J. Mol. Biol. (1987) **194**, 391–396

Statistical Test for the Comparison of Samples from Mutational Spectra

W. Thomas Adams† and Thomas R. Skopek

Chemical Industry Institute of Toxicology
Research Triangle Park, NC 27709, U.S.A.

(Received 11 June 1986, and in revised form 20 November 1986)

The Monte Carlo estimate of the p value of the hypergeometric test is described and advocated for the testing of the hypothesis that different treatments induce the same mutational spectrum. The hypergeometric test is a generalization of Fisher's "exact" test for tables with more than two rows and two columns. Use of the test is demonstrated by the analysis of data from the characterization of nonsense mutations in the *lac*I gene of *Escherichia coli*.

Unlike the chi-square test, the hypergeometric test remains valid when applied to sparse cross-classification tables. The hypergeometric test has the most discrimination power of any statistical test that could be employed routinely to compare samples from mutational spectra. Direct application of the hypergeometric test to large cross-classification tables is excessively computation intensive, but estimation of its p value *via* Monte Carlo techniques is practical.

1. Introduction

The analysis of the type and location of DNA alterations induced by mutagens has become a valuable and routine approach in the study of mutagenic mechanisms in *Escheria coli* (Miller, 1983). Several systems exist that permit the determination of the exact nucleotide change in DNA responsible for the appearance of a mutant phenotype. The analysis of a number of independent mutations induced by a given agent or treatment yields information on the relative probability of inducing each type of nucleotide change at each potential site in the DNA. This in turn can then be used to infer information about the mechanism of mutagenesis, to classify mutagens into categories according to the types of changes they induce, and to compare different mutagens or treatment regimens.

The most widely used assay is the *lac*I system in which *lac*I nonsense mutants are characterized genetically to infer the change in DNA sequence (Coulondre & Miller, 1977). Although samples of mutants from more than a dozen treatments have been characterized in this system, no systematic statistical hypothesis testing using this data has been reported.

We present here a statistical model for the analysis of mutational spectra. Based on this model, and some other considerations related to the sample size, we will advocate the use of a particular statistical procedure for testing the hypothesis that two treatment conditions have no differential effect on the spectrum of detectable mutants induced in a gene. Examples using data obtained from the *lac*I system will be presented.

2. A Statistical Model for Mutational Spectra

When a sample of mutant genes is categorized according to the specific DNA changes that have occurred, the result is a $1 \times N$ table where N is the number of different detectable modifications in the gene's DNA sequence. For instance, $N = 36$ in the case of the set of identifiable amber nonsense mutants in *lac*I. This table can be modeled as a sample drawn from a multinomial distribution (Hoel *et al.*, 1971) with parameters (S; p_1, p_2, ..., p_N) where p_i represents the probability that a mutant will fall into category i, and N represents the number of potential categories of mutation, and S represents the sample size.

The term *mutational spectra* has been used somewhat informally in the literature to refer both to the p_i values (i.e. the underlying reality) and to the classification tables resulting from experiments. We prefer to use this term to refer only to the underlying reality, or what statisticians call the population parameters. We will call the particular

† Present address: 12820 Beechwood Ct, Raleigh, NC 27614, U.S.A.

0022-2836/87/070391-06 $03.00/0

table of classified mutants that results from an experiment involving a single treatment a *spectral sample*. Of course, this spectral sample only roughly represents the mutational spectrum. The larger the number of mutants in the spectral sample, the more faithfully it represents the spectrum.

3. Possible Statistical Tests to Compare Spectral Samples

We wish to choose a statistical procedure to test the null hypothesis that two spectral samples were drawn from the same population. If the samples are collected from cultures treated under differing conditions, then this will constitute a test of the null hypothesis that the treatment conditions had no differential effect on the spectrum.

To test this hypothesis, we analyze the $M \times N$ cross-tabulation containing the mutants characterized for M treatments cross-classified by treatment and type/site of mutation. The factor that differs in the treatments could be the mutagenic agent, concentration of the agent, treatment protocol, strain, laboratory, or any other controllable factor of interest. The null hypothesis of no effect can be represented by a set of M identical multinomial distributions, possibly with M different sample sizes.

