surgeries and established cell lines. Twelve primary cancers are each represented by >600 samples in the database and to maintain large sample sizes we chose to partition by these categories only. These 12 cancers accounted for 9886 of the single-point mutation entries; the remaining 2873 cancers in the database caused by a single-point mutation were too rare for separate analyses and were instead analyzed

TABLE 1

Results of likelihood-ratio tests (LRTs) comparing the fit of eight nested models when applied to the p53 cancer mutation database

Models (parameters)	d.f.	2 log Δ
M_1 (Λ_N, Λ_M, κ) *vs.* M_0 (α_N, α_M, κ)	10	13,619.6
M_2 (Λ_N, Λ_M, $\underline{\kappa}$) *vs.* M_1 (Λ_N, Λ_M, κ)	5	378.3
M_3 (α_N, α_M, κ, λ_{CG}) *vs.* M_0 (α_N, α_M, κ)	1	7,633.4
M_4 (Λ_N, Λ_M, κ, λ_{CG}) *vs.* M_1 (Λ_N, Λ_M, κ)	1	6,950.2
M_5 (Λ_N, Λ_M, $\underline{\kappa}$, λ_{CG}) *vs.* M_2 (Λ_N, Λ_M, $\underline{\kappa}$)	1	6,977.0
M_6 (Λ_N, Λ_M, Λ_S, $\underline{\kappa}$, λ_{CG}, λ_{PY}) *vs.* M_5 (Λ_N, Λ_M, $\underline{\kappa}$, λ_{CG})	6	333.8
M_7 (Λ_N, Λ_M, $\underline{\kappa}$, λ_{CG}, λ_{PY}) *vs.* M_5 (Λ_N, Λ_M, $\underline{\kappa}$, λ_{CG})	1	93.8
M_8 ($\underline{\Lambda}_N$, $\underline{\Lambda}_M$, $\underline{\kappa}$, $\underline{\lambda}_{CG}$, $\underline{\lambda}_{PY}$) *vs.* M_5 (Λ_N, Λ_M, $\underline{\kappa}$, λ_{CG})	240	2,061.2

The test statistic 2 log Δ is approximately χ^2 distributed with the number of degrees of freedom equal to the difference in the number of free parameters between models. Note that Δ is the ratio of the probability of the observed data (maximized with respect to the free parameters) under the simple model *vs.* the more complex model. Each model was compared only to the submodel that minimized the difference of the number of free parameters between models. Models are described in the text. All comparisons were significant at the 0.000001 level.

collectively in a composite category labeled as "other cancers." The 12 cancer categories are listed in Tables 3 and 4.

RESULTS

The results of likelihood-ratio tests comparing all eight models are shown in Table 1. All of the increasingly complex models that we examined resulted in a significant improvement in the fit of the model to the p53 mutation data. The greatest improvements are obtained by partitioning rates according to functional domains and allowing higher rates of substitution at CG dinucleotides (see Figure 2 and Table 1). The most complex models considered (with or without constant rates across tumor tissues) are preferred over the remaining submodels for parameter estimation because all result in a significant improvement in the fit of the models to the data. We also used the Akaike information criterion (AIC; Akaike 1973) for model selection. Under this criterion, the model that minimizes the AIC (minus two times the likelihood of the data with maximum-likelihood estimates of the parameters plus twice the number of free parameters) is preferred. As the log-likelihood differences are much larger than the difference in the number of parameters (Table 1), use of this criterion also leads one to prefer the more complex models. Thus, we present here only the results for models M_6, M_7, and M_8. These results are found in Tables 2–4.

Table 2 shows the results for analyses of the p53 database using models M_6 and M_7. The only difference between these two models is that M_7 allows the relative rate of silent substitution, α_S, to vary across domains, whereas model M_6 assumes that it is constant. It is evident from the results of our analysis using model M_7 (see bottom half of Table 2) that α_S varies considerably across domains. Most strikingly, the silent substitution rate is at least an order of magnitude higher for the DNA-binding domain *vs.* the others. Because silent substitutions (by definition) do not affect the amino acid sequence, the potential functional significance of such changes is limited. Possible effects of silent substitutions might be an increase, or reduction, of the rate of translation, for example, if the relative abundance of tRNAs specific for each alternative codon varies. Although such a mechanism is a reasonable explanation for codon usage bias within a gene as a whole, it is not a likely explanation for the variation we observe in silent rates of substitution among functional domains within the p53 gene.

