Regulatory pattern identification in nucleic acid sequences

J.R.Sadler[1], M.S.Waterman[2] and T.F.Smith[3]

[1]Dept. Biochemistry/Biophysics/Genetics, Univ. Colorado Health Sciences Center, Denver, CO 80262, [2]Dept. Biochemistry and Biophysics, Sch. Med., Univ. California, San Francisco, CA 94121, and [3]Theoretical Biology, Los Alamos Natl. Lab., Los Alamos, NM 87545, USA

Received 1 September 1982; Revised 28 February 1983; Accepted 1 March 1983

ABSTRACT

In addition to the sequence homologies and statistical patterns identified among numerous genetic sequences, there are subtler classes of patterns for which most current computer search methods offer very limited utility. This class includes various presumptive eukaryotic regulatory sites. A critique of the often employed consensus and local homology methods suggests the need for new tools. In particular, such new methods should use the positional and structural data now becoming available on exactly what it is that is recognized in the DNA sequence by sequence-specific binding proteins.

INTRODUCTION

Several types of patterns have been identified within the current wealth of nucleic acid sequences. These include the statistical patterns such as nucleotide content and nearest neighbor frequencies which have been used to characterize large genetic domains (1). Sequence similarity measures have been the major tools employed in the identification of longer common patterns or homologies.* However, for several kinds of presumptive regulatory sequences which appear to be composed of two or more separated segments, these tools are of limited value. It is the aim of this study to examine these limitations in detail.

Well understood tools are available and have been successfully employed to identify the first two of the above pattern types. For example, the statistical characterization of the various genetic sequence domains is reviewed in a companion study (1), while Grantham et al. (2) have compared codon usage statistics over a wide taxonomic range. The literature is also replete with the identification of major regions of shared sequence homology or similarity. The success of such studies is best exemplified by our current knowledge on the evolution of the hemoglobins (3,4) and cytochrome-C's (5,6). The majority of the analytical methods used in these and similar studies are dynamic programming or matrix methods (7,8) originally introduced in 1970 by

Needleman and Wunsch (9).

Both the distinctive statistical patterns and the various sets of homologous sequences which have been identified give credence to the idea that even more general and/or more subtle patterns should be identifiable. The major functional patterns which appear to fall within such a class are the various presumptive regulatory sequences such as the three eukaryotic RNA polymerase initiation sites, the various RNA processing sites and the DNA replication origins. Previous methods for identifying such patterns have generally used a consensus sequence approach.

The problem of locating functional patterns is as difficult as it is lacking in definition. The focus of this paper is on elucidating these difficulties. The major example employed is the identification of the eukaryotic promoter sequences because, although it is an extremely difficult problem, it is well enough defined to be approached in an analytic fashion. Due to the nature of these functional patterns, they will be referred to as signals throughout this paper.

PUTATIVE SIGNAL PATTERNS

Numerous short sequence patterns have been identified within non-CDS (protein coding sequences) domains. Among these putative patterns are the poly (A) addition signal, AATAAA (10), the Goldberg-Hogness TATAAA sequence, the 5' leader CCAAT box (11) and the GT-AG intervening sequence boundaries (12). Such signals are identified with reference to a specific function (e.g., RNA or peptide initiation). Identification is obtained via a search for common nucleotides about the fixed functional site. In these cases one is often forced not to identify any common segment directly shared between sequences, but rather to identify shared similarities to a consensus sequence. The obvious problem is that such analyses require prior knowledge as to the location of the equivalent reference site. A critique of this approach (7), as applied to prokaryotic promoters, has shown that, even with the careful identification of assumed functionally equivalent reference sites, large ambiguities remain.

Signals, such as mentioned above, of less than seven or eight bases are unlikely to totally identify functional sites. For example, the most consistent signal, the poly (A) addition sequence AATAAA identified in the majority of some 83 vertebrate sequences encoding proteins, occurs 29 additional times within both the CDS and IVS (intervening sequences) domains of those same vertebrate sequences (see line 1 of Table 1).

