Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 4090–4094, June 1985
Biochemistry

Significance of similarities in patterns: An application to β interferon-related DNA on human chromosome 2

(nucleotide and amino acid sequence comparisons/Sellers TT algorithm/statistical evaluation/pattern score (Π)/human β_3 interferon locus)

LESTER T. MAY*†, FRANK R. LANDSBERGER‡, MASAYORI INOUYE*, AND PRAVINKUMAR B. SEHGAL‡§

‡The Rockefeller University, New York, NY 10021; and *Department of Biochemistry, State University of New York, Stony Brook, NY 11794

Communicated by Igor Tamm, March 12, 1985

ABSTRACT **The nucleotide sequence of a 14-kilobase (kb) region of the human β interferon (IFN-β)-related DNA locus on chromosome 2 (genomic DNA clone λB3) was determined and compared to that of the IFN-β_1 gene by using the Sellers TT algorithm. This algorithm aligns segments of one sequence with similar segments in a second sequence. A strategy was developed for assessing the significance of similarities between DNA sequences based on a scheme that recognizes patterns or runs of identities within an alignment. The pattern score (Π) thus obtained is an entropy-like measure. Numerically it is a reflection of the length of the second longest run of identity in an alignment plus a correction factor due to the other shorter identity runs in the alignment. When the IFN-β_1 gene is compared to a random nucleotide sequence, the distribution of Π scores in such comparisons fits a Gaussian function. This strategy has been used to identify seven segments along one strand of λB3 DNA that are related to segments in IFN-β_1; these seven alignments have Π scores ≥ 3 standard deviations above the mean score obtained in comparisons between IFN-β_1 and random nucleotide sequences. One of these alignments (section 7) has a Π score 8.02 standard deviations above this mean score. The likelihood of finding an alignment statement as good as that in section 7 in a random sequence the length of the human genome is approximately 10^{-7}. Furthermore, the λB3 DNA sequence in section 7 selects the human IFN-β_1 gene as the most significant alignment in computer searches of mammalian nucleotide sequence data bases.**

A number of algorithms have been devised in recent years that allow the detailed comparison of two genetic sequences (1). These include heuristic methods such as the dot-matrix (2, 3) as well as mathematical methods such as those based on the Needleman–Wunsch approach (4–8) or the Sellers approach (9–12). Among the mathematical methods, those based on the Needleman–Wunsch protocol essentially locate similarities by scoring matches between two sequences, whereas the Sellers algorithms are based on scoring mismatches between two sequences. The procedure whereby two sequences are compared by using a detailed mathematical analysis, termed "metric analysis" (13), has emerged as a powerful tool in identifying all patterns shared by two sequences that satisfy specific local and global criteria of similarity (13). For example, when used in metric analysis, the recent Sellers TT algorithm (11) provides a description of all patterns shared by two sequences that satisfy the length and mismatch density parameters preset by the investigator. The output of this algorithm can be presented as a two-dimensional path graph that readily shows the location and degree of similarity of common patterns. Each path in the output identifies an alignment or a set of alignments that are better than all alternative alignments in describing each specific common pattern.

After using Sellers TT algorithm or any other mathematical similarity search algorithm to compare, for example, two human sequences (or more generally, two sequences from a particular species), one must address a question of the following kind. What is the likelihood of finding a particular pattern by chance in the human genome?

The prevalent approach to this question (1, 4, 6, 7, 14–16) is to permute randomly one of the sequences in an alignment a large number of times (with refinements based on retention of nucleotide composition, dinucleotide frequency, and codon usage or of amino acid composition, dipeptide frequency, and amino acid characteristics), from which a mean identity score of the permuted sequences is calculated by the same scoring scheme. A standard deviation of the permuted scores is then calculated. If the score of the observation of the two original sequences is greater than the mean score of the permuted sequences by a predetermined number (3–5) of standard deviations, the particular alignment is considered to establish a statistically significant relationship between the two sequences (14–16).