Another possible effect of codon usage bias is on translational accuracy. Selection for translational accuracy might cause codon usage bias, and therefore silent substitution rates, to vary across functional domains. There is some evidence for such effects in Drosophila. Akashi (1994) showed that for 28 Drosophila proteins with DNA-binding domains, codon bias is greater in these domains. There is little evidence for codon usage bias causing silent rate variation in mammals, however. A more likely explanation for the observed variation in silent substitution rate is investigator sampling bias; some functional domains may be sequenced more often than others and therefore silent substitutions in those domains appear more often in the database (Levine *et al.* 1995). Because the central DNA-binding domain is widely perceived to be the most common target of mutation in p53, this domain is sequenced more often than other domains in studies of p53 mutations in cancers (Levine *et al.* 1995). Many studies have sequenced only exons 5–8 (Soussi and Beroud 2001). This ascertainment bias, if not properly taken into account, can lead to biased estimates of the relative substitution rates among functional domains. For example, under model M_6, the DNA-binding domain has the highest estimates of α_M, α_N, and κ, while under model M_7, which allows the silent

TABLE 2

Estimates of parameter values under models M_6 and M_7 using the combined p53 mutation database of 12,759 samples

	p53 functional domain					
Parameter	Linkers	Trans	Prol-R	DNA-B	TD	Regulatory
			Model M_6			
α_M	0.79	0.19	0.58	*8.3*	0.35	$\underline{0.11}$
α_N	10.5	$\underline{0.13}$	7.2	*12.5*	6.0	—
κ	1.8	$\underline{0.35}$	1.6	*4.4*	1.4	0.62
λ_{CG}	4.5					
λ_{PY}	1.3					
α_N/α_M	13.2	$\underline{0.70}$	11.7	1.5	*17.9*	—
			Model M_7			
α_M	1.1	2.7	2.8	*6.5*	4.3	$\underline{1.0}$
α_N	15.0	$\underline{1.9}$	33.1	9.9	*76.6*	—
α_S	—	$\underline{0.07}$	0.21	*2.0*	0.09	0.06
κ	2.2	1.8	3.3	4.4	3.0	—
λ_{CG}	4.5					
λ_{PY}	1.3					
α_N/α_M	13.3	$\underline{0.67}$	12.5	1.5	*17.5*	—

The model parameters are defined in the text. The six functional domains of the p53 gene are abbreviated as follows: Linker, linker sequences joining domains; Trans, transcription activation domain; Prol-R, proline-rich domain); DNA-B, DNA-binding domain; TD, tetramerization domain; and Regulatory, regulatory domain. These domains are further described in the text. The highest estimated value for each parameter is indicated in italics and the lowest is underlined. Missing elements in the table indicate that the same parameter value was assumed to apply across all domains. A dashed element indicates that standard errors were too large (data insufficient) to allow an estimate of the parameter.

rate to vary among domains, the DNA-binding domain retains the highest rate of missense substitution but now has one of the lowest rates of nonsense substitution (Table 2). Also under M_7, the oligomerization domain has a rate, α_N, which is roughly eight times higher than that of the DNA-binding domain. Moreover, under model M_6 the estimated values of α_N and/or α_M for several domains are <1, implying that silent mutations are more likely to cause cancer than are missense or nonsense mutations, which is not reasonable. Under model M_7 all relative substitution rates are >1. Another potential concern is that genes with multiple substitutions that violate our model assumptions will be ascertained into the sample because partial sequencing has revealed only one of the substitutions. If explicit information about the screening procedures used in each study were available, it might be possible to modify the model to correct for this potential source of bias.

