Subclass approach for mutational spectrum analysis

Igor B. Rogozin, Galina V. Glasko, Evgeniy I. Latkin*

Institute of Cytology and Genetics; 10.Lavrentyev Ave., Novosibirsk 630090, Russia
*RIMIBE at the Novosibirsk State University; 2. Pirogova St., Novosibirsk 630090, Russia
rogozin@cgi.nsk.su

Abstract

Analysis and comparison of mutational spectra represents a burning question in molecular biology. We report an algorithm based upon the SEM subclass approach (SEM - stochastique, estimations, maximizations). Any real mutational spectrum is regarded as a mixture of standard binomial distributions. The separation procedure is run by rounds. Each iteration includes simulation, maximization and estimation. The algorithm has been checked on random spectra with the preset parameters and on real mutational spectra. As has been shown, any real mutational spectrum can be represented as a mixture of two and more binomial distributions, of which one contains hotspots of mutation.

Keywords: *subclass approach, classification, mutational spectrum, SEM algorithm, maximization, likelihood function.*

Introduction

Novel gene engineering techniques have brought up a lot of information on the mutations observed in nucleotide sequences. The data of the sort were called "mutational spectra". Analysis of these spectra has shown spontaneous and induced mutations to be largely confined to certain regions of nucleotide sequences. Examination of such regions (mutational "hotspots") has provided evidence that the observed non random distribution of mutations along the sequence must be accounted for by some structural features of the hotspot subsequences (the DNA context). An example of a mutational spectrum is presented in Fig.1a. As is seen, a range of sites display elevated mutation frequencies (e.g., positions 84, 185, 202).

The idea that DNA context may affect mutability was first suggested by Benzer (1961). To date, evidence has been gathered that this is quite so. Investigation into these features may provide valuable information about the underlying molecular mechanisms. The influence of DNA context may be associated with base pair reactivity (Matters et al. 1986), with the secondary structure of DNA (Glickman and Ripley, 1984), with repair systems (Topal et al. 1986; Burns et al. 1986).
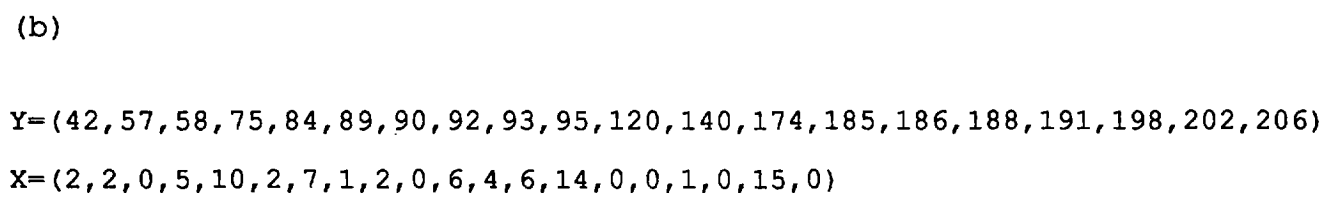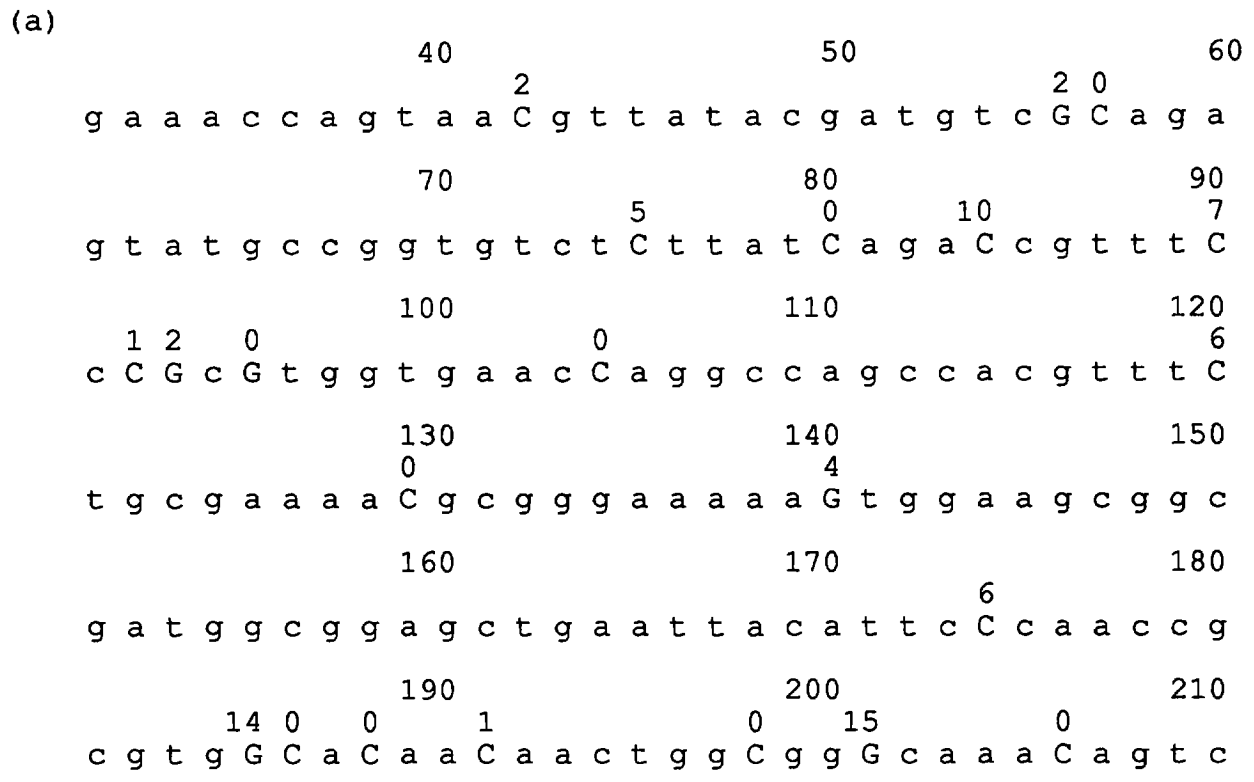
It should be emphasized that the context of hotspots can help specify the underlying molecular mechanism of mutagenesis (Horsfall et all. 1990).

