A statistical test for comparing several nucleotide sequences

Robert D. Sege* and Bo E.H. Saxberg**

Harvard - MIT Division of Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA

Received 25 September 1981

ABSTRACT

A statistical test useful for the simultaneous comparison of nucleotide sequences is presented. A likelihood ratio test is derived, examples demonstrated, and tables of critical values are provided.

INTRODUCTION

The advent of DNA sequencing technology made available vast amounts of data for the study of evolution and genetics. A number of statistical tools (1-3) and computer programs (4-9) have been devised which either compare two closely related sequences for regions of homology or look for symmetries within one sequence. Only those features which share a number of nearly consecutive bases within the region examined are considered statistically significant.

By contrast, a number of investigators have noted that the presence of just a few bases at a rigidly defined position in a nucleotide sequence is associated with a specific biological function. Two such examples are the RNA polymerase site (10) and the RNA splice sites in eukaryotes (11).

The statistical characterization of such observed consensus sequences requires a function capable of being applied to a single position in several related nucleotide sequences. By comparing data from many sequences, the statistical requirement that a number of consecutive bases be related is made to vanish.

The statistical function herein described allows the comparison of many sequences simultaneously. In this manner, consensus sequences of any size may be characterized. Those locations where a particular class of nucleotides (e.g. a purine) is required may also be located.

Demonstration of statistical significance does not demonstrate biological activity. However, the use of a statistical test may indicate which positions

have been conserved and provides a basis for the detection, quantification and comparison of any such conservation.

DERIVATION

The general problem addressed here is to determine the level of information contained in a group of N bases, i.e., to examine the distribution of bases at one location among N sequences, or at N locations along one sequence (or in fact in any group of N bases). For example, an investigator might find it useful to determine whether the frequency of occurrence of a base at a particular site would signify some special conservation, or could simply be the result of chance selection from the pool of bases available. A method will now be derived which will allow one to determine the level of confidence for rejecting the hypothesis that the observed data came by chance selection from the base pool.

The observed frequencies are given by n_i/N where n_i is the number of bases of type i in the group of N , $i=1$ to 4 representing the four bases. The null hypothesis which one seeks to reject is that the observed frequencies arose by chance selection from the pool of bases with pool frequencies θ_i . If one seeks to reject the hypothesis that the observed frequencies came from a pool with an equal distribution of bases, the null frequencies θ_i will be equal to $1/4$ for each base type i . To reject the null hypothesis, a test statistic is defined which will be a function of the observed frequencies, the null frequencies θ_i , and the total number of bases N .

Under the null hypothesis, H , the probability of seeing the observed data is given by the multinomial distribution,

$$1) \quad p_{\theta}(\underline{n}) = \frac{n!}{n_1!n_2!n_3!n_4!} \cdot \theta_1^{n_1} \theta_2^{n_2} \theta_3^{n_3} \theta_4^{n_4}$$

where $\underline{n}$ represents the vector (n_1, n_2, n_3, n_4) of the number of bases of each type. The likelihood ratio is given by

$$2) \quad \lambda_{\theta}(\underline{n}) = \frac{p_{\hat{\theta}}(\underline{n})}{p_{\theta}(\underline{n})}$$

where $\hat{\theta}_i$ is the maximum likelihood estimator for the frequency of the i -th base. $p_{\hat{\theta}}(\underline{n})$ represents the probability of seeing the data under that hypothesis which gives the maximum probability of seeing the data of all possible

hypotheses. Therefore

$$3) \quad p_{\theta}(\underline{n}) \geq p_{\theta}(\underline{n}) \quad \text{and} \quad \lambda_{\theta}(\underline{n}) \geq 1$$

Here $\theta_i = n_i/N$, so

$$4) \quad \lambda_{\theta}(\underline{n}) = \prod_{i=1}^4 \left(\frac{n_i}{N \cdot \theta_i} \right)^{n_i}$$

The larger λ_{θ} is, the smaller $p_{\theta}(\underline{n})$ is relative to $p_{\theta}(\underline{n})$; the null hypothesis gives a small probability of observing the data relative to the largest probability possible. The null hypothesis is therefore rejected when λ_{θ} is large, or equivalently when the statistic $s = 2 \log(\lambda_{\theta})$ is large:

$$5) \quad s = 2 \sum_{i=1}^4 n_i \cdot \log \left(\frac{n_i}{N \cdot \theta_i} \right)$$

For large values of N , the statistic s is approximated by the Pearson's χ^2 statistic, which here takes the form

$$6) \quad \chi^2 = \sum_{i=1}^4 \frac{(n_i - N \cdot \theta_i)^2}{N \cdot \theta_i}$$

which has a χ^2 distribution for large values of N (when $N \cdot \theta_i$ is at least about 5 seems to be a rule of thumb (12,13)). Assuming the approximations are valid, one rejects the null hypothesis for large values of χ^2 . The probability, p , of making an error in rejecting H should clearly be small; critical values $c(\alpha)$ are defined such that the null hypothesis is rejected when $\chi^2 \geq c(\alpha)$ with the probability of making an error in rejecting H equal to $p \leq \alpha$:

