Computational analysis of mutation spectra

Igor B. Rogozin, Vladimir N. Babenko, Luciano Milanesi and Youri I. Pavlov

Date received (in revised form): 27th June 2003

Igor B. Rogozin and Vladimir N. Babenko
are research fellows at the National Center for Biotechnology Information NLM/NIH (Bethesda, MD, USA) and senior research scientists at the Institute of Cytology and Genetics RAS (Novosibirsk, Russia).

Luciano Milanesi
is a group leader at the Istituto di Technologie Biomediche Avanzate CNR (Milano, Italy).

Youri I. Pavlov
is a research fellow at the National Institute of Environmental Health Sciences (Research Triangle Park, NC, USA).

Keywords: *hotspot, mutation spectra, classification, DNA sequence context, mutable motif, somatic hypermutation, correlation*

Abstract

Mutation frequencies vary along a nucleotide sequence, and nucleotide positions with an exceptionally high mutation frequency are called hotspots. Mutation hotspots in DNA often reflect intrinsic properties of the mutation process, such as the specificity with which mutagens interact with nucleic acids and the sequence-specificity of DNA repair/replication enzymes. They might also reflect structural and functional features of target protein or RNA sequences in which they occur. The determinants of mutation frequency and specificity are complex and there are many analytical methods for their study. This paper discusses computational approaches to analysing mutation spectra (distribution of mutations along the target genes) that include many detectable (mutable) positions. The following methods are reviewed: mutation hotspot prediction; pairwise and multiple comparisons of mutation spectra; derivation of a consensus sequence; and analysis of correlation between nucleotide sequence features and mutation spectra. Spectra of spontaneous and induced mutations are used for illustration of the complexities and pitfalls of such analyses. In general, the DNA sequence context of mutation hotspots is a fingerprint of interactions between DNA and DNA repair/replication/modification enzymes, and the analysis of hotspot context provides evidence of such interactions.

MUTATION SPECTRA AND MUTATION HOTSPOTS

The process of mutation is fundamental in biology, being an essential evolutionary factor that creates genetic variation. Understanding complex mechanisms by which mutations occur spontaneously or are induced by mutagens is an important goal of molecular biology. A mutation spectrum is a distribution of frequencies of every type of mutation along nucleotide sequences of a target gene. Mutations in target sequences are usually revealed by either phenotypic selection in experimental test systems or, in the case of disease-causing genes in humans, by clinical studies in which certain genes are sequenced in groups of patients and in control groups. Both the experimental test systems and the clinical studies rely on detectable (mutable) positions, which are sites where DNA sequence changes cause phenotypic changes. A standard representation of a mutation spectrum is a nucleotide sequence of a target gene with all changes detected put above this sequence. The base substitution mutation spectrum[1] in Figure 1(A) includes three principal elements: (1) the target sequence (Figure 1A; lower line of continuous DNA sequence); (2) the mutations in the target sequence (Figure 1A); and (3) a representation of all positions in the target sequence which can be detected using phenotypic selection (shown by lower case letters in Figure 1A). Mutations in DNA/RNA molecules are classified as point mutations, deletions/insertions, duplications, inversions and chromosomal rearrangements. Point mutations are subclassified as base pair substitutions, including transitions (purine (R) mutates to R or pyrimidine (Y) mutates to Y) and transversions (R mutates to Y or Y mutates to R), and +1 and −1 frameshifts (insertions and deletions of a single base pair). Complex mutations include combinations of several point mutations and are relatively rare.

Mutability varies significantly along

I. B. Rogozin,
NCBI/NLM/NIH,
8600 Rockville Pike,
Building 38A,
Bethesda, MD 20894, USA

Tel: +1 301 594 4271
Fax: +1 301 480 9241
E-mail: rogozin@ncbi.nlm.nih.gov

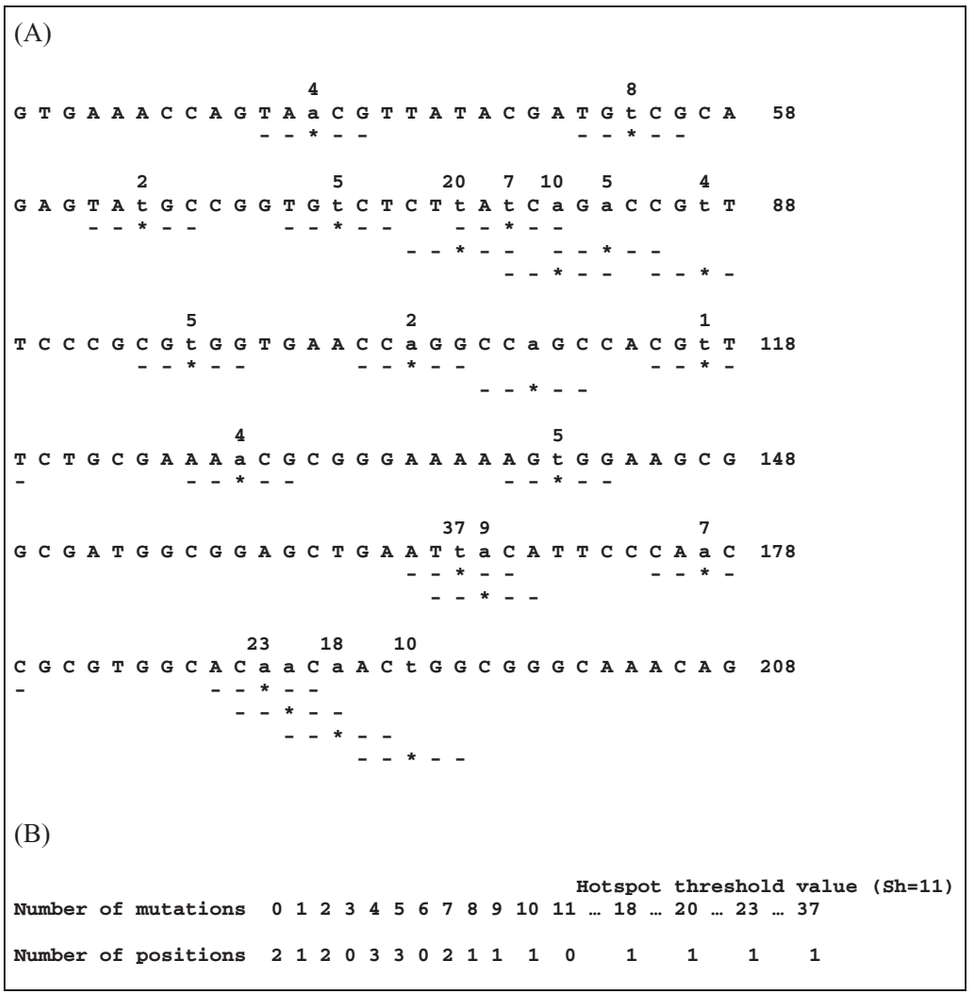

Figure 1: The sequence of the *lacI* gene with spontaneous A:T → C:G mutations revealed in the *mutT* strain of *E. coli*,[1] lower case letters denote the detectable positions, numbers above the sequence stand for the number of mutations at the position, detectable positions are asterisked, detectable sites are underlined (positions −2 to +2 are shown) (A); a distribution of observed mutation frequencies (B)

Mutation hotspots

nucleotide sequences: mutations, whether induced or spontaneous, occur at higher frequencies at certain positions of a nucleotide sequence (mutation 'hotspots').[2] Mutation hotspots often reflect a specific mechanism for generating mutations at a particular site and/or unusual properties of a phenotypic selection protocol. Thus, study of mutation hotspots can help reveal mutagenic mechanisms, or can reveal information about the functional domains of a target protein.[3–8] Some mutation hotspots are thought to depend on the nucleotide sequence and the mechanism of mutagenesis *per se*; these hotspots are called intrinsic mutation hotspots, however. In contrast, some hotspots may be due to expression and selection of a protein (RNA) molecule encoded by the target sequence,[5,6,8] for example hotspots in human *p53* might reflect both intrinsic mutability and selection for tumorgenesis.[6–9] This paper primarily discusses methods that are useful for analysis of intrinsic mutation hotspots.

