Volume 12 Number 1 1984 Nucleic Acids Research

On the statistical significance of nucleic acid similarities

D.J.Lipman and W.J.Wilbur,
Mathematical Research Branch, NIADDK, National Institutes of Health, Bethesda, MD 20205, and

T.F.Smith and M.S.Waterman
Departments of Mathematics and Biological Sciences, University of Southern California,
Los Angeles, CA 90089, USA

Received 22 July 1983

ABSTRACT

When evaluating sequence similarities among nucleic acids by the usual methods, statistical significance is often found when the biological significance of the similarity is dubious. We demonstrate that the known statistical properties of nucleic acid sequences strongly affect the statistical distribution of similarity values when calculated by standard procedures. We propose a series of models which account for some of these known statistical properties. The utility of the method is demonstrated in evaluating high relative similarity scores in four specific cases in which there is little biological context by which to judge the similarities. In two of the cases we identify the statistical properties which are responsible for the apparent similarity. In the other two cases the statistical significance of the similarity persists even when the known statistical properties of sequences are modelled. For one of these cases biological significance is likely while the other case remains an enigma.

INTRODUCTION

When apparently similar subsequences are detected between two nucleic acid sequences, an important question is: Does this degree of similarity imply an evolutionary or functional relationship between the subsequences? This question is particularly relevant when there is little or no biological context in which to judge an apparent similarity. Such a situation will increasingly arise with the advent of large sequence data banks and the development of rapid, sensitive methods with which to search for similarity (1).

Statistical methods are often useful in answering this question. Typically, a distribution of similarity scores among sequences which do not share an evolutionary or functional relationship is used to find a significance level associated with a computed similarity score. The distribution of similarity scores may be a theoretically determined distribution (e.g., the normal curve), or the distribution may be generated by Monte Carlo methods. The standard Monte Carlo method involves generating reference sequences which retain only the base composition and length of the original sequences.

However, using current statistical methods, one often encounters sequence similarities which are identified as statistically significant, but their biological significance is unclear, and perhaps, even dubious. A possible explanation for this situation may be that the set of reference sequences (the state space) sampled by the statistical method is not representative of the true statistical distribution from which the specific sequences were taken. More specifically, the similarity values between sequences in the reference set may be unrealistically low. Recent results on the statistical properties of sequenced nucleic acids suggests that some statistical patterns are common amoung sequences which do not have a verified functional or evolutionary relationship on the sequence level.

For example, Nussinov (2) has described universal rules on dinucleotide assymetry, which govern doublet frequencies. Fickett (3) and Smith, et al. (4) have detected statistical properties unique to coding sequences. Grantham, et al. (5,6) have compiled extensive data on codon usage clearly demonstrating a nonrandom use of synonymous codons, while, Lipman & Wilbur (7) have subsequently shown a contextual constraint on synonymous codon usage in eukaryotes. Data is also accumulating on the nonuniform base composition along nucleic acid sequences (i.e. adenine clusters, adenine & thymine rich regions) (8,9).

The statistical patterns are held to be largely the result of functional constraints acting on the evolution of sequences. These constraints may be virtually universal (2,8) or they may be different in coding versus noncoding regions (3). In either case, such statistical patterns might influence the distribution of similarity values between randomly chosen pairs of nucleic acid sequences. Recently, Fitch (10) proposed retaining the nearest neighbor frequencies of the original sequences in a Monte

Carlo method, however he did not, with the exception of two hypothetical examples, examine the effect of this modification.

We will show that two known statistical properties of sequences, the nearest neighbor frequencies, and the local fluctuations in base composition, can greatly affect the distribution of similarity between randomly generated sequences. We will also demonstrate that the choice of statistical model can affect conclusions regarding the statistical significance of similarities detected between specific pairs of sequences. We propose that consideration of the shared statistical properties of nucleic acid sequences will assist in understanding the relationship between statistical significance and biological significance.

