Volume 13 Number 2 1985 Nucleic Acids Research

The statistical distribution of nucleic acid similarities

Temple F.Smith[1], Michael S.Waterman[2] and Christian Burks[3]

[1]Dept. Biostatistics, Harvard Med. Sch., Harvard Univ., 44 Binney Street, Boston, MA 02115, [2]Depts. Mathematics and Biol. Sciences, Univ. Southern California, Los Angeles, CA 90089, and [3]Theoret. Biol. and Biophysics Group, Los Alamos Nat. Lab., Los Alamos, NM 87545, USA

Received 5 June 1984; Revised 23 October 1984; Accepted 11 December 1984

ABSTRACT

All pairs of a large set of known vertebrate DNA sequences were searched by computer for most similar segments. Analysis of this data shows that the computed similarity scores are distributed proportionally to the logarithm of the product of the lengths of the sequences involved. This distribution is closely related to recent results of Erdos and others on the longest run of heads in coin tossing. A simple rule is derived for determination of statistical significance of the similarity scores and to assist in relating statistical and biological significance.

INTRODUCTION

Identification and interpretation of molecular sequence similarities is a fundamental problem in molecular biology. An increasing amount of nucleic acid sequence data is becoming available in such data bases as GenBank in the U.S. and the EMBL data bank in Europe. A compendium of the data has appeared as a supplement to Nucleic Acids Research (1). These data can be analyzed for relationships, both functional and evolutionary, by a variety of techniques (briefly reviewed in (2)). The recent identification (3) of a simian sarcoma viral onc gene with a human growth factor is a good example of the utility of these data. Useful computer methods have been developed for this analysis, where, among other techniques, dynamic programming is employed to find best matching (most similar) regions of sequences (4-6). Sequence comparison methods are reviewed in (7). What has been, until now, lacking in such analyses is a completely valid test to assess the statistical significance of these similarity scores observed between DNA sequences. Though the literature abounds with sequence alignments, and biological arguments based on those alignments, there is very seldom any estimate provided of the statistical significance (given the lengths and compositions of the two sequences being compared) of those alignments. This article addresses the need to provide such an estimate.

Even if it were not possible to give a derivation of the statistical distribution of similarity scores from first principles, the existing nucleic acid sequence data are sufficient for an empirical investigation of the distribution. Such an investigation is important since all known heuristic and Monte Carlo techniques frequently assign statistical significance where unwarranted (8). The data can be divided into subsets of sequences having similar function and taxonomic classification. Different distributions might be anticipated for these subsets. For example, protein coding sequences might display higher similarity among themselves simply due to their similar statistical properties (base composition and nearest neighbor frequencies (8)). Our subsets include 204 vertebrate DNA protein coding sequences as well as eukaryotic structural RNA's, eukaryotic viruses, vertebrate non-coding sequences, and non-vertebrate eukaryotic sequences from GenBank (1). For example, we have compared vertebate DNA sequences (and their complements) and eukaryotic virus sequences to a set of 204 vertebrate DNA sequences.

We present both empirical evidence and theoretical justification for a specific statistical distribution of the similarity scores among biologically realized sequences. This leads to a simple rule for assessing statistical significance of similarities. The method developed in this paper is not only of practical value for nucleic acid sequence analyses, but is shown to be related to important recent developments in probability theory.

METHOD

All sequence data were from GenBank (1). The alignment algorithm employed in this study incorporates genetic transformations (base substitutions and deletion/insertions) and finds the most similar or highest scoring segments between two sequences; this algorithm has been described in detail previously (4). The similarity score of two aligned segments is the number of matches minus penalties for mismatches and gaps. The algorithm finds the maximum of the scores of all such aligned pairs of segments, therefore finding the best matching segments out of all possibilities. The algorithm is a generalization of the dynamic programming algorithm introduced by Needleman and Wunsch (9), and was designed for the specific nature of the data, which include many repeated (e.g. Alu sequences) and biologically related (e.g. mRNA and genomic sequences) segments.

