Available online at www.sciencedirect.com
SCIENCE DIRECT®
Mutation Research 544 (2003) 65–85
Reviews in Mutation Research
www.elsevier.com/locate/reviewsmr
Community address: www.elsevier.com/locate/mutres

Review

Theoretical analysis of mutation hotspots and their DNA sequence context specificity

Igor B. Rogozin [a,b], Youri I. Pavlov [c,*]

[a] *Institute of Cytology and Genetics, Russian Academy of Sciences, Novosibirsk, Russia*
[b] *National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA*
[c] *Laboratory of Molecular Genetics and Laboratory of Structural Biology, National Institute of Environmental Health Sciences, Research Triangle Park, Triangle Park, NC 27709, USA*

Received 12 November 2002; received in revised form 12 February 2003; accepted 3 April 2003

Abstract

Mutation frequencies vary significantly along nucleotide sequences such that mutations often concentrate at certain positions called hotspots. Mutation hotspots in DNA reflect intrinsic properties of the mutation process, such as sequence specificity, that manifests itself at the level of interaction between mutagens, DNA, and the action of the repair and replication machineries. The hotspots might also reflect structural and functional features of the respective DNA sequences. When mutations in a gene are identified using a particular experimental system, resulting hotspots could reflect the properties of the gene product and the mutant selection scheme. Analysis of the nucleotide sequence context of hotspots can provide information on the molecular mechanisms of mutagenesis. However, the determinants of mutation frequency and specificity are complex, and there are many analytical methods for their study. Here we review computational approaches for analyzing mutation spectra (distribution of mutations along the target genes) that include many mutable (detectable) positions. The following methods are reviewed: derivation of a consensus sequence, application of regression approaches to correlate nucleotide sequence features with mutation frequency, mutation hotspot prediction, analysis of oligonucleotide composition of regions containing mutations, pairwise comparison of mutation spectra, analysis of multiple spectra, and analysis of "context-free" characteristics. The advantages and pitfalls of these methods are discussed and illustrated by examples from the literature. The most reliable analyses were obtained when several methods were combined and information from theoretical analysis and experimental observations was considered simultaneously. Simple, robust approaches should be used with small samples of mutations, whereas combinations of simple and complex approaches may be required for large samples. We discuss several well-documented studies where analysis of mutation spectra has substantially contributed to the current understanding of molecular mechanisms of mutagenesis. The nucleotide sequence context of mutational hotspots is a fingerprint of interactions between DNA and DNA repair, replication, and modification enzymes, and the analysis of hotspot context provides evidence of such interactions.
Published by Elsevier B.V.

Keywords: DNA sequence context; Classification analysis; Mutable motif; Microsatellite; Hotspot; Direct repeat; Palindrome; Oligonucleotides; Mutation spectra

Abbreviations: AID, activation-induced cytidine deaminase; B, T or G or C; D, A or T or G; DSB, double-strand break; H, A or T or C; K, keto, G or T; M, amino, A or C; N, any nucleotide; Pol, polymerase; R, purine, A or G; S, strong, G or C; UV, ultraviolet; V, A or G or C; W, weak, A or T; Y, pyrimidine, T or C

* Corresponding author. Tel.: +1-919-541-4555; fax: +1-919-541-7613.
E-mail address: pavlov@niehs.nih.gov (Y.I. Pavlov).

1383-5742/$ – see front matter Published by Elsevier B.V.
doi:10.1016/S1383-5742(03)00032-2

1. Introduction

Mutations arise in nucleic acids as part of an important and essential biological process generating genetic variation that is required for a species to evolve. Thus, genomes are replicated at a level of fidelity that leads to a defined rate of spontaneous mutagenesis [1–3]. An increase in spontaneous mutation rate leads to an increase in genetic variation and may be associated with deleterious effects [4]. The mechanisms of spontaneous and induced mutagenesis are complex, and much research is devoted to understanding of these mechanisms and the factors that alter mutation rate. Deeper insight in mutagenic mechanisms can be achieved as more mutation spectra are collected and as sophisticated systems for studying mutagenesis are developed [5–8]. This article describes computational approaches for analyzing mutation spectra with a particular focus on nucleotide sequence context factors that influence mutation frequency.

1.1. Mutation spectra and experimental mutagenesis systems

A mutation spectrum is a set of data that includes the frequency of mutations in a target nucleotide sequence under defined conditions. Mutation spectra are often determined by applying phenotypic selection in an experimental mutagenesis system. Phenotypic selection restricts the mutation spectrum to detectable nucleotides in which a mutation leads to phenotypic changes. Alternatively, mutants are identified by random sequencing of DNA clones or PCR-amplified DNA molecules. A mutation spectrum is usually displayed with the target nucleotide sequence along a horizontal linear axis and each mutational variant listed vertically above the unmutated residue it replaces (e.g. Fig. 1A).

Several types of mutagenesis systems are used to generate mutation spectra. Phenotypic selection systems can be designed to select for reversion or forward mutation. In some reversion systems, only one nucleotide can mutate to produce the desired phenotype and the target sequence is one nucleotide in length. In these systems, the observed mutation frequencies obtained in different conditions can be readily compared (see [9–14]). In other reversion systems, there are multiple mutation mechanisms that revert the mutant to a wild-type phenotype. In this case, and in the case of forward mutation systems using very small target genes (i.e. *supF*, *SUP4-o*), a limited mutation spectrum is generated (see, for example [15–18]). However, the results can be biased due to the small number of detectable positions. In contrast, a forward mutation spectrum in a reasonably large RNA or protein-coding gene usually possesses many (more than 50) detectable positions. Such a mutation spectrum may include substitutions, deletions, insertions, inversions, transpositions and other changes including complex mutations. An example of mutation spectrum in the *lacZ* gene is shown in Fig. 1A [19].

The base substitution mutation spectrum in Fig. 1 includes three principal elements: (1) the target sequence (Fig. 1A; lower line of continuous DNA sequence); (2) the mutations generated by DNA polymerase η when replicating the target sequence (Fig. 1A); and (3) a representation of all positions in the target sequence which can be detected using phenotypic selection for white phage plaques (Fig. 1B) [20]. Pol η generates mutations in all target positions (Fig. 1A) due to its low level of fidelity; therefore, mutant frequency is about 60%, which is the theoretical maximum for this system predicted by reconstruction experiments [20], and multiple mutations are observed in almost all mutants [19].

Fig. 1C shows Pol η mutations that are detectable by phenotypic selection; this mutation spectrum is very different than the unselected spectrum in Fig. 1A, which includes all Pol η mutations, indicating that phenotypic selection limits information about mutation hotspots and the error specificity of an enzyme. Twice as many hotspots are observed in nucleotides 1–40 (*lacZ* promoter region) in the unselected (Fig. 1A) than in the selected spectrum (Fig. 1C). In the protein coding region, prominent hotspot positions 50, 56, 71, 77, 80 and 98 are present in the unselected but not in the selected spectrum. Phenotypic selection reveals only transitions at the first position of the initiator codon (Fig. 1C), however the complete spectrum shows that transversions are strongly favored at this site (Fig. 1A). Therefore, only sites where all base changes are detectable by the chosen selection protocol (e.g. positions 70, 73, 84) provide complete information on the mutation signature of the polymerase when phenotypic selection is applied.

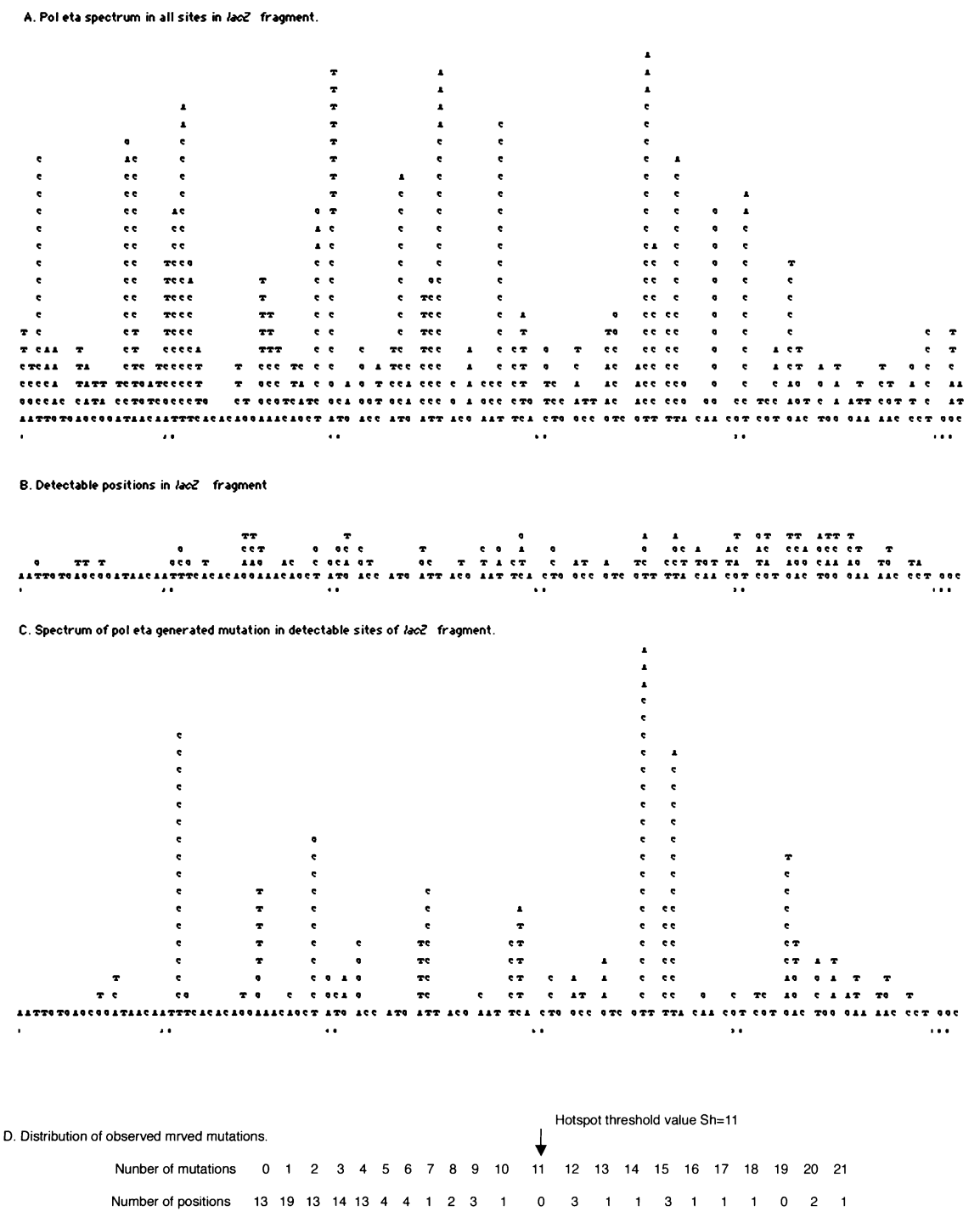

Fig. 1. Mutations in the *lacZ* gene. (A) Error spectrum of human DNA polymerase η without phenotypic selection [19]. (B) Detectable positions in the *lacZ* gene [20]. (C) Error spectrum of the human DNA polymerase η for detectable positions. (D) Mutation frequency distribution.