The traditional tool for testing this null hypothesis is Pearson's chi-square test (see, for instance, Upton, 1978). However, assuring the accuracy of the p value from this test requires certain minimum numbers of mutants in each cell of the cross-tabulation, because the test is based on an approximation that is not valid for small samples. Fienberg (1980) recommends an average cell size of at least ten counts. The small sample adequacy of the chi-square approximation is still an active research topic, and opinions differ on the minimum sample size below which the chi-square is too inaccurate to be useful. Fingleton (1984) indicates that an average cell count of five might be adequate, unless the counts are highly concentrated in a few categories. Simulations by Roscoe & Byars (1971) show that a 40% error can occur in the chi-square significance level when testing a 2×5 table with an average cell count of four in a situation where the counts tended to be concentrated in a few (1 to 3) of the five categories. Most of the spectral samples obtained from *lac*I nonsense mutants and all of the spectral samples obtained thus far by direct sequencing do not meet the most liberal sample size criterion for application of the chi-square test.

There is another statistical test, which we will call the *hypergeometric test*, that is suitable for testing this null hypothesis. The hypergeometric test is a generalization of the well known Fisher's "exact" test (two-tailed) to tables with more than two rows and columns. The hypergeometric test was described by Freeman & Halton (1951). Tocher (1951) proved that the hypergeometric test yields results almost identical with the uniformly most powerful unbiased (UMPU)† test, except when applied to extremely sparse tables. In this context, a cross-classification table would not be considered extremely sparse if each of the spectral samples to be compared had at least one category that contained four or more mutants. The fact that the hypergeometric test is almost identical with the UMPU test means that the hypergeometric test has virtually optimal discrimination power when used to compare spectral samples. (One might ask "Why not just use the UMPU test?" The UMPU test, although of theoretical importance, is not considered to be suitable for routine use because it yields odd results in certain degenerate cases.)

However, the hypergeometric test is extremely computation intensive when applied to tables with sample sizes as large as any of the *lac*I spectral samples thus far published ($n > 40$). A computer subroutine for the hypergeometric test is available in the International Mathematical and Statistical Library (1984). A much more efficient computer program for this test, based on an algorithm developed by Pagano & Halvorsen (1981), is available from Pagano.‡ Neither implementation is fast enough to be applied to cross-tabulations of most of the published spectral samples because of the computation-intensive nature of the algorithms.

Although the hypergeometric test cannot be applied directly, the p value resulting from this test can be estimated using Monte Carlo techniques (Agresti *et al.*, 1979). This estimation procedure is the one we have chosen to use to compare spectral samples.

The algorithm for Monte Carlo estimation of the p value of the hypergeometric test is as follows.

(1) Calculate the hypergeometric probability of the $N \times M$ table representing the mutants observed in the experiments cross-classified by the M treatments and the N types and sites of mutation.

The hypergeometric probability of the observed table, p, is given by the formula

$$p = \frac{\prod_{i=1}^{N} (R_i!) \prod_{j=1}^{M} (C_j!)}{T! \prod_{i=1}^{N} \prod_{j=1}^{M} (X_{ij}!)},$$

where the R_i and C_j values are the row and column marginal totals, T is the total number of observed mutants, and the X_{ij} values are frequencies of mutants in each cell.

For example, the components of this equation for Table 1 are as follows. N indicates the number of rows in the table, i.e. $N = 36$. R_1 through R_{36} would represent the combined number of mutants in each category from both of the spectral samples being compared; for instance, $R_{20} = 14$. $M = 2$,

† Abbreviations used: UMPU, uniformly most powerful unbiased; u.v., ultraviolet; CPPE, 3,4-epoxycyclopenta[*cd*]-pyrene.

‡ M. Pagano, Sidney Farber Cancer Institute, 677 Huntington Ave., Boston MA 02115, U.S.A.

Table 1

Distribution of lac*I amber nonsense mutations induced by CPPE in two strains of* E. coli

Base substitutions	Site	−pKM101	+pKM101
G·C → A·T	A5	0	0
	A6	1	0
	A9	1	0
	A15	0	0
	A16	0	0
	A19	1	0
	A21	1	0
	A23	0	0
	A24	0	0
	A26	0	0
	A31	0	0
	A33	0	0
	A34	2	3
	A35	0	0
G·C → T·A	A2	8	1
	A7	0	5
	A10	1	3
	A12	5	4
	A13	3	2
	A17	10	4
	A20	5	9
	A25	0	0
	A27	0	1
	A28	3	5
A·T → T·A	A11	1	0
	A18	2	1
	A32	0	3
	A36	0	0
A·T → C·G	A3	0	0
	A4	2	0
	A14	0	0
	A22	0	1
	A30	1	0
G·C → C·G	A1	1	0
	A8	0	0
	A29	0	0
Total	36 sites	48	42

Data from Eisenstadt *et al.* (1982).

since two spectral samples are tabulated for comparison; $C_1 = 48$ and $C_2 = 42$; $T = 90$. The X_{ij}, $(i = 1, \ldots, 36; j = 1, 2)$ are the incidence of mutants in each sample. For instance, $X_{20,1}$, which is the incidence of amber nonsense mutants at site 17 in the −pKM101 strain, is equal to 10.