Table 1

OLIGONUCLEOTIDE OCCURRENCES AMONG 83 VERTEBRATE SEQUENCES

Putative Regulatory Subsequences	Independent Occurrences Within IVS's (27K bp)	CDS's (35K bp)
AATAAA[1]	20	9
TATAAA[1]	14	4
TATATA[1]	19	2
ATATAA[1]	15	3
ATAAAA[1]	16	13
CCAATC[2]	2	4
AGGT(A/G)AGT[3]	0	0
$(T/C)_2X(C/T)AGG$[3]	3	2
Most Common Sixmers[6]		
TTTTTT	59 $\overline{48}$	.
GTGTGT	54 $\overline{4}$[4]	.
AAAAAA	48 $\overline{59}$	.
CTCCTG	.	76 $\overline{18}$[4]
CTGCTG	.	75 $\overline{35}$
CAAGAA	.	56 $\overline{7}$[4]
Minimum Length Least Common[5]		
ATAGG	.	0
ATCG	5	.

1) Alternate Goldberg-Hogness box sequences.
2) The promoter CAAT box.
3) CDS/IVS boundary consensus sequences (S. Mount (13)).
4) Sequences which contribute to strand asymmetry; overlining indicates complement occurrences.
5) Note, CGCG is not the least common minimum length oligonucleotide as might be expected from the nearest neighbor frequencies.
6) The mean expectation values for sixmers have a range from 7-16 depending on composition.

Therefore, additional information must be required for the cell hardware to correctly identify the poly-(A) addition point. For a more variable signal, such as the Goldberg-Hogness sequence presumed to locate RNA polymerase II

initiation, the occurrences in Table 1 (lines 1-5) of acceptable alternative sequences clearly show that this short sequence is also insufficient to uniquely identify this functional signal.

Longer consensus sequences, as expected, eliminate such false positives. For example, Mount (13) has given AG/GT($\frac{A}{G}$)AGT as the 5' CDS/IVS boundary junction consensus, which, in our limited data set of 83 vertebrate sequences, is found seven times and in all cases correctly identifies a splice junction. However, well over 100 other 5' CDS/IVS junctions went unidentified. Birnstiel (14) has likewise identified a longer signal, the histone 3'mRNA termination subsequence GGCYTTGTCAGRGCCA, which is so far unique for these sequences within Drosophila. Thus, while sufficiently long consensus sequences independent of an additional reference points probably can eliminate false positives, they are as likely to eliminate the identification of some true positives.

A limited degree of similarity among any set of nearly equivalent regulatory signals is to be expected. For example, while all RNA polymerase II sites are recognized by the same basic molecular complex, the requirements for differing rates or cell cycle times of transcription could impose considerable dissimilarities among these sites. In the case of RNA splicing regulation, a combination of common features and variation is probably involved (16), allowing for divergent as well as co-ordinate regulation of gene subsets). Variations such as these will confuse attempts at identifying any consensus without the analysis of data beyond that currently available.

SHORT SIGNALS IDENTIFIED BY SIMILARITY

Motivated initially by a search for potential regulation signals analogous to restriction endonuclease sites[+], a search for common subsequences among related (same gene) IVS's was carried out. This search for subsequences of local maximum similarity among the 83 vertebrate sequences used the modified Needleman-Wunsch algorithm (17) with the weights: +1.0, -1.75 and -(1.0 + 1.75 × length) for matches, mismatches and deletion-insertions respectively. This choice of high mismatch weight restricted consideration to contiguous subsequences of 62 percent base matching or higher. Table 2 contains a few examples of these subsequences.