Mutation spectra generated without the use of phenotypic selection are rare and usually restricted to a few specific positions (i.e. restriction enzyme cleavage sites). Advances in mutational spectrometry allow mutation spectra to be reconstructed without test systems [21–23]. This approach is not widely used and produces mutant fractions instead of mutation frequencies. Many of the statistical methods described in this paper are not directly applicable to this type of data. It is difficult to reconstruct a complete mutation spectrum without a test system, and it requires a very high frequency of mutation. One example is mutagenesis of immunoglobulin genes during somatic hypermutation. In this case, clones can be isolated and sequenced randomly, and it can be assumed that mutations are detectable at all positions [24,25] (note that clonal expansion of certain B-cells may create a bias). Mutation spectra can also be generated without phenotypic selection when mutations are made in vitro by inaccurate DNA polymerases such as Pol η (Fig. 1A), Pol κ and Dpo4 [19,26,27]. Another example is a phylogenetically reconstructed mutation spectrum in human mitochondrial DNA; in this case, all mutations were identified by direct sequencing [28].

For many mutation spectra determined using phenotypic selection, the list of detectable positions is not known. For some intensively studied genes (e.g. *hisD3052* and *lys2–ΔBgl2* for reversion and *lacI* and *lacZα* for forward mutation) detectable positions can be inferred with a high degree of accuracy. However, even more than 10,000 mutations in *lacI* have not saturated this sequence with base substitution mutations, since transversions are usually less frequent than transitions [8]. The most intensively used mutational target sequences are *lacI*, *lacZα*, *p53*, *CAN1*, *SUP4-o* and *supF* for forward mutations [8,18,29–33]. Reversion mutation spectra are often generated using *hisD3052* allele in *Salmonella* or *lys2–ΔBgl* allele in yeast [16,17].

1.2. Mutation types and mechanisms

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. Point mutations are often considered to be the direct result of mutagenesis, and other types of mutations are considered to be the result of genetic recombination. However, these assignments are not valid, because gene conversion between homologous sequences can create point mutations and, DNA polymerase slippage can create large deletions between direct repeats (see [34–36]).

Mutations are generally classified as induced or spontaneous; induced mutations are caused by exposure to exogenous mutagenic factors and spontaneous mutations occur in the absence of such exposure. Spontaneous mutations can arise due to errors in DNA replication, recombination or repair, or can reflect a basal level of endogenous or environmental DNA damage. The rate of mutagenesis during in vitro DNA synthesis is dependent on the concentration of deoxynucleotide substrates, the properties of the DNA polymerase and the interaction between the enzyme and the template/primer. In vivo mutagenesis is likely to be a complex multi-step process involving DNA target sequences and enzymes that play roles in DNA precursor metabolism, DNA replication, recombination and repair [37,38].

This paper focuses on computational methods for studying mutation spectra composed mainly of base pair substitutions. It is possible to apply these methods to spectra composed of other types of mutations, despite some technical differences.

1.3. Mutation hotspots

Mutation frequencies vary along a nucleotide sequence. Nucleotide positions with an exceptionally high mutation frequency are called mutation "hotspots" [5]. Mutation hotspots often reflect a specific mechanism of 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 [16,39–42]. Some mutation hotspots are thought to depend on the nucleotide sequence and the mechanism of

Table 1
Mutable motifs

Test system/mutagen/spectrum	Mutable motif	Comments	Reference
Spontaneous G·C → A·T mutations in human genome	CG	May result from the spontaneous deamination of 5-methylcytosine	[71]
Sn1-type alkylating agents, the *lacI* gene	RG	GG is more mutable compared to AG	[67]
Spontaneous mutations in the *lacI* gene	CCAGG	14 amber nonsense sites were studied	[84]
Triplet repeats associated with human disease	$(CAG)_n$; $(CGG)_n$; $(GAA)_n$	Amplification of repeats results in disease	[162]
Somatic mutations in immunoglobulin genes	RGYW; WA	AGYW is more mutable compared to GGYW; TA is more mutable compared to AA	[25,95,114]
Hotspots of errors produced by human DNA polymerase η	WA	In vitro gap-filling	[19]
Hotspots of errors produced by DNA polymerases in vitro	SM	Errors produced by DNA polymerases α, β and γ were merged in one spectrum	[107]
UV-induced mutations in the lambda *cI* gene	YY	Similar with sites of UV-induced photoproducts	[6]
Pyrimidine (6-4) pyrimidine photoproducts	YTC	In vitro DNA damages induced by UV	[163]
8-OxoG induced hotspots in vitro	GGA	This motif was found to be mutable in some human genes	[128]
AF2-induced mutations in the *lacI* gene	TGC	Characteristics of mutations are similar to those that are due to apurinic sites	[164]
Hotspot of frameshifts in *S. typhimurium*	$(CG)_4$	Spontaneous mutagenesis	[49]
Single-base deletions	YTG	In vitro DNA synthesis	[165]
Spontaneous A·T → T·A mutations in the *lacI* gene	GTGG	*MutD5* strain of *E. coli*	[166]
Spontaneous substitutions in the *supF* gene	GR	Mutations after transfection of monkey cells	[167]
Target signal of retroposable elements in mammals	TTAAAA	L1 reverse transcriptases show nicking in vitro with preference for similar targets	[168]
Signal of recombination in *Bacillus subtilis mal* gene	CATCGCTTRRT	Similar with gyrase binding sites	[169]
Hotspots of frameshifts by *Sulfolobus solfataricus* DNA polymerase IV	GC	In vitro DNA synthesis	[27]

Hotspot positions are underlined, for some motifs the exact location of hotspot positions can not be defined.

mutagenesis per se; these hotspots are called intrinsic mutation hotspots. In contrast, some hotspots may be due to preferential expansions of mutants with high fitness [41,42], e.g. some hotspots of somatic mutations in functional immunoglobulin genes [43]. It has been suggested that hotspots in human *p53* reflect both intrinsic mutability and selection for tumorogenesis [41,44]. This article discusses primarily methods that are useful for analysis of intrinsic mutation hotspots.

1.4. Nucleotide sequence context of mutation hotspots

Many studies have identified specific DNA sequence patterns associated with elevated mutation frequency (Table 1). For example, repetitive sequences such as homonucleotide runs, direct and inverted repeats and microsatellite repeats are involved in specific types of high frequency mutational events (reviewed in [36]). For these mutation hotspots, the

exact DNA sequence is not critical but only the fact that a sequence motif is repeated. Alternatively, mutation hotspots can depend on nucleotide sequence context (mutable motifs, subsequences). Both of these scenarios are discussed below.

In 1966, Streisinger et al. proposed that short deletions and insertions within homonucleotide or homopolymeric tracts arise by misalignment of DNA strands during replication [45]. This misalignment can lead to heterogeneity in the length of homopolymeric tracts; similar arguments apply to the more complex tandemly repeated structures of microsatellites (reviewed in [6,36,46–48]). A well-studied example of misalignment mutagenesis is the two base pair deletion in the CGCGCGCG region of *Salmonella typhimurium hisD3052* (Table 1) [49]. One base pair insertions and deletions are frequent in homonucleotide runs, and the longer the run, the higher the probability of mutation; this observation is consistent with the suggestion that the mutation rate increases as the frequency of misalignment increases in longer homonucleotide runs (see extensive review [48]). Dislocation mutagenesis is similar to misalignment mutagenesis, but involves transient misalignment of a homonucleotide run leading to a base substitution hotspot. This mechanism was proposed based on studies of the in vitro mutation spectra of Pol β [50,51] and HIV reverse transcriptase [52]. Dislocation mutagenesis may also play an important role in vivo generating base substitution hotspots in the control region of human mitochondrial DNA [28].

There is strong evidence that short direct repeats mediate deletions and duplications in DNA [53–55]. Two possible mechanisms for these events are: (1) recombination between short homologous repeats [56] or (2) DNA polymerase slippage between short repeated sequences [57]. In addition, if heteroduplexes form between imperfect direct repeats, repair of the mismatches could cause base substitutions and frameshift mutations [58] in a concerted manner [59]. This mechanism applies to *hisD3052* reversion [16] and has been suggested as a mechanism for some classes of somatic mutations in immunoglobulin genes [60,61] and spontaneous mutations in bacterial and eukaryotic genes [62,63].

Long inverted repeats (40–150 bases) are also particularly unstable in bacterial cells [64,65]. This instability is likely due to formation of hairpin structures in single-stranded DNA and/or DNA polymerase "jumps" (see Gordenin and Resnick [36]). Correction of a quasipalindrome to a perfect inverted repeat may occur by either inter- or intramolecular strand switch [58]. Many mutations of this type have been observed in bacteria, yeast and human cells [16,34,35,58,63]. Thus, this mechanism or a direct repeat-mediated mechanism may explain some classes of somatic mutations in immunoglobulin genes [60,61]. Although there is a significant correlation between substitutions and direct and inverted repeats in immunoglobulin genes [66], it is not clear whether a similar process contributes to base substitution hotspots in other contexts.

Repeated DNA sequences can be found using standard computer programs for DNA analysis which identify repeated elements. Specialized programs have also been developed to identify short direct and inverted repeats (i.e. http://wwwmgs.bionet.nsc.ru/mgs/programs/oligorep). Microsatellites can be identified using the Tandem Repeat Finder system (http://c3.biomath.mssm.edu/trf.html).

As mentioned above, some intrinsic mutation hotspots are caused by mutable motifs (reviewed in [6,41,42,67]). Sequence context effects can act over a significant distance: in one example, the mutation rate was altered by a change 80 bases away from actual site of mutation [18], and in another example, a single base pair change altered the mutation rate 12 bases away (eight-fold effect on 2-aminopurine induced mutagenesis [68]). The effect of sequence context on mutation rate is a well-studied phenomenon. For example, mutation spectra of Sn1 alkylating agents in *lacI* show that most of the induced mutations are G·C → A·T transitions [67]. Mutations at RG sites, where G is the mutable base (underlined), are several times more frequent than mutations at YG sites [67] (Table 1). However, this pattern does not apply for all alkylating agents [67,69,70].

CG dinucleotides are correlated with mutation hotspots in human genes (Table 1). The mutational mechanism for this effect is likely to involve deamination of 5-methylcytosine, which is frequently found at CG dinucleotides. Thus, C·G → T·A mutations occur at CG mutable motifs (hotspot bases are underlined) due to deamination of 5-methylcytosine [71,72] followed by replication of the resulting T·G mispair. It has been proposed that C·G → T·A mutations at

CG mutable motifs prevent genome instability due to recombination of repeated sequences [73]. Dipyrimidines containing 5-methylcytosine are also preferential targets of sunlight-induced mutagenesis in cultured mammalian cells; this observation might explain the large proportion of CG mutations in *p53* in skin tumors in vivo [74]. Many other nucleotide sequence context effects on mutation rate have been studied and characterized (Table 1). A compilation of recombination signals and mutable motifs is available at ftp.bionet.nsc.ru/pub/biology/mutan/RECOMB.ZIP.

2. Methods for analyzing mutation spectra

Computational methods for analyzing mutation spectra are reviewed below. The simpler methods are described in detail; more complicated methods are discussed briefly and sources for more detailed information are cited.