There are some well-recognized problems with this approach (4, 6, 7, 13, 17). It is difficult to obtain a relevant statistical interpretation of the permutation analysis (6, 7, 13). This test provides a measure only of the internal order of the aligned sequences and determines whether the observed match is a consequence only of nucleotide composition or amino acid composition. It does not allow any predictions to be made about sequences outside of the test alignment (see refs. 6, 7, and 13 for a detailed discussion). To extrapolate the information obtained from this permutation test to longer sequences (e.g., the human genome) is theoretically difficult. It is not clear how to estimate the number of comparisons that must be made between a given sequence and another longer sequence (e.g., a random sequence the length of the human genome) (see ref. 17 for a discussion of this problem).

We have compared, using Sellers TT algorithm, kilobase-long sections of computer-generated random nucleotide sequence with that of the human β_1 interferon (IFN-β_1) gene (18). Remarkably good alignment patterns were consistently observed when the IFN-β_1 gene was compared to computer-generated random nucleotide sequences or to kilobase-long sections of the human nucleotide sequence extracted "at random" out of the Los Alamos data base (19). The conventional "scores" (based on counting the number of matches/mismatches and gaps) of these patterns were up to 6.05 standard deviations better than that of the mean of appropriate permuted comparisons. These estimates, as well as signifi-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: kb, kilobase; IFN, interferon.

†Present address: The Rockefeller University, New York, NY 10021.
§To whom reprint requests should be addressed.

cance analyses based on another method described in the literature (17), leads the investigator to the conclusion that these alignment patterns are extremely unlikely to be chance observations. Intuitively, however, it seems that the interferon gene sequence cannot be significantly related to a random sequence.

In this paper we describe a strategy for the assessment of the significance of similarities uncovered between sequences using search algorithms such as the Sellers TT algorithm. While the present description is largely confined to similarities in nucleotide sequences identified by using the TT algorithm, the strategy appears to be of general applicability—even of applicability to pattern similarities other than those in nucleotide or amino acid sequences. The essence of the strategy is to look for and score patterns (e.g., runs of identities) within an alignment and not the actual number of matches or mismatches. This strategy is not a search-mode algorithm but is a method to assess the significance of alignment statements identified by metric alignment algorithms. We have used this strategy to obtain a description of the significant nucleotide sequence patterns common between the IFN-β-related DNA on human chromosome 2 (genomic DNA clone λ3 in Charon 4A) (20, 21) and the IFN-β_1 gene on chromosome 9 (22, 23).

Studies of the production of biologically active human IFN-β in cultures of human–rodent somatic cell hybrids induced with poly(I)·poly(C) or with appropriate viral inducers have indicated that functional IFN-β genes are located on human chromosomes 2, 5, and 9 (reviewed in ref. 21). The available data are consistent with the assignment of IFN-β mRNA species of different lengths to different human chromosomes (e.g., IFN-β_1 to chromosome 9, IFN-β_2 to chromosome 5, IFN-β_3 to chromosome 2). The well-characterized human IFN-β_1 gene has also been assigned to human chromosome 9 by using DNA blot-hybridization procedures (22). Experiments in which a human genomic DNA library in λ phage Charon 4A was screened with the IFN-β_1 cDNA insert led, in addition to the isolation of the cognate IFN-β_1 gene, to the isolation of two unusual human genomic DNA clones (λB3 and λB4) that hybridized the IFN-β_1 cDNA insert (20, 24). Blot-hybridization analyses of DNA samples obtained from human–rodent somatic cell hybrids have led to the assignment of λB3 to human chromosome 2 (the "IFN-β_3" locus) and that of λB4 to human chromosome 4 (20, 24). Clone λB3 contains a human DNA insert of length 15 kb. Approximately 14 kb of this insert has been sequenced. Where in the λB3 nucleotide sequence are the regions related to IFN-β_1? What is the likelihood of finding such nucleotide sequence relationships by chance in the human genome?

BIOCHEMICAL METHODS

The recombinant human genomic DNA clone λB3 contains a 15-kb section of human DNA inserted into the *Eco*RI site of λ phage Charon 4A (20, 24). This human DNA insert was cut into smaller fragments by using an array of restriction endonucleases, and these were subcloned into pBR322-derived vectors or directly into M13 sequencing vectors (25). Approximately 14 kb of the human DNA in λB3, encompassing DNA sequences that cross-hybridize IFN-β_1 cDNA, were sequenced in both directions by appropriate cloning into M13 phage vectors, followed by sequence determination by the dideoxy method (26).