Among the IVS comparisons made, similarity values greater than 9.0 were rare but did appear now and then between assumed unrelated sequences. Even among a set of twenty-four Monte Carlo generated sequences of "hemoglobin-

Table 2

SUBSEQUENCES OF MAXIMUM LOCAL SIMILARITY AMONG "GENE RELATED" IVS'S

Position in IVS	Alignment	Similarity Value
Chick ovalbumin IVS-3 (581 bp) versus IVS-4 (400 bp)		
561	TTTCTTTCTCT-TTGTATT	9.75
99	TTTCTTTTTCTCTTGTTTT	
383	TTACAGAAGAAAAACAGCACAAA	9.25
118	TTACAAATGAAAGAGAGGACAAA	
514	TACACTACGTAAA	9.25
238	TACACTAC-TAAA	
465	TGAAAAATAGTTTTGTAAAC	9.00
282	TGTAAAATAGCTTTTTACAC	
551	TAACAACATCTTTCTTTCTCTT	8.25
91	TAAGAAAATTTCTTTTTCTCTT	
Mouse Kappa I chain IVS-3 (284 bp) versus IVS-4 (299 bp)		
203	TTTTTCTCTGAACTTAGCCTATCTAACTGGA	14.5
214	TTTTCCTCTGAATTTGGCCCATCTAGTTGGA	
238	CAGGCAGGTTTTTGTAAAGGGGGGC	19.50
254	CAGGCAGGTTTTTGTAGAGAGGGGC	
Human δ-hemoglobin IVS-2 (890 bp) versus ε-hemoglobin IVS-1 (122 bp)		
441	ATGGAAAGGAA-AGTGAATATT	10.0
21	ATGGGAATGAAGGGTGAATATT	
Human γ-hemoglobin IVS-1 (122 bp) versus IVS-2 (866 bp)		
69	TCTCAGGATTTGTGG	9.50
736	TCTCAGGCTTTGAGG	
Human β-hemoglobin IVS-1 (128 bp) versus ε-hemoglobin IVS-1 (122 bp)		
69	GTTTCTGATA	7.75
106	GTATCTGATA	

like" base compositions, several subsequences with local maximum similarity greater than 9.0 were found, but only one with a value of ten, and none greater. Thus subsequence similarities such as those in Table 2 between the two IVS's of a single globin gene are of questionable significance. This same near-random similarity was observed (4) in overall similarity comparisons between the IVS-1's of ε-globin and β-globin in humans. This was somewhat

surprising since there is no question as to the homology of the neighboring CDS domains in the ε- and β-globin genes, or as to the similarities of the IVS-1's of other human globin genes. Thus whatever there may be about these sequences which is functionally equivalent its basis in sequence is highly degenerate. Any of the observed similarities between pairs of subsequences in the chick ovalbumin IVS's also might have occurred by chance; however, it is very improbable for all five subsequence pairs to have occurred by chance within this single gene's IVS's. There are some single statistically significant similarities that occur among these data. See for example, mouse Kappa IVS-3 and IVS-4 in Table 2. The two pairs of similar subsequences occur in the same order in the two IVS segments, suggesting true homology (of common origin) as well as function.

There are other interesting short subsequences of regulatory potential. For example, there is a strong consensus ninemer, CCAGCCTGG found throughout the Alu/B1/4.5s middle repetitive family. Whether this is related to the suggested role of the Alu sequence in replication (18) is unknown. As if to confuse the issue, this short segment is less variable (conserved?) than the rest of the sequences within this family which shows a complex set of overall sequence similarities or homologies (19).

A few other short sequences appear to be quite specific: the sequence TGGAAATAAAAC is observed only three times in the 83 vertebrate sequences examined, all in the mouse kappa light chain sequence and then always just 5' of the CDS/IVS splicing sequences. The sequences AGTAATA and TGTATTC appear rather specific to chick non-CDS domains, occurring there six times, while found only four other times throughout the 83 vertebrate sequences, compared to an average occurence of other A-T rich sixmers at 16 to 20.

A major difficulty in recognizing functional signals is due to the context dependent nature of the signal. A pattern which is a signal common to some set of IVS's in a particular taxon might occur in other taxons but not serve as a signal there. For example, the three terminator triplets are not absent in CDS regions but only from a given reading frame.