$$7) \quad c(\alpha) = \max \{ c \mid P_H(\chi^2 \geq c) \leq \alpha \}$$

When N is large, the χ^2 distribution can be used so that the critical values for the χ^2 statistic are given by the quantiles $\chi^2_{1-\alpha}$ defined such that

$$8) \quad P(\chi^2 \geq \chi^2_{1-\alpha}) = \alpha$$

H is then rejected if

$$9) \quad \chi^2 \geq \chi^2_{1-\alpha}$$

In practice, however, N may not be large (e.g., one may have only 10 sequences to compare). In this case, use of the likelihood ratio statistic, s ,

is possible, provided critical values for the various α levels can be derived:

$$10) \quad c(\alpha) = \max \{ c \mid P_H(s \geq c) \leq \alpha \}$$

The difficulty here is the calculation of the distribution of s under H (which specifies the θ_i) for $\underline{n}$ from a multinomial distribution. Note that earlier, when N was assumed large, the approximations used allow the critical values to be independent of both N and H ; in general this is not the case, and here it is no longer true for small N .

A computer program was created to explicitly calculate the cumulative distribution function of s' by summing the multinomial probabilities, $M(\underline{n}, \underline{\theta})$, of those vectors $\underline{n}$ which generated values of s less than s' :

$$11) \quad F(s') = P_H(s < s') \\ = \sum_{\underline{n} \in A} M(\underline{n}, \underline{\theta}) \quad \text{where } A = \{ \underline{n} \mid \sum_{i=1}^4 n_i = N \text{ \& } s(\underline{n}, \underline{\theta}) < s' \}$$

(Note that this equation can be used to compute the exact p value.) Since $F(s')$ is monotonically increasing with s' and $P_H(s \geq c) = 1 - P_H(s < c)$, c can be expressed as

$$12) \quad c(\alpha) = \min \{ c \mid P_H(s < c) \geq 1 - \alpha \} \\ = \min \{ c \mid F_H(c) \geq 1 - \alpha \}$$

The program determines $c(\alpha)$ for a specified level α by using a Newton-Raphson iteration to approach the desired level and then stepping to precisely the smallest allowed value of c . (F_H is a step function changing values at only certain discrete values of s which arise from the quantized nature of this problem, i.e., bases occur in whole numbers. Hence only c values coinciding with possible s values are of interest.)

These critical values depend on the number of bases in the group studied, N , and the null hypothesis frequency of bases, $\underline{\theta}$. In Tables I to III are listed the critical values calculated for the likelihood ratio statistic, s , for various values of N and for different null hypotheses. Note that as N gets larger, the critical values should approach those for a χ^2_3 distribution.

In summary, for large N the null hypothesis H can be rejected with probability of making an error equal to $p \leq \alpha$ if one rejects when

$$13) \quad \chi^2(\underline{n}, \underline{\theta}, N) \geq \chi^2_{(1-\alpha)}$$

with the critical value independent of N and H . For small N , the exact solution reported herein allows H to be rejected with $p \leq \alpha$ if one rejects when

$$14) \quad s(n, \theta, N) \geq c(\alpha, \theta, N)$$

where $c(\alpha, \theta, N)$ is either taken from one of the tables listed above or calculated for other null hypotheses using the developed computer program. The authors will make the program or appropriate tables available upon request.¹

USE OF THE STATISTIC

Calculation of the statistical significance of a given observation requires performance of the following steps: selection of an appropriate null hypothesis, calculation of the test statistic, and evaluation of the probability of the observation based on this statistic.

Proper selection of the null hypothesis is clearly important for interpreting the results of a statistical comparison. Here, the null hypothesis can be considered as the a priori distribution of nucleotides in a pool from which the sequences to be analyzed are drawn. Many alternative null hypotheses are possible, such as: (1) The abstract nucleotide pool is unlimited and therefore the distribution of nucleotides is effectively equal; (2) the sequences are drawn from a pool comprised of the nucleotide distribution typical of the species; or (3) the nucleotide pool for the class of sequences examined is well represented by the total distribution of nucleotides in the sequences themselves. Our examples show the use of each of these alternative hypotheses.

The calculation of the statistic was discussed above. Here, we merely repeat the formula for the statistic:

$$s = 2 \sum_{i=1}^4 n_i \log \left(\frac{n_i}{N \cdot \theta_i} \right)$$

Tables I to III present the critical values of the test statistic under three different null hypotheses. The p value is computed by first calculating the value of the statistic from the observed data and the null hypothesis frequencies; the p value is then that α whose corresponding critical value is the greatest critical value (for the size N of the group of bases being examined) less than or equal to the calculated value of the statistic. In the case of samples of large size (e.g., $N \geq 40$), the Pearson's χ^2 statistic can be used with a table of critical values for the χ^2 distribution to calculate the p values. Such critical values are found in many books of statistical tables; a few are listed here for convenience (entries are from

Table 1: C(α , N) for equal base frequencies:

