**REGRESSION ANALYSIS OF MUTATIONAL SPECTRA.**

ROGOZIN I.B.+, BERIKOV V.B.1

Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences 10 Lavrentyev Ave., Novosibisk 630090, Russia;
e-mail: rogozin@bionet.nsc.ru;

1Institute of Mathematics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia;
e-mail: berikov@math.nsc.ru;

+Corresponding author

Keywords: nucleotide sequences, mutational spectra, mutational hotspots, DNA context

Novel gene engineering techniques have revealed much information on the spontaneous and induced mutations observed in nucleotide sequences. This kind of data is called "mutational spectra". Analysis of these spectra has shown mutations to be largely confined to certain regions of nucleotide sequences. Examination of such regions (mutational "hotspots") provides evidence that these hotspots arise due to some structural features of hotspot subsequences (the DNA context). Thus the mutability varies significantly along nucleotide sequences. The idea that DNA context may affect mutability was first suggested by Benzer [1]. There are several context features that can influence DNA damage and repair: polytracts, specific motifs, potential zDNA structure, cruciform structures, etc. In many cases, mutational hotspots emerge due to neighbouring nucleotides. It should be emphasized that the context of hotspots can help to specify the underlying molecular mechanism of mutagenesis [2].

To investigate the influence of neighbouring bases on mutagenesis, it is necessary to analyze both the positions where mutations have been observed and the neighboring nucleotides. In this case, the position of a mutation with adjacent nucleotides will be considered a mutational site (Fig. 1).

 4 12 2

 G A A A C C A G T A A c G T T A T A C G A T G T C g c A G A

 - - \* - - - - \* - -

 - - \* - -

 2 0 3 1

 G T A T G C C G G T G T C T c T T A T c A G A c C G T T T c

 - - \* - - - - \* - - - - \*

 - - \* - -

 9 8 0 0 2

 C c g C g T G G T G A A C c A G G C C A G C C A C G T T T c

 - \* - - - - \* - - - - \*

 - - \* - -

 - - \* - -

 0 8

 T G C G A A A A c G C G G G A A A A A g T G G A A G C G G C

 - - - - \* - - - - \* - -

 10

 G A T G G C G G A G C T G A A T T A C A T T C c C A A C C G

 - - \* - -

 11 1 0 2 0 4 1

 C G T G g c A c A A c A A C T G G c G G g C A A A c A G T C

 - - \* - - - - \* - - - - \* - - - - \* - -

 - - \* - - - - \* - -

 - - \* - -

Figure 1. The sequence of the lacI gene with mutations induced by EMS is taken from [3]. Number above the sequence stands for the number of mutations at the position. 0 implies that no mutations were observed at the site. Sites detected previously in the lacI gene are shown by lowercase letters. Mutation sites are underlined (in this case, the site length is 5 nucleotides). Asterisk designates mutational position at each site.

Several mutational sites can be aligned with respect to the mutation position and represented as a set of aligned sequences (Fig. 2a). The usual way for analysis of a set of aligned

sequences is the construction of a consensus. However, one can use the consensus to describe the aligned sites given that the site under consideration is highly conservative at a majority of positions (the set of sites is homogeneous). If the set is heterogeneous, (that is, involves a number of homogenous subsets), using the consensus will lead to a rough averaging of positions. A method has been developed for constructing a set of consensus sequences [4]. The main purpose of the method is to construct a set of consensus sequences matching all hotspot sites (Fig. 2b). Each consensus matches some sites, and mutations should be distributed evenly throughout these sites (all differences between them should be owing to random reasons). This method was applied for analysis of somatic mutations in immunoglobulin V genes, for which two different consensus sequences, RgYW and TaA, for mutational hotspots were constructed [4].

In this work we describe a new approach for mutational spectra classification based on "regression trees". Although there are a lot of "classical" methods of regression analysis, regression tree [5,6] approach has a number of properties, which makes it very useful for mutational spectrum analysis:

* this approach allows to utilise mutational spectrum characteristics of heterogeneous nature: qualitative and quantitative;
* it makes possible to work under condition of high uncertainty (limited data size; absence of*a priori*information about distributions);
* a regression tree represents hierarchical logical-and-probability model of mutational process.

For fast analysis a modification of dynamic programming method for regression tree design was applied, which earlier had been suggested for the design of decision trees in the problem of pattern recognition [7].