METHODS

Similarity between two sequences was determined using the local similarity algorithm of Smith & Waterman (11). A similarity score is obtained by adding the number of base matches with penalties assessed for the number of mismatches, gaps and/or for each base in a gap. The algorithm considers all possible alignments of two sequences and finds the subsequences with the best possible score under the scoring rules. The scoring rules used in all comparisons were,

$$\text{match} = +1.0 \;;$$
$$\text{mismatch} = -0.9 \;;$$
$$\text{each base in a gap} = -2.0 \;.$$

A similarity score for a given alignment is the sum of all the matches, mismatches and gaps, weighted by the above factors. The algorithm will find the alignments which correspond to the highest possible score.

The computer simulations of various statistical models will be described in the Results. Computing was done on the DEC KL-10, and VAX 11/780 computers at the National Institutes of Health, and on the CRAY1 computer at Los Alamos National Laboratory. All sequences were taken from the National Institutes of Health DNA Data Bank of Los Alamos Scientific Laboratory (Genbank).

RESULTS

To examine the effect of the statistical properties of sequences on the distribution of similarity values, a set of 100 vertebrate nucleic acid sequences was randomly chosen from Genbank (the library set). The Genbank sequences had previously been screened to remove exceptionally long or short sequences. A subset of ten sequences (the query set) was also randomly chosen and each of its members were compared for local similarity with all 100 sequences in the library set. The mean, median and estimated standard deviation of the resulting 1000 similarity scores were computed and are in Table I.

Three subsequent sets of comparisons were also made with randomized sequences which retained specified statistical properties of the library set.

Randomized Set A (base composition preserved):

One hundred sequences were generated retaining the length distribution as well as the base composition of the library set. Two techniques exist to accomplish this task. The usual one is to generate a random sequence where the (independent) probability of a base in any

Table I

SET	MEAN	MEDIAN	E.S.D.[1]
The Library Set	15.0	14.8	4.44
Randomized Set A (base composition preserved)	13.3	13.3	2.81
Randomized Set B (nearest neighbors preserved)	14.7	14.7	3.25
Randomized Set C (local base composition preserved)	13.5	13.7	3.85

[1] Estimated Standard Deviation: calculated as .740 * (interquartile range), which is unbiased for the standard deviation if the distribution is normal.

position is proportional to its composition in the original sequence. Second is the shuffling technique where a random permutation of the original sequence is calculated (12). For sequences of length greater than 30, these two techniques are equivalent.

Randomized Set B (nearest neighbors preserved):

One hundred sequences were generated retaining the length distribution and nearest neighbor frequencies of the library set.

Randomized Set C (local base composition preserved):

One hundred sequences were generated retaining the length distribution and base composition of the library set as well as the pattern of fluctuations in base composition along each sequence (i.e. regions rich in adenine and thymine remained so). This was done by shuffling the sequence within 12 base blocks; the blocks overlapped each other by 3 bases. This choice of block size and overlap size were somewhat arbitrary. Small variations in block size and overlap size however have little effect on the final results.

The query set sequences were compared to each of the randomized library sets. 1000 pairwise comparisions were thus made for each case, the mean, median, and estimates of spread or standard deviation were calculated and are shown in Table I.

Figure 1 contains histograms of the different sets of pairwise comparisons. The upper histogram is based on the comparisons among the real sequences and the lower histograms are based on comparisons among the different sets of randomized sequences. Sequence comparisons resulting in similarity scores greater than 36 are not shown. For the real sequences, most pairs with scores greater than 36 represented known homologies (i.e. related protein coding regions or repetitive elements). Those pairs in the group below 36 undoubtedly share short homologies such as Pribnow boxes, or some heretofore unidentified, related repetitive elements. For the Set A comparisons (base composition preserved) and Set B comparisons (nearest neighbors preserved) there were no scores greater than