2.1. Hotspot prediction

A mutation spectrum (e.g. Fig. 1A) can be transformed into a distribution of observed mutation frequencies (Fig. 1D). This distribution has been approximated by a Poisson distribution assuming the uniform distribution of mutation frequencies along a target sequence [5,75]. However, this assumption is generally incorrect, since different nucleotide positions have different probabilities of mutation [76]. Thus, there are significant differences between observed and expected distributions (in accordance with the simple Poisson model). These differences may be explained by the presence of mutation hotspots and/or "coldspots", which may be revealed by comparing observed and expected distributions. Two steps are required to predict a distribution of mutation frequencies in a nucleotide sequence: (1) estimation of the parameters of distribution, which assumes *no significant deviation* from a standard distribution (Poisson or binomial); and (2) prediction of hotspots, which are sites *with significant deviation* from the standard distribution. An obvious methodological contradiction between steps 1 and 2 is the major problem of hotspot prediction using approximation by one standard Poisson distribution [41].

An alternative approach was suggested by Glazko et al. [77]. These authors propose to define mutation hotspots using a threshold (Sh) for the number of mutations at a detectable site (Fig. 1D). The threshold is established by analyzing the frequency distribution derived from a mutation spectrum using CLUSTERM program (http://www.itba.mi.cnr.it/webmutation) [41,77]. CLUSTERM identifies several homogeneous classes of sites from a mutation spectrum. Each class of sites is approximated by a binomial (or Poisson) distribution. The probability of mutation is the same for all sites in a class, so variation in mutation frequency for sites of the same class is random and not statistically significant. In contrast, differences in mutation frequency for sites from different classes are statistically significant. Classes with a very high mutation frequency include mutation hotspots. See Rogozin et al. [41] for detailed discussion of this approach and problems associated with its application.

2.2. Comparing mutation spectra

A common problem in mutagenesis is to compare mutation spectra generated under different conditions or by different compounds. This is not a simple problem and it requires statistical methods. One approach uses a contingency table. If there are "T" mutation spectra and "n" detectable sites, then the data are described by an $n \times T$ matrix (or contingency table). The number of mutations in site "i" of spectrum "j", is Y_{ij}. The total number of mutations in spectrum "j" (N_j), is assumed to be fixed and known. 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) [76,78]. The principle of the exact test is based on works of R.A. Fisher, who suggested that testing for homogeneity in an $n \times T$ matrix can be performed without the use of large sample distributions such as χ^2. An HG-PUBL program for such comparisons is available from the FTP site (ftp://sunsite.unc.edu/pub/academic/biology/dna-mutations/hyperg) [79]. The Kendall's tau correlation coefficient can be used as a complementary approach [80]. 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 CORR12 (ftp://ftp.bionet.nsc.ru/pub/biology/dbms/

CORR12.ZIP). Multiple and pairwise spectra comparisons are discussed by Piegorsch and Bailer [76], Khromov-Borisov et al. [69], Rogozin et al. [41] and Lewis and Parry [81]. New analytical strategies for mutational spectra comparisons were suggested recently, these approaches might be also useful for hotspot prediction and analysis [82,83].

2.3. Nucleotide sequence context of mutation hotspots

As described above, nucleotide sequence context influences mutation probability [5,25,42,67,71,72,84]. Several methods are available to analyze this phenomenon. For example, a set of aligned sites can be analyzed to derive a consensus sequence [75] (Table 2) using one of several available approaches as described by Day and McMorris [85,86]). 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 [28,87,88]. 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 2). The probability $P(N_{IJ}, M, F_I)$ to find 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 the significance level α. It is important to note that $\alpha = 0.05$ can not be used for rejecting or accepting a statistical hypothesis due to multiplicity of binomial tests; moreover these tests are strongly inter-dependent for each position. In order to estimate the significance level for $P(N_{IJ}, M, F_I)$, Malyarchuk et al. [28] 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. Statistical analysis described above was repeated for each sample, and the minimal value $P_{\mathrm{mr}}(N_{IJ}, M, F_I)$ was calculated for all positions. This procedure was repeated 10,000 times to calculate the significance level α that separates the right critical region of the distribution $P_{\mathrm{mr}}(N_{IJ}, M, F_I)$ at 5% level of significance, α may be significantly less than 0.05 (for example, $\alpha = 0.005$ for the HVS1 spectrum) [28].

Table 2
Putative DNA polymerase η mutation hotspots in *lacZ* [19]

Sequence	Hotspot position	Type of changes	Number of mutations
CA<u>A</u>TT	3	A → G, T, C	15, 1, 1
TT<u>A</u>TC	14	A → G, C, T	14, 1, 1
GT<u>T</u>AT	15	T → G, A	10, 5
AA<u>A</u>TT	20	A → G, T	11, 1
GA<u>A</u>AT	21	A → G, T	16, 2
AT<u>A</u>GC	38	A → G, T, C	9, 2, 1
CA<u>T</u>AG	39	T → G, A, C	9, 9, 2
TC<u>A</u>TG	46	A → G, T	13, 1
GT<u>A</u>AT	50	A → G, T	16, 4
GA<u>A</u>TT	56	A → G	17
AA<u>A</u>CG	70	A → G, T	18, 3
GT<u>A</u>AA	73	A → G, T	14, 1
CG<u>T</u>TG	77	T → C, G	12
CG<u>A</u>CG	80	A → G, T	11, 2
W<u>A</u>	Consensus		

Hotspot positions are underlined. The spectrum, part of which is shown in Fig. 1, was converted to the complementary strand.

2.4. Regression analysis of nucleotide sequence context effects

Multiple regression models can be used for simultaneous analysis of how several neighboring positions influence mutation frequency. The purpose of multiple regression analysis is to learn more about the relationship between several independent (or predictor) variables X_i and a dependent (or criterion) variable *Y*. In general, multiple regression procedures estimate a linear equation of the form $Y = A + B_1X_1 + B_2X_2 + \cdots B_nX_n$ where "*A*" is a constant and "B_i"s are regression coefficients which represent the contributions of each independent variable to the dependent variable. In other words, X_i is correlated with *Y*, after controlling for all other independent variables. Stormo et al. [89] used multiple linear regression analysis to see how nucleotide sequence context affects 2-aminopurine mutagenesis in the *lacI* gene. The data indicate that two nucleotides immediately preceding the mutable base strongly affect the frequency of mutation. However,

the method assumes a direct linear correlation between the frequency of mutations in detectable positions and factors attributable to nucleotide sequence context, and that the factors are distributed normally; in general, these assumptions are not valid for experimental mutation spectra.

Rogozin and Kolchanov [25] employed a heuristic classification approach and a Monte Carlo procedure to build hotspot consensus sequences. This procedure assesses the non-randomness of nucleotides adjacent to or near a mutation hotspot. Somatic mutation hotspots in immunoglobulin genes were analyzed using this approach, which revealed the statistically significant consensus sequences RGYW and TAA [25].

Regression trees have also been used to analyze the effect of nucleotide sequence context on mutation frequency [90]. Regression trees are mathematically tenable, do not restrain the number of variables (as do heuristic methods) and are recommended for study of simulated and real mutation spectra [90]. However, these approaches are based on complex assumptions and need large datasets (http://www.stat.umn.edu/users/FIRM/firm-info.html).

2.5. Oligonucleotide composition

Nucleotide sequence context of mutation hotspots has also been analyzed by focusing on local mutable motifs. For example, Smith et al. [91] analyzed 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 [91–96]. However, this method neglects the influence of positions other than +1 and −1 on mutation frequency. Milstein et al. [92] suggested joining the most highly mutable triplets in longer consensus sequences, and used this method to study somatic mutation in immunoglobulin genes. The consensus sequences they deduced, G-A-G/a-C/t-T/A and T-A-T/C/G/a [92], are generally consistent with the results of previous studies (i.e. RGYW and TAA [25]). 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 (reviewed in [36,46,48]). In general, this approach requires a large number of detectable sites (hundreds of sites for trinucleotide motifs) in a target sequence. Estimated frequencies of mutations in oligonucleotides can be used for prediction of mutability of sites in any nucleotide sequence [91–96].

2.6. Statistical analysis of 5′ and 3′ neighboring bases

A commonly used approach for analysis of neighboring bases is to calculate the number of times a given base is next to a mutated base, immediately in the 5′ or 3′ direction (positions −1 and +1). A significance of deviation from the expected numbers can be estimated by using various statistical tests [19,33,42,97]. The following procedure can be used for such analysis. For each type of substitutions X → Z, the total number of mutations in sites AX, CX, GX and TX is calculated [M(AX), M(CX), M(GX) and M(TX), respectively]. The number of AX, CX, GX and TX target dinucleotides are calculated as N(AX), N(CX), N(GX) and N(TX). The expected number of mutations E(AX) in AX sites is estimated as

$$E(\mathrm{AX}) = \frac{N(\mathrm{AX}) \times (M(\mathrm{AX}) + M(\mathrm{CX}) + M(\mathrm{GX}) + M(\mathrm{TX}))}{N(\mathrm{AX}) + N(\mathrm{CX}) + N(\mathrm{GX}) + N(\mathrm{TX})}$$

and the P-value is determined using a standard χ^2-test with three degrees of freedom ($P(\chi^2)$). A Bonferroni correction for multiple comparisons can be used to estimate the significance level α ($\alpha = 0.05/\mathrm{Nt} = 0.004$, where $\mathrm{Nt} = 12$ is the total number of statistical tests used) [19]. The same procedure is repeated for XA, XC, XG and XT sites. In general, this method does not require prior analysis of mutation hotspots and may be extended to positions beyond +1 and −1. Krawczak et al. [72] analyzed nearest-neighbor 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-neighbor effects were developed by Zavolan and Kepler [42].

2.7. Correlation between nucleotide sequence features and mutation spectra

Nucleotide sequence features can be correlated with a mutation spectrum and the correlation can be tested for statistical significance. This approach is discussed

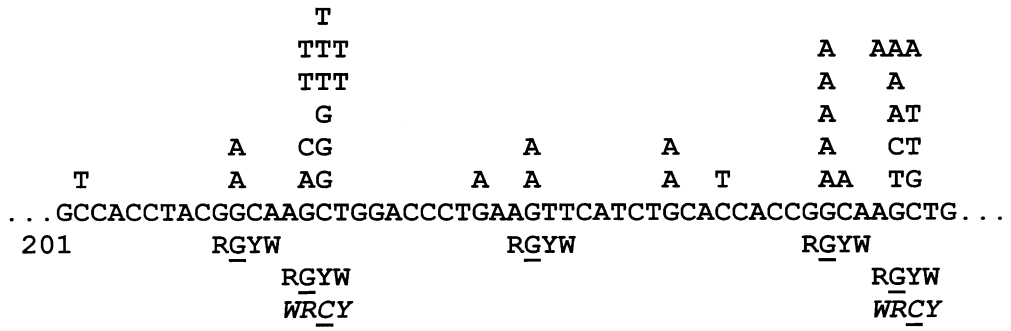

Fig. 2. Somatic hypermutation spectrum in an artificial *GFP* substrate [98]. The RGYW and complementary variant WRCY motifs (Table 1) are shown under the nucleotide sequence. Mutable motifs in the complementary strand are in italics.

here with regard to analysis of multiple somatic mutations in an artificial *GFP* substrate in non-immune system type cells, murine fibroblasts [98] (Fig. 2). The RGYW mutable motif is a signature of somatic hypermutation (Table 1), and may play a role in somatic mutations observed in *GFP* (Fig. 2) (see Section 3.2). The number of mutations (MM) and the number of target sequence positions, NP, included in this motif (or by the mutable position in this motif) are calculated [98,99]. The Fisher exact test or a Monte Carlo modification of the exact test [76,78] 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. The input numbers are MM, MA − MM, NP, NA − NP, where MA is the number of substitutions and NA is the number of positions in the target sequence. (The Fisher exact test is available at http://www.matforsk.no/ola/fisher.htm, MM, MA − MM in the first row and NP, NA − NP in the second row; a Monte Carlo modification of the exact test is discussed in Section 2.2). When the one of the numbers is large (>300), the χ^2-test with one degree of freedom can be applied instead of the exact test. In this case, the expected number of mutations in mutable motifs E(MM) and all other positions of the target sequence E(MA − MM) are calculated, $E(\text{MM}) = \text{MA}(\text{NP}/\text{NA})$ and $E(\text{MA} - \text{MM}) = \text{MA}[(\text{NA} - \text{NP})/\text{NA}]$. In the case of mutations in *GFP* (Fig. 2), the correlation is statistically significant ($P(\chi^2) < 0.01$).