THEORETICAL METHODS

Random nucleotide sequences were generated by using a computer program written in MBASIC. The BMDP statistical software package (27) was used in deriving frequency histograms, and a Fortran program (28) was written for curve-fitting. Listings of the BASIC and Fortran programs used are available on request.

The Sellers TT algorithm (11) locates all patterns common between two sequences that meet specific criteria that can be preset by the investigator by adjusting three numbers in the TT "filter." These are an upper limit r for mismatch density, a lower limit s for the adjusted length of the pattern, and a penalty n for each inserted or deleted nucleotide. A completely random alignment has a mismatch density of 0.75. Thus, the useful range is $0.5 \leq r \leq 0.7$. We have used $r = 0.5$ or 0.6 in various experiments. We have used $s = 10$ and $1 \leq n \leq 9$ in our experiments. The usual search mode settings for the TT filter were $r = 0.5$, $s = 10$, and $n = 1$.

A simple strategy can be developed to describe the degree of similarity shared by two sequences. The arguments presented will explicitly refer to DNA sequences but can be readily generalized to amino acid sequences, etc. It is assumed that the best alignment has been found by using any metric algorithm. Several simple assumptions can be made. (*i*) The likelihood of a match (mismatch) occurring at a given position i is independent of whether a match (mismatch) is present at positions $i - 1$ or $i + 1$; (*ii*) all nucleotides are equally likely at a given position; and (*iii*) a run of identical residues is terminated when a run as short as a single mismatch is detected. Since there are four possible nucleotides, a single identity (a run of length 1) is scored as 4^1, a run of length 2 is scored as 4^2, a run of length 3 is scored as 4^3, and a run of length m is scored as 4^m. If one assumes that a sequence comparison is characterized by N runs of lengths m_a, m_b, m_c, etc., the comparison is scored by the sum $\Sigma 4^{m_i}$. To smooth the large fluctuations, it is preferable to define the pattern score Π as the $\log_4$ of the sum such that

$$\Pi = \log_4 (\Sigma\, 4^{m_i}) \qquad [1]$$

To simplify the use of Eq. 1, the lengths of the runs are arranged in order of decreasing size such that $m_1 \geq m_2 \geq m_3 \geq \ldots$. The longest run m_1 is excluded from the sum because (*i*) a sequence comparison is only scored when there is at least one run of identity (there are only $N - 1$ "degrees of freedom" available), and (*ii*) the one run that therefore needs to be excluded has to be m_1 because that is the only step that maximizes the residual uncertainty or entropy.

Eq. 1 can be significantly simplified to reveal some of its basic properties. By factoring out 4^{m_2}, one obtains

$$\Pi = \log_4 [4^{m_2} (1 + \Sigma\, 4^{(m_i - m_2)})],$$

where the sum now extends over $i = 3, 4, \ldots, N$. Further simplification by using a conventional Taylor series expansion yields the expression

$$\Pi = m_2 + (1.3863)^{-1} \Sigma\, 4^{(m_i - m_2)} \qquad [2]$$

It is interesting to note that the consequence of rewriting Eq. 1 in the form of Eq. 2 shows that Π is largely determined by m_2 and that additional identity runs contribute a correction factor to Π. For example, if $m_2 = 7$, $m_3 = 6$, and all other $m_i = 0$, then the size of the correction factor is 0.161. Eq. 2 also implies that a single run of seven matches is equivalent to four runs of six matches each. It should be noted that if $m_i = 0$ for $i = 2, 3, \ldots N$, then $\Pi = 0$. This situation is the trivial case since it corresponds to a single identity run whose significance can be judged by inspection. The usefulness of Π comes into play when there are a number of identity runs in a comparison.

```
                  1        10        20        30        40        50        60        70        80        90       100       111
No. of alignments : ---11:::::::::::::::::--3:::::----8:::::::----------------------4::-----3:::---------8:::::----------------6:::---------
Random sequence   : TCCGCGTCTGAAC--CTGGGGCTG--CTCTTACT--AGAGGATGCGTGTGCAATC-GCAGTAT-AGGTTTACATTA-TAATGGCTTCATTCGTAATCCACTATGAAACCCCACACTAGAA
Common nucleotides: TCC  G CTG  C  CTGGG CTG  C  TT CT  A AG AT C T T CAA C GCAG AT   GTTTA  T A T ATGGCT  AT  GTA T CA TATGAAA    ACACTAGAA
IFN-beta 1 gene   : TCCTAGCCTGTGCCTCTGGGACTGGACAATTGCTTCA-AGCATTC-T-T-CAACCAGCAG-ATGCTGTTTAAGTGACTGATGGCTA-AT--GTACTGCA-TATGAAAGG--ACACTAGAA
                    645  650       660       670       680       690       700       710       720       730       740       753
```