NUCLEOTIDE CONTEXT OF MUTATION HOTSPOTS

Various DNA context features can influence mutagenesis

The examination of mutation hotspots provides evidence that they arise due to some structural features of hotspot subsequences (the local DNA sequence context of hotspots). There are several general DNA context features that can influence mutagenesis: homonucleotide runs, sites of potential Z-DNA formation, direct and inverted repeats, microsatellites, etc.[6,10–14]

Mutable motifs

In many cases, mutation hotspots emerge due to the influence of neighbouring nucleotides.[6,10,11,13,14] Examples of mutable DNA contexts are shown in Table 1. Sequence context effects can act over a significant distance: in one example, a single base pair change altered the mutation rate 12 bases away (8-fold effect on 2-aminopurine induced mutagenesis).[26] It was suggested that sequence context effects act as far as 80 bases away from a site of mutations.[27]

CpG hotspots

Local DNA sequence environment has been shown to be an important determinant of rates of base substitutions in human germinal cells. CpG is the usual context of mutation hotspots in human genes (Table 1), in which C•G → T•A mutations are thought to be the result of deamination of methylated cytosine.[13,15] A more complex pattern of hotspots was found in the human dystrophin gene, however: the frequency of CpG mutations was found to be lower than reported for other human genes, while the motif of TGRRGA (sometimes referred as DNA polymerase α reaction termination site) was found to be associated with >50 per cent of single-base mutations.[28] It was found that dipyrimidines that contain 5-methylcytosine are preferential targets for sunlight-induced (but not UVC-induced) mutagenesis in the methylated *lacI* transgene in cultured mammalian cells, which might be relevant to the observation of the large proportion of mutations in these sites among *p53* mutations found in skin tumours *in vivo*.[29] Observations of this kind provide clues for understanding of molecular mechanisms of mutations and require further experimental and computational analysis.[14,30]

Site-specific illegitimate recombination is associated with specific motifs recognised by specialised enzymes. A site-specific joining of immunoglobulin V-J and V-D-J segments is mediated by CACAGTG and ACAAAAACC sequences with a short spacer between them.[31] Interestingly, RAG1 and RAG2 proteins interacting with these sites might

Table 1: Examples of mutable motifs

Test system/mutagen	Mutable motif	Comments
Sn1-type alkylating agents, the *lacI* gene	R<u>G</u>	G<u>G</u> is more mutable than A<u>G</u>[11]
Spontaneous G•C → A•T mutations in mammalian genomes	<u>CG</u>	May result from the spontaneous deamination of 5-methylcytosine[15]
Somatic mutations in immunoglobulin V genes	R<u>G</u>YW	A<u>G</u>YW is more mutable than G<u>G</u>YW[16]
	W<u>A</u>	T<u>A</u> is more mutable than A<u>A</u>[17,18]
Hotspots of errors produced by DNA polymerase η	W<u>A</u>	*In vitro* experiment[19]
8-oxoG induced hotspots *in vitro* and *in vivo*	R<u>G</u>R	This motif was found to be mutable in some human genes[20]
UV-induced mutations in the phage lambda *cI* gene	<u>YY</u>	*In vivo* experiments[10]
Pyrimidine (6–4) pyrimidone photoproducts	YTCA	*In vitro* DNA damages induced by UV[21]
Cyclobutane dimers photoproducts	YTT	*In vitro* DNA damages induced by UV[21]
Single-base deletions	YTG	*In vitro* experiment[22]
Context of complex mutations in human disease genes	GTAAGT	Spontaneous mutations[23]
Target signal of retroposable elements	TTAAAA	LINEs and SINEs[24]
Signal of recombination in the *B.subtilis* mal gene	CATCGCTTRRT	Similar to gyrase binding sites[25]

Hotspot positions are underlined. R = A or G; Y = T or C; S = G or C; W = A or T; K = G or T; M = A or C; B = T, C or G; H = A, T or C; V = A, C or G; D = A, T or G.

be evolutionary descendants of proteins encoded by a mobile element, and the whole joining recombination system may be derived from a selfish repetitive element.[32] Incomplete collections of motifs which mediate site-specific recombination can be found in Badge *et al.*[33] and in a compilation of recombination signals and mutable motifs (Table 2).

Homonucleotide runs and microsatellites

Repetitive sequences

In 1966 Streisinger *et al.*[34] proposed that short deletions and insertions within runs of a same base (homonucleotide runs, homopolymeric tracts) arise by misalignment of DNA strands during replication. This type of misalignment can lead to length heterogeneity within homopolymeric tracts and more complex tandemly repeated structures called microsatellites (for example, runs of a di- and trinucleotides).[10,12,34–36] It was demonstrated that one base pair insertions and deletions are frequent in homonucleotide runs: the longer the run is, the higher is the mutation rate, which can be precisely explained by misalignment.[37] Variation of this mechanism, called dislocation mutagenesis, involves transient misalignment in homonucleotide runs (Figure 2) and may be responsible for some substitution hotspots. This mechanism was first found in the spectrum errors produced by DNA polymerase-β during *in vitro* DNA synthesis.[38] Dislocation mutagenesis causing substitution hotspots was suggested to play an important role *in vivo* at the control region of human mitochondrial DNA.[39]

Direct repeats

Short direct repeats have been long known to mediate deletions and duplications.[40,41] The mechanism could involve illegitimate recombination within short regions of similarity (<20 bases) or DNA polymerase slippage between repeated sequences. Recombination frequency increases with the length and GC content of repeated sequences and decreases with the length of the spacer between repeated sequences.[42] When a DNA sequence flanked by short direct repeats could form a palindrome, recombination between them occurs at a higher frequency.[43] In addition it was suggested that the repair of heteroduplexes formed by direct repeats may result in base substitutions and frameshifts. This mechanism was suggested for some classes of spontaneous mutations in bacterial and eukaryotic genes.[21,44]

Inverted repeats

Long inverted repeats (40–150 bases) are particularly unstable in bacterial DNA.[45] The proposed mechanism of deletions simulated by inverted repeats includes formation of hairpin structures in single

Table 2: List of available databases and integrated software

Databases	
Human Genome Mutation Database	http://www.uwcm.ac.uk/uwcm/mg/hgmd0.html
List of databases	http://ariel.ucs.unimelb.edu.au:80/~cotton/mdi.htm
List of databases	http://info.med.yale.edu/mutbase/
List of databases	http://darwin.ceh.uvic.ca/bigblue/bigblue.htm
Compilation of recombination signals and mutable motifs	ftp://ftp.bionet.nsc.ru/pub/biology/mutan/RECOMB.ZIP
Integrated software	
Integrated software for the analysis of mutations	http://sunsite.unc.edu/dnam/mainpage.html
UMD (Universal mutation database)	http://www.umd.necker.fr/
MutPlus software	http://www.cs.brown.edu/people/gq/mutplus_home.html

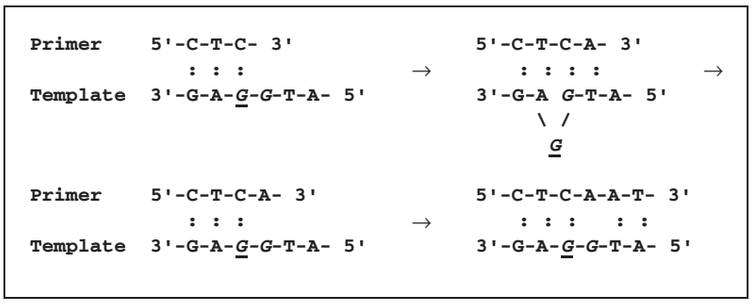

Figure 2: A model of the template strand dislocation. Dislocated bases are underlined, and homonucleotide runs are italicised

Z-DNA

strand DNA. Inverted repeats elevate intra- and interchromosomal recombination (reviewed by Gordenin and Resnick[12]).

It was proposed that correction of a quasipalindrome to a perfect inverted repeat occurred by either inter- or intramolecular strand switch. Many mutations of this type had been observed in bacteria, yeast and human cells.[36] This mechanism was suggested, in addition to direct repeat mechanism, to explain some classes of somatic mutations in immunoglobulin genes.[46,47] An example of hypothetical correction of the imperfect hairpin formed by the quasipalindrome is shown in Figure 3. Although a significant correlation between substitutions and direct and inverted repeats was detected in immunoglobulin genes,[48] any impact of this mechanism and relevance to base substitution hotspots is not clear.

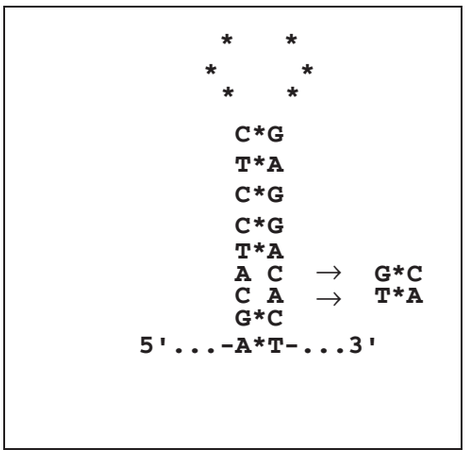

Figure 3: Hypothetical correction of the imperfect hairpin formed by the quasipalindrome[48]

Z-DNA

Several chemicals were found to react preferably with Z-DNA formed by regions of alternating purine–pyrimidine sequences. For example, hydroxylamine reacts very strongly with cytosines at junctions between B- and Z-DNA and in Z–Z-DNA junctions.[49] It was also suggested that repair is less effective in Z-DNA regions.[50]

Local mononucleotide composition

There are indications that GC-rich segments of DNA might be subject to more errors than AT-rich segments during replication.[51] A higher local AT content might increase the frequency of ultraviolet (UV) induced mutations in bacteria.[10]

DNA conformation and oligonucleotide content

Curved DNA is a sequence-directed curvature of the helix axis of double-stranded DNA, which is determined by the oligonucleotide content of a target sequence.[52] It was suggested that curved DNA has an important influence on various genomic rearrangements including deletions mediated by short direct repeats.[53,54] For bending and curvature calculation, the BEND_TRI program[55] can be used; it was shown that this program produces a good approximation of bending and curvature values.[56] Another example of possible involvement of oligonucleotide content in recombination was suggested by Konopka:[57] some trinucleotides are preferentially cleaved by topoisomerase I and thus may facilitate illegitimate recombination nearby.