A FORTRAN program was developed to implement the above similarity algorithm on a CRAY-1 computer system. By utilizing the vector architecture of this computer it is possible to investigate comparisons among very large numbers of nucleic acid sequences in reasonable execution time. All pairwise comparisons among 204 vertebrate sequences (including the complement strands) were carried out in approximately 170 minutes, at a rate of over 240 sequence comparisons per minute with an average sequence length of 800 nucleotides.

To simplify the problem of comparing these results, the algorithm parameters were held constant. While the ability to identify overall sequence homology among a given set of sequences is dependent on the algorithm parameters (10) and the statistical characteristics (8) of the genetic domains involved, the identification of maximal segment homologies appears to be less sensitive. The parameter values used in this study -- matches equal 1.0, mismatches equal -0.90 and gaps (single base deletion/insertions) equal -2.0 -- were chosen because they allow a high proportion of the known segment homologies among hemoglobin protein coding regions to be identified. In cases where previously identified hemoglobin homologies were not reproduced exactly with these parameters, the differences involved only a slight rearrangement of neighboring gaps. This was true even for previously studied non-protein encoding sequences such as the ribosomal RNAs (thus increasing confidence in the employment of these particular parameter values). The percentage of matched bases (including gaps) and the ratio of implied transitions to transversions among the aligned mismatches were also calculated. Among previously identified homologies the percentage is generally greater than sixty-eight and the ratio greater than two thirds.

The set of best similarity scores resulting from comparison of each given sequence with all sequences in the vertebrate data set was used to generate a frequency distribution. Representative examples of these distributions appear in Figure 1. The sequence being compared with the other sequences in a data set will be referred to as the query sequence.

Although similarity scores of 40 or larger are considered outliers and are easily identified as statistically significant, assessment of lower scores requires a deeper analysis. It is natural to ask whether these frequency distributions, or some subset of them, are normally distributed. A Lilliefor's test (11) of normality was run on these distributions where