A number of theoretical methods has been developed for mutational spectra analysis. Several approaches have been applied for comparison of two or several mutational spectra (Adams & Skopek 1986; Piegorsch & Bailer 1994). Benigni et al. (1992) used the standard multivariate method. Multiple regression analysis have been applied for construction of the best model for context effects on the 2AP insertion (Stormo et al. 1986). However, the problem of identification of hotspots in the mutational spectra remains unresolved. The mutational spectrum induced by O6-methylguanine was approximated by the Poisson distribution (Topal et al. 1986). By way of that approach, the content of cold spots of mutation (positions with lower mutation frequencies) was analyzed. However, with this approach, one can go as far as to reveal discrepancies between a real and expected mutational spectra, but fail to uncover the causes of these discrepancies (hotspots, coldspots).

To solve the problem, we decided to approximate the real mutational spectrum by a mixture of binomial distributions. We applied the standard classification method, slightly modified though. The program that we have developed was tested on real mutational spectra. As was shown, the mutational positions of the spectra can be grouped into two or more classes, one of which contains hotspots.

Statistical model of the experiment

Consider two sets $Y=(y1,...,yn)$, $X=(x1,...,xn)$ of integers, where yi stands for the number of the position in the sequence at which a mutation(s) can occur, and xi stands for the number of mutations at this position; n - total number of positions. Fig.1 presents a mutational spectrum and the corresponding two sets Y and X.

(a)

```
                40                              50                      60
                 2                               2 0
   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

                70                              80                      90
                           5                     0      10              7
   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

               100                             110                     120
    1 2   0                     0                                       6
   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

               130                             140                     150
               0                               4
   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

               160                             170                     180
                                                        6
   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

               190                             200                     210
      14 0     0     1                   0     15              0
   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
```


(b)

Y=(42,57,58,75,84,89,90,92,93,95,120,140,174,185,186,188,191,198,202,206)

X=(2,2,0,5,10,2,7,1,2,0,6,4,6,14,0,0,1,0,15,0)

FIGURE 1. The sequence of lacI gene with mutations induced by NMAM (Burns et al., 1986) (a) and corresponding X and Y arrays (b). Number above sequence means the number of mutations in the position. 0 means that no mutations were observed in the site. Sites detected previously in lacI gene (Gordon et al., 1988) are shown by uppercase letters.

Let ξ be a random value denoting the number of mutations at any site of the sequence. In that case X is the sampling of values for ξ.

The experiment is briefly described as follows: The known number, N, of copies of the sequence is affected by a mutagene. One or less mutations occurs in each copy. The number of mutations at each positions is summed over all copies and the sum represents the result of the experiment.