Global factors

Many factors influence mutation frequency in a particular nucleotide sequence. In most cases, however, only local nucleotide sequence context was

Global factors

studied. It is likely that other higher-level features of gene or chromatin structure also have significant influence on mutation frequency of a mutable motif at a specific site. For example, AGTA is more mutable in complementarity-determined regions than in framework regions of immunoglobulin genes.[58] Another factor could be the rate of DNA repair. DNA repair rates vary for transcribed and non-transcribed strands of the same gene and for more and less highly-expressed genes.[59] Inherent asymmetry between the two DNA strands at the replication fork could also influence mutation frequency and specificity.[60,61] Other potential factors include asymmetric base composition[62] or higher-order chromatin structure (reviewed by Boulikas[49]).

METHODS FOR MUTATION SPECTRA ANALYSIS

Comparison of mutation spectra

Comparison of mutation spectra is the most common approach when studying two or more spectra induced differently (e.g. by different mutagens) in the same gene. Piegorsch and Bailer described statistical methods to compare two spectra based on an exact or pseudo-probability test (a Monte Carlo modification of the exact test).[63,64] HG-PUBL and COMP12 programs for such comparisons were developed (Table 3).

In a case of a multiple spectrum, two types of analysis can be done: a test of overall homogeneity, and pairwise comparisons between mutation spectra. There are a number of programs available:

Table 3: List of available methods and programs

Analysed features	Program	Method description	URL/comments
Single spectrum			
Mutation hotspot prediction	CLUSTERM	Decomposition of a mutation spectrum into several homogeneous site classes	http://www.itb.cnr.it/webmutation/ ftp://ftp.bionet.nsc.ru/pub/biology/dbms/CLUSTERM.ZIP
Context of mutation hotspots	–	Several approaches including heuristic and binomial test-based reconstructions of consensus sequences	Methods that rely on arbitrary discrimination between informative and non-informative positions may lead to unreliable results
Correlation between nucleotide sequence features and mutation spectra	–	The exact test can be used to test the null hypothesis that mutations are equally probable in mutable motifs and all other positions in the target sequence	The input numbers are MM, MA-MM, NP, NA-NP, where MM is the number of mutations in mutable motifs, NP is the number of such motifs, MA is the number of mutations and NA is the number of positions in the target sequence
Oligonucleotide composition	–	Correlation between the oligonucleotide content and mutation frequencies	This approach requires a large number of detectable sites
Context-free mutations features analysis	–	Substitutions frequency, mutations and hotspots clustering, periodicity	This kind of information is important for the understanding of molecular mechanisms of mutagenesis
Multiple spectra			
Pairwise spectra comparison	HG-PUBL	Test of heterogeneity based on $2 \times N$ contingency table	ftp://sunsite.unc.edu/pub/academic/biology/dna-mutations/hyperg
Pairwise spectra comparison	COMP12	Test of heterogeneity based on $2 \times N$ contingency table	ftp://ftp.bionet.nsc.ru/pub/biology/dbms/comp12.zip
Pairwise correlation analysis	CORR12	Correlation test based on the $2 \times N$ contingency table	ftp://ftp.bionet.nsc.ru/pub/biology/dbms/CORR12.ZIP
Multiple spectra comparison	COLLAPSE	Test of heterogeneity based on $R \times M$ contingency table	ftp://ftp.bionet.nsc.ru/pub/biology/dbms/COLLAPSE.ZIP
Multiple spectra comparison	ARLEQUIN	Test of heterogeneity based on $R \times M$ contingency table	http://anthropologie.unige.ch/arlequin.htm
Multiple spectra comparison	GENEPOP	Test of heterogeneity based on $R \times M$ contingency table	ftp://ftp.cefe.cnrs-mop.fr/pub/pc/msdos/genepop

Tests of homogeneity

ARLEQUIN and GENEPOP (Table 3) provide the exact test of the overall homogeneity in $n \times T$ matrices and the exact tests for the pairwise comparison between spectra as well. The COLLAPSE program[65] (Table 3) allows collapsing (grouping, combining) of rows and columns of $n \times T$ matrices. It calculates expected frequencies, squared standardised residuals as well as Zelterman and Cressie-Read statistics, which were recommended for the analysis of mutation spectra when data are sparse.[64] A new analytical strategy for mutation spectra comparisons was suggested recently, this approach is also useful for hotspot prediction and analysis.[66] This strategy is based on comparison of mutation frequencies in each site of two studied spectra.[67] The problem of pairwise and multiple spectra comparisons was discussed by Piegorsch and Bailer,[64] Khromov-Borisov *et al.*,[65] Rogozin *et al.*[6] and Lewis and Parry.[67]

Correlation analysis

The Kendall's tau correlation coefficient can be used as a complementary approach.[6,68] If two mutation spectra are not significantly different, they may be assumed to be significantly similar only if a significant correlation is found between these two spectra, as shown by analysis with the CORR12 program (Table 3).[6,68] The simultaneous application of exact and correlation tests is complementary, underlining different factors affecting the structures of mutation spectra. For example, Babenko and Rogozin[68] identified a spectrum for which both tests produced statistically significant results. Based on this observation, two context factors affecting the spectrum in different ways have been determined.

Multiple linear regression analysis was successfully applied for comparison of three mutation spectra in the same mouse immunoglobulin target sequence.[69]

Revealing hotspots

Hotspot prediction

A mutation spectrum can be transformed into a distribution of observed mutation frequencies along a gene sequence (Figure 1B) which should be regarded as a sample of multinomial distribution (discussed by Piegorsch and Bailer[64]). Analysis of such distribution can be performed using a simulation, expectation, maximisation (SEM) classification approach.[70] A general principle of mutation hotspot prediction in this approach is based on a threshold (Sh) value for the number of mutations in a mutable site. All sites with the number of mutations greater than or equal to Sh are defined as hotspots (Figure 1B). The threshold and resulting hotspot sites are defined for each mutation spectrum separately based on results of classification analysis.[70] The CLUSTERM program (Table 3) is used for this purpose. This program decomposes a mutation spectrum into several homogeneous classes of sites. Each class is approximated by a binomial (or Poisson) distribution. Variations in mutation frequencies among sites of the same class are due to random reasons, since mutation probability is the same for all sites in one class. Differences between mutation frequencies among sites from different classes are statistically significant, however. A class (or classes) with the highest mutation frequency is called a hotspot class(es). Problems of hotspot prediction were discussed by Rogozin *et al.*[6] and Fijal *et al.*[71].

Analysis of neighbouring bases

Nucleotide sequence context influences mutation probability.[2,5,11,13,15,16] Several methods have been developed to analyse this phenomenon. A commonly used approach for analysis of neighbouring bases is to calculate the number of times a given base occurs next to a mutated base, immediately in the 5′ or 3′ direction (positions −1 and +1). A significant deviation from the expected numbers can be estimated by using various statistical tests.[5,19,72,73] Krawczak *et al.*[13] analysed nearest-neighbour effects with correction for codon usage and for different probability of detecting different amino acid substitutions in a clinical study, which may be useful in studying human disease susceptibility genes. Maximum

likelihood estimates of nearest-neighbour effects were developed by Zavolan and Kepler.[5]

Consensus sequences

A set of aligned hotspot sites (Figure 4) can be analysed separately to derive a consensus sequence[74] using one of several available approaches.[75] Methods that rely on arbitrary discrimination between informative and non-informative positions may lead to controversial and/or unreliable results. Simple consensus sequences can be misleading especially when the data set is small; however, they can be reconstructed using any mutation spectrum and any subset of positions. The binomial test can also be used to study consensus sequences at or near mutation hotspots.[16,39] In this method, a number (N_{IJ}) of a nucleotide I is calculated in each position J in a set of M aligned mutation hotspot sequences (Table 3). The probability $P(N_{IJ},M,F_I)$ of finding N_{IJ} or more nucleotides I in a position J is calculated taking a frequency F_I of a nucleotide I in a target sequence as an expected number of the nucleotide I in the position J. A nucleotide with the lowest probability $P(N_{IJ},M,F_I)$ among all possible nucleotides in a position J is accepted as a consensus nucleotide for this position if $P(N_{IJ},M,F_I)$ for this nucleotide is below a threshold value P^*. It is important to note that the estimate of $P(N_{IJ},M,F_I)$ cannot be used for rejecting or accepting a statistical hypothesis owing to a multiplicity of binomial tests; moreover these tests are strongly inter-dependent for each position. To estimate a significance level for $P(N_{IJ},M,F_I)$, Malyarchuk *et al.*[39] developed a resampling procedure. In this procedure, M sites were randomly chosen from a target sequence. Thus, each 'random' sample was a mixture of hotspots and non-hotspots. The statistical analysis described above was repeated for each sample, and the minimal value $P_{mr}(N_{IJ},M,F_I)$ was calculated for all positions. This procedure was repeated 10,000 times to calculate a significance level P^* that separates the right critical region of the distribution $P_{mr}(N_{IJ},M,F_I)$ at 5 per cent level of significance, P^* may be significantly less than 0.05 (for example, $P = 0.005$ for the HVS1 spectrum).[39]