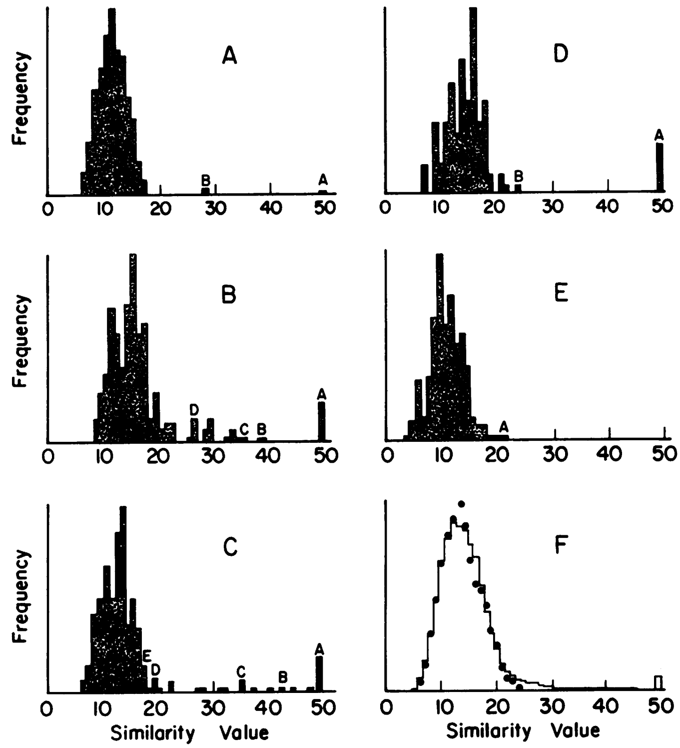

Figure 1. A-E are similarity score histograms of observed maximum similarities of a single query sequence to the members of a reference set of sequences. F contains two composite histograms. All values higher than 50 were recorded at 50. A) Histogram obtained from the query of 204 vertebrate sequences using the chicken x-gene (32); AA represents homology with chicken ovalbumin (33) while AB represents the suspected (10) homology with the primate alpha-1 antitrypsin. B) Histogram obtained from the query of 423 eukaryotics and viral sequences using the mouse alpha hemoglobin pseudogene (19); BA represents homologies with seven other vertebrate alpha globins; BB represents the least similar alpha globin, the human pseudogene (34); BC-BD represent the other hemoglobins ranging from the X.laevis beta globin (20) to the rabbit beta globins (35). C) Histogram obtained from the query of 204 vertebrate sequences using one of the mouse B1 ubiquitous repeat (21) sequences; CA-CB represent the other mouse B1's (21), two Chinese Hamster equivalents (36) and two human Alus (22) that neighbor the epsilon globin and preproinsulin genes; CC represents mouse and hamster RNAs (36), presumably arising from B1-like repeat transcription; between CC and CD are all the other unequivocal Alu/B1-like sequences including those from rat, human, and mouse; CE includes a number of apparently unrelated short sequence similarities, but also includes the most distant previously identified hamster Alu-like sequence, 250 close (36). D) Histogram obtained from the query of 160 vertebrate protein coding (spliced) sequences using the bovine growth hormone, presomatotropin (37); DA represents four other somatotropin sequences from human (38) and rat (39);

DB represents the next most similar sequence found in a mouse immunoglobulin heavy chain constant region (40). E) Histogram from the query of the vertebrate non-protein encoding sequences using the same coding sequence for a query as in D above; EA represents the most similar sequence within this data set, a rat tRNA cluster. F) the sum of 423 eukaryotic similarity histograms, solid line; and the sum of 100 similarity histograms for random sequences having nearest neighbor frequencies identical to those found in vertebrate coding regions (8), solid circles.

all similarity scores larger than 40 were trimmed from the distribution. A one percent level test resulted in rejection of normality in 98 percent of the cases (see Fig. 1A for an example of one of the few distributions passing this test). These results clearly indicate that statistical significance should not be assigned by standard normal distribution techniques.

Earlier attempts to perform analysis of the distributions of matches for comparison of random sequences have provided few results directly useful for sequence analysis. Chvatal and Sankoff (12) began studies of the distribution of the number of matches in random sequences where gaps and mismatches receive no penalty. Their problem, known as the longest common subsequence problem, has attracted a good deal of attention but nothing directly applicable to the more general problem of molecular sequence comparison. The difficulty of this problem seems to leave little hope for a complete distribution theory.

Deriving the probability distribution of the length of the longest run of heads in a sequence of n independent coin tosses is a problem with a long history of solutions difficult to do computations with (13). In 1970, however, Erdos and Renyi found the longest run of heads to be, in the limit with probability one, log(n) where the logarithm is to base 1/p, p=P(Heads) (14,15). The technical statements of these and related results, known as the Erdos-Renyi law, are involved and precise formulations appear in the references (14,15).

The coin tossing problem is related to sequence matching problems in the following way. Two random sequences of length n are aligned by

$$\begin{matrix} a_1 \; a_2 \; \cdots \; a_n \\ b_1 \; b_2 \; \cdots \; b_n \end{matrix}$$

We now convert the alignment to a sequence of H's and T's. If $a_i = b_i$, an "H" replaces $\begin{smallmatrix} a_i \\ b_i \end{smallmatrix}$; otherwise a "T" does. This replaces the alignment by a sequence of heads and tails. The length of the longest run of matches in the alignment is equal to the length of the longest run of heads in the associated coin tossing sequence, and therefore follows the Erdos-Renyi law.

The algorithm employed in the present study gives the best matching region for all possible alignments, motivating the following formulation. Let R_i be the longest run of uninterrupted matches for the particular alignment

$$\begin{array}{l} a_1 a_2 \cdots a_i \; a_{i+1} \cdots a_n \\ \qquad\qquad\quad b_1 \; b_2 \quad \cdots b_{n-i+1} \cdots b_n \end{array}$$