A final concern is that "investigator sampling bias" may be enhanced by p53 germline polymorphisms in the general population. In our analysis, we treated the "reference" germline p53 sequence as fixed. In reality, p53 nucleotide polymorphisms exist in the human population that could influence whether a tumor is included in our analysis (*e.g.*, has a single-nucleotide substitution) or excluded (*e.g.*, has two, or more, nucleotide substitutions). More detailed models (and more detailed information) regarding the tumor sampling (and sequencing) process are needed to fully address such issues.

In contrast with the nonsense and missense rates relative to the silent rate, the nonsense/missense rate ratio is effectively independent of the investigator sampling bias. The variance of the estimated α_N/α_M ratio for each domain is influenced by investigator sampling bias (because this sampling bias reduces the sample size for some domains and not others) but the estimates are not biased by this effect (compare estimates of α_N/α_M between models M_6 and M_7 in Table 2). If we consider the ratio α_N/α_M, the DNA-binding domain displays a constant ratio of 1.5 under either model M_6 or M_7; this is dramatically lower than that for all other domains, apart from the transactivation domain (ratio of $\sim$0.7). The largest ratio is observed for the oligomerization domain (ranging from 17.5 to 17.9, depending on which model is used).

The striking differences that we observe in the rates of nonsense *vs.* missense substitutions among domains have a direct biological interpretation: the structural regions (linkers and proline-rich and oligomerization domains) may be largely unaffected by missense mutations because the precise residues found in such regions are often unimportant for p53 function; the specific residues of the DNA-binding and transactivation domains, on the other hand, may have a more important effect

TABLE 3

Estimates of parameter values obtained by applying model M_8 to the p53 cancer mutation database and partitioning by 12 primary cancers (accounting for 9886 samples), each represented by at least 600 samples in the database, and a composite of the remaining cancers (accounting for 2873 samples)

	Parameter estimates					
	α_N/α_M: p53 functional domain					
Primary tumor	Linkers	Prol-R	DNA-B	TD	λ_{CG}	λ_{PY}
Bladder	5.9	4.6	2.0	13.3	3.1	0.98
Brain	12.5	3.8	0.43	20.4	6.3	*1.4*
Breast	10.4	26.3	1.9	22.3	3.9	0.86
Colon	11.1	—	1.3	—	9.5	0.94
Esophagus	34.2	—	2.0	19.4	4.9	0.85
Hematopoietic	4.5	26.2	1.1	13.4	4.2	0.98
Liver	15.0	—	0.95	14.5	2.6	0.32
Lung	17.4	*81.8*	1.6	9.1	3.5	0.61
Ovary	—	6.8	1.2	—	4.1	0.79
Rectum	*109.3*	—	1.4	—	*10.8*	0.86
Skin	13.64	4.0	*2.7*	*68.7*	2.8	0.91
Stomach	1.9	22.0	1.7	12.9	6.4	0.93
Other cancers	15.9	13.3	1.6	19.3	3.7	0.93

The first four columns show estimates of the ratio of nonsense to missense substitutions α_N/α_M for four functional domains labeled as follows: Linkers, linker sequences joining domains; Prol-R, proline-rich domain; DNA-B, DNA-binding domain; and TD, tetramerization domain. These domains are further described in the text. The highest estimated value for each parameter is indicated in italics and the lowest is underlined. The last two columns show estimates of the relative mutation rate for CpG dinucleotides, λ_{CG}, and dipyrimidines, λ_{PY}, for each cancer type. A dashed element in the table indicates that standard errors were too large (data insufficient) to allow an estimate of the parameter.

on function, and missense or nonsense substitutions in these domains thus contribute nearly equally to tumor development.

The low estimated rates of missense substitutions for the transactivation and oligomerization domains are likely due to the nonspecific nature of those domains. Studies suggest that a single-point mutation in those domains is generally not able to completely abolish the protein function (Lin *et al.* 1994; Pietenpol *et al.* 1994; Jeffrey *et al.* 1995; Waterman *et al.* 1995).