Position	Site sequence	Number of mutations	
	- - * - -		
77	A T a A G	20	
167	G T a A T	37	Hotspot sites
189	A C a A C	23	
192	A C a A C	18	

	N N a A N		
41	T A a C G	4	
54	C G a C A	8	
64	G C a T A	2	
72	A G a C A	5	
79	T G a T A	7	
81	T C a G A	10	
83	A G a C C	5	
87	A A a C G	4	
96	C C a C G	5	
105	C C a G G	2	
110	C C a G C	0	
117	A A a C G	1	
128	A A a C G	4	
141	C C a C T	5	
168	T T a C A	9	
177	C A a C C	7	
190	C A a C A	0	
195	C C a G T	10	

	N N a B N		

Figure 4: An alignment of putative hotspot and non-hotspot sites for the mutation spectrum described in Figure 1. The detectable position is displayed as a purine, using the appropriate DNA strand

Correlation between nucleotide sequence features and mutation spectra

In cases when a context factor influencing the frequency of mutations is known *a priori*, a correlation between distributions of this factor (for example, a mutable motif) and mutations along a target sequence may be measured. This approach is discussed here with regard to analysis of multiple somatic mutations in the 5′ flanking region of *BCL-6* (Figure 5), a proto-oncogene encoding for a transcriptional repressor from B-cell lymphomas.[76] The RGYW mutable

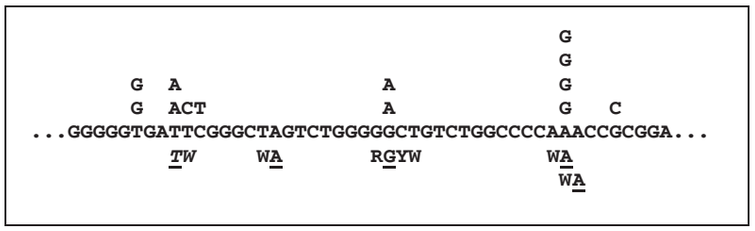

Figure 5: Somatic mutation spectrum in the human *BCL-6* gene.[76] Hotspots in the RGYW, WA and complementary variant TW motifs (Table 1) are shown under the primary nucleotide sequence of the gene. Sequence motifs that represent hotspot consensus sequences in the complementary strand are in italics

Targeting of mutations to mutable motifs

motif (the hotspot base is underlined) is a signature of somatic hypermutation (Table 1), and may play a role in somatic mutations observed in *BCL-6*. The Fisher exact test for analyses of 2×2 tables can be used to test the null hypothesis that mutations are equally probable in mutable motifs and all other positions of the target sequence.[14] When one of the numbers is large (>300), the χ^2 test with one degree of freedom can be applied instead of the exact test.[14] In the case of mutations in *BCL-6*[76] (Fig. 5), the correlation is statistically significant ($P(\chi^2) < 0.05$).

Statistical significance can be also estimated using a modified Monte Carlo procedure.[16] This approach takes into account frequency of mutations in A, T, G and C bases, the presence of several mutations in a site and context properties of the target sequence. Weight Wj of site j was defined as the number of mutations in a mutable motif; however, more complex definitions were also used (discussed below). A distribution of statistical weights W_{random} was calculated for 10,000 computationally generated groups of random sites. Each group contained the observed number of mutations distributed similarly in all sites. The distribution in W_{random} was used to calculate probability $P(W \leqslant W_{\text{random}})$. This probability is equal to the number of groups of random mutations in which W_{random} is the same or higher than W. Small probability values ($P(W \leqslant W_{\text{random}}) \leqslant 0.05$) indicate a significant correlation between mutable motif and mutation frequency. Modified versions of this approach were used to analyse a dislocation model in human mitochondrial DNA,[39] mutability of direct and inverted repeats in immunoglobulin genes[48] and gene conversion in immunoglobulin genes.[77] A similar approach was used to analyse illegitimate recombination events.[78] Theoretically, this method can be applied to any data where a reliable correlation measure (the weight W_j) between mutation/recombination events and the nucleotide sequence context can be derived.

Other approaches

Various multiple regression models can be used for simultaneous analysis of how several neighbouring positions influence mutation frequency.[16,79,80] Nucleotide sequence context of mutation hotspots can also be analysed by focusing on oligonucleotides. For example, Smith *et al.*[81] analysed the relative frequency of somatic mutations in 16 dinucleotide and 64 trinucleotide motifs. This approach revealed that the mutation frequencies in different di- or trinucleotides were significantly different.[18,81–83] A meta-analysis of complex mutations causing human genetic diseases revealed a hotspot motif GTAAGT over-represented in the vicinity of these mutations.[23] A local oligonucleotide composition is also the focus of studies on frameshift mutations in microsatellites. These mutation hotspots are affected by length and base composition of the microsatellite repeat. In general, both approaches require a large number of detectable sites (eg hundreds of sites for trinucleotide motifs) and mutations in a target sequence.

Several aspects of a mutation spectrum, including frequency of substitutions, clustering of mutations and hotspots, and periodicity of mutation can be considered as 'context-free' characteristics of the spectrum. These approaches were discussed by Rogozin and Pavlov.[14]

In general, various computational approaches can be used to analyse aligned

sequences of mutation hotspots. Many techniques have been developed for analysis of functional signals including information content, weight matrices, perceptron, *k*-tuple frequencies, discriminant analysis, hidden Markov models, linguistic approaches and neural network models.[84–86] These methods are well established and have been tested on different types of data, but all of these methods require large data sets.

Mutation databases and integrated software

Databases

There are many mutation databases freely available via the internet.[87,88] It is relatively easy to find various databases of mutation spectra on the internet, starting from the web servers listed in Table 2. Various information about mutation databases are regularly published in the first January issue of the journal *Nucleic Acids Research*; another valuable source of such information is the journal *Human Mutation*. Database formats vary significantly among databases, and a lot of work is required to provide and perform their analysis. Designing standard formats of mutation databases can help a lot in development and application of computational tools. An example of such an initiative is the Human Gene Mutation Database (HGMD). HGMD constitutes a comprehensive core collection of data on germ-line mutations in nuclear genes associated with human inherited diseases (Table 2). Each mutation is entered into HGMD only once in order to avoid confusion between recurrent and identical-by-descent lesions (discussed in the 'Unsolved problems, perspectives and conclusions' section).

Databases and software applications for the analysis of DNA mutations at the human *p53* gene, the human *hprt* gene and both the rodent transgenic *lacI* and *lacZ* loci have been developed.[73] The databases are stand-alone dBASE files, and the software for analysis of these databases runs on PCs (Table 2). A piece of genetic software called the Universal Mutation Database (UMD, Table 2) allows development of locus-specific databases. The software includes an optimised structure to assist data entry and allows the input of various clinical data.[89] A MutPlus program (Table 2) is another attempt to integrate different analytical tools and mutation data sets.

EXAMPLES OF MUTATION SPECTRA ANALYSIS

Mutation spectra induced by ultraviolet light

UV-induced mutations

Ultraviolet (UV) light generates a complex spectrum of mutations including base substitutions, frameshifts, complex mutations, large deletions and duplications.[10,90–94] Mutation spectra of UV light were studied in numerous experimental systems. These studies were conducted most frequently with short-wave UV in the range 180–280 nm (UVC), however, and may have somewhat limited relevance to the effects of sunlight itself, which is a mixture of wavelengths.[30] Principal damage to DNA leading to UV-induced mutations are pyrimidine–pyrimidine cyclobutane dimers and pyrimidine (6–4) pyrimidone photoproducts.[95,96] These lesions block replicative DNA polymerases and should be either repaired or bypassed by specialised mechanisms to escape cell death, mutagenesis and cancer.