Here $R_i \leq n-i+1$ and, for n-i large, the Erdos-Renyi law holds for R_i. The best of all these R_i is R where

$$R = \max_{-n<i<n} R_i$$

It is possible to prove (18) that the limit law of R is equivalent to an Erdos-Renyi law with a different constant, that is, 2.0 multiplied by $\log(n)/\log(1/p)$, where $p = P(\text{Match}) = p_A^2 + p_T^2 + p_G^2 + p_C^2$. For sequences of length n and m, the expected value of R, allowing k mismatches, is

$$E(R) = (\log(nm) + k \log\log(nm) + (k+1) \log(1-p) - \log(k!))/\log(1/p) + k + \gamma/\lambda - \tfrac{1}{2}$$

and the variance is

$$\sigma^2 = \pi^2/6\lambda^2 + 1/12$$

where $\gamma = .577\ldots$ is the Euler-Mascheroni constant and $\lambda = \ln(1/p)$. These results, with a complete error analysis, appear in a paper by Arratia et al. (16). Karlin et al. (17) announced a related result with $k = 0$ (no mismatches allowed) and slightly different constants. Surprisingly, the variance does not grow with n. There are mathematical reasons that lead one to believe that this feature of essentially constant variance also holds for a reasonable number of mismatches, deletions, and insertions (18).

RESULTS AND DISCUSSION

To study these data from this viewpoint, the similarity scores were plotted versus the logarithm of the products of the sequence length, where log is again to the base 1/p, p = P(Match). A strong linear trend is observed, with essentially constant variance, and the data are shown in Figure 2. The possible influence of the biological properties of the sequences on these results was tested by comparison with Monte Carlo simulations of sequences with the same nearest neighbor frequencies as the biological sequences (8); this test resulted in linear trends with slopes

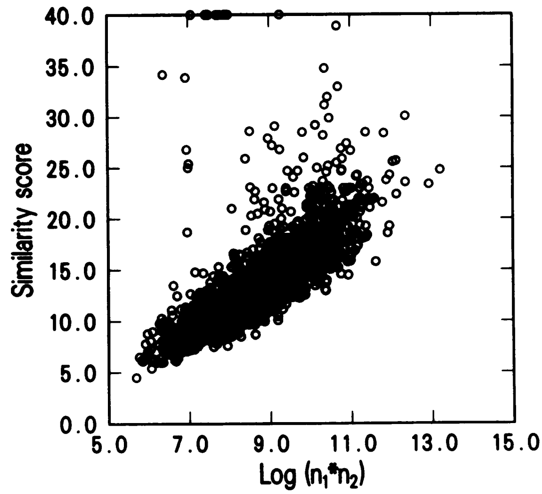

Figure 2. Similarity scores of vertebrate DNA sequences and their complements and eukaryotic virus sequences (a total of 20,706 data points) compared with a set of 204 vertebrate DNA sequences plotted against $\log(n_1n_2)$ where n_1 and n_2 are sequence lengths and log is to the base $1/p$, where $p = P(\text{Match})$. Points plotted on the upper horizontal axis represent similarity scores ≥ 40.0. Each point represents the best similarity score found in comparing the corresponding query sequence to the 204 vertebrate sequences.

close but not identical to the slopes resulting from the biological sequences.

We also studied the results of querying two clearly biologically disjoint data sets — vertebrate protein coding and non-protein coding (see examples in Fig. 1D and 1E). The general statistical properties of the resulting frequency distributions for similarity scores were quite close to each other and to those generated by querying the full vertebrate data set. While there was a slight (constant) increase in the distribution mean when querying the protein coding data set with protein coding sequences (as compared to querying the non-protein coding data set with protein coding sequences), the linear relationship was retained with approximately identical slope. Sensitivity to the algorithm parameters was explored by varying the algorithm weights as well as the form of the gap weights (10). The linear trends persist, with the slope decreasing as the mismatch and gap penalties increase.

To estimate statistical significance, there are two approaches. We calculate how many standard deviations a similarity score is above the mean and judge it significant if it is more than, say, 2σ above the mean. It is possible to take a much more cautious approach with Chebyshev's inequality (13): The probability that a random variable exceeds its mean by more than λ is less than or equal to $(\sigma/\lambda)^2$. Our $\hat{\sigma} \approx 1.5$, calculated from the data in Figure 2, so that a similarity score exceeds its (estimated) mean by more than 4.5 with probability less than or equal $1/9 \approx .111....$, by more than 6.0 with probability less than or equal $1/16 \approx 0.0625$, etc. Both of these procedures are useful and conservative. We use the first method and calculate the number of $\hat{\sigma}$'s a similarity score exceeds the mean.