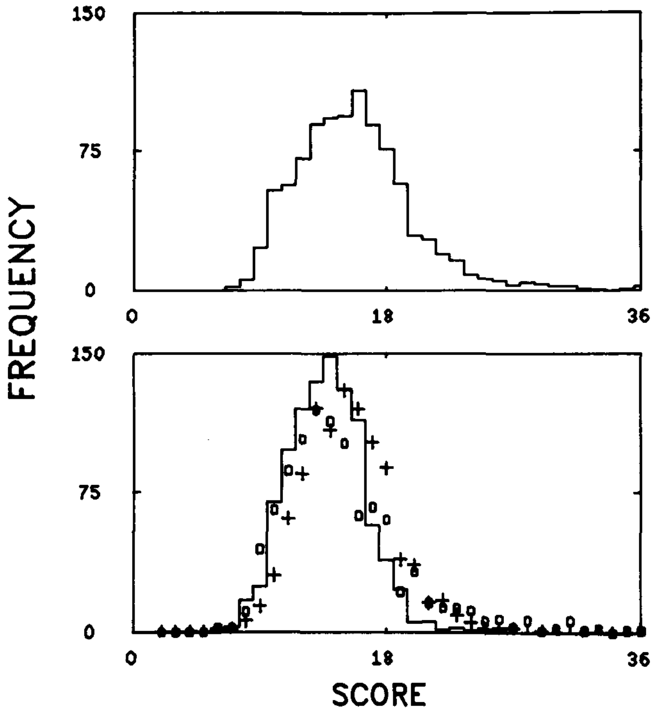

Figure 1:

Similarity comparisons of Query Set with indicated reference sets. Upper Histogram: Query Set versus Library Set. Lower Histogram: Query Set versus Randomized Set A (base composition preserved), solid line; Randomized Set B (nearest neighbors preserved), "+"; Randomized Set C (local base composition preserved), "o".

30. For Set C comparisons (local base composition preserved) there were 15 scores greater than 36.

Consistent with the means and standard deviations shown on Table I, the area under the right hand tail of the histogram is least in the comparisons between Set A sequences (base composition preserved). In addition, the histograms for the Set B, and the Set C comparisons are more similar to the histogram for the real sequences than is the histogram for the Set A comparisons. Thus it is clear that the distribution of similarity values is dependent on the statistical distributions of the ensemble of sequences being considered.

We now examine how the choice of statistical model can affect conclusions regarding the statistical significance of similarity values of specific pairs of sequences. Four particular comparisons were chosen because of their relatively high respective similarity scores, and because of the lack of biological context in which to judge the significance of the similarities found. (The choice of pairs was not influenced by whether or not a member of a pair coded for protein.) The pairs are listed in Table II.

We will evaluate the statistical significance of the similarities detected with respect to three models.

Model A (base composition preserved)

The degree of similarity found between two sequences is due to global similarity in the base composition of the sequences.

Table II

Pair #1 (similarity score=31.5, 3.8 s.d. above mean[1])
Human alpha-2 globin gene:
The subsequence encompassed the 3' flanking region, the first exon and approximately the first half of the first intervening sequence.
Xenopus laevis ribosomal RNA genes:
The subsequence between the 18s ribosomal RNA gene and the 5.8s ribosomal RNA gene.
Pair #2 (similarity score=37.6, 5.1 s.d. above mean)
Human leukocyte (lambda alpha-2) cDNA:
The 5' flanking region.
Human delta globin gene:
The 3' end of the second intervening sequence.
Pair #3 (similarity score=37.7, 5.2 s.d. above mean)
Mouse B1 repetitive cDNA, copy A:
The subsequence encompassing the B1 consensus sequence.

Rat preprolactin gene;
The subsequence within the first intervening sequence.
Pair #4 (similarity score=29.4, 3.3 s.d. above mean)
Mouse H2 transplantation antigen cDNA:
The subsequence within first third of cDNA.
Rat preproinsulin gene:
The subsequence from 3' end of second intervening sequence through coding region #2.

[1] The number of standard deviations above mean, using the values on Table I.

Model B (nearest neighbors preserved)

The degree of similarity found between two sequences is due to similarity in the nearest neighbor frequencies of the two sequences.

Model C (local base composition preserved)

The degree of similarity found between two sequences is due to local similarities in the base composition of the two sequences.

To indicate the statistical significance of a similarity under each of the above models, a mean and standard deviation is determined for 100 randomized sequence pairs; each sequence in a pair retaining either the overall base composition (Model A), nearest neighbor frequencies (Model B), or local base composition (Model C) of its respective original sequence. To maintain doublet frequencies, we used an algorithm similar to that of Fitch (1983) (Our algorithm does not have the possibility of nonuniform sampling of the state space and will be reported on elsewhere.). To maintain local base composition, the bases within nonoverlapping four base blocks of a sequence are shuffled. This is the smallest blocksize which insures adequate randomization. The number of standard deviations above the mean for each model is displayed in Table III.