Examining the context specificity of UV mutagenesis helps to determine which photoproducts are important premutational lesions, since neighbouring DNA sequence and the resulting local DNA conformation play a critical role in formation of these photoproducts.[10,30] In bacteria, TC, TCC, CC and CCC are hotspots of UV-induced mutations; however, both cyclobutane dimers and pyrimidine (6–4) pyrimidone photoproducts are frequent in these mutable motifs.[10,30] Early analysis of UVC-induced mutations did not reveal transitions in CT dinucleotides, suggesting that pyrimidine (6–4) pyrimidone photoproducts rather than cyclobutane dimers are primarily responsible for targeting mutations in

bacteria, since pyrimidine (6–4) pyrimidone photoproducts are rare at CT sequences, whereas cyclobutane dimers are quite abundant.[10,95,96]

Damage bypass DNA polymerases

Later it was shown that pyrimidine (6–4) pyrimidone photoproducts are more mutagenic than cyclobutane dimers in yeast and mammalian cells.[97] It was not clear, however, why TT dinucleotides appear as frequent sites of transitions in *E. coli*, since pyrimidine (6–4) pyrimidone photoproduct formation is sharply reduced at TT dinucleotides relative to TC and CC sequences.[10] Since then it has been discovered that the same UV photolesions have different mutagenic potential in different organisms, for example bacteria and yeast.[97] The complexity of UV-induced spectra may be illustrated using results of a correlation analysis between photoproduct hotspot consensus sequences and UV-induced mutation spectra (Table 4).[21]

Alkylating agents

Apparent discrepancies between different experiments (Table 4) are not surprising taking into account that various repair/replication enzymes involved in UV mutagenesis and final UV-induced mutations in different species[30,97–99] and the precise mechanisms of interspecies differences were not fully understood until recently, when damage bypass DNA polymerases were discovered. It was found that the superfamily Y of DNA polymerases with relaxed active centre exists to deal with UV- and chemically-induced DNA adducts. Translesion synthesis is a two-step process of insertion of bases opposite damaged site and further extension, that may require several DNA polymerases — sometimes belonging to different superfamilies.[98–100] Mutagenesis depends on a lesion bypass event, including incorporation opposite the damage and extension from the thus formed DNA 3′ end. The family Y DNA polymerases are much less accurate than replicative DNA polymerases;[19] however, even low fidelity of incorporation would substantially reduce the probability of mutation generation when replicating opposite the damage. The difference in mutagenic potential of photoproducts in different experiments (Table 4) may be easily explained by the involvement of polymerases with different properties (DNA polymerase V and DNA polymerase η, respectively) in this initial bypass event.[98–100]

Mutations induced by alkylating agents

SN1 alkylating agents preferentially generate G•C → A•T transitions at RG sites, while SN2 agents do not.[11] A comparative analysis of mutation spectra induced by SN2 alkylating agents using regression analysis, revealed various mutable motifs specific for SN2 alkylating agents.[101] For the mutation spectra induced by *N*-methyl-*N*-nitrosourea (MNU) in the *lacI* and *gpt* genes[102,103] different mutable motifs were revealed:

Table 4: A non-random correlation ($P(W < W_{random}) < 0.05$) between hotspots of UV-induced substitutions and consensus sequences of two major UV-induced photoproducts

Gene	Species	Comments	cyclobutane dimers	Py(6–4)Pyo photoproduct
lacI	E. coli	Wild type[91]	+	+
lacI	E. coli	UvrB-strain[91]	+	+
SUP4-o	Yeast	UVB-induced[93]	+	
SUP4-o	Yeast	UVC-induced[93]	+	
supF	Monkey	Cell line[90]		+
gpt	Hamster	CHO cells[92]		+
supF	Human	XPV cell line[94]	+	

Consensus sequences for pyrimidine (6–4) pyrimidone [Py(6–4)Pyo] and cyclobutane dimers photoproducts are YTCA and YTT, respectively,[21] consensus sequences were derived using regression analysis of pyrimidine (6–4) pyrimidone[93] and cyclobutane dimers photoproduct spectra.[95,96] The significance of correlations between the distribution of mutable motifs and mutations along a target sequence was estimated using a Monte Carlo approach.[16]

three mutable motifs GGC, AGNG and GGKM for the *lacI* spectrum and one mutable motif GG for the *gpt* spectrum.[101] The correlation between the former set of motifs and distribution of MNU-induced mutations in the *gpt* gene, however, was statistically insignificant[101] — a Monte Carlo test[16] (described above) was used. Sets of mutable motifs thus revealed were ambiguous and cannot be recommended as a signature of MNU-induced mutagenesis. In general, the analysis suffered from the small number of detectable positions (22 G•C → A•T detectable sites) and the small number of mutations.

Somatic mutations in immunoglobulin genes

Mutable motifs suggested a two-step mechanism of somatic hypermutation

The wide variety of immunoglobulins in vertebrates results from the combinatorial joining of different variable (V), diversity (D) and joining (J) gene segments to create the primary antigen–receptor repertoire, followed by somatic hypermutation of variable (V) regions. These mutations are introduced at a rate estimated to be about six orders of magnitude greater than the normal rate of spontaneous mutations in the genome.[31] A number of different models of somatic hypermutation have been proposed.[104] Most models postulate involvement of mutator polymerases to account for the high frequency of mutagenesis in V regions. One important feature of somatic hypermutation in V regions is the non-random distribution of mutations. Somatic mutation hotspots in V regions occur primarily within two DNA sequence motifs. RGYW hotspots[16,17] are found in both strands and WA hotspots preferentially are found in only one strand.[16–18,82] A candidate for a principal RGYW mutator is activation-induced cytidine deaminase (AID), which converts cytosine in DNA into uracil.[105–107]

Sample size is a major problem in analysis

Analysis of mutation spectra of errors made by various DNA polymerases during *in vitro* DNA synthesis provided clues on which polymerase could operate during somatic hypermutation. A correlation between the WA motif and the error specificity of human DNA polymerase η[17,108] and lack of A–T mutations in XP-V patients deficient in DNA polymerase η,[109] suggested that this polymerase may contribute to the WA hotspots. Additional analysis of this correlation using the same mouse immunoglobulin target sequence for *in vivo* and *in vitro* spectrum generation, combined with studies of mutable motifs and frequencies of substitutions, greatly improved the power of comparisons, allowing use of different statistical methods.[69] It was found that two DNA polymerase η error spectra, determined while this polymerase synthesises the transcribed or non-transcribed strands, correlate in a mosaic fashion with a spectrum of somatic mutations *in vivo*. This suggested that this polymerase contributes to somatic hypermutation in mice during short patch DNA synthesis on alternating DNA strands.[69]

In general, analysis of mutable motifs became a standard procedure in studies of somatic hypermutation of immunoglobulin genes.

UNSOLVED PROBLEMS, PERSPECTIVES AND CONCLUSIONS

Many factors influence mutation frequency in a particular nucleotide sequence. In most cases, however, analytical methods only attempt to characterise factors related to local nucleotide sequence context. It is likely that other higher-level features of gene or chromatin structure also have significant influence on mutation frequency of a mutable motif at a specific site. Analyses of these features require large data sets obtained under different experimental conditions; small sample size is a major problem in analysis of mutation spectra. Even if the number of mutations is large, the number of mutation hotspots is likely to be small (eg Figure 4). A few approaches can be robust with small data sets (ie hotspot prediction, comparing mutation spectra, correlation between nucleotide sequence features and

mutation spectra). Other methods may not be reliable when applied to small data sets.

Mutation assays

It is very important to identify the detectable positions in a target sequence before analysing its mutation spectrum. This can be achieved by analysing large numbers of mutants[14,19,37] or by systematic site-directed mutagenesis of all amino acids in the target gene.[3,6] Non-phenotypic assays are rarely used since they are usually restricted to a few specific positions (e.g. using sites of restriction enzymes to detect mutations) or they require a very high frequency of mutation.[14,19] In some tests a list of detectable positions is known *a priori*. This is the case for mutation spectra in amber (resulting in UAG stop-codons) and ochre (resulting in UAA stop-codons) nonsense sites. All such nonsense sites can be easily found in protein coding sequences. Unfortunately, for many other spectra the list of detectable positions is the most problematic component of a mutation spectrum.[6] Population polymorphism becomes another important issue when the mutated sequences from one individual are compared with non-mutated sequences from another individual (such an approach is used sometimes for studies of somatic mutations in immunoglobulin genes). In such cases, each polymorphic position will be counted as a mutation, which biases mutation spectra. It is possible to misassign a functional mutation at a specific site even if a data set is carefully collected. This can occur in cases of multiple mutations when an unidentified distal mutation alters gene function, and the mutation in the assigned site does not have a functional effect. Thus, only well-characterised detectable sites, in which several independent mutations have been observed, should be used when a mutation spectrum is analysed.

Another problem of mutation spectra analysis, which is restricted only to the case of locus specific mutation databases, is the problem of repetitions (ie mutations that are identical by descent). Such repetitions should be counted as a single mutation. It is not always possible to detect them, however.[110–112] In general, mutation spectra revealed by clinical studies do not represent random samples of all arising mutations. Rather, the vast majority of them are middling to highly deleterious mutations, whereas advantageous, neutral or slightly deleterious mutations are hardly or not at all represented. This is because for a mutation to be represented in a database, it was to come to clinical attention, otherwise it remains undetectable. The same complication applies to phenotypic selection systems. The non-random sampling of mutations may systematically bias context properties of detected mutation hotspots.[14,110–112]

Hotspot context and molecular mechanisms of mutagenesis

It should be emphasised that the context of hotspots may be very helpful in deep understanding of underlying molecular mechanisms of mutagenesis;[6,10,11,14,30] however, the determinants of mutation frequency and specificity are complex and there are many analytical methods for their study. The most reliable results can be obtained if several methods are combined or used sequentially and if many different sources of information are considered. Simple, and thus robust, approaches should be used with small mutation samples, while combinations of simple and complex approaches can be used for large samples. Complex approaches are needed because mutation spectra reflect the simultaneous influence of multiple diverse local and global factors. It is a challenging task to analyse mutation spectra and, in some cases, the effort will be primarily descriptive in nature. In several well-documented studies, however, the analysis of mutation spectra has contributed substantially to understanding molecular mechanisms of mutagenesis. As analytical methods continue to be developed, more theoretical and experimental studies will contribute insights into the complex process of mutagenesis.