FIG. 1. An alignment between segments of the human IFN-β_1 gene and a computer-generated random sequence. The IFN-β_1 gene was compared with a random DNA sequence 1 kb long by using the TT algorithm at $r = 0.5$, $s = 10$, and $n = 1$. The alignment with the highest Π score (see Eq. 1) is illustrated. The observed Π score (7.254) is only 2.79 standard deviations above the mean of 5.184 (Fig. 2*A*). The "No. of alignments" line shows a dash for each univalent alignment and a field consisting of a number followed by a string of colons (e.g., 3:::::) to indicate the extent of a multivalent region and the number (in this example, 3) of equivalent metric alignments possible in that region (see ref. 18). In such regions only one of the several equivalent alignments is shown.

RESULTS AND DISCUSSION

After using the TT algorithm to find the alignment patterns, it was necessary to determine the likelihood of finding the observed alignment patterns by chance in the human genome. We estimated the likelihood of finding the observed patterns in a comparison between the IFN-β_1 gene and a random nucleotide sequence the length of the haploid human genome (approximately 3×10^9 nucleotides; ref. 29). The approach to this problem consists of the following steps: (*i*) find the most stringent settings for *r*, *s*, and *n* in the TT algorithm that allow the specific alignment to survive in a TT path graph ("threshold settings"); (*ii*) compare at threshold settings the IFN-β_1 DNA sequence with five separate kilobase-long sections of computer-generated random DNA sequence (probability of finding a particular nucleotide in any given position is 0.25 in these sequences); (*iii*) estimate the number of alignments per kilobase of random sequence that appear in the path graph and estimate the distribution of the pattern (Π) scores observed in step *ii*; and (*iv*) estimate the probability of finding an alignment of a given Π score in a random sequence the length of the human genome by using the information obtained in step *iii*. In practice, when steps *ii* and *iii* are carried out with randomly selected kilobase-long sections of human DNA sequence corresponding to "single-copy" DNA obtained from the GenBank data base, the results obtained are equivalent to those obtained with computer-generated random sequences. The statistic in step 4 underestimates the true significance of a particular alignment because one-quarter to one-third of human DNA consists of highly repetitive sequences (30).

Comparison of the IFN-β_1 Gene with Computer-Generated Random DNA Sequences. The 840-nucleotide-long IFN-β_1 DNA sequence (23) was compared with five separate 1000-nucleotide-long computer-generated random DNA sequence samples by using the TT algorithm at three different "threshold" settings that were judged useful to analyze. The key difference in these settings was that the penalty for a missing nucleotide or "null" was the equivalent of one, two, or nine mismatches. (In the latter case the algorithm does not allow any insertions or deletions to be made in the alignment). Fig. 1 illustrates one of the better alignments observed in null = 1. The Π score of this alignment is 7.254. Alignments as good as this were observed in every comparison.

Fig. 2 is a description of the total number of alignments observed and the distribution of the respective observed Π scores in comparisons between the IFN-β_1 gene and a total of 5 kb of random DNA sequence at three different "threshold" settings in the TT algorithm. For example, at null = 9 (Fig. 2*C*) 158 alignment events appeared in the TT output per 5 kb of random sequence (approximately 32 per kb), and these had a mean Π score of 3.995. The observed frequency distribution of Π reveals a good fit to a Gaussian curve. The likelihood that this distribution deviates from a Gaussian distribution is in the range of 10^{-4} to 10^{-5} by the χ^2 test. Fig. 2 represents a test of the hypothesis that the length of the second longest run (m_2) in an alignment is distributed at random (see Eq. 2). Thus, it is reasonable to use the Gaussian function as a description of the distribution of the Π scores.