Example of mutational spectrum classification is shown in Fig.2b. One can see that Each consensus matches some sites, and mutations should be distributed evenly throughout these sites (all differences between them should be owing to random reasons). Homogeneity was assessed with the Monte Carlo test with X2 statistic [8]. For example, the differences between the number of mutations in the sites corresponding to the SgSR consensus (12, 11, 10, 9 and 4) can be attributed to chance (Fig.2b). Each consensus is characterised by the average number of mutation, and "hotter" consensus sequences correspond to larger average number of mutations.

a) b)

Posi- Site Number of Posi- Site Number of

tion sequence mutations tion sequence mutations

 - - \* - - in the site - - \* - - in the site

 42 A C g T T 4 56 T C g C A 12

 56 T C g C A 12 92 G C g G G 9

 57 C T g C G 2 93 C C g C G 8

 75 A A g A G 2 174 T G g G A 10

 80 C T g A T 0 185 T G g C A 11

 84 C G g T C 3 201 G G g C A 4

 90 G G g A A 1 ------------> --------- ---

 92 G C g G G 9 Consensus C1 N S g S R 9.0

 93 C C g C G 8 set

 95 G C g T G 0 construction

104 C T g G T 0

120 C A g A A 2 75 A A g A G 2

129 G C g T T 0 140 A A g T G 8

140 A A g T G 8 --------- ---

174 T G g G A 10 C2 N A g N G 5.0

185 T G g C A 11

186 G T g C C 1

188 T T g T G 0

191 T T g T T 2

198 C C g C C 0

201 G G g C A 4

206 C T g T T 1

Figure 2. Example of applying the classification approach to revealing non random hotspot consensus sequences. Positions with more than 7 mutations are hotspots. The sequence with mutations is shown in Fig. 1.

a) A sample of sites aligned with respect to mutational positions. Number of mutations in the site stands for the number of mutations observed in the central position of the site (phasing position); mutations in other positions of the given site are omitted from consideration.

b) The set of two consensus sequences (C1 and C2) constructed for the sample of mutation sites. Mutation position in consensus is shown by small letter.

To test real biological data, 11 mutational spectra induced by Sn1 alkylating agents in the lacI gene of*Escherichia coli*were analyzed. The role of the context in the incidence of these mutations is well known. Most of the induced mutations occurred at the G:C positions (transitions G:C->A:T). Mutations at the positions with the RG sites (R is A or G, G in the second position is the mutation position) occurred several times as frequently as mutations at the YG sites (Y is T or C, G in the second position is the mutation position) [9]. The analysis of real and generated data has proven the efficiency of the algorithm suggested for analysis of mutational spectra. The results of the real data are in good relation with existing knowledge about context specificity of mutations.

This work was supported by grants from the Russian Fund of Fundamental Research (grant N 96-04-49957) and the Russian State Program "Frontiers in Genetics".

**References**

1. S. Benzer, "On the topology of the genetic fine structure" Proc. Natl. Acad. Sci. USA 47, 403 (1961).
2. T. Boulinkas, "Evolutionary consequences of nonrandom damage and repair of chromatin domains" J. Mol. Evol. 35, 156 (1992).
3. J.A. Halliday, M. Zielenska, S.S. Awadallah and B.W. Glickman, "Colony hybridisation in Escherichia coli: a rapid procedure for determining the distribution of specific classes of mutations among a number of preselected sites" Envirom. Mol. Mutagen. 16, 143 (1990).
4. I.B. Rogozin and N.A.Kolchanov, "Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis" Biochim. Biophys. Acta 1171, 11 (1992).
5. L. Breiman, J. Friedman, R. Olshen and C. Stone, "Classification and regression trees" (Wardsworth Int. Group, Belmont, California, 1984).
6. P. Chou, "Optimal partitioning for classification and regression trees" IEEE Trans. Pattern Anal. and Mach. Intell. 13, 340 (1991).
7. V. Berikov and G. Lbov, "Recursive Method of Formation of the Recognition Decision Rule in the Class of Logical Functions" Pattern Recognition and Image Analysis 3, 428 (1993).
8. W.W. Piegorsch and A.J.Bailer, "Statistical approaches for analyzing mutational spectra: some recomendations for categorial data" Genetics 136, 403 (1994).
9. M.J. Horsfall., M.J., A.J.E. Gordon, P.A. Burns, M. Zielenska, G.M.E. van der Vliet and B.W. Glickman, (1990) Mutational specificity of alkylating agents and the influence of DNA repair, Envirom. Mol. Mutagen., 15, 107-122 (1990).