Acknowledgments
We thank B. A. Rogozin, N. A. Kolchanov, E. A. Vasunina and N. N. Khromov-Borisov for helpful discussions and G. V. Glazko and anonymous referees for helpful comments on the manuscript. This work was partially supported by the Russian Foundation of Basic Research (grants 96-04-49957, 99-04-49535 and 02-04-48342), the EC grant IST 2001 32688 'ORIEL', 'Functional Genomics', 'Molecular Genomics' MIUR CNR and CISI Projects.

References

1. Fowler, R. G. and Schaaper, R. M. (1997), 'The role of the *mutT* gene of *Escherichia coli* in maintaining replication fidelity', *FEMS Microbiol. Rev.*, Vol. 21, pp. 43–54.
2. Benzer, S. (1961), 'On the topography of the genetic fine structure', *Proc. Natl. Acad. Sci. USA*, Vol. 47, pp. 403–415.
3. Suckow, J., Markiewicz, P., Kleina, L. G. *et al.* (1996), 'Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure', *J. Mol. Biol.*, Vol. 261, pp. 509–523.
4. Walker, D. R., Bond, J. P., Tarone, R. E. *et al.* (1999), 'Evolutionary conservation and somatic mutation hotspot maps of *p53*: correlation with *p53* protein structural and functional features', *Oncogene*, Vol. 18, pp. 211–218.
5. Zavolan, M. and Kepler, T. B. (2001), 'Statistical inference of sequence-dependent mutation rates', *Curr. Opin. Genet. Dev.*, Vol. 11, pp. 612–615.
6. Rogozin, I. B., Kondrashov, F. A. and Glazko, G. V. (2001), 'Use of mutation spectra analysis software', *Human Mut.*, Vol. 17, pp. 83–102.
7. Wacey, A. I., Cooper, D. N., Liney, D. *et al.* (1999), 'Disentangling the perturbational effects of amino acid substitutions in the DNA-binding domain of *p53*', *Human Genet.*, Vol. 104, pp. 15–22.
8. Betz, A. G., Neuberger, M. S. and Milstein, C. (1993), 'Discriminating intrinsic and antigen-selected mutational hotspots in immunoglobulin V genes', *Immunol. Today*, Vol. 14, pp. 405–411.
9. Krawczak, M., Smith-Sorensen, B., Schmidtke, J. *et al.* (1995), 'Somatic spectrum of cancer-associated single basepair substitutions in the *TP53* gene is determined mainly by endogenous mechanisms of mutation and by selection', *Human Mutat.*, Vol. 5, pp. 48–57.
10. Miller, J. H. (1983), 'Mutational specificity in bacteria', *Annu. Rev. Genet.*, Vol. 17, pp. 215–238.
11. Horsfall, M. J., Gordon, A. J., Burns, P. A. *et al.* (1990), 'Mutational specificity of alkylating agents and the influence of DNA repair', *Environ. Mol. Mutagen.*, Vol. 15, pp. 107–122.
12. Gordenin, D. A. and Resnick, M. A. (1998), 'Yeast ARMs (DNA at-risk motifs) can reveal sources of genome instability', *Mutat. Res.*, Vol. 400, pp. 45–58.
13. Krawczak, M., Ball, E. V. and Cooper, D. N. (1998), 'Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes', *Amer. J. Hum. Genet.*, Vol. 63, pp. 474–488.
14. Rogozin, I. B. and Pavlov, Y. I. (2003), 'Theoretical analysis of mutation hotspots and their DNA sequence context specificity', *Mutat. Res.*, Vol. 544(1), 65–85.
15. Cooper, D. N. and Youssoufian, H. (1988), 'The CpG dinucleotide and human genetic disease', *Human Genet.*, Vol. 78, pp. 151–155.
16. Rogozin, I. B. and Kolchanov, N. A. (1992), 'Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis', *Biochim. Biophys. Acta*, Vol. 1171, pp. 11–18.
17. Rogozin, I. B., Pavlov, Y. I., Bebenek, K. *et al.* (2001), 'Somatic mutation hotspots correlate with DNA polymerase eta error spectrum', *Nat. Immunol.*, Vol. 2, pp. 530–536.
18. Oprea, M., Cowell, L. G. and Kepler, T. B. (2001), 'The targeting of somatic hypermutation closely resembles that of meiotic mutation', *J. Immunol.*, Vol. 166, pp. 892–899.
19. Matsuda, T., Bebenek, K., Masutani, C. *et al.* (2001), 'Error rate and specificity of human and murine DNA polymerase η', *J. Mol. Biol.*, Vol. 312, pp. 335–346.
20. Watanabe, T., Nunoshiba, T., Kawata, M. and Yamamoto, K. (2001), 'An *in vivo* approach to identifying sequence context of 8-oxoguanine mutagenesis', *Biochem. Biophys. Res. Commun.*, Vol. 284, pp. 179–184.
21. Kolchanov, N. A. and Rogozin, I. B. (1994) 'Contribution of nucleotide context to spontaneous and induced mutations', in Kolchanov, N. A. and Lim, H. A. (Eds.), 'Computer Analysis of Genetic Macromolecules', World Scientific, Singapore, pp. 278–288.
22. Papanicolaou, C. and Ripley, L. S. (1989), 'Polymerase-specific differences in the DNA intermediates of frameshift mutagenesis. *In vitro* synthesis errors of *Escherichia coli* DNA polymerase I and its large fragment

derivative', *J. Mol. Biol.*, Vol. 207, pp. 335–353.