In the above analysis, statistical significance can be also estimated using a modified Monte Carlo procedure [25]. This approach takes into account frequency of substitutions in A, T, G and C bases, the presence of several mutations in a site and the nucleotide sequence of the target sequence. Weight "W_j" of site "j" is defined as the number of substitutions in a mutable motif. 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 $P(W \leq W_{\text{random}})$. 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 ($P(W \leq W_{\text{random}}) \leq 0.05$) indicate a significant correlation between mutable motif and mutation frequency. Modified versions of this approach were used to analyze a dislocation model in human mitochondrial DNA [28], mutability of direct and inverted repeats in immunoglobulin genes [66] and gene conversion in immunoglobulin genes [100]. A similar approach was used to analyze illegitimate recombination events [101]. 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. This approach is different from a Monte Carlo modification of the exact test (Section 2.2). A hypothesis of no correlation between context factors and mutations was tested instead of a hypothesis of contingency table homogeneity.

2.8. Analysis of aligned sites

Theoretically, various computational approaches can be used to analyze 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 (reviewed in [102–106]. These methods are well established and have been tested on different types of data, but all of these methods require large datasets.

2.9. "Context-free" characteristics of a mutation spectrum

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. This kind of information can be used to understand molecular mechanisms of mutagenesis. Some statistical approaches for analyzing context-free characteristics are described in [24,107–111]. For example, context-free analysis was combined with analysis of mutable motifs in a recent analysis of somatic hypermutation in immunoglobulin genes [110].

3. Examples of mutation spectra analysis

The information on parameters of mutation spectra and, specifically, mutation hotspots provides insight on mechanism of mutagenesis and could be invaluable for evolutionary biologists (see [6,42,81,112–114]). For example, it has been often assumed that a nine base pair deletion between COII and tRNA(Lys) genes in human mitochondrial genome have arisen just once in human populations, thus this deletion has been used for tracing migration patterns of various populations. However, newer data suggested that the nine base pair deletion originated multiple times in populations of Africa, Europe, South India and China, multiple origins of the deletion is supported by coalescence and phylogenetic analyses [115,116]. This intergenic region contains a nine base pair tandem repeat (Fig. 3)

```
CACCCCCTCTACCCCCTCTAGAGCC
-------->------->
```

Fig. 3. The intergenic region between the COII and tRNA(Lys) genes in the human mitochondrial genome (positions 8269–8293; AF382000). The polymorphic tandem direct repeat is underlined.

and therefore may be a hotspot for deletions (see Section 1.4). We will further illustrate the importance of a "hotspot context" approach for mutagenesis using two additional examples.

3.1. Analysis of mutations in the mutT deficient strain of Escherichia coli

Chemical agents, ionizing radiation and oxidative stress cause DNA oxidation [117,118]. 8-Oxoguanine (8-oxoG) is one of the most prominent base oxidation products and has been implicated in mutagenesis, carcinogenesis and aging [119]. It has been shown to cause G·C → T·A and A·T → C·G mutations in vivo and in vitro, depending whether guanine is oxidized in DNA or in the DNA precursor pools, respectively [120,121]. To counter the mutagenic effects of 8-oxoG, *E. coli* has an effective repair system containing three genes, *mutT*, *mutM* and *mutY* [122]. *mutM* and *mutY* are involved in repair of 8-oxoG in DNA and *mutT* codes for an enzyme that converts 8-oxoGTP in the nucleotide pool to 8-oxoGMP, preventing the incorporation of 8-oxoG into DNA [123].

A spontaneous mutation spectrum in the *mutT* deficient *E. coli* strain is composed, almost exclusively, of A·T → C·G transversions which is in general consistent with mutagenic properties of 8-oxoGTP [124,125]. Hotspot context analysis of these transversions using regression trees (Section 2.5) revealed AA mutable sequence [90]. Comparison of the *mutT*$^{-}$ spectrum and A·T → C·T transversions in a spectrum of spontaneous mutations in the *lacI* gene (*lacI*$^{-d}$ test system) [126] did not reveal significant differences between them (Table 3). Furthermore, a highly significant positive correlation was found (Table 3). This result suggested that a fraction of spontaneous A·T → C·T mutations in *E. coli* may be caused by 8-oxoGTP in nucleotide pools. Reconstructed spontaneous mutations in human pseudogenes [88,127] were also analyzed, and the frequencies of nucleotides surrounding A·T → C·T transversions are shown in Table 4. Notably, AA and TT are the most frequent dinucleotide combinations. Such excess is statistically significant ($P(\chi^2) < 0.01$) as compared to dinucleotide frequencies in reconstructed ancestral sequences [88] (Table 4).

A strong influence of neighboring bases was also revealed for G·C → T·A transversions, another hallmark

Table 3
Comparison of A·T → C·G transversion in *lacI* gene from *mutT*⁻ and wild-type strains of *E. coli*

Position																			
41	81	72	64	87	168	79	189	192	195	167	83	117	96	128	177	77	141	105	54
A·T → C·G mutations in *mutT*⁻ strain																			
4	10	5	2	4	9	7	23	18	10	37	5	1	5	4	7	20	5	2	8
A·T → C·G spontaneous mutations																			
2	3	2	2	1	2	2	8	1	6	10	0	1	0	0	3	4	0	1	3

Results of direct comparison between spectra (Section 2.2): probability that these two spectra are different $P(\chi^2) = 0.69$ [79], Kendell's ran tau correlation coefficient = 0.65 ($P < 0.01$) [80]. Positions of AA mutable motifs are underlined.

of the 8-oxoG-dependent mutagenesis. A consensus mutable sequence GGA was derived for this type of error made in vitro by T4 DNA polymerase replicating 8-oxoG containing oligonucleotides [128]. It was found that the mutable motif G(8-oxoG)A not only was more prone to direct misincorporation of A opposite the template 8-oxoG, but also allowed relatively a higher efficiency of incorporation of C. One implication of this finding is that this nucleotide context of the 8-oxoG lesion induced less distortion of the DNA structure [128]. Quite remarkably, the same GGA context for spontaneous G·C → T·A mutations in the *lacI* gene in *E. coli* was very prominent, even though DNA was replicated in vivo by a different replicative complex. More, G·C → T·A mutations in the same context were over-represented in the collection of *p53* mutations in humans [128]. These results suggested that mutagenesis due to 8-oxoG is significantly influenced by nearest neighboring bases and the context is quite evolutionarily stable.

Table 4
Frequencies of bases in position +1 and −1 in a set of spontaneous A·T → C·G transversions found in human pseudogenes [88]

Mutation	Position −1				Position +1			
	A	T	G	C	A	T	G	C
A → C					0.35	0.24	0.17	0.24
T → G	0.25	0.32	0.21	0.22				
Expected	0.25	0.22	0.22	0.31	0.26	0.23	0.29	0.22

Expected values (frequencies of AN and NT dinucleotides) were calculated in ancestral sequences used for reconstruction of spontaneous mutations [88,127]. The differences between observed and expected frequencies of AA and TT dinucleotides were statistically significant ($P(\chi^2) < 0.01$).

3.2. Analysis of somatic mutations in immunoglobulin genes

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 [129,130]. Immunologists have been investigating possible mechanisms of somatic hypermutation for more than 20 years and a number of different models have been proposed [131,132]. Most models postulate involvement of mutator polymerases to account for high frequency of mutagenesis in V regions [133]. 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 [25,114] are found in both strands and WA hotspots preferentially are found in only one strand [25,92,95,114,134,135] (Table 5).

Analysis of mutation spectra of errors made by various DNA polymerases during in vitro DNA synthesis provided clues on what polymerase could operate during somatic hypermutation. A correlation between the WA motif and the error specificity of human Pol η and lack of A–T mutations in XP-V patients deficient in Pol η suggested that this polymerase may contribute to the WA hotspots [114,135,136]. The error specificity of Pol η does not correlate with SHM at G–C base pairs in the RGYW sequence motif. This suggests that SHM may involve more than one DNA

Table 5
Somatic mutation hotspots in the VkOx1 transgene

Exact match with RGYW	One mismatch with RGYW	Exact match with TA	Exact match with AA	Other hotspots
CAGCT	AtGCA	GTACC	GAAGG	GCAGT
AAGTT	GtGTA	GTAGT	GAAGT	GCATG
GAGCT	CtGCT	GTAAG	AAAAG	
CAGCA	GAGgT	TTACA	CAATC	
TGGTA	CAGTg	TTATG	GAAGA	
TGGCT	GtGCA	TTACT		
CAGCA	CtGCA	GTAAC		
CAGCA	GAGaT	GTAAG		
CAGCA	TtGCT			
GAGTA				
TAGTA				
GGGTT				
GAGTA				
GAGCA				
CAGTT				
TGGTA				
TGGCA				

Hotspots were predicted by CLUSTERM [77]. Data was taken from [92]. The mutable base is displayed as a purine, using the appropriate DNA strand. Mismatches are indicated by lower-case letters, hotspot positions are underlined.

transaction and more than one DNA polymerase [114] what is consistent with the two-phase model of SHM proposed earlier [134,137]. Additional analysis of this correlation using the same mouse immunoglobulin target sequence for in vivo and in vitro spectrum generation (described in Section 2.2) combined with studies of mutable motifs and frequencies of substitutions greatly improved the power of comparisons, allowing use of different statistical methods [138]. It was found that two Pol η error spectra determined while it synthesizes 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. Interestingly, in *Xenopus* somatic hypermutation is strongly biased toward alterations in G–C pairs (with a strong preference for RGYW motifs) suggesting that WA-mutator not always has a significant role in mutagenesis [139].

It was suggested recently that double-strand breaks (DSBs) in V genes are associated with this mutable motif and thus may initiate the somatic hypermutation [140,141]. However, a careful analysis of DSB hotspots and their correlation with RGYW and WA motifs suggested that two different mutator processes might produce closely spaced mismatches that yield DSBs, owing to overlapping excision tracts during subsequent processing [142]. Thus, DSBs might be a consequence rather than a cause of somatic hypermutation in immunoglobulin V genes.