Estimates of λ_{CG} suggest that the rate of mutation at CG dinucleotides is more than fourfold the rate at non-CG sites. Estimates of parameter λ_{PY}, on the other hand, are close to one, indicating only a slight increase of the mutation rates at dipyrimidine sites. The estimated transition/transversion ratio, κ, varies from 1.8 to 4.4 under model M_7, which is within the range of values observed in evolutionary studies. The values of κ are biased downward when variation in silent substitition rates is not accounted for (*i.e.*, compare estimates of κ under models M_6 and M_7 in Table 2).

The results of our analyses using model M_8, which allows parameters to vary across primary tumor types, as well as functional domains, are shown in Tables 3 and 4. First, we consider the substitution process; there is considerable variation in α_N/α_M among tumors, but some domains show much greater variation than others. Results are shown in Table 3 for only four of the six domains because too few observations were available to reliably estimate α_N/α_M for the transactivation and regulatory domains using the partitioned datasets. The least variation of α_N/α_M across tumor tissues is observed for the DNA-binding domain, with the ratio varying from a low of 0.43 (in brain cancers) to a high of 2.7 (in skin cancers). The most variation of α_N/α_M across tumor tissues is observed for the linkers with the ratio varying from a low of 1.9 (in stomach cancers) to a high of 109.3 (in rectal cancers). There are also some clear trends across tumor types: bladder and brain cancers appear to have the lowest average α_N/α_M ratio (averaged across domains) and lung, rectum, and skin cancers have the highest. These differences in substitution rates are very pronounced and it is likely that they are indicators of fundamental underlying differences in the biological role of p53 in cancer initiation and progression in these different tissues.

We also studied the mutation process in different tumor types by examining estimates of λ_{CG}, λ_{PY}, and κ (Table 3). Parameter λ_{CG} varies widely among tumor types with brain, colon, stomach, and rectum having the highest values (ranging from 6.3 to 10.8) and bladder, liver, lung, and skin having the lowest values (ranging from 2.6 to 3.5). This is likely a reflection of the influence of exogenous *vs.* endogenous mutagenic influences in the different

TABLE 4

Estimates of the ratio of transitions to transversions for each functional domain obtained by applying model M_8 to the p53 cancer mutation database and partitioning by 12 primary cancers (accounting for 9886 samples), each represented by at least 600 samples in the database, and a composite of the remaining cancers (accounting for 2873 samples)

		Transition/transversion ratio (κ): p53 functional domain					
Primary tumor	Parameter estimates:	Linkers	Trans	Prol-R	DNA-B	TD	Regulatory
Bladder		2.2	0.63	2.2	5.2	0.44	0.73
Brain		4.4	1.2	*7.2*	7.6	1.9	3.0
Breast		3.3	0.29	2.0	5.9	1.9	0.98
Colon		7.3	*3.1*	—	*9.0*	*6.0*	*9.5*
Esophagus		1.5	0.84	—	4.0	1.9	—
Hematopoietic		1.5	0.96	1.5	4.7	0.97	—
Liver		1.1	—	—	1.6	1.7	—
Lung		0.54	0.19	0.62	2.1	0.67	—
Ovary		*11.2*	0.86	1.9	5.1	4.6	2.7
Rectum		3.4	—	—	8.2	2.3	—
Skin		3.4	—	4.0	6.2	5.3	—
Stomach		2.0	—	1.1	6.7	0.70	—
Other cancers		1.3	0.10	1.2	4.4	1.3	—

The highest estimated value for each parameter is indicated in italics and the lowest is underlined. A dashed element in the table indicates that standard errors were too large (data insufficient) to allow an estimate of the parameter.

organs. Mutations in p53 from bladder, liver, lung, and skin may be more heavily influenced by exogenous factors, while mutations from brain, colon, stomach, and rectum may be most heavily influenced by endogenous factors such as primary sequence. The dipyrimidine mutation rate parameter, λ_{PY}, is much less variable among primary tumor types and is quite close to 1 in most cases (ranging from a low of 0.32 in liver to a high of 1.4 in brain). This suggests that there is little difference in mutation rates as a consequence of a dipyrimidine in the primary sequence. Because at least one mechanism of dipyrimidine mutation (conversion of CC to TT by UV; Brash *et al.* 1991) results in two nucleotide substitutions, this effect would not be detectable in our analysis, which focuses on single-nucleotide substitutions.