Comparison of the IFN-β_1 Gene with λB3 DNA Sequence. Both strands of the λB3 DNA sequence were compared with that of the IFN-β_1 gene by using relaxed search-mode settings in the TT algorithm. Significant alignment patterns were found primarily along one strand. The particular section illustrated in Fig. 3 attracted attention because it did not contain any insertion/deletion and was part of an open reading frame that exhibited extensive amino acid homology with the amino-terminal section of mature IFN-β_1. Furthermore,

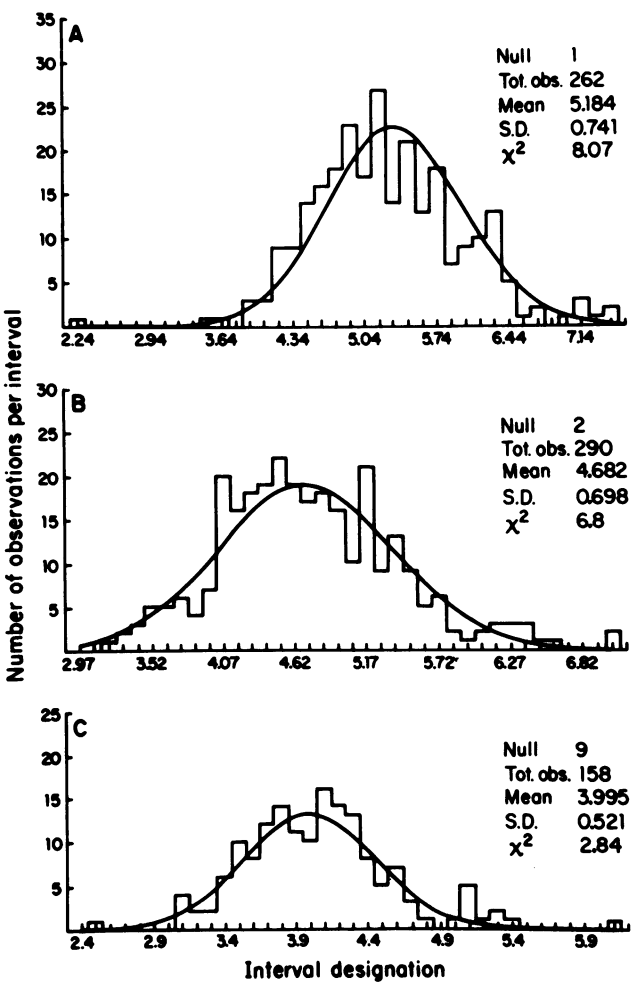

FIG. 2. Distribution of the pattern (Π) scores in comparisons between IFN-β_1 and computer-generated random sequences. The IFN-β_1 gene was compared with five separate sections of random nucleotide sequence, each 1 kb long, at different settings of the TT filter. The total number of alignment events observed, the distribution of the Π scores, and the mean and standard deviations of these scores are described. The observed distributions were plotted as frequency histograms by using the BMDP statistical package and were fit to a Gaussian function by a least-squares curve-fitting program written in Fortran. The χ^2 values obtained are also indicated. The TT filter settings used were $r = 0.5$, $s = 10$, and $n = 1$ in *A* (same settings as in Fig. 1); $r = 0.6$, $s = 10$, and $n = 2$ in *B*; and $r = 0.6$, $s = 10$, and $n = 9$ in *C* (same as in Fig. 3).

```
                :      1          5          10         15         20         25      28
lambda B3 section 7:   AlaLeuSerGlyHisGluArgSerLeuThrArgPheGlnArgAsnTyrAspIlePheLeuGlnLysLeuLeuMetGlnMetAsn
No. of alignments :    ----------------------------------------------------------------------------------
                :      1      10         20         30         40         50         60         70        84
lambda B3 section 7:   GCTCTTTCAGGGCACGAAAGGTCTCTCACCCGATTTCAAAGAAACTATGATATCTTCCTTCAGAAGCTGCTCATGCAAATGAAT
common nucleotides :   GCTCTTTC   G  C A A  T  CT        T CAAAGAA C    AT T      TCAGAAGCT CT   GCAA TGAAT
IFN beta 1 gene   :    GCTCTTTCCATGAGCTACAACTTGCTTGGATTCCTACAAAGAAGCAGCAATTTTCAGTGTCAGAAGCTCCTGTGGCAATTGAAT
                :    130       140        150        160        170        180        190        200       213
codon phase       :    <---in (+0)----------------------------------------------------------------------->
IFN beta 1 protein :   AlaLeuSerMetSerTyrAsnLeuLeuGlyPheLeuGlnArgSerSerAsnPheGlnCysGlnLysLeuLeuTrpGlnLeuAsn
                :      -3         1          5          10         15         20         25
common amino acids :   AlaLeuSer             Leu        GlnArg                GlnLysLeuLeu   Gln   Asn
```