Formally, we may assume that N experiments have been made over one sequence. Here pi is the probability of a mutation occurring at the i-th site over one experiment. For the i-th site, N experiments present a succession of independent Bernoulli's probes; and the probability of k successes and (N-k) failures over N experiments is

$$P(\xi=k) = C_N^k \, pi^k \, (1 - pi)^{N-k}$$

The sample of values x1,...,xn for ξ may be regarded as a mixture of l binomial distributions with the different probabilities of mutations for any distribution of the mixture. Thus any distribution is responsible for the operation of certain mechanism of mutagenesis.

By virtue of the formula for the total probability, the probability of xj mutations occurring at the j-th site is

$$P(\xi=xj) = \sum_{i=1}^{l} P(\xi=xj \mid Si) * P(Si)$$

where S_i is the i-th group of sites at each of which a mutation occurs one experiment with the probability p_i; $P(S_i)$ is the probability of the S_i group being selected from among all the groups; l is the number of groups.

Thus the density of distribution is

$$f(x) = \sum_{i=1}^{l} \lambda_i \ C_N^x \ p_i^x \ (1 - p_i)^{N-x}$$

λ_i is the probability of S_i being selected.

In order to assess the number of components and parameters of the mixture of distributions, it necessary to separate the distributions from the mixture.

Algorithm of separation

We used the algorithm SEM (stochastique - estimations - maximizations) (Celeux & Riebolt 1984) for assessing the number of distributions of a mixture with the unknown parameters and weights of each component.

To start with, we set the initial number of distributions in the mixture, kmax, which should be higher than the true number of the component distributions. We also set up the matrix of the posterior probabilities $g_{i1},...,g_{ikmax}$ (i=1,n) such that

$$g_{ij}=1/kmax, \ \sum_{i} g_{ij}=1, \ 0<g_{ij}<1$$

Here g_{ij} is the posterior probability of x_i belonging to classes j=1,2,....,kmax.

Three steps of the algorithm follow in succession.

(i) Simulation

Here we classify the members $x_1,...,x_n$ of the initial sample. By virtue of the matrix G={g_{ij}} (i=1,...,n; j=1,...,k) the matrix E is built; k - number of classes.

$$e_{ij}= \begin{cases} 1 \text{ with the probability } g_{ij} \\ 0 \text{ with the probability } 1-g_{ij} \end{cases}$$

As result, each member of the sample $x_1,...,x_n$ was assigned to one or more classes depending on g_{ij} randomly.

It means that the rows of the matrix E={e_{ij}} are represented by polynomially distributed random values (with the parameters $g_{i1},...,g_{ik}$). The matrix E sets up the following rule by which to classify the members of the sample:

$$S_j=\{x_i : e_j(x_i)=1\}, \text{ at each } j=1...k$$

Over the first iteration, a member of the sample may co-occur in several classes. The class which does not contain any member is omitted from consideration, and the number of classes decreases. The matrix G is recalculated for the new number of classes:

$$g_{ij}=g_{ij} / \sum_{j=1}^{k} g_{ij}$$

(ii) Maximization

Now evaluating the parameters of the classes. First, the weights of the distributions are defined by the following formula:

$$p_j = 1/n \sum_{i=1}^{n} e_j(x_i)$$

Then we evaluate the parameters of the classes by maximization of the likelihood function

$$\sum_{x \in S_j} \ln f(x_i,\theta_j) \rightarrow \sup_{\theta_j} \quad j=1,...,k$$

(iii) Estimation

The Bayes' reestimation of probabilities g_{ij} is based upon the estimates obtained from the preceding iteration

$$\hat{g}_{ij} = \hat{p}_j \ f(x_i,\hat{\theta}_j) \ / \ \sum_{l=1}^{k} \hat{p}_l \ f(x_i,\hat{\theta}_j)$$

The iteration procedure terminates when all members have been classified. If there are members with a high probability of membership in two classes (and therefore no classification is possible for them), the procedure terminates when the difference between the values of the parameters of the i-th and (i+1)-th iteration has fallen short of accuracy.

As was shown by Shlezinger (1964), the algorithm converges to the stationary points of the logarithmic likelihood function used.

Results

We checked our version of the algorithm on random binomial distributions with various numbers of components and parameters. A standard random generator was used.