N	0.4	0.3	0.2	A:0.250	G:0.250	0.10	C:0.250	0.05	T:0.250	0.01	0.001
5	7.1328239	7.1328239	7.1328239	8.5859190	8.5859190	8.5859190	13.8629436	13.8629436	13.8629436	16.6355286	16.6355286
6	6.2247713	6.2247713	6.2247713	8.3177662	8.3177662	8.3177662	11.2287883	11.2287883	11.2287883	19.4081116	19.4081116
7	5.3487196	6.0283206	8.2597561	8.2597561	8.2597561	8.2597561	9.8474035	9.8474035	9.8474035	13.6664896	13.6664896
8	4.8655796	5.0040245	6.5916729	7.7766132	7.7766132	7.7766132	10.4107533	10.4107533	10.4107533	18.1833420	22.1807098
9	4.2702196	5.1782713	6.9043665	7.0428081	8.0893030	8.0893030	9.6769550	9.6769550	9.6769550	12.6459885	18.6743126
10	3.8548889	5.9478827	6.6274774	7.1328239	8.7204747	8.7204747	8.9169111	8.9169111	8.9169111	13.8629436	17.7178380
11	4.7910709	4.7910709	6.5171585	7.025010	7.7601089	7.7601089	9.9335289	9.9335289	9.9335289	13.7884178	17.2948151
12	3.5906219	4.986820	5.820558	6.9043636	7.4677200	7.4677200	9.6769571	9.6769571	9.6769571	12.4495440	17.9947195
13	4.0394266	4.3163080	4.9154768	6.9043636	7.1353916	7.1353916	9.7230425	9.7230425	9.7230425	12.6920843	18.1792517
14	3.7030578	4.3022294	5.3487263	6.0283184	7.9248524	7.9248524	9.2840528	9.2840528	9.2840528	13.6664867	17.5659995
15	3.5459967	4.255926	5.2720900	6.1221313	6.8017378	6.8017378	9.2616282	9.2616282	9.2616282	12.9906940	17.4893742
16	3.9217186	4.5208788	5.2324800	5.5673752	7.1550236	7.1550236	9.3284521	9.3284521	9.3284521	12.8966522	17.7585222
17	3.5061283	3.8508406	5.232209	6.1402769	6.7618656	6.7618656	8.9932976	8.9932976	8.9932976	12.9226990	17.5975304
18	3.3979783	4.0775738	5.1782780	5.8578734	6.9303665	6.9303665	8.5404401	8.5404401	8.5404401	12.6459980	17.4738665
19	3.3392879	4.0188819	5.2938385	5.7991811	6.8494720	6.8494720	8.4333286	8.4333286	8.4333286	12.5651045	17.8178892
20	3.2913055	3.9709015	5.5005417	5.9478884	6.6274843	6.6274843	8.1511927	8.1511927	8.1511927	12.9422674	17.7178335
21	3.3719616	3.9935593	4.8657988	5.6838431	6.6276999	6.6276999	8.5667538	8.5667538	8.5667538	12.9066677	17.2349043
22	3.3215504	4.0369606	4.9092016	5.7272367	7.0059958	7.0059958	8.6101580	8.6101580	8.6101580	13.0343170	17.2948160
23	3.2740086	4.0376129	4.8616569	5.5615625	6.6941881	6.6941881	8.4522820	8.4522820	8.4522820	12.1532326	17.7041043
24	3.2237196	4.2344055	5.0243254	5.6836166	6.5016470	6.5016470	8.3177624	8.3177624	8.3177624	12.1628475	17.4177628
25	3.2564899	4.0165209	5.1530285	5.4657289	6.5606518	6.5606518	8.4347620	8.4347620	8.4347620	12.2218466	17.3822212
26	3.2934608	4.0534971	4.8434229	5.5831509	6.4912031	6.4912031	8.5381038	8.5381038	8.5381038	11.9239442	17.7421751
27	3.2895441	4.0135475	5.0901585	5.7494411	6.6701107	6.6701107	8.0533989	8.0533989	8.0533989	11.9026022	17.7962856
28	3.2142239	3.9635286	4.8598814	5.8184681	6.6238990	6.6238990	8.2017403	8.2017403	8.2017403	11.7199337	17.5659943
29	3.2951088	3.8362780	4.9610728	5.4984188	6.7472897	6.7472897	8.1809921	8.1809921	8.1809921	11.7288895	17.8520230
30	3.2469091	3.8006802	4.9933081	5.6287018	6.6892776	6.6892776	8.0582876	8.0582876	8.0582876	12.0158224	17.8436480
31	3.2606573	4.0371885	4.823019	5.6828451	6.6432095	6.6432095	8.2744868	8.2744868	8.2744868	11.8292954	17.5225000
32	3.3166323	3.8493398	5.0146008	5.6475821	6.4919291	6.4919291	8.3719673	8.3719673	8.3719673	11.9731760	17.0795126
33	3.2508430	4.0108709	4.864905	5.650579	6.3583221	6.3583221	8.0942230	8.0942230	8.0942230	11.7853069	17.0015612
34	3.1671648	3.9212542	4.8113565	5.5489597	6.5093279	6.5093279	8.0365582	8.0365582	8.0365582	11.8819103	16.7417898
35	3.1488093	3.9088354	4.9469928	5.3943243	6.5532569	6.5532569	8.0508871	8.0508871	8.0508871	11.6511116	17.1382723
36	3.1336994	3.9056368	4.7695465	5.6035976	6.5092487	6.5092487	7.9899936	7.9899936	7.9899936	11.7157383	16.8297720
37	3.0563145	3.7626417	4.9132517	5.4606197	6.5046835	6.5046835	8.1488791	8.1488791	8.1488791	11.6455145	16.8318918
38	3.1163578	3.9281912	4.9145374	5.5604630	6.5401331	6.5401331	8.0377636	8.0377636	8.0377636	11.7212925	16.8124481
39	3.1850710	3.7947396	4.8808594	5.4927044	6.4589329	6.4589329	8.1547470	8.1547470	8.1547470	11.7797049	16.8696842
40	3.1268654	3.9167690	4.8149481	5.4328022	6.4666004	6.4666004	8.1404467	8.1404467	8.1404467	11.7348900	16.8995476