FIG. 3. Comparison of a section of λB3 DNA with IFN-β_1. The TT filter settings used in this comparison were the same as those in Fig. 2*C*. This section corresponds to section 7 in Fig. 4. The Π score (8.174) is 8.02 standard deviations above the mean of 3.995 (Fig. 2*C*). The "codon phase" line shows that the reading frame of the upper sequence is in phase throughout with that of the lower sequence.

the λB3 DNA sequence in this section selected the human IFN-β_1 gene as the most significant alignment in computer searches of both the Dayhoff and the Los Alamos GenBank mammalian nucleotide sequence data bases. The alignment in Fig. 3 survives threshold settings in the TT algorithm of null = 9 (Fig. 2*C*). Its Π score is 8.174. This is 8.02 standard deviations above the mean Π score of the distribution described in Fig. 2*C*. Since under the settings used in Fig. 2*C* approximately 32 alignment events appear per kilobase of random sequence, we expect to find approximately 10^8 alignment events in a comparison between the IFN-β_1 gene and a random sequence the length of the human genome ($32 \times 3 \times 10^6$). This leads to the estimate that the likelihood of finding an alignment statement with a Π score 8.02 standard deviations above the mean in a random sequence the length of the human genome is approximately 10^{-7} (31). Thus, this is a statistically significant observation.

The location of the section described in Fig. 3 in λB3 DNA is illustrated in Fig. 4. Section 7 in Fig. 4 corresponds to the stretch of DNA shown in Fig. 3. An inspection of Fig. 1 (a Π score of 2.79 standard deviations above the mean in Fig. 2*A*) suggests that a Π score of 3 standard deviations above the mean may be a useful cut-off to identify interesting alignments. We have used the strategy described above to identify six additional alignments between sections of the IFN-β_1 gene and λB3 DNA that have Π scores 3 or more standard deviations above the mean. Although these scores by themselves, one at a time, are not significant on a human genome-wide basis (to meet this test the Π score should be approximately 5.5–6 standard deviations above the mean), the clustering of these sections in λB3 DNA is significant. It is apparent that there is an interesting segmental relationship between the IFN-β_1 gene and λB3 DNA sequence.

Comparison of the λB3 DNA Sequence with an *Alu*-Like Repetitive DNA Sequence. We also have used the strategy described above to compare the λB3 DNA sequence in both directions with a 333-nucleotide-long stretch of itself that was identified as an internally duplicated *Alu*-like repetitive DNA element (30). The bold arrows in Fig. 4 show that λB3 DNA is rich in these repetitive DNA elements. The Π scores of alignments in regions other than those indicated by the bold arrows were similar to the distributions described in Fig. 2. The Π scores of all of the observed alignment statements that appeared in the TT output in regions between the *Alu*-like elements were well below the 3 standard deviation cutoff.