Fig.2 presents the dependence of variance and the expectation of p_1 for a one- component mixture on sample volume. The original value of p was 0.01. Fifty
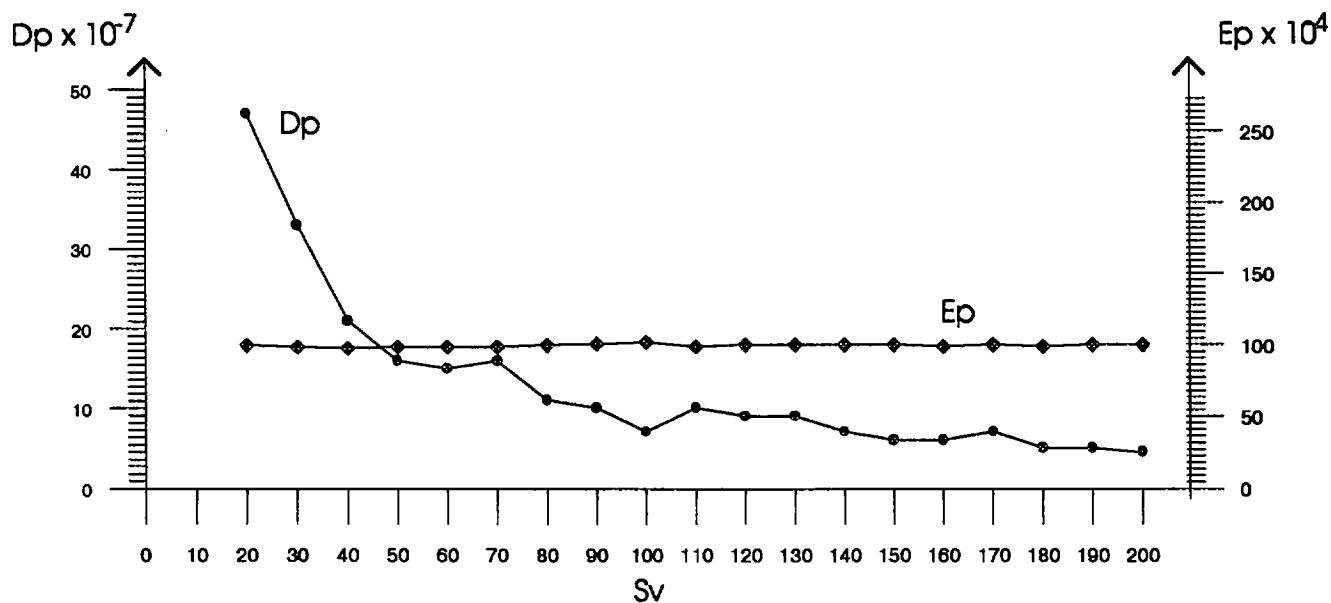
FIGURE 2. The dependence of the variance (Dp*) and the expectation (Ep*) of p1 for a one- component mixture on sample volume Sv. The values of Ep* are given near the corresponding points of the plot.

random distributions were generated for sample volumes of 20 to 200 mutations. For each of the distributions generated, except for four, only one-component distribution was obtained. The four incorrect distributions were omitted from the calculation of the mean and variance. For each sample volume, variance (Dp*) and expectation (Ep*) were evaluated. As is seen, the variance of p1 decreases as $1/\sqrt{Sv}$ with an increase of sample volume Sv.

The algorithm was also tested on a mixture of two binomial distributions at the following parameter values: sample volume N = 100; l1 = l2 = 0.5; p1=0.09, p2 ranged from 0.01 through 0.08. The algorithm was unable to separate two distributions if the difference between P1 and P2 was less than 0.02. In that case, the two distributions were "mingled" and could not be separated significantly.

The algorithm was also tested on real mutational spectra induced by alkylating agents MNAM and MNNG in the E.coli genome. The role of context in the incidence of these mutations is well known. The overwhelming majority of the mutations occurs at the G:C positions. Mutations at the positions with the RG context (G is the mutation position) occur several times as frequently as mutations at the YG positions. Thus, these mutational spectra should represent a mixture of not less that two distributions. The mutational spectrum induced by the mutagene MNAM (Horsfall & Glickman 1988) is presented in Fig.1a. Analysis of this spectrum by the algorithm as described suggests two distributions with the

parameter values N = 100, p1 = 0.005, λ1 = 0.64, p2 = 0.08, λ2 = 0.36. One of the distributions included only the sites with the YG context (number of mutations < 3, positions 42, 57, 58, 92, 93, 186, 188, 191, 198, 206), the other included the spectra with the RG context (number of mutations > 3, positions 75, 84, 90, 120, 140, 174, 185, 202). The mutational spectra induced by MNNG (Burns et al. 1987) was revealed within the 1- 580 bp region of the lacI gene. However, the results obtained are quite similar with what we had obtained for MNAM (within 1-232 bp): one of the distributions included only the sites with the YG context, the other included the spectra with the RG context.The parameter values were evaluated as N = 100, p1 = 0.01, λ1 = 0.61, p2 = 0.1, λ2 = 0.39.