A candidate for a principal RGYW mutator is activation-induced cytidine deaminase (AID), converting cytosine in DNA into uracil [111,132,143–145]. Indeed, overexpression of AID in murine fibroblasts was mutagenic and mutations occurred in RGYW motifs (for example see [98]). Current models of somatic hypermutation in two mutable motifs were discussed by Kunkel et al. [146].

4. Problems in analysis of mutation spectra

4.1. Defining hotspots

Defining hotspots in a mutation spectrum is a non-trivial task. Hotspots are not simply the CLUSTERM class with the highest frequency of mutation (i.e. a hotspot can not be assigned arbitrarily based on relative mutation frequency). A mutation hotspot should be assignable as a hotspot with $\geq$0.95 probability [41]). The problem becomes more complicated as the number of CLUSTERM classes increases. Some empirical rules based on CLUSTERM out-

put have been suggested to deal with these issues [41].

Cold spots also present difficulties in analysis of mutation spectra. A cold spot is a position where mutations are not observed or are observed with low frequency. Cold spots are difficult because it is very difficult to estimate low mutation frequency, and it is not technically possible to characterize a cold spot if the mutation rate can not be accurately measured.

4.2. Small sample size

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. A few approaches can be robust with small data sets (i.e. hotspot prediction, comparing mutation spectra, correlation between nucleotide sequence features and mutation spectra). Other methods may not be reliable when applied to small datasets. In some cases, a combination of two approaches can be used. For example, a consensus sequence can be constructed and correlation between the consensus sequence and a mutation spectrum can be analyzed. This procedure may be repeated several times until a consensus sequence is derived that has a small value of $P(W \leq W_{\text{random}})$ or $P(\chi^2)$. When two (or more) mutation spectra with similar nucleotide sequence features are analyzed, one spectrum can be used to derive a consensus (Section 2.3) and the other can be used for correlation analysis (Section 2.7).

4.3. Complexity of mutation spectra

Mutation spectra include different types of base substitutions occurring with different frequencies at different sites. If these substitutions are further divided into subgroups, the uni-dimensional set "Y" ("Y_i" is the number of mutations in site "i") becomes two-dimensional or multidimensional requiring more complex analytical methods (see [19,69]). The approaches described above do not take this level of complexity into account. With very extensive spectra, it is possible to analyze various types of mutations separately, which is one way to deal with the complexity of the data. For example, A·T → C·G mutations could be analyzed without including other transition and transversion substitutions in the analysis (Section 3.1).

4.4. Global factors

Many factors influence mutation frequency in a particular nucleotide sequence. However, in most cases, analytical methods only attempt to characterize 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. For example, AGTA is more mutable in CDR regions than in FR regions of immunoglobulin genes [147]. 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 [148,149]. Inherent asymmetry between the two DNA strands at the replication fork could also influence mutation frequency and specificity (see [13,14,150–152]). Other potential factors include asymmetric base composition [153] or higher order chromatin structure (reviewed by Boulikas [154]).

Theoretically, it is possible to analyze the correlation between all factors affecting mutation frequency and an observed mutation spectrum. However, a large number of correlations would be tested in such an analysis, and it is expected that correlations with $P < 0.05$ would occur by chance at a rate of 5 per 100 analyses. Moreover, interdependent factors could bias the results of such exhaustive searches (i.e. base composition and frequency of homonucleotide runs). The only way to address this problem is to systematically collect and analyze mutation spectra for the same mutational target under different experimental conditions. This approach has been undertaken for *lacI* and *supF* (see [8,18,81,155]).

4.5. Consensus sequence discrepancies

A consensus sequence for a mutation hotspot is rarely definitive and can therefore have several variants. For example, mutation hotspots associated with somatic hypermutation in immunoglobulin genes have been reported as RGYW and TAA [25] or G-A-G/a-C/t-T/A and T-A-T/C/G/a [92]. Other variants have also been reported [91,95,134,156]. Rogozin et al. proposed a method to evaluate the relative merit of different consensus sequences [114]. All possible pentanucleotides NNNNN were analyzed as potential consensus sequences. Some consensus sequences

included the ambiguous positions R, Y, W, S, K and M. A Monte Carlo procedure was used to test for correlation between mutations in 15 spectra and the distribution of each motif in the target. Small probability values ($P(W \leq W_{\text{random}}) \leq 0.05$) indicate a significant correlation between mutation frequency and a mutable motif. The number of spectra where $P(W \leq W_{\text{random}})$ was ≤ 0.05 divided by the number of spectra that include a motif defined the score for the motif. The sequences RGYW, AGYW, WA and TA had highest scores. All these motifs were used for further analysis of errors made by DNA polymerases in vitro [114].

4.6. Parameterization of theoretical methods

Methods for analyzing mutation spectra vary greatly in their complexity. However, a higher degree of complexity does not guarantee better results. Indeed, the result depends on the quality of the data and the ability to correctly define the relevant sequence context features and the theoretical models. When the number of parameters in a method increase (for example, in regression models), the method may behave unpredictably with small data sets or if several context factors simultaneously influence mutation frequency. Thus, complex models can be used only for large data sets. It is better to analyze different types of mutations separately (e.g. mutations associated with repetitive sequences or a mutable motif) [28], which reduces model complexity.

4.7. Identification of detectable positions

It is very important to identify the detectable positions in a target sequence before analyzing its mutation spectrum. For example, mutations in the human *p53* gene are not 100% accurate [157], and the database may include false detectable positions (due to sequencing errors and DNA polymorphism). In general, the molecular basis for the relationship between the mutations in *p53* and human cancer is not always clear (reviewed in [158] and [159]) and the meaning of "a detectable position" cannot be clearly defined for this gene. It is difficult to interpret the significance of results obtained for mutations in *p53* (for example, results of spectra comparisons [41]). Population polymorphism becomes an 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 may bias mutation spectra. It is possible to misassign a functional mutation at a specific site even if a dataset 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 characterized detectable sites, in which several independent mutations have been observed, should be used when a mutation spectrum is analyzed.

4.8. Mutation hotspots and phylogenetic analysis

On one hand, phylogenetic analysis may be used for reconstruction of mutation spectra [28,88,127], however these spectra could contain errors due to unforeseen problems with alignments and phylogenetic reconstructions [28]. On the other hand, hotspot context analysis may be important for phylogenetic reconstruction, which is based on models of mutational process [160]. In general, spontaneous mutations are influenced by selection and specificity of mutagenesis. An important role of selection is obvious (e.g. the rate of synonymous substitutions is much higher than nonsynonymous substitutions for most protein-coding genes). However, it was suggested that mutational bias in the introduction of novelty strongly influences the course of evolution [161]. Thus, information about context specificity of mutations might improve phylogenetic studies, however this would dramatically increase complexity of substitutions models [160] and currently is not used. Instead hotspot sites (for example, CG mutable motifs) may be simply removed from phylogenetic analysis. Alternatively, substitutions models, which accounts for substitution rate variation among sites (e.g. gamma distances [160]) can be used.

5. Conclusion

This article reviews computational methods for analyzing context specificity of mutation spectra and discusses common problems in such analyses. The goal of these methods is ultimately to increase understanding

of molecular mechanisms of mutagenesis. To this end, 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, robust approaches should be used with small samples of mutations (see Section 4.2), whereas combinations of simple and complex approaches can be used for large samples. Complex approaches are needed because mutation spectra reflect the influence of multiple diverse local and global factors. It is a challenging task to analyze mutation spectra, and in some cases, the effort will be primarily descriptive in nature. However, in several well-documented studies, the analysis of mutation spectra has contributed substantially to understanding molecular mechanisms of mutagenesis. As analytical methods continue to be developed and/or improved, more studies will contribute insights into the complex process of mutagenesis.

Acknowledgements

This work was partially supported by RFBR (grant nos. 96-04-49957, 99-04-49535 and 02-04-48342). We thank B.A. Rogozin, N.N. Khromov-Borisov, N.A. Kolchanov, G.V. Glazko, O.I. Sinitsina, V.V. Solovyev, V.N. Babenko and A.S. Kondrashov for helpful discussions and P.V. Shcherbakova, V.N. Babenko, E.A. Vasunina, T.A. Kunkel, W.C. Copeland and anonymous referees for helpful comments on the manuscript. Miriam Sander (Page One Editorial Services) is acknowledged for professional scientific editorial work.

References

[1] J.W. Drake, B. Charlesworth, D. Charlesworth, J.F. Crow, Rates of spontaneous mutation, Genetics 148 (1998) 1667–1686.
[2] M. Radman, Enzymes of evolutionary change, Nature 401 (1999) 866–867.
[3] S.M. Rosenberg, Evolving responsively: adaptive mutation, Nat. Rev. Genet. 2 (2001) 504–515.
[4] J.F. Crow, The high spontaneous mutation rate: is it a health risk? Proc. Natl. Acad. Sci. U.S.A. 94 (1997) 8380–8386.
[5] S. Benzer, On the topography of the genetic fine structure, Proc. Natl. Acad. Sci. U.S.A. 47 (1961) 403–415.
[6] J.H. Miller, Mutational specificity in bacteria, Annu. Rev. Genet. 17 (1983) 215–238.
[7] F. Hutchinson, J.E. Donnellan Jr., A general database for DNA sequence changes induced by mutagenesis of several bacterial and mammalian genes, Nucleic Acids Res. 24 (1996) 172–176.
[8] J.G. De Boer, B.W. Glickman, The *lacI* gene as a target for mutation in transgenic rodents and *Escherichia coli*, Genetics 148 (1998) 1441–1451.
[9] B.N. Ames, F.D. Lee, W.E. Durston, An improved bacterial test system for the detection and classification of mutagens and carcinogens, Proc. Natl. Acad. Sci. U.S.A. 70 (1973) 782–786.
[10] C.G. Cupples, J.H. Miller, A set of *lacZ* mutations in *E. coli* that allow rapid detection of each of the six base substitutions, Proc. Natl. Acad. Sci. U.S.A. 86 (1989) 5345–5349.
[11] C.G. Cupples, M. Cabrera, C. Cruz, J.H. Miller, A set of *lacZ* mutations in *E. coli* that allow rapid detection of specific frameshift mutations, Genetics 125 (1990) 275–280.
[12] G. Maenhaut-Michel, R. Janel-Bintz, R.P. Fuchs, A umuDC-independent SOS pathway for frameshift mutagenesis, Mol. Gen. Genet. 235 (1992) 373–380.
[13] P.V. Shcherbakova, Y.I. Pavlov, $3' \rightarrow 5'$ exonucleases of DNA polymerases ε and δ correct base analog induced DNA replication errors on opposite DNA strands in *Saccharomyces cerevisiae*, Genetics 142 (1996) 717–726.
[14] Y.I. Pavlov, C.S. Newlon, T.A. Kunkel, Yeast origins establish a strand bias for replicational mutagenesis, Mol. Cell 10 (2002) 207–213.
[15] L. Kohalmi, B.A. Kunz, In vitro mutagenesis of the yeast *SUP4-o* gene to identify all substitutions that can be detected in vivo with the *SUP4-o* system, Environ. Mol. Mutagen. 19 (1992) 282–287.
[16] D.M. DeMarini, M.L. Shelton, A. Abu-Shakra, A. Szakmary, J.G. Levine, Spectra of spontaneous frameshift mutations at the *hisD3052* allele of *Salmonella typhimurium* in four DNA repair backgrounds, Genetics 149 (1998) 17–36.
[17] B.D. Harfe, S. Jinks-Robertson, Removal of frameshift intermediates by mismatch repair proteins in *Saccharomyces cerevisiae*, Mol. Cell. Biol. 19 (1999) 4766–4773.
[18] K.A. Canella, M.M. Seidman, Mutation spectra in *supF*: approaches to elucidating sequence context effects, Mutat. Res. 450 (2000) 61–73.
[19] T. Matsuda, K. Bebenek, C. Masutani, I.B. Rogozin, F. Hanaoka, T.A. Kunkel, Error rate and specificity of human and murine DNA polymerase η, J. Mol. Biol. 312 (2001) 335–346.
[20] K. Bebenek, T.A. Kunkel, Analyzing fidelity of DNA polymerases, Methods Enzymol. 262 (1995) 217–232.
[21] K. Khrapko, H. Coller, P. Andre, X.C. Li, F. Foret, A. Belenky, B.L. Karger, W.G. Thilly, Mutational spectrometry without phenotypic selection: human mitochondrial DNA, Nucleic Acids Res. 25 (1997) 685–693.
[22] A. Tomita-Mitchell, A.G. Kat, L.A. Marcelino, X.C. Li-Sucholeiki, J. Goodluck-Griffith, W.G. Thilly, Mismatch repair deficient human cells: spontaneous and MNNG-induced mutational spectra in the *HPRT* gene, Mutat. Res. 450 (2000) 125–138.