In summary, we have determined that there is a statistically significant relationship at the nucleotide sequence level between λB3 DNA located on human chromosome 2 and the IFN-β_1 gene located on human chromosome 9. We arrived at this inference through the development of a strategy for assessing the significance of similarities between DNA sequences; this strategy is based on a scoring scheme that recognizes patterns or runs of identities within an alignment and not the actual number of matches or mismatches. The Π score is an entropy-like measure of the uncertainty in accessible "identity-space" (cf. ref. 32 and references cited therein) and numerically corresponds to the length of the second longest run in the alignment plus a correction factor. Since it can be reasonably expected that the length of the second longest run in an alignment between the IFN-β_1 gene and a random sequence is likely to be random, it is reassuring to observe experimentally that the distribution of Π scores in such comparisons follows a Gaussian distribution. It is important to note that Π is a statistic of length and not of the precise sequence within a run of identity. It then becomes a

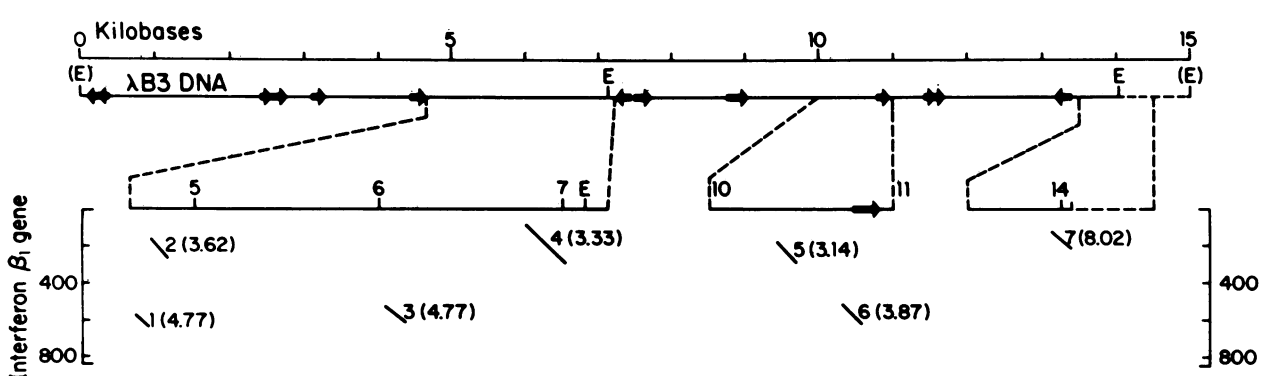

FIG. 4. Location in λB3 DNA of sections related to IFN-β_1 and to *Alu*-like repetitive DNA. The seven sections related to segments in IFN-β_1 are shown in the lower portion of the figure and are numbered 1–7. Numbers in parentheses indicate the difference between the Π scores and the appropriate mean scores in standard deviations. Section 2 is described in detail in figure 4 of ref. 24, and section 7 is described in Fig. 3 in this paper. The locations and orientations of *Alu*-like repetitive DNA elements are indicated by the bold arrows. E, *Eco*RI sites in human DNA; (E), *Eco*RI sites artificially created by synthetic linkers used in the cloning process.

simple matter to assign meaningful probability values to specific alignment patterns with specific Π scores.

Note Added in Proof. Additional nucleotide sequence data (in the region indicated by the dashed line in Fig. 4 located between the two right-hand *Eco*RI sites) allows us to extend the conserved domain identified in Fig. 3 to include approximately one-fifth of the interferon molecule. This domain now includes Cys-31 in the IFN-β_1 sequence as an amino acid conserved between it and the λB3 locus. It is interesting to note that of the two Cys residues in IFN-β_1 now included in the domain identified in Fig. 3, Cys-17 is not conserved, whereas Cys-31 is conserved between IFN-β_1 and the λB3 sequence. It is known that Cys-17 is dispensable, whereas Cys-31 (sulfhydryl bonded to Cys-141) is essential for the biological activity of the interferon molecule.

We thank Dr. Peter H. Sellers for extensive instruction in metric analysis of genetic sequences; Drs. Igor Tamm and James S. Murphy for numerous insightful discussions; and Ms. Yvonne Buhler and Joy Graham for excellent technical assistance. P.B.S. and L.T.M. are supported by Research Grant AI 16262 from the National Institutes of Allergy and Infectious Diseases, an Irma T. Hirschl Award, an Established Investigatorship from the American Heart Association, and grants from Enzo Biochem, Inc., and the National Foundation for Cancer Research. F.R.L. is an Andrew W. Mellon Foundation Fellow and is supported by Research Grants GM 31790 from the National Institutes of General Medical Sciences and PCM 8409213 from the National Science Foundation.