For pair #1, the three models cannot be discriminated between. The relative similarity between this pair is adequately accounted for by base composition alone (Model A). While Models A and B do not account for the similarity value of pair #2, Model C clearly does. Thus the relative similarity between pair #2 appears due to similar, nonrandom local fluctuations in base

Table III

Significance[1] under each Model

	Pair			
	#1	#2	#3	#4
Model A	<0	4.4	23.5	12.2
Model B	<0	6.3	12.1	7.6
Model C	<0	0	5.7	5.9

[1] The number of standard deviations from the mean.

composition. Without this analysis, simply running a Model A Monte Carlo gives the impression of high statistically significant similarity. Pairs #3 and # 4 require a more detailed examination which is given in the discussion.

DISCUSSION

Table I and Figure 1 show how strongly statistical properties known to be present in biological sequences affect the distribution of similarity values. As an example, consider a similarity score of 24. 1000 comparisons among Set A sequences (base composition preserved), which lack statistical properties present in real sequences, produced a score of > 24 only three times. Randomized Set B (nearest neighbors preserved) produced a score of > 24 thirteen times, while Randomized Set C (local base composition preserved) produced a score > 24 sixty three times. Thus, although the sequences in sets A,B, and C are random, the distributions of their respective similarity scores are clearly different. A given similarity score would be considered more significant under Model A then under Model C. What then should the reference set be?

Perhaps the ideal similarity score distribution for evaluating the evolutionary significance of an apparent similarity would be determined from a large, representative sample of all naturally occuring nucleic acid sequences from the same taxon and encoding the same kind of information (protein, structural RNA, regulatory signals, etc.), which do not have a known evolutionary relationship at the sequence level. The upper histogram of Figure 1 would seem to be just such a distribution. Its most important failing is that the current nucleic acid sequence data bank does not yet contain a representative sample of all naturally occuring nucleic acid sequence types. Thus, while this ideal reference set would be desirable, it is not a realistic option with the present data base.

An available alternative is the combination of Monte Carlo methods we used to evaluate specific examples of Table II. Pair #1 provides an excellent example of the weakness of using a representative sample of the present limited data base as a reference set. Although the similarity score is 3.8 standard deviations above the mean for this set, one would not reject Model A for this pair. The relatively high similarity score was due to a long stretch of biased base composition, apparently only rarely seen in the data bank. If, however, the data bank did contain a truly representative sample of all naturally occuring, unrelated sequences, then a high relative similarity score based solely on a global similarity in base composition may be sufficient to suggest relatedness. For Pair #2 the relatively high similarity score was explained by similarities in local base composition, Model C, but not by Models A or B. In this case, the subsequences matched were extremely guanine and cytosine rich. Rejecting Models A and B but not Model C does not necessarily imply that there is no functional or evolutionary relationship between the two subsequences. It however establishes that the similarity is random within local regions and is therefore a useful analytical tool. We have observed instances of known biological similarities where the relationship appears random at the local level. In these cases, an overall architecture of large fluctuations in base composition appears important and not a particular sequence of bases.

Although the level of significance for pair #3 decreases dramatically from Model A to Model C, one would conclude that the similarity detected was statistically significant. This would suggest that there is a relationship between a subsequence within an intervening sequence of the rat prolactin gene and a mouse B1 sequence (an Alu like repetitive sequence). Gubbins, et al. (13) in their analysis of the rat prolactin sequence reported that the intervening sequences of the rat prolactin gene contained sequences repeated elsewhere in the rat genome. This observation and the rigorously established, statistically significant similarity between the mouse B1 sequence and a subsequence within the first intervening sequence of the rat prolactin gene, strongly suggest that this rat subsequence is an Alu like repetitive sequence.

All three models can be rejected for pair #4 and thus one can conclude that there is a statistically significant similarity between the two matched subsequences. The biological significance of this similarity is however unclear. The rat preproinsulin subsequence spans an intervening sequence and a protein coding region while the mouse transplantation antigen subsequence is entirely protein coding. Because of the locations of gaps and mismatches in the alignment, the level of similarity between protein coding regions decreased when translated into the amino acid sequence. This unexpected similarity remains an enigma and is evidence of the distinction between biological and statistical significance.