(14) except those marked with a *, which are from (15)):

α	.4	.3	.2	.10	.05	.01	.001
c(α)	2.946	3.665	4.642	6.251	7.815	11.345*	16.266*

The authors will make available a PL/I computer program which produces tables similar to those reported here for null hypothesis frequencies defined by the user.¹ Tables I through III were produced on the IBM 370/168 at M.I.T.

EXAMPLES

Figure 1 represents the sites surrounding positions of intramolecular integration in murine leukemia viruses (17). The sequences surrounding recombination events in viral clones were compared. A possible consensus sequence is also listed (calculated from $p3 \leq .10$), which differs in a number of positions from the consensus originally reported. The use of three different null hypotheses is also illustrated in Figure 1: $p2$ and $s2$ were calculated from the experimental pool (v. i.), while $p1$ and $p3$ are p values from Tables I and III, calculated with statistics $s1$ and $s3$ from the appropriate null hypothesis.

Figure 2 presents, in summary form, the RNA sequence data reviewed by Sharp (18). These sequences occur at the boundaries of RNA splice sites. Note the extremely high s values here; these s values can be used to define the actual boundary of the consensus sequence. Table III was used as representing a typical eukaryote to define the null hypothesis for the comparison.

DISCUSSION

As stated above, the choice of base frequencies for the null hypothesis can make a significant difference in the p value for rejecting the null hypothesis. The investigator must consider carefully the level of information the test should select for. The base frequencies of the null hypothesis can be considered to be the frequencies of the four bases in a "virtual pool". This virtual pool of bases is that pool from which the bases would have been drawn by chance (under the null hypothesis) to make up the group of bases under study. By rejecting the null hypothesis one is stating it is unlikely ($p \leq \alpha$) that the group under study was drawn by chance from the virtual pool. Note that different virtual pools will usually result in different critical levels and statistic (s) values. For a given group of bases, different p values for rejecting the null hypothesis may therefore be obtained for different virtual pools.

For example, suppose a set of A-T rich sequences are being examined for

Table 2: C(α , N) For E. coli K12 Base Frequencies(16):