Table 4 shows the variation of the transition/transversion rate ratio, κ, across tumor types and across functional domains. The average value of κ is highest for the DNA-binding domain and lowest for the transactivation domain. These results may be biased, however, because we have not corrected for investigator sampling bias (variation of α_S across domains) in this analysis. More reliable is the variation of the average κ values across primary tumor types. The highest average value of κ is observed for tumors of the brain, colon, and ovary. The lowest is observed for tumors of the bladder, liver, and lung. Once again, this is likely to reflect differences in exogenous *vs.* endogenous mutational influences: the most pervasive endogenous factor influencing rates of mutation is the presence of CpG sites; this increases the rates of transitions *vs.* transversions whereas many exogenous mutagens have the opposite effect.

DISCUSSION

Large-scale databases that compile the frequencies of somatic mutations at particular nucleotides of cancer genes from tumors are an important new resource for studying the role of somatic mutation in cancer development and progression. In this article, we have developed a general parametric framework aimed at modeling the spectrum of mutations in cancer genes and facilitating estimation of biologically relevant parameters. It is shown (by examining the p53 mutation database) that an important parameter to consider is the relative rate of substitution of nonsense *vs.* missense mutations (*i.e.*, the ratio of nonsense to missense substitution rates), α_N/α_M, in different functional domains and primary cancer types. A ratio close to 1 was observed for the DNA-binding domain, indicating that missense and nonsense mutations were about equally likely to produce cancer in this domain. The remaining domains, which are primarily involved in protein structure, displayed ratios $\gg$1 (100-fold greater in some tumor types), indicating that these domains can tolerate a much higher level of missense mutation without producing cancer. A codon-based model, such as we have developed, is needed to extract this information because it depends critically on the probabilities that particular codons produce missense or nonsense changes. The overall frequency of missense mutations is much higher in all domains (Levine *et al.* 1995), swamping the effect of selection on the substitution process if codon usage is not explicitly taken into account.

Another finding in our analysis of the p53 mutation database is that estimates of λ_{CG}, the parameter that describes the effect of CG dinucleotides on mutation rates, vary greatly among primary cancer types; this likely reflects the differing importance of endogenous and exogenous factors on the mutational spectrum in these organs. Similarly, the ratio of the transition rate to the transversion rate, κ, varies dramatically across domains and across primary tumor types; this is also likely to reflect an underlying heterogeneity of the mutation process in different organs, at least partially due to differing environmental influences. Both the effects of environment (on the spectrum of mutations) and the influence of selection acting on cells carrying particular mutations (on the spectrum of substitutions) can be detected using our models. Selection acting during the substitution process likely accounts for much of the variation in substitution rates among domains, and endogenous and exogenous factors influencing the mutation process likely account for much of the variation among tumor types. It is important to try to tease apart the effects of these different influences.

By examining the relative rates of silent substitution, α_S, among domains we find strong evidence supporting the conjecture of Levine *et al.* (1995) that sampling bias exists in the p53 mutation database, with the DNA-binding domain sequenced more often than other domains. If we concentrate inferences on the relative rates of nonsense *vs.* missense substitution, this ratio is insensitive to sampling bias (*i.e.*, estimates of the ratio are not biased by the differential sampling). However, the sampling bias does reduce the effective sample size for some domains, increasing the variance of estimates and preventing reliable parameter estimation. This was the case for the p53 regulatory domain in our analyses. An important question not addressed in this article is how variable substitution rates are among nucleotides within functional domains. Studies of the overall frequency of mutation at particular sites suggest that mutational "hotspots" may exist. However, because codon-based models were not used in these analyses it is conceivable that some hotspots may be an artifact of codon usage patterns within a domain. Ideally, Bayesian methods for predicting the distribution of mutational hotspots (and testing whether hotspots in fact exist) should be developed in the context of a codon-based model.