SEGMENTED SIGNALS

The "split" form of the eukaryotic promoter(s)--one for each eukaryotic RNA polymerase--represents an additional pattern form requiring recognition. These patterns contain more than one subsequence showing varying degrees of consensus and varying amounts of separation, from one gene to another (20). The intercalated subsequences generally have little compositional similarity.

Eukaryotic Promoter II

Mammalian Consensus

CCAAT ...(X)$^{41-79}$...[T]ATA[A/T]A...(X)$^{24-25}$...ACXXT.....
.....AXXCTT....
+1

Histone H3/H2A/H2B Consensus

[C/A]CAAT...(X)$^{39-60}$.....TATAAA...(X)$^{22-29}$.....[C]ATTC.....
+1

Eukaryotic Promoter III

tRNA Consensus

?(TATA)?...RXXT.....TGGCXXRRXGG...(X)$^{29-34}$.....GTTCRAXXC.....
+1 +8

Prokaryotic Promoter

E.coli Lac Z gene

-38 -12 +1
...CTTTTACACA.....TATAATGT.....A.....
(initial polymerase binding) (polymerase melt-in) (mRNA start)

Fig. 1: Presumptive RNApolymerase consensus patterns. The "X"s represent any of the four bases while "R"s represent purines, and bases in square brackets represent common alternatives.

The promoter II consensus specified in Fig. 1 even with a total of 12 determined bases will still identify a false positives unless restricted á priori to a known 5' mRNA location. The most interesting case is found in the rat pancreatic amylase gene, about 185 codons into the gene. This presumptive false promoter, CCAAT(X)^{45}ATAAA(X)^{25}ACTAT, is even followed 54 bases later by an in-phase ATG.

Other promoter-like sequences are found, for example, 55 bases three prime of the known poly (A) attachment site for the human fibroblast interferon message except that the initial segment is CCAAGT rather than CCAAT. There is even an interesting open reading frame 14 bases three prime of the possible ACCTT cap site, having the form: ATG(X)^{3}TGA(X)^{6}GTG(X)^{207}TAA. Thus, if this promoter is accessible to RNA polymerase II, a potential 70 amino acid peptide is possible if initiation can begin at the GTG. It is worth noting that no poly (A) attachment signal is observed in the next 275 bases and that the statistical properties of this open reading frame are not typical of coding domains (1). Finally in the Moloney murine Leukemia

virus terminal repeat there is the promoter-like sequence $CAAAT(X)^{69}AATAAAA(X)^{24}ACTCT$ in which, the 24 base sequence between the second and third segments contains a contiguous (A,T) run of eleven bases.

The TATA box or Goldberg-Hogness signal appears to be the most obvious common feature, including its great similarity to the Pribnow box in the prokaryotic promoter. However, only the dinucleotide TA is shared by all eukaryotic structural gene sequences found 30 or so nucleotides up stream of the mRNA initiation site, as was noted in earlier work on the human hemoglobins (4). To date it is only the chicken lysozyme which does not contain at least an ATA in the proper position. It is of interest to note that both a human and two mouse hemoglobin pseudo-genes contain a G in positions four and five respectively in their ATAAA sequences.

The mRNA cap site emphasizes the heterogeneous nature of these promoters. In general the cap site begins with an A and is followed by a C and one or more T's within the next five bases, although as the histone sequence shows, not necessarily in that order. In all non-histone eukaryotic structural genes examined there are exactly 30 or 31 bases between the first or second T in the [T]ATAAA region and the first in the cap site. In the histone genes this spacing is more variable, as is the spacing between the two identified subsequences in the promoter for polymerase III.

Finally, there are other patterns often segmented which are expected within genetic sequences arising from rather simple structural-symmetry considerations. These include the short inverted repeats. They are generated (required) by such structural constraints as those imposed by RNA secondary structure or the common identical-subunit structure of DNA binding proteins. The search for potential RNA secondary structure rests on the idea of measuring a distance between a sequence and itself in units of free energy, where one recognizes that regions of homology between different genetic sequences are analogous to RNA paired-helical regions, non-homologous regions to internal RNA secondary structure loops, and deletions or insertions to unpaired bulges (25). While such algorithms exist for their identification, the same problems discussed above no doubt apply, i.e., the occurence of such symmetries alone is probably insufficient to imply any given significance or function, without reference to other information.