1. Sankoff, D. & Kruskal, J. B., eds. (1983) *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparisons* (Addison-Wesley, Reading, MA).
2. Heiter, P. M., Max, E. E., Seidman, J. G., Maizel, J. V. & Leder, P. (1980) *Cell* **22,** 197–207.
3. Korn, L. J., Queen, C. L. & Wegman, M. N. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 4401–4405.
4. Needleman, S. B. & Wunsch, C. D. (1970) *J. Mol. Biol.* **48,** 443–453.
5. Goad, W. B. & Kanehisa, M. I. (1982) *Nucleic Acids Res.* **10,** 247–263.
6. Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 726–730.
7. Fitch, W. M. & Smith, T. F. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 1382–1386.
8. Waterman, M. S., Smith, T. F. & Beyer, W. A. (1976) *Adv. Math.* **20,** 367–387.
9. Sellers, P. H. (1974) *SIAM J. Appl. Math.* Suppl. A, **16,** 253–258.
10. Sellers, P. H. (1980) *J. Algorithms* **1,** 359–373.
11. Sellers, P. H. (1984) *Bull. Math. Biol.* **46,** 501–514.
12. Wagner, R. A. & Fischer, M. J. (1974) *J. A. C. M.* **21,** 168–183.
13. Erickson, B. W. & Sellers, P. H. (1983) in *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparisons*, eds. Sankoff, D. & Kruskal, J. B. (Addison-Wesley, Reading, MA), pp. 55–91.
14. Doolittle, R. F. (1981) *Science* **214,** 149–159.
15. Barker, W. C. & Dayhoff, M. O. (1972) in *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Washington, D.C.) Vol. 5, pp. 101–110.
16. Church, W. R., Jernigan, R. L., Toole, J., Hewick, R. M., Knopf, J., Knutson, G. J., Nesheim, M. E., Mann, K. G. & Fass, D. N. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 6934–6937.
17. Reich, J. G., Drabsch, H. & Daumler, A. (1984) *Nucleic Acids Res.* **12,** 5529–5543.
18. Erickson, B. W., May, L. T. & Sehgal, P. B. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 7171–7175.
19. Kanehisa, M. I. (1982) *Nucleic Acids Res.* **10,** 183–196.
20. Sehgal, P. B., May, L. T., Sagar, A. D., LaForge, K. S. & Inouye, M. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 3632–3636.
21. Sagar, A. D., Sehgal, P. B., May, L. T., Inouye, M., Slate, D. L., Shulman, L. & Ruddle, F. H. (1984) *Science* **223,** 1312–1315.
22. Owerbach, D., Rutter, W. J., Shows, T. B., Gray, P., Goeddel, D. V. & Lawn, R. M. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 3123–3127.
23. Ohno, S. & Taniguchi, T. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 5305–5309.
24. Sehgal, P. B. & May, L. T. (1985) in *The Biology of the Interferon System 1984*, eds. Kirchner, H. & Schellekens, H. (Elsevier, Amsterdam), pp. 27–32.
25. Messing, J. & Vieira, J. (1982) *Gene* **19,** 269–276.
26. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
27. Dixon, W. J., ed. (1983) *BMDP Statistical Software* (Univ. of California Press, Berkeley, CA).
28. Bevington, P. R. (1969) *Data Reduction and Error Analysis for the Physical Sciences* (McGraw-Hill).
29. Altman, P. L. & Katz, D. D., eds. (1976) *Cell Biology*, Biological Handbooks (Fed. Am. Soc. Exp. Biol., Bethesda, MD), Vol. 1.
30. Schmid, C. W. & Jelinek, W. R. (1982) *Science* **216,** 1065–1070.
31. Abramowitz, M. & Stegan, I. A. (1964) *Handbook of Mathematical Functions*, National Bureau of Standards Applied Mathematics Series (Govt. Printing Office, Washington, D.C.), Vol. 55.
32. Papoulis, A. (1984) *Probability, Random Variables, and Stochastic Processes* (McGraw-Hill, New York), 2nd Ed.