23. Chuzhanova, N. A., Anassis, E. J., Ball, E. V. *et al.* (2003), 'Meta-analysis of indels causing human genetic disease: Mechanisms of mutagenesis and the role of local DNA sequence complexity', *Human Mutat.*, Vol. 21, pp. 28–44.
24. Jurka, J. and Klonowski, P. (1996), 'Integration of retroposable elements in mammals: Selection of target sites', *J. Mol. Evol.*, Vol. 43, pp. 685–689.
25. Lopez, P., Espinosa, M., Greenberg, B. and Lacks, S. A. (1984), 'Generation of deletions in pneumococcal *mal* genes cloned in *Bacillus subtilis*', *Proc. Natl Acad. Sci. USA*, Vol. 81, pp. 5189–5193.
26. Sugino, A. and Drake, J. W. (1984), 'Modulation of mutation rates in bacteriophage T4 by a base-pair change a dozen nucleotides removed', *J. Mol. Biol.*, Vol. 176, pp. 239–249.
27. Canella, K. A. and Seidman, M. M. (2000), 'Mutation spectra in supF: Approaches to elucidating sequence context effects', *Mutat. Res.*, Vol. 450, pp. 61–73.
28. Todorova, A. and Danieli, G. A. (1997), 'Large majority of single-nucleotide mutations along the dystrophin gene can be explained by more than one mechanism of mutagenesis', *Human Mutat.*, Vol. 9, pp. 537–547.
29. You, Y. H., Li, C. and Pfeifer, G. P. (1999), 'Involvement of 5-methylcytosine in sunlight-induced mutagenesis', *J. Mol. Biol.*, Vol. 293, pp. 493–503.
30. Dogliotti, E., Hainaut, P., Hernandez, T. *et al.* (1998), 'Mutation spectra resulting from carcinogenic exposure: From model systems to cancer-related genes', *Recent Results Cancer. Res.*, Vol. 154, pp. 97–124.
31. Tonegawa, S. (1983), 'Somatic generation of antibody diversity', *Nature*, Vol. 302, pp. 575–581.
32. Bowen, N. J. and Jordan, I. K. (2002), 'Transposable elements and the evolution of eukaryotic complexity', *Curr. Issues Mol. Biol.*, Vol. 4, pp. 65–76.
33. Badge, R. M., Yardley, J., Jeffreys, A. J. and Armour, J. A. (2000), 'Crossover breakpoint mapping identifies a subtelomeric hotspot for male meiotic recombination', *Human Mol. Genet.*, Vol. 9, pp. 1239–1244.
34. Streisinger, G., Okada, Y., Emrich, J. *et al.* (1966), 'Frameshift mutations and the genetic code', *Cold Spring Harbor Symp. Quant. Biol.*, Vol. 31, pp. 77–84.
35. Strauss, B. S. (1999), 'Frameshift mutation, microsatellites and mismatch repair', *Mutat. Res.*, Vol. 437, pp. 195–203.
36. Ripley, L. S. (1990), 'Frameshift mutation: Determinants of specificity', *Annu. Rev. Genet.*, Vol. 24, pp. 189–213.
37. Kunkel, T. A. and Bebenek, K. (2000), 'DNA replication fidelity', *Annu. Rev. Biochem.*, Vol. 69, pp. 497–529.
38. Kunkel, T. A. (1985), 'The mutational specificity of DNA polymerase-β during *in vitro* DNA synthesis. Production of frameshift, base substitution, and deletion mutations', *J. Biol. Chem.*, Vol. 260, pp. 5787–5796.
39. Malyarchuk, B. A., Rogozin, I. B., Berikov, V. B. and Derenko, M. V. (2002), 'Analysis of phylogenetically reconstructed mutational spectra in human mitochondrial DNA control region', *Human Genet.*, Vol. 111, pp. 46–53.
40. Albertini, A. M., Hofer, M., Calos, M. P. and Miller, J. H. (1982), 'On the formation of spontaneous deletions: The importance of short sequence homologies in the generation of large deletions', *Cell*, Vol. 29, pp. 319–328.
41. Singer, B. S. and Westlye, J. (1988), 'Deletion formation in bacteriophage T4', *J. Mol. Biol.*, Vol. 202, pp. 233–243.
42. Chedin, F., Dervyn, E., Dervyn, R. *et al.* (1994), 'Frequency of deletion formation decreases exponentially with distance between short direct repeats', *Mol. Microbiol.*, Vol. 12, pp. 561–569.
43. Foster, T. J., Lundblad, V., Hanley-Way, S. *et al.* (1981), 'Three Tn10-associated excision events: Relationship to transposition and role of direct and inverted repeats', *Cell*, Vol. 23, pp. 215–227.
44. Golding, G. B. and Glickman, B. W. (1985), 'Sequence-directed mutagenesis: Evidence from a phylogenetic history of human alpha-interferon genes', *Proc. Natl. Acad. Sci. USA*, Vol. 82, pp. 8577–8581.
45. Lilley, D. M. (1981), '*In vivo* consequences of plasmid topology', *Nature*, Vol. 292, pp. 380–382.
46. Golding, G. B., Gearhart, P. J. and Glickman, B. W. (1987), 'Patterns of somatic mutations in immunoglobulin variable genes', *Genetics*, Vol. 115, pp. 169–176.
47. Kolchanov, N. A., Solovyov, V. V. and Rogozin, I. B. (1987), 'Peculiarities of immunoglobulin gene structures as a basis for somatic mutation emergence', *FEBS Lett.*, Vol. 214, pp. 87–91.
48. Rogozin, I. B., Solovyov, V. V. and Kolchanov, N. A. (1991), 'Somatic hypermutagenesis in immunoglobulin genes. I. Correlation between somatic mutations and repeats. Somatic mutation properties and clonal selection', *Biochim. Biophys. Acta*, Vol. 1089, pp. 175–182.

49. Boulikas, T. (1992), 'Evolutionary consequences of nonrandom damage and repair of chromatin domains', *J. Mol. Evol.*, Vol. 35, pp. 156–180.

50. Boiteux, S., Costa de Oliveira, R. and Laval, J. (1985), 'The *Escherichia coli* O6-methylguanine-DNA methyltransferase does not repair promutagenic O6-methylguanine residues when present in Z-DNA', *J. Biol. Chem.*, Vol. 260, pp. 8711–8715.

51. Modrich, P. (1987), 'DNA mismatch correction', *Annu. Rev. Biochem.*, Vol. 56, pp. 435–466.

52. Eckdahl, T. T. and Anderson, J. N. (1990), 'Conserved DNA structures in origins of replication', *Nucleic Acids Res.*, Vol. 18, pp. 1609–1612.

53. Stary, A. and Sarasin, A. (1992), 'Molecular analysis of DNA junctions produced by illegitimate recombination in human cells', *Nucleic Acids Res.*, Vol. 20, pp. 4269–4274.

54. Milot, E., Belmaaza, A., Wallenburg, J. C. *et al.* (1992), 'Chromosomal illegitimate recombination in mammalian cells is associated with intrinsically bent DNA elements', *EMBO J.*, Vol. 11, pp. 5063–5070.

55. Goodsell, D. S. and Dickerson, R. E. (1994), 'Bending and curvature calculations in B-DNA', *Nucleic Acids Res.*, Vol. 22, pp. 5497–5503.

56. Gabrielian, A. E., Landsman, D. and Bolshoy, A. (1999), 'Curved DNA in promoter sequences', *In Silico Biol.*, Vol. 1, pp. 183–196.

57. Konopka, A. K. (1988), 'Compilation of DNA strand exchange sites for non-homologous recombination in somatic cells', *Nucleic Acids Res.*, Vol. 16, pp. 1739–1758.

58. Bachl, J., Steinberg, C. and Wabl, M. (1997), 'Critical test of hot spot motifs for immunoglobulin hypermutation', *Eur. J. Immunol.*, Vol. 27, pp. 3398–3403.

59. Hanawalt, P. C. (1987), 'Preferential DNA repair in expressed genes', *Environ. Health Perspect.*, Vol. 76, pp. 9–14.

60. Veaute, X. and Fuchs, R. P. (1993), 'Greater susceptibility to mutations in lagging strand of DNA replication in *Escherichia coli* than in leading strand', *Science*, Vol. 261, pp. 598–600.

61. Shcherbakova, P. V., Noskov, V. N., Pshenichnov, M. R. and Pavlov, Y. I. (1996), 'Base analog 6-*N*-hydroxylaminopurine mutagenesis in the yeast *Saccharomyces cerevisiae* is controlled by replicative DNA polymerases', *Mutat. Res.*, Vol. 369, pp. 33–44.

62. Lobry, J. R. (1996), 'Asymmetric substitution patterns in the two DNA strands of bacteria', *Mol. Biol. Evol.*, Vol. 13, pp. 660–665.

63. Adams, W. T. and Skopek, T. R. (1987), 'Statistical test for the comparison of samples from mutational spectra', *J. Mol. Biol.*, Vol. 194, pp. 391–396.

64. Piegorsch, W. W. and Bailer, A. J. (1994), 'Statistical approaches for analyzing mutational spectra: Some recommendations for categorical data', *Genetics*, Vol. 136, pp. 403–416.

65. Khromov-Borisov, N. N., Rogozin, I. B., Pegas Henriques, J. A. and de Serres, F. J. (1999), 'Similarity pattern analysis in mutational distributions', *Mutat. Res.*, Vol. 430, pp. 55–74.

66. Piegorsch, W. W. and Richwine, K. A. (2001), 'Large-sample pairwise comparisons among multinomial proportions with an application to analysis of mutant spectra', *J. Agric. Biol. Envir.*, Vol. 6, pp. 305–325.

67. Lewis, P. D. and Parry, J. M. (2002), 'An exploratory analysis of multiple mutation spectra', *Mutat. Res.*, Vol. 518, pp. 163–180.

68. Babenko, V. N. and Rogozin, I. B. (1999), 'Use of a rank correlation coefficient for comparing mutational spectra', *Biofizika*, Vol. 44, pp. 632–638.

69. Pavlov, Y. I., Rogozin, I. B., Galkin, A. P. *et al.* (2002), 'Correlation of somatic hypermutation specificity and A-T base pair substitution errors by DNA polymerase η during copying of a mouse immunoglobulin κ light chain transgene', *Proc. Natl. Acad. Sci. USA*, Vol. 99, pp. 9954–9959.

70. Glazko, G. V., Milanesi, L. and Rogozin, I. B. (1998), 'The subclass approach for mutational spectrum analysis: Application of the SEM algorithm', *J. Theor. Biol.*, Vol. 192, pp. 475–487.

71. Fijal, B. A., Idury, R. M. and Witte, J. S. (2002), 'Analysis of mutational spectra: Locating hotspots and clusters of mutations using recursive segmentation', *Stat. Med.*, Vol. 21, pp. 1867–1885.

72. Blake, R. D., Hess, S. T. and Nicholson-Tuell, J. (1992), 'The influence of nearest neighbors on the rate and pattern of spontaneous point mutations', *J. Mol. Evol.*, Vol. 34, pp. 189–200.

73. Cariello, N. F., Douglas, G. R., Gorelick, N. J. *et al.* (1998), 'Databases and software for the analysis of mutations in the human *p53* gene, human *hprt* gene and both the *lacI* and *lacZ* gene in transgenic rodents', *Nucleic Acids Res.*, Vol. 26, pp. 198–199.

74. Topal, M. D., Eadie, J. S. and Conrad, M. (1986), 'O6-methylguanine mutation and repair is nonuniform. Selection for DNA most interactive with O6-methylguanine', *J. Biol. Chem.*, Vol. 261, pp. 9879–9885.