(2) Simulate the drawing of tables at random from the same hypergeometric distribution (i.e. the hypergeometric distribution parameterized by the row and column marginal totals of the observed table) using the method presented by Agresti *et al.* (1979, p. 77). The hypergeometric probability of each simulated table is calculated.

(3) The proportion of the simulated tables as improbable or more improbable than the observed table is the estimate of the p value of the observed table under the null hypothesis. This estimate will be close to the p value of the hypergeometric test if the number of simulated samples is sufficiently large.

4. Comparison of Spectral Samples

We have surveyed the literature to find spectral samples to compare using the Monte Carlo hypergeometric test. Most of the differential responses of mutational spectra to treatments in the *lac*I system are obvious without statistical analysis. The sole fact that these treatments had differential effects on the mutational spectra could have been established with much smaller samples. We will present three historical comparisons of published spectral samples in order to demonstrate the usefulness of the hypergeometric test. We will consider a p value less than 0·05 significant. A total of 1500 simulated samples was used to obtain the estimated p values. The basis for calculating confidence intervals for this estimated p value was presented by Agresti *et al.* (1979).

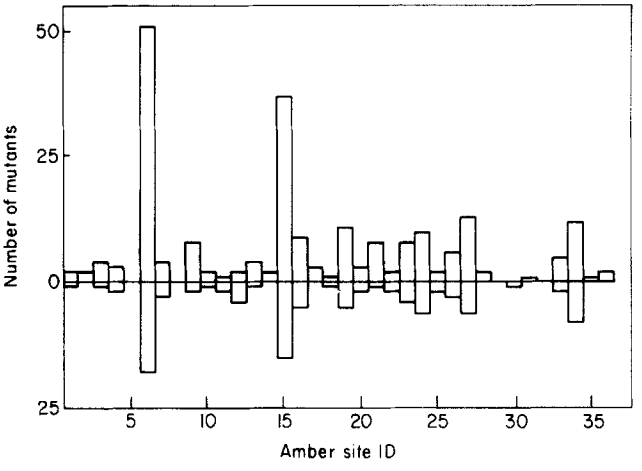

Figure 1. Frequency histogram of spontaneously occurring amber nonsense mutations in the *lac*I gene in *E. coli*, characterized by Coulondre & Miller (1977) (above) and Glickman *et al.* (1980) (below, inverted). No significant difference.

Coulondre & Miller (1977) and Glickman *et al.* (1980) have performed independent experiments in different laboratories to characterize samples of spontaneous *lac*I nonsense mutations. The 2×36 cross-tabulation of amber mutants from these two experiments, presented graphically here (Fig. 1), was tested with the Monte Carlo hypergeometric test. No significant difference was found. Also, no significant difference was found when the ochre samples from these same experiments were compared.

Coulondre & Miller (1977) and Todd & Glickman (1982) characterized nonsense mutations induced by similar doses of u.v. in two independent experiments. The experiment of Todd & Glickman was not a reproduction of the original Miller protocol. In the Todd & Glickman experiment, the bacteria were grown in minimal media after exposure to u.v. light. In the Coulondre & Miller experiment, the bacteria were allowed to resume replication after exposure in a rich media. The 2×36 table of amber and ochre mutations occurring in these experiments were compared with the Monte Carlo hypergeometric test. The amber spectral samples (Fig. 2) were found to be significantly different. The p value from this test is an

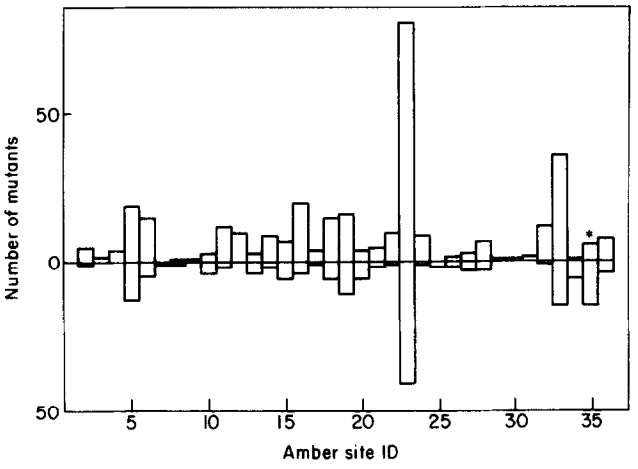

Figure 2. Frequency histogram of u.v.-induced amber nonsense mutations in the *lac*I gene in *E. coli* characterized by Coulondre & Miller (1977) (above) and Todd & Glickman (1982) (below, inverted). The spectra are different ($p < 0{\cdot}01$).