[23] B.P. Muniappan, W.G. Thilly, The DNA polymerase β replication error spectrum in the adenomatous polyposis coli gene contains human colon tumor mutational hotspots, Cancer Res. 62 (2002) 3271–3275.
[24] P.J. Gearhart, D.F. Bogenhagen, Clusters of point mutations are found exclusively around rearranged antibody variable genes, Proc. Natl. Acad. Sci. U.S.A. 80 (1983) 3439–3443.
[25] I.B. Rogozin, N.A. Kolchanov, Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis, Biochim. Biophys. Acta 1171 (1992) 11–18.
[26] E. Ohashi, K. Bebenek, T. Matsuda, W.J. Feaver, V.L. Gerlach, E.C. Friedberg, H. Ohmori, T.A. Kunkel, Fidelity and processivity of DNA synthesis by DNA polymerase κ, the product of the human *DINB1* gene, J. Biol. Chem. 275 (2000) 39678–39684.
[27] R.J. Kokoska, K. Bebenek, F. Boudsocq, R. Woodgate, T.A. Kunkel, Low fidelity DNA synthesis by a Y family DNA polymerase due to misalignment in the active site, J. Biol. Chem. 277 (2002) 19633–19638.
[28] B.A. Malyarchuk, I.B. Rogozin, V.B. Berikov, M.V. Derenko, Analysis of phylogenetically reconstructed mutational spectra in human mitochondrial DNA control region, Hum. Genet. 111 (2002) 46–53.
[29] B.A. Kunz, K. Ramachandran, E.J. Vonarx, DNA sequence analysis of spontaneous mutagenesis in *Saccharomyces cerevisiae*, Genetics 148 (1998) 1491–1505.
[30] J.D. Roberts, T.A. Kunkel, Eukaryotic DNA replication fidelity, in: M.D. Pamphilis (Ed.), DNA Replication in Eukaryotic Cells: Concepts, Enzymes and Systems, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1996, pp. 217–247.
[31] D.X. Tishkoff, N. Filosi, G.M. Gaida, R.D. Kolodner, A novel mutation avoidance mechanism dependent on *S. cerevisiae RAD27* is distinct from DNA mismatch repair, Cell 88 (1997) 253–263.
[32] T. Soussi, K. Dehouche, C. Beroud, *p53* website and analysis of *p53* gene mutations in human cancer: forging a link between epidemiology and carcinogenesis, Hum. Mutat. 15 (2000) 105–113.
[33] N.F. Cariello, G.R. Douglas, N.J. Gorelick, D.W. Hart, J.D. Wilson, T. Soussi, 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. 26 (1998) 198–199.
[34] L.S. Ripley, B.W. Glickman, Unique self-complementarity of palindromic sequences provides DNA structural intermediates for mutation, Cold Spring Harb. Symp. Quant. Biol. 47 (Part 2) (1983) 851–861.
[35] R.I. Salganik, G.L. Dianov, O.A. Medvedev, Cluster of point mutations predetermined by a quasipalindromic nucleotide sequence in plasmid pBR322 DNA, FEBS Lett. 261 (1990) 28–30.
[36] D.A. Gordenin, M.A. Resnick, Yeast ARMs (DNA at-risk motifs) can reveal sources of genome instability, Mutat. Res. 400 (1998) 45–58.
[37] J.W. Drake, R.H. Baltz, The biochemistry of mutagenesis, Annu. Rev. Biochem. 45 (1976) 11–37.
[38] H. Maki, Origins of spontaneous mutations: specificity and directionality of base-substitution, frameshift, and sequence-substitution mutageneses, Annu. Rev. Genet. 36 (2002) 279–303.
[39] J. Suckow, P. Markiewicz, L.G. Kleina, J. Miller, B. Kisters-Woike, B. Muller-Hill, 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. 261 (1996) 509–523.
[40] D.R. Walker, J.P. Bond, R.E. Tarone, C.C. Harris, W. Makalowski, M.S. Boguski, M.S. Greenblatt, Evolutionary conservation and somatic mutation hotspot maps of *p53*: correlation with p53 protein structural and functional features, Oncogene 18 (1999) 211–218.
[41] I.B. Rogozin, F.A. Kondrashov, G.V. Glazko, Use of mutation spectra analysis software, Hum. Mutat. 17 (2001) 83–102.
[42] M. Zavolan, T.B. Kepler, Statistical inference of sequence-dependent mutation rates, Curr. Opin. Genet. Dev. 11 (2001) 612–615.
[43] A.G. Betz, M.S. Neuberger, C. Milstein, Discriminating intrinsic and antigen-selected mutational hotspots in immunoglobulin V genes, Immunol. Today 14 (1993) 405–411.
[44] M. Krawczak, B. Smith-Sorensen, J. Schmidtke, V.V. Kakkar, D.N. Cooper, E. Hovig, Somatic spectrum of cancer-associated single basepair substitutions in the *TP53* gene is determined mainly by endogenous mechanisms of mutation and by selection, Hum. Mutat. 5 (1995) 48–57.
[45] G. Streisinger, Y. Okada, J. Emrich, J. Newton, A. Tsugita, E. Terzaghi, M. Inouye, Frameshift mutations and the genetic code, Cold Spring Harb. Symp. Quant. Biol. 31 (1966) 77–84.
[46] L.S. Ripley, Frameshift mutation: determinants of specificity, Annu. Rev. Genet. 24 (1990) 189–213.
[47] B.S. Strauss, Frameshift mutation, microsatellites and mismatch repair, Mutat. Res. 437 (1999) 195–203.
[48] K. Bebenek, T.A. Kunkel, Streisinger revisited: DNA synthesis errors mediated by substrate misalignments, Cold Spring Harb. Symp. Quant. Biol. 65 (2000) 81–91.
[49] K. Isono, J. Yourno, Chemical carcinogens as frameshift mutagens: *Salmonella* DNA sequence sensitive to mutagenesis by polycyclic carcinogens, Proc. Natl. Acad. Sci. U.S.A. 71 (1974) 1612–1617.
[50] T.A. Kunkel, The mutational specificity of DNA polymerase-β during in vitro DNA synthesis. Production of frameshift, base substitution, and deletion mutations, J. Biol. Chem. 260 (1985) 5787–5796.
[51] T.A. Kunkel, A. Soni, Mutagenesis by transient misalignment, J. Biol. Chem. 263 (1988) 14784–14789.
[52] K. Bebenek, J. Abbotts, J.D. Roberts, S.H. Wilson, T.A. Kunkel, Specificity and mechanism of error-prone replication by human immunodeficiency virus-1 reverse transcriptase, J. Biol. Chem. 264 (1989) 16948–16956.
[53] A. Efstratiadis, J.W. Posakony, T. Maniatis, R.M. Lawn, C. O'connell, R.A. Spritz, J.K. Deriel, B.G. Forget, S.M. Weissman, J.L. Slightom, A.E. Blechl, O. Smithies, F.E.

Baralle, C.C. Shoulders, N.J. Proudfoot, The structure and evolution of the human β-globin gene family, Cell 21 (1980) 653–668.

[54] A.M. Albertini, M. Hofer, M.P. Calos, J.H. Miller, On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions, Cell 29 (1982) 319–328.

[55] A.M. Albertini, M. Hofer, M.P. Calos, T.D. Tlsty, J.H. Miller, Analysis of spontaneous deletions and gene amplification in the *lac* region of *E. coli*, Cold Spring Harb. Symp. Quant Biol. 47 (Part 2) (1983) 841–850.

[56] S.D. Ehrlich, H. Bierne, E. D'alencon, D. Vilette, M. Petranovic, P. Noirot, B. Michel, Mechanisms of illegitimate recombination, Gene 135 (1993) 161–166.

[57] H.T. Tran, N.P. Degtyareva, N.N. Koloteva, A. Sugino, H. Masumoto, D.A. Gordenin, M.A. Resnick, Replication slippage between distant short repeats in *Saccharomyces cerevisiae* depends on the direction of replication and the *RAD50* and *RAD52* genes, Mol. Cell. Biol. 15 (1995) 5607–5617.

[58] L.S. Ripley, Model for the participation of quasi-palindromic DNA sequences in frameshift mutation, Proc. Natl. Acad. Sci. U.S.A. 79 (1982) 4128–4132.

[59] L.S. Ripley, Concerted mutagenesis:its potential impact on interpretation of evolutionary relationships, in: J. Klein, D. Klein (Eds.), Molecular Evolution of the Major Histocompatibility Complex, Springer-Verlag, Berlin, 1991, pp. 63–94.

[60] N.A. Kolchanov, V.V. Solovyov, I.B. Rogozin, Peculiarities of immunoglobulin gene structures as a basis for somatic mutation emergence, FEBS Lett. 214 (1987) 87–91.

[61] G.B. Golding, P.J. Gearhart, B.W. Glickman, Patterns of somatic mutations in immunoglobulin variable genes, Genetics 115 (1987) 169–176.

[62] G.B. Golding, B.W. Glickman, Sequence-directed mutagenesis: evidence from a phylogenetic history of human α-interferon genes, Proc. Natl. Acad. Sci. U.S.A. 82 (1985) 8577–8581.

[63] N.A. Kolchanov, I.B. Rogozin, V.V. Solovyev, Theoretical analysis of mechanisms of spontaneous and induced mutations in DNA based on repeated sequences, Genetika 25 (1989) 1690–1698.

[64] D.M. Lilley, In vivo consequences of plasmid topology, Nature 292 (1981) 380–382.

[65] R.R. Sinden, G.X. Zheng, R.G. Brankamp, K.N. Allen, On the deletion of inverted repeated DNA in *E. coli*: effects of length, thermal stability, and cruciform formation in vivo, Genetics 129 (1991) 991–1005.