75. Day, W. H. and McMorris, F. R. (1992), 'Threshold consensus methods for molecular sequences', *J. Theor. Biol.*, Vol. 159, pp. 481–489.

76. Zan, H., Li, Z., Yamaji, K. *et al.* (2000), 'B cell receptor engagement and T cell contact induce Bcl-6 somatic hypermutation in human B cells: Identity with Ig hypermutation', *J. Immunol.*, Vol. 165, pp. 830–839.

77. Rogozin, I. B., Sredneva, N. E. and Kolchanov, N. A. (1996), 'Somatic hypermutagenesis in immunoglobulin genes. III. Somatic mutations in the chicken light chain locus', *Biochim. Biophys. Acta*, Vol. 1306, pp. 171–178.

78. Hasson, J. F., Mougneau, E., Cuzin, F. and Yaniv, M. (1984), 'Simian virus 40 illegitimate recombination occurs near short direct repeats', *J. Mol. Biol.*, Vol. 177, pp. 53–68.

79. Stormo, G. D., Schneider, T. D. and Gold, L. (1986), 'Quantitative analysis of the relationship between nucleotide sequence and functional activity', *Nucleic Acids Res.*, Vol. 14, pp. 6661–6679.

80. Berikov, V. B. and Rogozin, I. B. (1999), 'Regression trees for analysis of mutational spectra in nucleotide sequences', *Bioinformatics*, Vol. 15, pp. 553–562.

81. Smith, D. S., Creadon, G., Jena, P. K. *et al.* (1996), 'Di- and trinucleotide target preferences of somatic mutagenesis in normal and autoreactive B cells', *J. Immunol.*, Vol. 156, pp. 2642–2652.

82. Milstein, C., Neuberger, M. S. and Staden, R. (1998), 'Both DNA strands of antibody genes are hypermutation targets', *Proc. Natl. Acad. Sci. USA*, Vol. 95, pp. 8791–8794.

83. Kepler, T. B. and Bartl, S. (1998), 'Plasticity under somatic mutation in antigen receptors', *Curr. Top. Microbiol. Immunol.*, Vol. 229, pp. 149–162.

84. Milanesi, L. and Rogozin, I. B. (1998), 'Prediction of human gene structure', in Bishop, M. J. (Ed.), 'Guide to Human Genome Computing', Academic Press, Cambridge, UK, pp. 215–259.

85. Gelfand, M. S. (1995), 'Prediction of function in DNA sequence analysis', *J. Comput. Biol.*, Vol. 2, pp. 87–115.

86. Pesole, G., Attimonelli, M. and Saccone, C. (1996), 'Linguistic analysis of nucleotide sequences: Algorithms for pattern recognition and analysis of codon strategy', *Methods Enzymol.*, Vol. 266, pp. 281–294.

87. Lehvaslaiho, H. (2000), 'Human sequence variation and mutation databases', *Brief. Bioinform.*, Vol. 1, pp. 161–166.

88. Wacey, A. I. and Tuddenham, E. G. (1998), 'Mutation databases on the Web', *J. Med. Genet.*, Vol. 35, pp. 529–533.

89. Beroud, C., Collod-Beroud, G., Boileau, C. *et al.* (2000), 'UMD (Universal Mutation Database): A generic software to build and analyze locus-specific databases', *Human Mutat.*, Vol. 15, pp. 86–94.

90. Hauser, J., Seidman, M. M., Sidur, K. and Dixon, K. (1986), 'Sequence specificity of point mutations induced during passage of a UV-irradiated shuttle vector plasmid in monkey cells', *Mol. Cell. Biol.*, Vol. 6, pp. 277–285.

91. Schaaper, R. M., Dunn, R. L. and Glickman, B. W. (1987), 'Mechanisms of ultraviolet-induced mutation. Mutational spectra in the *Escherichia coli lacI* gene for a wild-type and an excision-repair-deficient strain', *J. Mol. Biol.*, Vol. 198, pp. 187–202.

92. Romac, S., Leong, P., Sockett, H. and Hutchinson, F. (1989), 'DNA base sequence changes induced by ultraviolet light mutagenesis of a gene on a chromosome in Chinese hamster ovary cells', *J. Mol. Biol.*, Vol. 209, pp. 195–204.

93. Armstrong, J. D. and Kunz, B. A. (1990), 'Site and strand specificity of UVB mutagenesis in the *SUP4-o* gene of yeast', *Proc. Natl. Acad. Sci. USA*, Vol. 87, pp. 9005–9009.

94. Maher, V. M. and McCormick, J. J. (1986), 'Role of DNA lesions and DNA repair in mutagenesis by carcinogens in diploid human fibroblasts', *Prog. Clin. Biol. Res.*, Vol. 209A, pp. 245–253.

95. Brash, D. E. and Haseltine, W. A. (1982), 'UV-induced mutation hotspots occur at DNA damage hotspots', *Nature*, Vol. 298, pp. 189–192.

96. Pfeifer, G. P., Drouin, R., Riggs, A. D. and Holmquist, G. P. (1991), '*In vivo* mapping of a DNA adduct at nucleotide resolution: Detection of pyrimidine (6–4) pyrimidone photoproducts by ligation-mediated polymerase chain reaction', *Proc. Natl. Acad. Sci. USA*, Vol. 88, pp. 1374–1378.

97. Gibbs, P. E., Borden, A. and Lawrence, C. W. (1995), 'The T-T pyrimidine (6–4) pyrimidinone UV photoproduct is much less mutagenic in yeast than in *Escherichia coli*', *Nucleic Acids Res.*, Vol. 23, pp. 1919–1922.

98. Goodman, M. F. (2002), 'Error-prone repair DNA polymerases in prokaryotes and eukaryotes', *Annu. Rev. Biochem.*, Vol. 71, pp. 17–50.

99. Friedberg, E. C., Feaver, W. J. and Gerlach, V. L. (2000), 'The many faces of DNA polymerases: Strategies for mutagenesis and for mutational avoidance', *Proc. Natl. Acad. Sci. USA*, Vol. 97, pp. 5681–5683.

100. Prakash, S. and Prakash, L. (2002),

'Translesion DNA synthesis in eukaryotes: A one- or two-polymerase affair', *Genes Dev.*, Vol. 16, pp. 1872–1883.

101. Rogozin, I. B., Berikov, V. B., Vasiunina, E. A. and Sinitsina, O. I. (2001), 'Study of the DNA primary structure effect on induction of mutations by alkylating agents', *Genetika*, Vol. 37, pp. 854–861.

102. Burns, P. A., Gordon, A. J. and Glickman, B. W. (1988), 'Mutational specificity of *N*-methyl-*N*-nitrosourea in the *lacI* gene of *Escherichia coli*', *Carcinogenesis*, Vol. 9, pp. 1607–1610.

103. Richardson, K. K., Richardson, F. C., Crosby, R. M. *et al.* (1987), 'DNA base changes and alkylation following *in vivo* exposure of *Escherichia coli* to *N*-methyl-*N*-nitrosourea or *N*-ethyl-*N*-nitrosourea', *Proc. Natl. Acad. Sci. USA*, Vol. 84, pp. 344–348.

104. Storb, U. (1998), 'Progress in understanding the mechanism and consequences of somatic hypermutation', *Immunol. Rev.*, Vol. 162, pp. 5–11.

105. Martin, A., Bardwell, P. D., Woo, C. J. *et al.* (2002), 'Activation-induced cytidine deaminase turns on somatic hypermutation in hybridomas', *Nature*, Vol. 415, pp. 802–806.

106. Faili, A., Aoufouchi, S., Gueranger, Q. *et al.* (2002), 'AID-dependent somatic hypermutation occurs as a DNA single-strand event in the BL2 cell line', *Nat. Immunol.*, Vol. 3, pp. 815–821.

107. Petersen-Mahrt, S. K., Harris, R. S. and Neuberger, M. S. (2002), 'AID mutates *E. coli* suggesting a DNA deamination mechanism for antibody diversification', *Nature*, Vol. 418, pp. 99–103.

108. Rogozin, I. B., Pavlov, Y. I. and Kunkel, T. A. (2001), 'Response 1 to "Smaller role for pol η?"', *Nat. Immunol.*, Vol. 2, pp. 983–984.

109. Zeng, X., Winter, D. B., Kasmer, C. *et al.* (2001), 'DNA polymerase η is an A-T mutator in somatic hypermutation of immunoglobulin variable genes', *Nat. Immunol.*, Vol. 2, pp. 537–541.

110. Cooper, D. N. (2000), 'Human gene mutation in pathology and evolution', *J. Inherit. Metab. Dis.*, Vol. 25, pp. 157–182.

111. Giglia-Mari, G. and Sarasin, A. (2003), 'TP53 mutations in human skin cancers', *Human Mutat.*, Vol. 21, pp. 217–228.

112. Heddle, J. A. (1999), 'On clonal expansion and its effects on mutant frequencies, mutation spectra and statistics for somatic mutations *in vivo*', *Mutagenesis*, Vol. 14, pp. 257–260.