Another intriguing possibility is that the boundaries of functional domains might be initially identified, or further refined, by examining the spectrum of somatic mutations. One could choose the gene boundaries as part of a Bayesian or maximum-likelihood analysis. Our analyses suggest that it is very important to partition analyses according to both tumor type and functional domain. However, this greatly reduces the power of the analyses because there may be only a few hundred observations in each tumor type category *vs.* thousands of observations in the database as a whole. The general model that we have developed can be extended to account for additional complexities; given that all the models we considered provided a highly significant improvement in the fit of the model to the p53 mutation database it is very likely that yet more complex models can be proposed that will further improve the fit to the data. Our models should be viewed as only an initial step toward the development of a realistic parametric framework for modeling the spectrum of mutations in cancer genes.

The program oncSpectrum, written in the C language, implements maximum-likelihood estimation of parameters for all the models described in this article. It is intended for use with data from a cancer mutation database such as the p53 database. The program can be downloaded from http://rannala.org.

This research was supported by grants from the Alberta Heritage Foundation for Medical Research (AHFMR), the Peter Lougheed Foundation/Canadian Institutes of Health Research (MOP44064, PLS47851), and the National Institutes of Health (HG01988) to B. Rannala and by grants from the Biotechnology and Biological Sciences Research Council and Human Frontier Science Program to Z. Yang. AHFMR provided funds for Z. Yang to visit the University of Alberta in 2001.

LITERATURE CITED

Akaike, H., 1973 Information theory and an extension of maximum likelihood principle, pp. 267–281 in *2nd International Symposium on Information Theory*. Akademia Kiado, Budapest.

Akashi, H., 1994 Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics **136:** 927–935.

Balint, E., and K. H. Vousden, 2001 Activation and activities of the p53 tumor suppressor protein. Br. J. Cancer **85:** 1813–1823.

Bennett, W. P., S. P. Hussain, K. H. Vahakangas, M. A. Khan and P. G. Shields, 1999 Molecular epidemiology of human cancer risk: gene-environment interactions and p53 mutation spectrum in human lung cancer. J. Pathol. **187:** 8–18.

Bishop, J. M., 1991 Molecular themes in oncogenesis. Cell **64:** 235–248.

Brash, D. E., J. A. Rudolph, J. A. Simon, A. Lin, G. J. McKenna *et al.*, 1991 A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. Proc. Natl. Acad. Sci. USA **88:** 10124–10128.

Chéne, P., 2001 The role of tetramerization in p53 function. Oncogene **20:** 2611–2617.

Cooper, D. N., and H. Youssoufian, 1988 The CpG dinucleotide and human genetic disease. Hum. Genet. **87:** 151–155.

Deleo, A. B., G. Jay, E. Appella, G. C. Dubois, L. W. Law *et al.*, 1979 Detection of a transformation-related antigen in chemically induced sarcomas and other transformed cells of the mouse. Proc. Natl. Acad. Sci. USA **76:** 2420–2424.

Hanahan, D., and R. A. Weinberg, 2000 The hallmarks of cancer. Cell **100:** 57–70.

Hollstein, M., K. Rice, M. S. Greenblatt, T. Soussi, R. Fuchs *et al.*, 1994 Database of p53 gene somatic mutations in human tumors and cell lines. Nucleic Acids Res. **22:** 3551–3555.

Hussain, S. P., and C. C. Harris, 1999 p53 mutation spectrum and load: the generation of hypotheses linking the exposure of endogenous or exogenous carcinogens to human cancer. Mutat. Res. **428:** 23–32.

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

Jiang, M., T. Axe, R. Holgate, C. P. Rubbi, A. L. Okorokov *et al.*, 2001 p53 binds the nuclear matrix in normal cells: binding involves the proline-rich domain of p53 and increases following genotoxic stress. Oncogene **20:** 5449–5458.

Laird, P. W., and R. Jaenisch, 1996 The role of DNA methylation in cancer genetics and epigenetics. Annu. Rev. Genet. **30:** 441–464.