N	α	0.4	0.3	0.2	0.15	A:O.268	G:O.241	C:O.246	T:O.245	0.01	0.001
5	4.0486715	4.4157339	4.4157339	4.4157339	6.9514999	7.3762589	7.3762589	8.8489041	9.1926374	14.2295837	14.2295837
6	4.4151245	4.4151245	4.8317374	4.8317374	6.2796026	6.5907507	6.5907507	8.8482919	11.4304992	16.8290710	16.8290710
7	4.3086271	5.1838131	5.7058954	5.7058954	6.0991182	6.3444211	6.3444211	8.6990623	11.3402946	14.1387177	14.1387177
8	4.4547625	5.1725931	5.7126494	5.7126494	6.5000896	7.0384645	7.0384645	8.0850145	11.9769773	16.2961569	16.2961569
9	3.3600969	4.9791088	5.9287860	5.9287860	6.5750256	7.5312319	7.5312319	8.4381151	12.7829504	15.7659318	15.7659318
10	3.3755171	4.2030363	6.1565313	6.1565313	6.6400394	7.2767372	7.2767372	9.0698404	11.9651127	16.0114174	16.0114174
11	3.1017013	3.6085986	5.7276979	5.7276979	6.6684422	7.5770550	7.5770550	8.8687981	11.7075983	16.1632229	16.1632229
12	3.2744728	3.9152901	4.9771883	4.9771883	6.6951284	7.4883299	7.4883299	8.8852196	11.9557495	17.0439348	17.0439348
13	3.1762619	4.0019979	4.8857250	4.8857250	6.1042337	7.6021080	7.6021080	9.1461973	12.2959738	17.3190165	17.3190165
14	3.2993031	3.8868418	4.8373499	4.8373499	5.6868687	7.5859480	7.5859480	9.0996585	12.2308235	16.4168530	16.4168530
15	3.3780317	4.0002078	4.7586651	4.7586651	5.7404580	6.7753043	6.7753043	9.2889309	12.2600012	16.4325728	16.4325728
16	3.0260782	4.0261955	4.9661369	4.9661369	5.5233884	6.7005588	6.7005588	9.2422562	12.3823586	16.7930326	16.7930326
17	3.0193024	3.8921499	4.9656717	4.9656717	5.6030111	6.5111151	6.5111151	8.7436924	12.4875631	17.1178465	17.1178465
18	3.0811530	3.8178997	5.0114660	5.0114660	5.6544580	6.4920816	6.4920816	8.4478741	12.5567284	17.0824094	17.0824094
19	3.1146183	3.830231	5.0430126	5.0430126	5.6646900	6.4635396	6.4635396	8.1327753	12.6474004	17.0892878	17.0892878
20	3.0722480	3.7982297	4.9147406	4.9147406	5.7617922	6.6113110	6.6113110	8.1138116	12.6660337	17.0739956	17.0739956
21	3.1455278	3.8828268	4.7958190	4.7958190	5.6466208	6.6113148	6.6113148	8.1260786	12.6762981	17.2098384	17.2098384
22	3.0642519	3.8502344	4.8959160	4.8959160	5.5237799	6.6891947	6.6891947	8.1946110	12.6263676	17.2883244	17.2883244
23	3.0629691	3.9458570	4.8622026	4.8622026	5.4592404	6.6224041	6.6224041	8.1793164	12.3360119	17.3789968	17.3789968
24	3.0350113	3.8667507	4.9406376	4.9406376	5.4976978	6.4285831	6.4285831	8.1621590	12.0424004	17.3070602	17.3070602
25	3.1071100	3.7081251	4.8379955	4.8379955	5.5665913	6.4433546	6.4433546	8.2245560	11.8240834	17.4496555	17.4496555
26	3.0810064	3.7850924	4.8486545	4.8486545	5.5056992	6.4822559	6.4822559	8.3016481	11.8010950	17.4430285	17.4430285
27	3.0570078	3.8252640	4.7885046	4.7885046	5.5256406	6.4261845	6.4261845	8.1620150	11.7416700	17.5652695	17.5652695
28	3.0631094	3.7289998	4.8036070	4.8036070	5.4610230	6.5168734	6.5168734	8.0865847	11.7803144	17.5428743	17.5428743
29	2.9785775	3.8581810	4.8203726	4.8203726	5.4107310	6.4857683	6.4857683	7.9763267	11.7220011	17.4574585	17.4574585
30	2.9343395	3.7959948	4.7213774	4.7213774	5.4752054	6.4926834	6.4926834	8.1098042	11.7552156	17.5123291	17.5123291
31	3.0379447	3.7236652	4.843008	4.843008	5.4091568	6.4720097	6.4720097	8.0676517	11.7782516	17.5248327	17.5248327
32	3.0241461	3.7270269	4.7783594	4.7783594	5.5032387	6.4156370	6.4156370	8.0965485	11.7862160	17.2959604	17.2959604
33	3.0708656	3.7064819	4.7889721	4.7889721	5.4645512	6.4376841	6.4376841	8.0960982	11.7906590	16.9655962	16.9655962
34	3.0248795	3.7906494	4.7839251	4.7839251	5.4881394	6.3835074	6.3835074	8.0802641	11.7003288	16.8682241	16.8682241
35	3.0634099	3.8015957	4.7029581	4.7029581	5.4565306	6.4114170	6.4114170	7.9896488	11.7368212	16.7633905	16.7633905
36	3.0200537	3.8136597	4.7825332	4.7825332	5.4471553	6.4710283	6.4710283	8.0089464	11.6191149	16.8338712	16.8338712
37	3.0078578	3.7357883	4.7538795	4.7538795	5.4709435	6.4635458	6.4635458	8.0146198	11.6130247	16.7960815	16.7960815
38	3.0747166	3.7249947	4.7816289	4.7816289	5.4191950	6.4073181	6.4073181	7.9881897	11.6231985	16.7546167	16.7546167
39	3.0067139	3.7901554	4.7428312	4.7428312	5.4816818	6.3974161	6.3974161	7.9523058	11.6242018	16.7152643	16.7152643
40	3.0003705	3.7008972	4.7421513	4.7421513	5.4684715	6.3576531	6.3576531	8.0469379	11.6950998	16.7983770	16.7983770

Table 3: $c(\alpha, N)$ For Human Base Frequencies (16):