So far we have demonstrated that the standard Monte Carlo methods of assessing statistical significance, by ignoring the statistical properties of biological sequences, may give unrealistic results. The statistical methods used by Korn, et al. (14), Brutlag, et al. (15) and Goad & Kanehisa (16), may give unrealistic estimates of significance because they are based on assumptions which also ignore the statistical properties of biological sequences. The mathematical expression developed by Goad & Kanehisa (16) has additional problems in that it can be shown that it does not sum to unity over all the possible states, in fact it can result in probabilities > 1. All of these methods may be helpful at times but are clearly deficient when the estimation of statistical significance is critical in evaluating an apparent similarity.

In those instances when the estimation of statistical significance is critical, the approach we have employed should be useful. By testing for statistical significance in this series of models, one can be assured that a statistically significant similarity is the result of a higher level (though not necessarily biologically significant) relationship between the sequences than the shared, nonspecific constraints which, to a large extent, determine the statistical properties of biological sequences. When the similarity values fit a model, one can determine which statistical properties are responsible for the apparent similarity. This information will be useful in giving a perspective on the relationship of the two sequences, particularly as we learn more about the forces which determine the statistical properties of sequences.

A disadvantage of this method is that it requires moderate computation. A useful observation therefore is that although it is clear the doublet frequencies affect the distribution of similarity and controlling for the doublets greatly reduced the significance of most similarities, we have not yet found an example in which the conclusion of statistical significance was altered by imposing Model B (nearest neighbor frequencies preserved), although we assume that such examples exist. Therefore, in many instances, it may not be necessary to include this test.

Even when imposing some of the important statistical properties of sequences, apparent similarities (such as that between pair #4, the rat preproinsulin gene and the mouse H2 transplantation antigen) may be found which have a poorly understood biological basis. Such enigmas may eventually be explained by experiment or by further improvements of the statistical model.

Acknowledgements

The work of the second two authors (T.F.S. & M.S.W.) was partially supported by System Development Foundation. Computation was provided, in part, by Los Alamos National Laboratory.

REFERENCES

1. Wilbur,W.J. & Lipman,D.J. (1983). Proc.Natl.Acad.Sci.,U.S.A.
2. Nussinov,R. (1980a). Nuc.Ac.Res.,8,4545-4562.
3. Fickett,J.W. (1982). Nuc.Ac.Res. 10,5303-5313.
4. Smith,T.F.,Waterman,M.S. & Sadler,J.R. (1983). Nuc.Ac.Res.,11, 2205-2220.
5. Grantham,R.,Gautier,C.,Gouy,M.,Mercier,R. & Pave,A. (1980a). Nuc.Ac.Res., 8, r49-r62.
6. Grantham,R.,Gautier,C. & Gouy,M. (1980b). Nuc. Ac.Res.,8,1893-1912.
7. Lipman,D.J. & Wilbur,W.J. (1983). J.Mol.Biol., 163,363-376.
8. Nussinov,R. (1980b). J.Theor.Biol.,85,285-291.
9. Moreau,J.,Marcaud,L.,Maschat,F.,Kejzlarova-Lepesant,J.,Lepesant,J.A. & Scherrer,K. (1982). Nature, 295,260-262.
10. Fitch,W. (1983). J.Mol.Biol., 163,171-176.
11. Smith,T.F. & Waterman,M.S. (1981). J.Mol.Biol., 147,195-197.
12. Knuth,D.E. (1969). The Art of Computer Programming Volume 2, Addison-Wesley.
13. Gubbins,E.J., Maurer,R.A., Lagrimini,M.,Erwin,C.R. & Donelson,J.E. (1980). J.Biol.Chem.,255,8655-8662.
14. Korn,L.J.,Queen,C.L. & Wegman,M.N. (1977). Proc.Natl.Acad.Sci.,U.S.A.,74,4401-4405.
15. Brutlag,D.L., Clayton,J., Friedland, P. & Kedes, L.H. (1982).Proc. Natl. Acad. Sci., U.S.A., 74, 4401-4405.
16. Goad,W.B. & Kanehisa,M.I. (1982). Nuc.Ac.Res., 10,247-264.