The asterisk indicates that a difference in the relative mutation rate at site 35 would fully account for the overall difference.

estimate that has a confidence interval. The p value was estimated to be 0 with a 90% confidence interval of 0·002 to 0. The ochre spectral samples from the same experiments were not significantly different. It was found that the difference in the amber spectral samples could be explained by a change in the relative mutation rate at amber site 35. When the data for amber site 35 are set aside and the remaining 2×35 cross-tabulation was tested, there was no significant difference. No other explanation based on a mutation rate change at a single site can account for difference in these spectral samples.

Eisenstadt *et al.* (1982) characterized nonsense mutations induced in the *lac*I gene by 3,4-epoxycyclopenta[*cd*]-pyrene (CPPE) in two strains of *E. coli*, one with and one without the pKM101 mutation-enhancing plasmid. The estimate of the p value of the hypergeometric test of the 2×36 table (Table 1) of CPPE-induced amber mutations in +pKM101 and −pKM101 strains was 0·019, with a 90% confidence interval of 0·026 to 0·013. The test of the ochre mutants showed no significant difference. The Bonferroni correction should be used to correct for the fact that two hypothesis tests (i.e. the comparison of the amber samples and the independent comparison of the ochre samples) were performed. The correction consists of multiplying each p value by the total number of hypothesis tests performed. This results in a p value of 0·038. So, the spectral samples are significantly different.

5. Discussion

The comparison of the spontaneous spectral samples shows a good level of interlaboratory reproducibility for the spontaneous spectrum. Of course, the hypothesis test does not prove that the null hypothesis is true. Larger samples would provide more discrimination power. The result shows good operational reproducibility up to a sample size of 100, at least. It would be interesting to see if relative probabilities of nonsense mutations in the *lac*I gene have such a high level of reproducibility for other treatments. The results of the comparison of the u.v.-induced spectral samples leads to the conclusion that the different media in which the bacteria were grown after exposure or some other unreported difference in the experiments had an effect on the relative mutation rates at some sites. In the study of CPPE-induced mutants, there was a 15-fold increase in the number of *lac*I nonsense mutants in the +pKM101 strain. Considering this, it is surprising that the spectral samples are so similar. The difference in the spectra could be an indirect effect of this large increase in induced mutagenesis. In an attempt to interpret the results of the comparison of the CPPE-induced spectral samples, the 2×36 table of amber mutants classified by site was collapsed to a 2×5 table classified only by type, and this table was tested. No significant difference was found. This means that the result cannot be explained as a change in the type-specific mutation rates. The statistical analysis was not taken further, but inspection of the G·C → T·A base substitutions classified by site suggests that the difference in the CPPE-induced spectral samples could be explained by changes in the relative mutation rates among the sites where G·C → T·A base substitutions can be detected.

In the experiments analyzed above, each spontaneous mutant was taken from a different culture. However, groups of the u.v.-induced and CPPE-induced mutants were taken from the same treated cultures. This means that the observed distribution of u.v.-induced and CPPE-induced mutants would not reflect any interculture variability that may exist in these experiments. We have assumed that the microbe is the experimental unit in our analysis, but this is appropriate only if there is no interculture variability in spectra. All the positive results we have reported could be explained by excess interculture variance. In the next section we will present an experimental design for the researcher who is willing to go to some extra effort to avoid having to make the assumption that no interculture variability exists.

6. The Design of Experiments for the Comparison of Spectral Samples

In this section we describe an experimental design that will test the null hypothesis that a single factor has no effect on the spectrum of detectable mutations in a gene. This design is applicable to both the *lac*I nonsense mutation characterization and to the classification of mutations by direct sequence analysis.

Treatment 1 and treatment 2 are assumed to

differ by only the factor of interest. S is the sample size in each treatment.

(1) Prepare $2 \times S$ tubes of media.
(2) Inoculate the tubes with bacteria.
(3) Select S of the cultures at random. Apply treament 1 to these selected S cultures. Apply treatment 2 to the remaining cultures.
(4) Characterize one mutant per culture.
(5) Cross-tabulate the results by treatment *versus* type of modification to the DNA sequence.
(6) Estimate the p value under the null hypothesis using Monte Carlo estimation of the p value of the hypergeometric test.

It is assumed that the cultures are handled in an equivalent manner.