$N \backslash \alpha$	0.4	0.3	0.2	0.15	G:0.195	C:0.199	T:0.303	0.001
5	3.9948116	4.8366690	5.8415505	6.9362035	6.9362035	7.8545074	11.1810896	16.1445007
6	4.5895886	4.7952074	5.6398364	6.4769220	6.4769220	8.5490444	11.1774759	19.3733978
7	4.1773769	5.0550680	5.9365383	6.4088120	7.1556015	8.4628782	11.8155370	16.2634830
8	4.1836100	4.4336276	5.9946957	6.4631937	7.1758634	8.5193434	11.8699291	15.3680315
9	3.8692868	4.9195480	5.6322203	6.7230899	7.4864979	8.7792451	12.3457146	16.0948973
10	3.8247738	4.6656277	5.7629894	6.3079081	7.4916738	8.9105228	11.9406672	15.6960365
11	3.2995508	4.8097486	5.6506046	6.1339211	7.0245705	8.7028333	12.0328627	16.2852993
12	3.3428268	4.1836832	5.6618373	6.2572279	7.3887146	8.8702049	12.2228689	16.7927971
13	3.2180561	4.2073205	5.6465611	6.1834320	7.1995773	8.8250132	12.2226915	16.9992771
14	3.2041155	3.8181696	5.1681345	6.3278999	7.1974320	8.6579491	12.2843523	16.6810623
15	3.1064746	3.8088884	5.1502819	6.2005409	7.0665879	8.7205219	12.2675171	16.5689321
16	3.2316740	3.8798462	4.8413506	5.9204540	7.2999926	8.4645470	12.3728931	16.8557806
17	3.1195450	3.9431705	4.9404564	5.6617271	7.0733049	8.6695232	12.5658512	17.3978100
18	3.2621945	3.9022470	4.8463600	5.6455774	6.9100423	8.7014170	12.3150165	16.9628146
19	3.1950974	3.9936848	4.8464542	5.5028344	6.8885784	8.7580562	12.1408997	17.1119747
20	3.1433174	3.8911773	4.9278154	5.5309896	6.5025005	8.7523508	12.0150871	17.1194372
21	3.1109155	4.0022346	4.8837051	5.5541897	6.6107475	8.7843943	12.3483200	17.1910383
22	3.1422872	3.8353090	4.9568195	5.5571632	6.5057115	8.2804260	12.1237693	17.3140525
23	3.0643692	3.8911963	4.9127962	5.5729189	6.5626232	8.2230570	12.1757536	17.2863474
24	3.0169450	3.8334464	4.9443474	5.4866559	6.5763779	8.1927508	12.2555408	17.4388275
25	3.0780019	3.7713870	4.8723240	5.5712337	6.4105797	8.0240030	12.2823009	17.0992756
26	3.1535033	3.7941426	4.8352193	5.6700926	6.4674978	8.1536732	12.2908669	17.1489630
27	3.1413091	3.7618179	4.7711013	5.5339298	6.5514002	8.0895567	12.4021959	17.0584593
28	3.0351565	3.8952999	4.8065662	5.4636908	6.4969951	8.0291348	12.2612581	16.9312143
29	3.0147104	3.8373059	4.7282027	5.4136496	6.4981753	8.0415878	12.1654129	17.0595865
30	3.0055647	3.7645922	4.8465147	5.4242659	6.4851828	8.0328779	11.9201546	17.0847785
31	3.0483646	3.7676325	4.7573214	5.5466379	6.4379301	8.2068606	11.8666935	17.0778875
32	3.0187051	3.8271160	4.8690882	5.5074949	6.4755149	8.2230115	11.8037348	17.1645298
33	3.0136032	3.7864609	4.8639258	5.5222270	6.4714745	8.1735039	11.7910690	17.1507483
34	3.0706130	3.6879392	4.7436094	5.4998074	6.4581518	8.0814748	11.7325993	17.2381039
35	3.0400620	3.7804670	4.7632685	5.4853652	6.3934774	8.0238132	11.5548019	17.1907787
36	3.0002499	3.8179455	4.7027888	5.5023918	6.4058486	8.0473719	11.6528788	17.1944160
37	3.0128593	3.8166000	4.7803838	5.4541874	6.4660757	8.0654327	11.7045412	17.1309204
38	3.0319446	3.7636814	4.7236490	5.4788429	6.4865808	8.0163565	11.6845121	17.1873977
39	3.0816949	3.7219362	4.7547579	5.4249023	6.4343414	8.0106411	11.7227545	17.2596846
40	2.9678621	3.7599505	4.7770109	5.4302226	6.4166193	7.9954720	11.6715091	17.0568104

	Sequences	N	s1	s2	s3	p1	p2	p3	p3 $\leq$.10
-4	TTTATCCAC	9	5.858	3.837	4.920	.3	>	.4	
-3	CACCCCTTCT	9	8.089	5.372	8.833	.10	.3	.05	C
-2	TCTCTTTTA	9	9.677	6.763	7.898	.05	.15	.10	T
-1	AACCCAAGG	9	5.858	7.556	6.682	.3	.10	.20	
α	CCCGGCCTC	9	9.677	8.263	13.024	.05	.10	.01	C
β	ATTTCACAG	9	0.313	0.291	0.296	>	>	>	
γ	TGGTTTTAT	9	9.677	8.697	7.979	.05	.05	.10	T
δ	GGAGTTTCT	9	3.085	3.203	3.110	>	>	>	
1	CGTACT T	8	2.773	1.322	2.259	>	>	>	
2	AAACATG A	8	5.004	6.796	3.650	.3	.15	>	
3	GGCTAGG T	8	2.773	4.353	4.063	>	.4	>	
4	CAGCGCG A	8	4.866	6.372	6.956	.4	.15	.15	

Figure 1: Site of intramolecular integration in MuLV (17). Sequences are read down. The data are evaluated under three different null hypotheses to illustrate the differences in the resulting pvalues.