[66] I.B. Rogozin, V.V. Solovyov, N.A. Kolchanov, Somatic hypermutagenesis in immunoglobulin genes. I. Correlation between somatic mutations and repeats. Somatic mutation properties and clonal selection, Biochim. Biophys. Acta 1089 (1991) 175–182.

[67] M.J. Horsfall, A.J. Gordon, P.A. Burns, M. Zielenska, G.M. Van Der Vliet, B.W. Glickman, Mutational specificity of alkylating agents and the influence of DNA repair, Environ. Mol. Mutagen. 15 (1990) 107–122.

[68] A. Sugino, J.W. Drake, Modulation of mutation rates in bacteriophage T4 by a base-pair change a dozen nucleotides removed, J. Mol. Biol. 176 (1984) 239–249.

[69] N.N. Khromov-Borisov, I.B. Rogozin, J.A. Pegas Henriques, F.J. De Serres, Similarity pattern analysis in mutational distributions, Mutat. Res. 430 (1999) 55–74.

[70] I.B. Rogozin, V.B. Berikov, E.A. Vasiunina, O.I. Sinitsina, Study of the DNA primary structure effect on induction of mutations by alkylating agents, Genetika 37 (2001) 854–861.

[71] D.N. Cooper, H. Youssoufian, The CpG dinucleotide and human genetic disease, Hum. Genet. 78 (1988) 151–155.

[72] M. Krawczak, E.V. Ball, D.N. Cooper, Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes, Am. J. Hum. Genet. 63 (1998) 474–488.

[73] M.C. Kricker, J.W. Drake, M. Radman, Duplication-targeted DNA methylation and mutagenesis in the evolution of eukaryotic chromosomes, Proc. Natl. Acad. Sci. U.S.A. 89 (1992) 1075–1079.

[74] Y.H. You, C. Li, G.P. Pfeifer, Involvement of 5-methylcytosine in sunlight-induced mutagenesis, J. Mol. Biol. 293 (1999) 493–503.

[75] M.D. Topal, J.S. Eadie, M. Conrad, O^6-Methylguanine mutation and repair is nonuniform. Selection for DNA most interactive with O^6-methylguanine, J. Biol. Chem. 261 (1986) 9879–9885.

[76] W.W. Piegorsch, A.J. Bailer, Statistical approaches for analyzing mutational spectra: some recommendations for categorical data, Genetics 136 (1994) 403–416.

[77] G.B. Glazko, L. Milanesi, I.B. Rogozin, The subclass approach for mutational spectrum analysis: application of the SEM algorithm, J. Theor. Biol. 192 (1998) 475–487.

[78] W.T. Adams, T.R. Skopek, Statistical test for the comparison of samples from mutational spectra, J. Mol. Biol. 194 (1987) 391–396.

[79] N.F. Cariello, W.W. Piegorsch, W.T. Adams, T.R. Skopek, Computer program for the analysis of mutational spectra: application to *p53* mutations, Carcinogenesis 15 (1994) 2281–2285.

[80] V.N. Babenko, I.B. Rogozin, Use of a rank correlation coefficient for comparing mutational spectra, Biofizika 44 (1999) 632–638.

[81] P.D. Lewis, J.M. Parry, An exploratory analysis of multiple mutation spectra, Mutat. Res. 518 (2002) 163–180.

[82] D.B. Dunson, K.R. Tindall, Bayesian analysis of mutational spectra, Genetics 156 (2000) 1411–1418.

[83] W.W. Piegorsch, K.A. Richwine, Large-sample pairwise comparisons among multinomial proportions with an application to analysis of mutant spectra, J. Agric. Biol. Environ. 6 (2001) 305–325.

[84] C. Coulondre, J.H. Miller, P.J. Farabaugh, W. Gilbert, Molecular basis of base substitution hotspots in *Escherichia coli*, Nature 274 (1978) 775–780.

[85] W.H. Day, F.R. McMorris, Threshold consensus methods for molecular sequences, J. Theor. Biol. 159 (1992) 481–489.

[86] W.H. Day, F.R. McMorris, Critical comparison of consensus methods for molecular sequences, Nucleic Acids Res. 20 (1992) 1093–1099.

[87] I.B. Rogozin, N.E. Sredneva, N.A. Kolchanov, Computer system for analysis of molecular mechanisms of mutagenesis, in: N.A. Kolchanov, H.A. Lim (Eds.), Computer Analysis of Genetic Macromolecules, World Scientific, Singapore, 1994, pp. 265–288.
[88] M.A. Pozdniakov, I.B. Rogozin, V.N. Babenko, N.A. Kolchanov, Neighboring base effect on emergence of spontaneous mutations in human pseudogenes, Dokl. Akad. Nauk. 356 (1997) 566–568.
[89] G.D. Stormo, T.D. Schneider, L. Gold, Quantitative analysis of the relationship between nucleotide sequence and functional activity, Nucleic Acids Res. 14 (1986) 6661–6679.
[90] V.B. Berikov, I.B. Rogozin, Regression trees for analysis of mutational spectra in nucleotide sequences, Bioinformatics 15 (1999) 553–562.
[91] D.S. Smith, G. Creadon, P.K. Jena, J.P. Portanova, B.L. Kotzin, L.J. Wysocki, Di- and trinucleotide target preferences of somatic mutagenesis in normal and autoreactive B cells, J. Immunol. 156 (1996) 2642–2652.
[92] C. Milstein, M.S. Neuberger, R. Staden, Both DNA strands of antibody genes are hypermutation targets, Proc. Natl. Acad. Sci. U.S.A. 95 (1998) 8791–8794.
[93] T.B. Kepler, S. Bartl, Plasticity under somatic mutation in antigen receptors, Curr. Top. Microbiol. Immunol. 229 (1998) 149–162.
[94] G.S. Shapiro, K. Aviszus, D. Ikle, L.J. Wysocki, Predicting regional mutability in antibody V genes based solely on di- and trinucleotide sequence composition, J. Immunol. 163 (1999) 259–268.
[95] M. Oprea, L.G. Cowell, T.B. Kepler, The targeting of somatic hypermutation closely resembles that of meiotic mutation, J. Immunol. 166 (2001) 892–899.
[96] M. Diaz, J. Velez, M. Singh, J. Cerny, M.F. Flajnik, Mutational pattern of the nurse shark antigen receptor gene (*NAR*) is similar to that of mammalian Ig genes and to spontaneous mutations in evolution: the translesion synthesis model of somatic hypermutation, Int. Immunol. 11 (1999) 825–833.
[97] R.D. Blake, S.T. Hess, J. Nicholson-Tuell, The influence of nearest neighbors on the rate and pattern of spontaneous point mutations, J. Mol. Evol. 34 (1992) 189–200.
[98] K. Yoshikawa, I.M. Okazaki, T. Eto, K. Kinoshita, M. Muramatsu, H. Nagaoka, T. Honjo, AID enzyme-induced hypermutation in an actively transcribed gene in fibroblasts, Science 296 (2002) 2033–2036.
[99] L. Pasqualucci, A. Migliazza, N. Fracchiolla, C. William, A. Neri, L. Baldini, R.S. Chaganti, U. Klein, R. Kuppers, K. Rajewsky, R. Dalla-Favera, *BCL-6* mutations in normal germinal center B cells: evidence of somatic hypermutation acting outside Ig loci, Proc. Natl. Acad. Sci. U.S.A. 95 (1998) 11816–11821.
[100] I.B. Rogozin, N.E. Sredneva, N.A. Kolchanov, Somatic hypermutagenesis in immunoglobulin genes. III. Somatic mutations in the chicken light chain locus, Biochim. Biophys. Acta 1306 (1996) 171–178.
[101] J.F. Hasson, E. Mougneau, F. Cuzin, M. Yaniv, Simian virus 40 illegitimate recombination occurs near short direct repeats, J. Mol. Biol. 177 (1984) 53–68.
[102] M.S. Gelfand, Prediction of function in DNA sequence analysis, J. Comput. Biol. 2 (1995) 87–115.
[103] L. Milanesi, I.B. Rogozin, Prediction of human gene structure, in: M.J. Bishop (Ed.), Guide to Human Genome Computing, Academic Press, Cambridge, 1998, pp. 215–259.
[104] G. Pesole, M. Attimonelli, C. Saccone, Linguistic analysis of nucleotide sequences: algorithms for pattern recognition and analysis of codon strategy, Methods Enzymol. 266 (1996) 281–294.
[105] R. Staden, Searching for patterns in protein and nucleic acid sequences, Methods Enzymol. 183 (1990) 193–211.
[106] G.D. Stormo, Computer methods for analyzing sequence recognition of nucleic acids, Annu. Rev. Biophys. Biophys. Chem. 17 (1988) 241–263.
[107] N.A. Kolchanov, I.B. Rogozin, Contribution of nucleotide context to spontaneous and induced mutations, in: N.A. Kolchanov, H.A. Lim (Eds.), Computer Analysis of Genetic Macromolecules, World Scientific, Singapore, 1994, pp. 278–288.
[108] H. Tang, R.C. Lewontin, Locating regions of differential variability in DNA and protein sequences, Genetics 153 (1999) 485–495.
[109] P. Morozov, T. Sitnikova, G. Churchill, F.J. Ayala, A. Rzhetsky, A new method for characterizing replacement rate variation in molecular sequences. Application of the Fourier and wavelet models to *Drosophila* and mammalian proteins, Genetics 154 (2000) 381–395.
[110] N. Michael, T.E. Martin, D. Nicolae, N. Kim, K. Padjen, P. Zhan, H. Nguyen, C. Pinkert, U. Storb, Effects of sequence and structure on the hypermutability of immunoglobulin genes, Immunity 16 (2002) 123–134.
[111] A. Faili, S. Aoufouchi, Q. Gueranger, C. Zober, A. Leon, B. Bertocci, J.C. Weill, C.A. Reynaud, AID-dependent somatic hypermutation occurs as a DNA single-strand event in the BL2 cell line, Nat. Immunol. 3 (2002) 815–821.
[112] E. Dogliotti, P. Hainaut, T. Hernandez, M. D'errico, D.M. DeMarini, Mutation spectra resulting from carcinogenic exposure: from model systems to cancer-related genes, Recent Results Cancer Res. 154 (1998) 97–124.
[113] U. Storb, E.L. Klotz, J. Hackett, K. Kage, G. Bozek, T.E. Martin, A hypermutable insert in an immunoglobulin transgene contains hotspots of somatic mutation and sequences predicting highly stable structures in the RNA transcript, J. Exp. Med. 188 (1998) 689–698.
[114] I.B. Rogozin, Y.I. Pavlov, K. Bebenek, T. Matsuda, T.A. Kunkel, Somatic mutation hotspots correlate with DNA polymerase η error spectrum, Nat. Immunol. 2 (2001) 530–536.
[115] W.S. Watkins, M. Bamshad, M.E. Dixon, B. Bhaskara Rao, J.M. Naidu, P.G. Reddy, B.V. Prasad, P.K. Das, P.C. Reddy, P.B. Gai, A. Bhanu, Y.S. Kusuma, J.K. Lum, P. Fischer, L.B. Jorde, Multiple origins of the mtDNA 9-bp deletion in populations of South India, Am. J. Phys. Anthropol. 109 (1999) 147–158.
[116] Y.G. Yao, W.S. Watkins, Y.P. Zhang, Evolutionary history of the mtDNA 9-bp deletion in Chinese populations and its