We propose this experimental design for the following reasons.

(1) Two mutants selected from the same culture may have arisen from the same mutagenic event. The possibility of this can be reduced by applying the treatment for less than a full cell cycle at the end of the culture's growth, if the treatment is mutagenic enough to render the spontaneous background negligible. It can also be reduced by selecting only a small percentage of the mutants generated in each culture. The possibility is eliminated by selecting a single mutant per culture.
(2) This protocol is not much more labor intensive than some protocols used in past experiments, e.g. the selection of a single mutant per culture. The proposed protocol would require, in addition, that each culture be individually treated.
(3) We have presented a statistical method that can detect smaller differences in spectral sample than have heretofore been revealed. It is appropriate to concern ourselves with the elimination of subtle confounders of the results.
(4) This experimental design may result in better interlaboratory reproducibility of spectral samples. Any interculture variance will be maximized within, rather than between, spectral samples.
(5) The characterization of the selected mutants is labor intensive, particularly when direct sequencing is employed. Using this experimental design will not significantly increase the total work involved in the project.
(6) Use of this protocol ensures that the chosen significance level represents the probability of incorrectly finding a significant difference between spectral samples when the treatments in fact have no differential effect on the mutational spectra.

Researchers who use experimental designs that involve the selection of multiple mutants per culture should retain and make available the classification of the mutants by culture as well as by type/site and treatment. This information is essential for a thorough statistical analysis of spectral samples collected in this manner.

7. Conclusion

The hypothesis that two treatments induce the same mutational spectra can be tested by the application of the Monte Carlo estimate of the p value of the hypergeometric test to the results of properly designed experiments. This provides a means of assessing the interlaboratory and intralaboratory reproducibility of spectral samples, and early results concerning reproducibility are good. Also, researchers will be able to distinguish the differences in mutation spectra with smaller sets of mutants.

We have not presented a general method for forming an assessment of what the differences are in spectra induced by two treatments. The difference in the u.v. spectra was a special, and easy, case. A general method for statistical assessment of the theories about mutational spectra is a matter for future research.

A computer program that implements the Monte Carlo estimation of the p value of the hypergeometric test is available from the authors. The VAX-11 FORTRAN-77 program utilizes subroutines from the International Mathematical and Statistical Library.

We gratefully acknowledge helpful discussions with Barry Margolin, Kerrie Boyle, Tom Starr, Ray Buck and Nancy Adams. We thank Jeffrey Miller and Barry Glickman for clarifying the differences in the experiments that yielded the u.v.-induced spectral samples.

The description of the experimental design we have presented drew some wording and concepts from the description of the fluctuation test by Collings *et al.* (1981).

References

Agresti, A., Wackerly, D. & Boyett, J. (1979). *Psychometrika*, **44**, 75–83.

Collings, B. J., Margolin, B. H. & Oehlert, G. W. (1981). *Biometrics*, **37**, 775–794.

Coulondre, C. & Miller, J. H. (1977). *J. Mol. Biol.* **117**, 577–606.

Eisenstadt, E., Warren, A. J., Porter, J., Atkins, D. & Miller, J. H. (1982). *Proc. Nat. Acad. Sci., U.S.A.* **79**, 1945–1949.

Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*, 2nd edit., MIT Press.

Fingleton, B. (1984). *Models of Category Counts*, Cambridge University Press.

Freeman, G. H. & Halton, J. H. (1951). *Biometrika*, **38**, 141–149.

Glickman, B. W., Rietveld, K. & Aaron, C. S. (1980). *Mutat. Res.* **69**, 1–12.

Hoel, P. G., Port, S. C. & Stone, C. J. (1971). *Introduction to Probability Theory*, Houghton Mifflin Co., Boston.

International Mathematical and Statistical Library (1984). *IMSL Reference Manual, Version 7*, CTPR-1-CTPR-3, Houston, TX.

Miller, J. H. (1983). *Annu. Rev. Genet.* **17**, 215–238.
Pagano, M. & Halvorsen, K. (1981). *J. Amer. Stat. Assoc.* **76**, 931–934.
Roscoe, J. T. & Byars, J. A. (1971). *J. Amer. Stat. Assoc.* **66**, 755–759.
Tocher, K. (1951). *Biometrika*, **37**, 130–144.
Todd, P. A. & Glickman, B. W. (1982). *Proc. Nat. Acad. Sci., U.S.A.* **79**, 4123–4127.
Upton, G. J. G. (1978). *The Analysis of Cross-Tabulated Data*, pp. 7–10, Wiley, New York.

Edited by M. Gottesman