	A	U	C	G	N	S	P
	8	10	10	10	38	2.939	>
	9	11	15	3	38	10.171	.05
	10	11	10	7	38	1.090	>
	13	3	18	4	38	21.943	.001
	24	5	4	5	38	17.845	.001
5' SS	4	5	1	28	38	52.503	.001
	0	0	0	38	38	122.698	.001
	0	38	0	0	38	90.746	.001
	20	3	2	13	38	22.865	.001
	27	5	2	4	38	27.348	.001
	3	3	0	32	38	76.187	.001
	8	20	6	4	38	8.633	.05
	15	5	8	10	38	6.408	.15
. . .	9	6	11	12	38	7.518	.10
↓	↓	↓	↓	↓	↓	↓	↓
1	17	16	3	37	29.420	.001	
7	12	13	6	38	5.864	.15	
4	22	12	0	38	31.601	.001	
1	19	17	1	38	38.333	.001	
2	22	12	2	38	27.737	.001	
15	11	4	8	38	2.898	>	
3	5	28	2	38	52.714	.001	
37	0	1	0	38	82.378	.001	
3' SS	0	0	0	38	38	122.698	.001
	9	4	5	20	38	22.078	.001
	4	15	4	15	38	15.092	.01
	8	6	12	12	38	9.005	.05
	5	7	15	11	38	14.093	.01
	12	10	12	4	38	4.648	.3
	11	6	10	11	38	5.413	.2

Figure 2: mRNA splice sites (18). The data are summarized, and the values of s and p evaluated as described in the text. For null values, see table III.

conservation at a particular position. If the virtual pool is selected as one with equal frequencies of each base (the even distribution), then the presence of many A's and T's and a few G's and C's at a position may give a low p value. Thus, in a comparison of A-T rich sequences, many locations may show a significant p value when compared with the even virtual pool, as one would expect.

But suppose the investigator is not interested in just the A-T richness compared with the even distribution, but rather is interested in locations within the A-T rich region that have been conserved to an even higher level than the background A-T richness. This higher level of conservation may not be revealed in a significantly smaller p value with the even virtual pool. One way to detect this "hidden conservation" is to define a new virtual pool with base frequencies being those in the entire group of examined data: the "experimental virtual pool".

In this example, the selection of A-T rich sequences for study will cause the experimental virtual pool to have high frequencies for A and T, and low frequencies for G and C. Hence a significant p value will indicate either a special A-T richness over that of the background, detecting "hidden conservation", or conversely, a special relative G-C richness which may signify a position of "hidden freedom" compared with that of the background.

It should be clear by now that the selection of the virtual pool to be used is an important step in assigning any meaning to the resulting p values. The virtual pool selected will be a function of the question posed by the experimenter and the level of information desired. Clearly the most readily interpreted virtual pools are the even virtual pool (frequencies = 0.25) and the "species/organism" virtual pool (average frequencies of bases for the species/organism). These virtual pools should serve as standards unless the investigator has sufficient reason to warrant another type (e.g., experimental virtual pool). When a non-standard virtual pool is used, its justification and the meaning of the resulting p values must be carefully considered. Of course, p values should always be reported along with the type of virtual pool used (or if a χ^2 approximation is used, it should be so stated).

There are cases where the sensitivity of the p values to variations in the virtual pool frequencies is of interest. For example, in evolutionary studies, one often compares sequences from species with differing species virtual pool frequencies. The question then arises of how to select an appropriate virtual pool for a cross species study. Another problem may be that the investigator does not have available a table of critical values for

the exact virtual pool frequencies specified by some a priori knowledge. The question in this case is whether a table of critical values for a slightly different set of virtual pool frequencies may be used.

Since the values of the statistic s and the critical values will increase or decrease together for small fluctuations in the virtual pool frequencies, and since one is normally interested in only classifying p values within certain ranges (e.g., $p \leq .05$), one can expect to be able to meaningfully use slightly different virtual pool frequencies for which a table of critical values exists, i.e., one modifies the virtual pool composition slightly in constructing the null hypothesis and statistic s . (If the a priori virtual pool frequencies were used in the calculation of the statistic, the connection between the s and critical values in response to frequency fluctuations would be lost.) In addition, as the number of bases in the group analyzed becomes larger, the sensitivity of the p values to fluctuations in the virtual pool frequencies decreases due to the fact that the critical values must approach the χ^2_3 distribution. Indeed, the "rule of thumb" described above would indicate that for frequencies around $1/4$, a group of greater than 20 or so can be analyzed with the χ^2_3 distribution. The tables listed in this paper contain the critical values for up to groups of size 40, so that possible error through the χ^2_3 approximation is reduced to a small amount relative to the p values (Yarnold indicates that for even distribution frequencies = $1/4$, the χ^2_3 approximation causes a maximum error in p value of $-.05$ over the whole range of p values, and a maximum error in p value of only $-.008$ over the range $.01 \leq p \leq .1$ (13)).

Results of analysis of the first example presented in this paper with two virtual pools (even and human) used to calculate the critical values in Tables I and III showed that the resulting p values were nearly always the same, except for segments which were rich in G's and C's where the human virtual pool gave at times the next lower p value. It therefore appears possible to use the human virtual pool for general eukaryotic calculations and the E. Coli virtual pool for prokaryotic calculations, as well as for evolutionary comparisons restricted to each class (eukaryotes or prokaryotes), as fluctuations in organism virtual pools are not usually large within each class, and p values are not likely to change as a result. Note that when the size of the group is sufficiently large the critical values are independent of the particular virtual pool chosen-- one needs only a χ^2_3 upper tail probability table to handle all large groups. (Of course, the virtual pool frequencies are still needed to evaluate the Pearson's χ^2 statistic.)