A fit of the data displayed in Figure 2, using robust techniques for handling outliers, resulted in the equation

[Eq. 1.1] $$\hat{S} = 2.55 \frac{\log(nm)}{\log(1/p)} - 8.99,$$

where $\hat{S}$ is the mean best similarity, n and m are the sequence lengths and p $= P(\text{Match}) = P_A^2 + P_T^2 + P_C^2 + P_G^2$. The estimate of σ from the data is

[Eq. 1.2] $$\hat{\sigma} = 1.78.$$

These values are now used to examine certain comparisons.

These values for $\hat{S}$ and $\hat{\sigma}$ can then be compared to the similarity score for any actual alignment, and thus provide a criterium for appraising statistical significance. The following paragraphs provide several examples of the calculation of statistical significance. These examples also illustrate the fact that statistical significance and biological significance are closely related but not identical.

In many cases, alignments indicating high statistical significance are the result of comparing two sequences already known to be biologically related (homologous). In Figure 1B a query of the data using a mouse alpha hemoglobin (19) produced a wide spread in similarity scores. Since the hemoglobins form a large and divergent family dating from before the origins of the vertebrates, the nearly continuous range of observed similarities is not surprising. Even the distant beta globin of the African toad, X. laevis (20), has a similarity score with the mouse alpha globin corresponding to 6 $\hat{\sigma}$'s above the mean, indicating statistical significance. Since the mouse sequence has n = 1441, the toad sequence has m = 600, and p = .248, the mean was estimated by

$$2.55 \frac{\log((600)(1441))}{\log(1/.248)} - 8.99 = 16.01.$$

The score was 27.00 so that 27.00 - 16.01 = 10.99 = (6.17)(1.78) = 6.17$\hat{\sigma}$.

Using a highly repetitive sequence such as mouse B1 (21) also generates a wide spread of similarity scores (Fig. 1C), but in this case statistical and biological signifance can be confused. Similarity scores between 16.5 and 20.0 are contributed both by apparently biologically unrelated sequence segments and by a previously identified hamster Alu-like repetitive segment (36). In such cases, statistically significant similarity scores may not reflect true homology (clear biological function or taxanomic relatedness), but merely compositional or pattern restrictions common to the compared sequences. An extreme example is the CCRCC (R = purine) repeat found in the winter flounder (P. americanus) antifreeze protein gene (23). This compositional restriction leads to high similarity to other sequences with regions rich in C (or complements of sequences rich in G). Similarly, the $(TC)_{24}$ region at the end of three rat tRNA genes (24) matches the complement of the mouse immunogloblin γ-1 intron (25), which contains a $(GAGAG)_{15}$ region, with a similarity score of 54.90. In this last example the mean is estimated by

$$2.55\,\frac{\log\,((764)(2109))}{\log(1/.265)} - 8.99 = 18.42.$$

The similarity score, 54.90, exceeds the mean by 20.49$\hat{\sigma}$.

There are a few cases where similarities are equal to or greater than four σ's above the mean and for which no reasonable biological justification yet exists. The best example observed in these data is obtained from comparison (see Fig. 1E) of the 18S rRNA of X. laevis (26) and an intron in the IE gene of Herpes simplex virus (27), which yielded a similarity score of 37.20. Here the mean is estimated by

$$2.55\,\frac{\log(2948)(431)}{\log(1/.279)} - 8.99 = 19.11.$$

S = 37.20 exceeds 19.11 by 10.16$\hat{\sigma}$.