relevance to the peopling of east and southeast Asia, Hum. Genet. 107 (2000) 504–512.
[117] R.Y. Denq, I. Fridovich, Formation of endonuclease III-sensitive sites as a consequence of oxygen radical attack on DNA, Free Radic. Biol. Med. 6 (1989) 123–129.
[118] R. Adelman, R.L. Saul, B.N. Ames, Oxidative damage to DNA: relation to species metabolic rate and life span, Proc. Natl. Acad. Sci. U.S.A. 85 (1988) 2706–2708.
[119] B.N. Ames, Endogenous DNA damage as related to cancer and aging, Mutat. Res. 214 (1989) 41–46.
[120] M.L. Michaels, J. Tchou, A.P. Grollman, J.H. Miller, A repair system for 8-oxo-7,8-dihydrodeoxyguanine, Biochemistry 31 (1992) 10964–10968.
[121] Y.I. Pavlov, D.T. Minnick, S. Izuta, T.A. Kunkel, DNA replication fidelity with 8-oxodeoxyguanosine triphosphate, Biochemistry 33 (1994) 4695–4701.
[122] M.L. Michaels, J.H. Miller, The GO system protects organisms from the mutagenic effect of the spontaneous lesion 8-hydroxyguanine (7,8-dihydro-8-oxoguanine), J. Bacteriol. 174 (1992) 6321–6325.
[123] H. Maki, M. Sekiguchi, MutT protein specifically hydrolyses a potent mutagenic substrate for DNA synthesis, Nature 355 (1992) 273–275.
[124] R.G. Fowler, R.M. Schaaper, The role of the *mutT* gene of *E. coli* in maintaining replication fidelity, FEMS Microbiol. Rev. 21 (1997) 43–54.
[125] M.L. Tassotto, C.K. Mathews, Assessing the metabolic function of the *MutT* 8-oxodeoxyguanosine triphosphatase in *E. coli* by nucleotide pool analysis, J. Biol. Chem. 277 (2002) 15807–15812.
[126] A.R. Oller, R.M. Schaaper, Spontaneous mutation in *E. coli* containing the *dnaE911* DNA polymerase antimutator allele, Genetics 138 (1994) 263–270.
[127] T. Gojobori, W.H. Li, D. Graur, Patterns of nucleotide substitution in pseudogenes and functional genes, J. Mol. Evol. 18 (1982) 360–369.
[128] Z. Hatahet, M. Zhou, L.J. Reha-Krantz, S.W. Morrical, S.S. Wallace, In search of a mutational hotspot, Proc. Natl. Acad. Sci. U.S.A. 95 (1998) 8556–8561.
[129] S. Tonegawa, Somatic generation of antibody diversity, Nature 302 (1983) 575–581.
[130] M.S. Neuberger, C. Milstein, Somatic hypermutation, Curr. Opin. Immunol. 7 (1995) 248–254.
[131] U. Storb, Progress in understanding the mechanism and consequences of somatic hypermutation, Immunol. Rev. 162 (1998) 5–11.
[132] C. Rada, G.T. Williams, H. Nilsen, D.E. Barnes, T. Lindahl, M.S. Neuberger, Immunoglobulin isotype switching is inhibited and somatic hypermutation perturbed in *UNG*-deficient mice, Curr. Biol. 12 (2002) 1748–1755.
[133] P.J. Gearhart, R.D. Wood, Emerging links between hypermutation of antibody genes and DNA polymerases, Nat. Rev. Immunol. 1 (2001) 187–192.
[134] J. Spencer, M. Dunn, D.K. Dunn-Walters, Characteristics of sequences around individual nucleotide substitutions in IgVH genes suggest different GC and AT mutators, J. Immunol. 162 (1999) 6596–6601.
[135] I.B. Rogozin, Y.I. Pavlov, T.A. Kunkel, Response 1 to "smaller role for pol η?", Nat. Immunol. 2 (2001) 983–984.
[136] X. Zeng, D.B. Winter, C. Kasmer, K.H. Kraemer, A.R. Lehmann, P.J. Gearhart, DNA polymerase η is an A–T mutator in somatic hypermutation of immunoglobulin variable genes, Nat. Immunol. 2 (2001) 537–541.
[137] C. Rada, M.R. Ehrenstein, M.S. Neuberger, C. Milstein, Hotspot focusing of somatic hypermutation in *MSH2*-deficient mice suggests two stages of mutational targeting, Immunity 9 (1998) 135–141.
[138] Y.I. Pavlov, I.B. Rogozin, A.P. Galkin, A.Y. Aksenova, F. Hanaoka, C. Rada, T.A. Kunkel, 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. U.S.A. 99 (2002) 9954–9959.
[139] E. Hsu, Mutation, selection, and memory in B lymphocytes of exothermic vertebrates, Immunol. Rev. 162 (1998) 25–36.
[140] L. Bross, Y. Fukita, F. Mcblane, C. Demolliere, K. Rajewsky, H. Jacobs, DNA double-strand breaks in immunoglobulin genes undergoing somatic hypermutation, Immunity 13 (2000) 589–597.
[141] F.N. Papavasiliou, D.G. Schatz, Cell-cycle-regulated DNA double-stranded breaks in somatic hypermutation of immunoglobulin genes, Nature 408 (2000) 216–221.
[142] I.B. Rogozin, T.A. Kunkel, Y.I. Pavlov, Double-strand breaks in DNA during somatic hypermutation of Ig genes: cause or consequence? Trends Immunol. 23 (2002) 12–13.
[143] V. Poltoratsky, M.F. Goodman, M.D. Scharff, Error-prone candidates vie for somatic mutation, J. Exp. Med. 192 (2000) F27–F30.
[144] A. Martin, P.D. Bardwell, C.J. Woo, M. Fan, M.J. Shulman, M.D. Scharff, Activation-induced cytidine deaminase turns on somatic hypermutation in hybridomas, Nature 415 (2002) 802–806.
[145] S.K. Petersen-Mahrt, R.S. Harris, M.S. Neuberger, AID mutates *E. coli* suggesting a DNA deamination mechanism for antibody diversification, Nature 418 (2002) 99–103.
[146] T.A. Kunkel, Y.I. Pavlov, K. Bebenek, Functions of human DNA polymerases η, κ and ι suggested by their properties, including fidelity with undamaged templates, DNA Repair 2 (2003) 135–149.
[147] J. Bachl, C. Steinberg, M. Wabl, Critical test of hotspot motifs for immunoglobulin hypermutation, Eur. J. Immunol. 27 (1997) 3398–3403.
[148] P.C. Hanawalt, Preferential DNA repair in expressed genes, Environ. Health Perspect. 76 (1987) 9–14.
[149] I. Mellon, G. Spivak, P.C. Hanawalt, Selective removal of transcription-blocking DNA damage from the transcribed strand of the mammalian *DHFR* gene, Cell 51 (1987) 241–249.
[150] X. Veaute, R.P. Fuchs, Greater susceptibility to mutations in lagging strand of DNA replication in *E. coli* than in leading strand, Science 261 (1993) 598–600.
[151] I.J. Fijalkowska, P. Jonczyk, M.M. Tkaczyk, M. Bialoskorska, R.M. Schaaper, Unequal fidelity of leading strand and lagging strand DNA replication on the *E. coli*

chromosome, Proc. Natl. Acad. Sci. U.S.A. 95 (1998) 10020–10025.
[152] M. Radman, DNA replication: one strand may be more equal, Proc. Natl. Acad. Sci. U.S.A. 95 (1998) 9718–9719.
[153] J.R. Lobry, Asymmetric substitution patterns in the two DNA strands of bacteria, Mol. Biol. Evol. 13 (1996) 660–665.
[154] T. Boulikas, Evolutionary consequences of nonrandom damage and repair of chromatin domains, J. Mol. Evol. 35 (1992) 156–180.
[155] S. Zhang, B.W. Glickman, J.G. De Boer, Spontaneous mutation of the *lacI* transgene in rodents: absence of species, strain, and insertion-site influence, Environ. Mol. Mutagen. 37 (2001) 141–146.
[156] S.J. Foster, T. Dorner, P.E. Lipsky, Somatic hypermutation of VκJκ rearrangements: targeting of RGYW motifs on both DNA strands and preferential selection of mutated codons within RGYW motifs, Eur. J. Immunol. 29 (1999) 4011–4021.
[157] S.A. Ahrendt, S. Halachmi, J.T. Chow, L. Wu, N. Halachmi, S.C. Yang, S. Wehage, J. Jen, D. Sidransky, Rapid *p53* sequence analysis in primary lung cancer using an oligonucleotide probe array, Proc. Natl. Acad. Sci. U.S.A. 96 (1999) 7382–7387.
[158] A.J. Levine, p53, the cellular gatekeeper for growth and division, Cell 88 (1997) 323–331.
[159] D.P. Guimaraes, P. Hainaut, *TP53*: a key gene in human cancer, Biochimie 84 (2002) 83–93.
[160] M. Nei, S. Kumar, Molecular Evolution and Phylogenetics, Oxford University, Oxford, 2001.
[161] L.Y. Yampolsky, A. Stoltzfus, Bias in the introduction of variation as an orienting factor in evolution, Evol. Dev. 3 (2001) 73–83.
[162] M. Mitas, Trinucleotide repeats associated with human disease, Nucleic Acids Res. 25 (1997) 2245–2254.
[163] G. Kotturi, J.G. De Boer, B.F. Koop, B.W. Glickman, Correlation of UV-induced mutational spectra and the in vitro damage distribution at the human *hprt* gene, Mutat. Res. 403 (1998) 237–248.
[164] I.B. Lambert, T.A. Chin, D.W. Bryant, A.J. Gordon, B.W. Glickman, D.R. Mccalla, The mutational specificity of 2-(2-furyl)-3-(5-nitro-2-furyl)-acrylamide (AF2) in the *lacI* gene of *E. coli*, Carcinogenesis 12 (1991) 29–34.
[165] C. Papanicolaou, L.S. Ripley, Polymerase-specific differences in the DNA intermediates of frameshift mutagenesis. In vitro synthesis errors of *E. coli* DNA polymerase I and its large fragment derivative, J. Mol. Biol. 207 (1989) 335–353.
[166] R.M. Schaaper, Mechanisms of mutagenesis in the *E. coli* mutator *mutD5*: role of DNA mismatch repair, Proc. Natl. Acad. Sci. U.S.A. 85 (1988) 8126–8130.
[167] J. Hauser, A.S. Levine, K. Dixon, Unique pattern of point mutations arising after gene transfer into mammalian cells, EMBO J. 6 (1987) 63–67.
[168] J. Jurka, P. Klonowski, Integration of retroposable elements in mammals: selection of target sites, J. Mol. Evol. 43 (1996) 685–689.
[169] P. Lopez, M. Espinosa, B. Greenberg, S.A. Lacks, Generation of deletions in pneumococcal *mal* genes cloned in *Bacillus subtilis*, Proc. Natl. Acad. Sci. U.S.A. 81 (1984) 5189–5193.