Lane, D. P., and L. V. Crawford, 1979 T antigen is bound to a host protein in sv40-transformed cells. Nature **278:** 261–263.

Levine, A. J., 1997 p53, the cellular gatekeeper for growth and division. Cell **88:** 323–331.

Levine, A. J., M. C. Wu, A. Chang, A. Silver, E. A. Attiyeh *et al.*, 1995 The spectrum of mutations at the p53 locus. Ann. NY Acad. Sci. **768:** 111–128.

Lin, J., J. Chen, B. Elenbaas and A. J. Levine, 1994 Several hydrophobic amino acids in the p53 amino-terminal domain are required for transcriptional activation, binding to mdm-2 and the adenovirus 5 E1B 55-kD protein. Genes Dev. **8:** 1235–1246.

Linzer, D. I., and A. J. Levine, 1979 Characterization of a 54K dalton cellular SV40 tumor antigen present in SV40-transformed cells and uninfected embryonal carcinoma cells. Cell **17:** 43–52.

May, P., and E. May, 1999 Twenty years of p53 research: structural and functional aspects of the p53 protein. Oncogene **18:** 7621–7636.

Ozoren, N., and W. S. El-Deiry, 2000 Introduction to cancer genes and growth control, pp. 3–43 in *DNA Alterations in Cancer: Genetics and Epigenetic Changes*, edited by M. Ehrlich. Eaton Pressing, Natick, MA.

Pietenpol, J. A., T. Tokino, S. Thiagalingam, W. S. el-Deiry, K. W. Kinzler *et al.*, 1994 Sequence-specific transcriptional activation is essential for growth suppression by p53. Proc. Natl. Acad. Sci. USA **91:** 1998–2002.

Ponder, B. A., 2001 Cancer genetics. Nature **411:** 336–341.

Rodin, S. N., and A. S. Rodin, 2000 Human lung cancer and p53: the interplay between mutagenesis and selection. Proc. Natl. Acad. Sci. USA **97:** 12244–12249.

Roemer, K., 1999 Mutant p53: gain-of-function oncoproteins and wild-type p53 inactivators. Biol. Chem. **380:** 879–887.

Slatkin, M., and B. Rannala, 1997 The sampling distribution of disease-associated alleles. Genetics **147:** 1855–1861.

Soussi, T., and C. Beroud, 2001 Assessing TP53 status in human tumours to evaluate clinical outcome. Nat. Rev. Cancer **1:** 233–240.

Soussi, T., C. Caron de Fromentel and P. May, 1990 Structural aspects of the p53 protein in relation to gene evolution. Oncogene **5:** 945–952.

Taylor, H. M., and S. Karlin, 1984 *An Introduction to Stochastic Modeling*. Academic Press, New York.

Venkitaraman, A. R., 2002 Cancer susceptibility and the functions of BRCA1 and BRCA2. Cell **108:** 171–182.

Vogelstein, B., E. R. Fearon, S. R. Hamilton, S. E. Kern, A. C. Preisinger *et al.*, 1988 Genetic alterations during colorectal tumor development. N. Engl. J. Med. **319:** 525–532.

Walker, D. R., J. P. Bond, R. E. Tarone, C. C. Harris, W. Makalowski *et al.*, 1999 Evolutionary conservation and somatic mutation hotspot maps of p53: correlation with p53 protein structural and functional features. Oncogene **18:** 211–218.

Waterman, J. L., J. L. Shenk and T. D. Halazonetis, 1995 The dihedral symmetry of the p53 tetramerization domain mandates a conformational switch upon DNA binding. EMBO J. **14:** 512–519.

Weinberg, R. A., 1995 The retinoblastoma protein and cell cycle control. Cell **81:** 323–330.

Yang, Z., 2001 Adaptive molecular evolution, pp. 327–348 in *Handbook of Statistical Genetics*, edited by D. J. Balding, C. Cannings and M. Bishop. John Wiley & Sons, New York.

Communicating editor: S. Tavaré