Discussion

In all, the analysis of real and generated data has proven the efficiency of the algorithm suggested for analysis of mutational spectra. For any real spectrum studied, one of the classes contains hotspot sites. Thus, the algorithm may be applied for mutational spectra analysis.

A lot of mutational spectra has now been obtained. For only part of them, the mechanisms of mutagenesis has been revealed. The algorithm reported allows more detailed analysis of any spectrum to be performed. It can be successful at the study of correlations between hotspots and the nucleotide context features within any class of sites revealed. These prospects are pursued by the system for analysis of mutational spectra MutAn (mutation

analysis) (Rogozin et al. 1992) which is currently under development.

The program has been written in C and tested on a PC (MS DOS) computer. The program is freely available from the first author (E.mail: rogozin@cgi.nsk.su).

Acknowledgements

References

Adams, W. T., and Scopek, T. R. 1987. Statistical test for the comparison of samples from mutational spectra. J.Mol.Biol.194: 391-396.

Begnini, R., Palombo, F., and Dogliotti, E. 1992. Multivariate statistical analysis of mutational spectra of alkylating agents. Mutat. Res. 267: 77-88.

Benzer, S. 1961. On the topology of the genetic fine structure. Proc. Natl. Acad. Sci. USA 47: 403-415.

Bolshoy, A., McNamura, P., Harrington, R. E., and Trifonov, E. N. 1991. Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. Proc. Natl. Acad. Sci. USA 88: 2312-2318.

Burns P. A., Gordon, A. J. E., and Glickman, B. W. 1987. Influence of neighbouring base sequence on N-methyl-N'-nitro-N-nitro-soguanidine mutagenesis in the lacI gene of Escherichia coli. J.Mol.Biol. 194: 385-390.

Celeux, Q., and Riebolt, J. 1984. Reconnaissance de melange de densite et classification. Un algorithme d'apprentissage probabiliste: l'algorithme SEM. Rapports de Recerche de l'INRIA. Centre de Rocquencort.

Glickman, B. W., and Ripley, L. S. 1984. Structural intermediates of deletion mutagenesis: a role for palindromic DNA. Proc. Natl. Acad. Sci. USA 81: 512-516.

Horsfall, M. J., and Glickman, B. W. 1988. Mutation site specificity of N-nitroso-N-methyl-N-alpha-acetoxybenzyl-amine: a model derivative of an esophageal carcinogen. Carcinogenesis 9: 1529-1532.

Horsfall., M. J., Gordon, A. J. E., Burns, P. A., Zielenska, M., van der Vliet, G. M. E., and Glickman, B. W. 1990. Mutational specificity of alkylating agents and the influence of DNA repair. Envirom. Mol. Mutagen. 15: 107-122.

Matters, W. B., Hartley, J. A., and Kohn, K. W. 1986. DNA sequence selectivity of guantine-N7 alkytron by nitrogen mustards. Nucleic Acids Res. 14: 2971-2987.

Piegorsch, W. W., and Bailer, A. J. 1994. Statistical approaches for analyzing mutational spectra: some recomendations for categorial data. Genetics 136: 403-416.

Rogozin, I. B., Sredneva, N. E., and Kolchanov, N. A. 1992. Intelligent system of mutational analysis. In Modelling and Computer Methods in Molecular Biology and Genetics (Ratner, V.A., and Kolchanov, N.A., eds.), 63-72. Nova Sience Publishers, Inc., NY.

Shlezinger, M. I. On the spontaneous image recognition. Reading automats. 1965, Naychnaia Misl', Kiev, 38-45 (in Russian).

Stormo, G. D., Schneider, T. D., and Gold, L. 1986. Quantative analysis of the relationship between nucleotide sequence and functional activity. Nucleic Acids Res. 14(16): 6661-6679.

Topal, M. D., Eadie, J. S., and Conrad M. 1986. O6-methylguanine mutation and repair is nonuniform, J. Biol. Chem. 261: 9879-9885.