Numerous segment similarities can now be clearly identified as statistically insignificant in spite of appearances. For example the algorithm aligns the following segments between the protein encoding regions of yeast actin (28) and mouse alpha-fetoprotein (29).

```
GTTCTGGCATGCTGCAAAGCTG
GTTCTGGTATG-TGTAAAGCGG
```

The expected score is

$$2.55\,\frac{\log(2012)(1750)}{\log(1/.251)} - 8.99 = 18.83.$$

The actual similarity score of 13.3 = 18.0(18 matches) - 2.7(3 mismatches) - 2.0(1 deletion) therefore supports chance rather than biology, even though there are 87 percent matches and two out of the three implied point mutations are transitions. Note that the gap is not a multiple of three as expected for homologous coding regions.

Biological and/or experimental information can explain what might otherwise by surprisingly significant alignments. Statistically significant similarities were often found when the query sequences were the complements of the 'sense' or published strands. As expected, the similarity value distributions were on the average equivalent to those generated by the original sequence. Structural rRNA's have the interesting property that they are more similar to their complements than to any other complemented sequence. This is no doubt the result of the secondary structure motifs in these molecules. Unexpectedly, a few cDNA sequences (from mRNA) were found to be highly self-complementary as well. For example, cDNA from rat preprorelaxin mRNA (30) shows a weak imperfect reflected repeat in the first and second thirds of the B peptide. An even stronger example appears in the published cDNA sequence from the human enkephalin precursor mRNA sequence (31). Here the first 113 bases of the presumptive mRNA leader are found repeated exactly as a reverse complement some seven hundred bases downstream. The fact that the repeat is perfect and comprises one of the termini of the cDNA suggests that it may have arisen as a reverse transcriptase error. The statistical significance of the resulting similarity value draws our attention in this case to a potential experimental complication rather than a historical biological event.

In summary, Equations 1.1 and 1.2 provide a quick method of estimating the statistical significance of sequence alignments. For alignment algorithms employing the weighting parameters used here (match = 1.0, mismatch = -0.9, deletion = -2.0) the constant values in these equations are good as they stand; for alternative parameter weights, Eqs. 1.1 and 1.2 can be rederived using a fitting procedure for data such as that in Fig. 2. Finally, we stress that the affirmation (or negation) of the biological significance of a given found similarity should be based in part, though not entirely, on the statistical significance.

ACKNOWLEDGEMENTS

The authors T.F.S. and M.S.W. were supported by a grant from the System Development Foundation, and C.B. was supported under the auspices of the U.S. Department of Energy. Computer support was provided through Los Alamos National Laboratory. GenBank is funded by NIH contract NO1-GM-2-2127 with joint support from the NIGMS, the NCI, the NIAID, the DRR of the NIH, the NSF, the DOE and the DOD. We are grateful for the graphical expertise of B. Atencio.