The statistic generated for the case of four bases is a special case of a larger family of such statistics. The methods used to derive and evaluate this function can also be used to derive and evaluate functions which look at sequences of other than four different elements which are also of biological interest. For example, the distribution of bases including 5-me Cyt could be calculated for related sequences and the statistic

$$s = \sum_{i=1}^5 2 \log \left(\frac{\hat{\theta}_i}{\theta_i} \right)$$

could be evaluated for each position. The 5-nomial distribution would then be summed to compute the corresponding $c(\infty)$ and p values in strict analogy to the derivation above. As the frequency of 5-me C is actually quite low, it would be necessary to use the exact statistic even for rather large N (5-me C accounts for 1% of all bases in the mouse(16); therefore $0.01N \geq 5$ or $N \geq 500$ sequences would have to be analyzed to reliably use the standard χ^2_4 distribution).

Similar extensions could be used to analyze dinucleotides, trinucleotides or amino acids. The distribution of codons coding for a particular amino acid could also be analyzed in this manner.

So far only the simultaneous comparison of many sequences at a particular position has been considered. However, as the derivation of the statistic makes clear, it may be applied to any group of N bases. Thus groups of bases chosen from a set of sequences by an arbitrary selection rule may be analyzed for their deviation from the null distribution. By way of example, the bases thought to surround nucleosomes may be selected and compared to the total sequence base composition (the virtual pool is here the sequence itself) to determine if sequence-specific nucleosome formation is a possibility. In this regard, it should be noted that deviation from a null hypothesis need not necessarily implicate the existence of a "consensus" sequence. Once statistical divergence has been demonstrated, the question of exactly what rule governs that divergence remains to be examined.

This latter approach also leads to a possible computer algorithm for sequence alignment. Individual sequences may be analyzed according to a family of rules. Each sequence would then be characterized by its particular set of significant autocorrelation rules. A group of sequences which possess similar sets of rules could only be aligned in such a way that the rules are "in phase" between sequences. The algorithm would then select those alignments from among the set of "in phase" alignments which minimize the p values. These rules

could be applied to relatively short subsequences to allow for insertions, deletions, and recombinations. This type of analysis may be helpful in determining the possible genetic/evolutionary history of a gene family.

ACKNOWLEDGMENT

We would like to express our gratitude to Dr. David Baltimore for generously allowing us the use of his computer facilities and access to his unpublished data. We would also like to thank Dr. Allan Maxam for his critical insight. The authors were supported by the Harvard Medical School Medical Scientist Training Program.

* and Department of Biology, MIT

**and Department of Physics, MIT

FOOTNOTES

Please address requests for the program to the authors at the Division of Health Sciences and Technology, MIT, Bldg 16- 512 Cambridge, Massachusetts 02146 USA. Printed copies of the program will be sent out unless other arrangements are made in advance.

REFERENCES

1. Brezinski, D. P. (1975) Nature 253, 128-130
2. Korn, L.J., Queen, C.L. and Wegman, M.N. (1977) Proc. Nat. Acad. Sci. 74, 4401-4405
3. Nussinov, R. and Jacobson, A. B. (1980) Proc. Nat. Acad. Sci. 77, 6309-6313
4. Sege, R.D., Soll, D., Ruddle, F.H. and Queen, C. (1981) Nucl. Acids Res. 9, 437-444
5. Queen, C.L., and Korn, L.J. (1980) Methods in Enzymology 65, 595-609.
6. Staden, R. (1977) Nucl. Acids Res. 4, 4037-4051
7. Staden, R. (1978) Nucl. Acids Res. 5, 1013-1015
8. Staden, R. (1980) Nucl. Acids Res. 8, 817-825
9. Gingeras, T. K. and Roberts, R. J. (1980) Science 209, 1322-1325
10. Pribnow, D. (1975) Proc. Nat. Acad. Sci. 72, 784-788
11. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. and Chambon, P. (1978) Proc. Nat. Acad. Sci. 75, 4853-4857
12. Bickel, P. J. and Doksum, K. A. (1977) Mathematical Statistics: Basic Ideas and Selected Topics, pp. 312-332, Holden-Day, San Francisco.
13. Yarnold, J. K. (1970) J. Amer. Stat. Assoc. 65, 864-886
14. Isaacs, G. L., Christ, D. E., Novick, M. R., and Jackson, P. H. (1974)

- Tables of Bayesian Statistics, Department of Statistics, University College London (as No. 31 in the series "Tracts for Computers").
15. Weast, R. C. ed. (1968) Handbook of Tables for Probability and Statistics, 2nd edn., CRC Press, Cleveland, Ohio.
 16. Sober, H. A. ed. (1970) Handbook of Biochemistry Selected Data for Molecular Biology, 2nd edn., CRC Press, Cleveland, Ohio.
 17. Shoemaker, C., Hoffman, J., Goff, S. P., and Baltimore, D. (1981) J. Virology (in press)
 18. Sharp, P. A. (1981) Cell 23, 643-646