REFERENCES

1. Andersen, J.S., Anderson, J., Atencio, E., Bergman, B.E., Bilofsky, H.S., Brown, L., Burks, C., Cameron, G.N., Channin, D.S., Elbe, U., England, C., Fickett, J., Goad, W., Hamm, G.H., Hayter, J.A., Kay, L., Kanehisa, M., Koile, K., Lennon, G., Linder, R., Lewitter, F.I., McLeod, M., Melone, D.L., Myers, G., Nelson, B., Nelson, D., Nial, J.L., Perry, H.M., Rindone, W.P., Rudloff, A., Simon, S., Smith, T.F., Stoesser, G., Stueber, K. (1984). Nucleotide sequences 1984, IRL Press, Oxford.
2. Smith, T.F. and Burks, C. (1983) Nature 301, 194
3. Doolittle, R.F., Hunkapiller, M.W., Hood, L.E., Devare, S.G., Robbins, K.C., Aaronson, S.A. and Antoniades, H.N. (1983) Science 221, 275-276
4. Smith, T.F. and Waterman, M.S. (1981) J. Mol. Biol. 141, 195-197
5. Wilbur, W. and Lipman, D. (1983) PNAS (USA) 80, 726-730
6. Goad, W.B. and Kanehisa, M. (1982) Nucl. Acids Res. 10, 247-264
7. Waterman, M.S. (1984) Bulletin Math. Biol.
8. Smith, T.F., Waterman, M.S. and Sadler, J.R. (1983) Nucl. Acids Res. 11, 2205-2220
9. Needleman, S.B. and Wunsch, C.D. (1970) J. Mol. Biol. 48, 443-453
10. Fitch, W.M. and Smith, T.F. (1983) PNAS (USA) 80, 1382-1386
11. Lilliefors, H.W. (1967) J. Amer. Statist. Assoc. 62, 399-402
12. Chvatal, V. and Sankoff, D. (1975) J. Appl. Prob. 12, 306-315
13. Feller, W. (1968) An Introduction to Probability Theory and its Applications, 3rd Ed., New York
14. Erdos, P. and Renyi, A. (1970) J. Analyse Math. 22, 103-111
15. Erdos, P. and Revesz, P. (1975) Topics in Information Theory, Math. Soc. J. Bolyai 16, pp. 219-228
16. Arratia, R., Gordon, L., and Waterman, M.S. (1984). Manuscript.
17. Karlin, S., Ghandour, G., Ost, F., Tavare, S. and Korn, L.J. (1983) PNAS (USA) 80, 5660-5664
18. Arratia, R. and Waterman, M.S. (1984) Adv. Math., in press
19. Vanin, E.F. and Goldberg, G.I., Tucker, P.W., Smithies, O. (1980) Nature 286, 222-226
20. Williams, J.G., Kay, R.M. and Patient, R.K. (1980) Nucl. Acids Res. 8, 4247-4258
21. Krayev, A.S., Kramerov, D.A., Skryabin, K.G., Ryskov, A.P., Bayev, A.A. and Georgiev, G.P. (1980) Nucl. Acids Res. 8, 1201-1215
22. Pan, J., Elder, J.T., Duncan, C.H. and Weissman, M.S. (1981) Nucl. Acids Res. 9, 1151-1170
23. Lin, Y. and Gross, J.K. (1981) PNAS (USA) 78, 2825-2829
24. Sekiya, T., Kuchino, Y. and Mishimura, S. (1981) Nucl. Acids Res. 9, 2239-2250
25. Kataoaka, T., Kawakami, T., Takahashi, N. and Honjo, T. (1980) PNAS (USA) 77, 919-923
26. Salim, M. and Madden, B.E.H. (1981) Nature 291, 205-208
27. Watson, R.J., Umene, K. and Enquist, L.W. (1981) Nucl. Acids Res. 9, 4189-4199

28. Ng, R. and Abelson, J. (1980) PNAS (USA) 77, 3912-3916
29. Law, S.W. and Dugaiczyk, A. (1981) Nature 291, 201-205
30. Hudson, P., Haley, J., Cronk, M., Shine, J. and Niall, H. (1981) Nature 291, 127-131
31. Comb, M., Seeburg, P.H., Adelman, J., Eiden, L. and Herbert, E. (1982) Nature 295, 663-666
32. Heilig, R., Perrin, F., Gannon, F., Mandel, J.L. and Chambon, P. (1980) Cell 20, 625-637
33. Woo, S.L.C., Beattie, W.G., Catterall, J.F., Dugaiczyk, A., Staden, R., Brownlee, G.G. and O'Malley, B.W. (1981) Biochemistry (USA) 20, 6437-6446
34. Proudfoot, N.J. and Maniatis, T. (1980) Cell 21, 537-544
35. Hardison, R.C., Butler, E.T., Lacy, E., Maniatis, T., Rosenthal, N. and Efstratiadis, A. (1979) Cell 18, 1285
36. Haynes, S.R. and Jelinek, W.R. (1981) PNAS (USA) 78, 6130-6134
37. Miller, W.L., Martial, J.A. and Baxter, J.D. (1980) J. Bio. Chem. 255, 7521-7524
38. Roskam, W.G. and Rougeon, F. (1979) Nucl. Acids Res. 7, 305-320
39. Seeburg, P.H., Shine, J., Martial, J.A., Baxter, J.D. and Goodman, H.M. (1977) Nature 270, 486-494
40. Ollo, R., Auffrey, C., Morchamps, C. and Rougeon, F. (1981) PNAS (USA) 78, 2442-2446