CONCLUSIONS

The various consensus and homology sequences that have been identified give credence to the idea that even more general pattern recognition tools

should provide new insight. Statistical methods like those reviewed in the companion paper to this study (1) are clearly able to identify the rather gross genome domains, such as the vertebrate CDS's, IVS's, mitochondrial sequences, and the "AT rich spacers" as noted by Moreau (20). However, patterns which are more difficult to define precisely, such as promoters, require new and more subtle methods.

There appear to be two basic problems in identifying, for example, promoter elements. One is the variablility in the sequences of known promoters. The second is the existence of many "false" promoter sequences. The first problem has two aspects: one is the need for differing promoter strengths which presumably accounts for some of the variability. The second aspect relates to our ignorance concerning what is recognized or recognizable in DNA by RNA polymerase II. For instance an A/T or a G/C base pair may provide equivalent contacts for a protein, if the functional group touched is the N-7 atom of the purine in the major groove or the pyrimidine keto in the minor. A C/G may be equivalent to a G/C if the functional group is the 2-amino group of Guanine, whose position remains constant in B-DNA (see Fig. 7 of Ref. 22). In Table III we have summarized the chemical site equivalences of the different base pairs. The identification of such sites as likely points of protein contact has been recently carried out (26-29) on the CI and cro repressor binding sites in λ phage. A few known Lac operator O^c (30) and λ cro (29) binding site mutations, along with some known restriction enzyme recognition site degeneracies which display these equivalences are also given in Table III.

To our knowledge such functional chemical group equivalences have not been explored relative to promoter sites. Coupled with the compositional determination of local twist angles (23), DNA conformation (23,24) and local thermal stability such equivalences may provide the missing recognition. The second problem of presumably false positives could reflect the non-accessibility of many such sites within the chromatin due, for example, to local nucleosome phasing. The total solution will require much more data as to what is potentially "recognizable" within nucleic acid sequences (26) and at least one clear test example. The current work on histone binding in eukaryotes, and repressor binding prokaryotes (27-29) is providing information on the first of these requirements. And just possibly, the eukaryotic promoter(s) will provide the second requirement, a test example for more general pattern recognition.

As we learn more about the regulatory nature of the various genetic

	A	T	G	C
A	Major Groove N_7-amino environment unique Minor Groove N_3-keto environment unique	Major Groove C_6 amino position only slightly shifted (but keto orientation reversed) Minor Groove No Amino group in groove	Major Groove N_7 position equivalence Minor Groove Cytosine/Thymine keto equivalence	Major Groove Amino-keto position only slightly shifted (same orientation) Minor Groove No equivalence
T	ECOR II CC(A/T)GG Ava II GG(A/T)CC	Major Groove Methyl C equivalent to T otherwise unique	SAME AS A <-> C	SAME AS A <-> G
G	Hind II GTYRAC Lac weak O^C G_9 → A Lac weak O^C T_{14} → C	Acc I GT(A/C T/G)AC Lac weak O^C G_9 → T λ cro alternative binding site G_{13} → T	Major Groove N_7-keto environment unique Minor Groove N_3-amino environment unique	Major Groove No equivalence Minor Groove Guanin C_2 amine position constant
C	Acc I GT(A/C T/G)AC	Hind II GTYRAC	Nci I CC(C/G)GG	Major Groove C_4 amino unique

Table III. DNA recognizable site degeneracies. Upper right half contains potential site equivalences, diagonal contains unique sites and the lower left half contains restriction enzyme sites and weak operator mutants displaying such equivalences. [Helene and Lancelot (1982) Prog. Biophys. Molec. Biol. 39, 1-68.]

functional domains, and more about what it is that is recognized within those domains by the cellular hardware, our comparative sequence and pattern identification algorithms will have to be generalized to incorporate these data.

* Homology is employed here in the strict taxonomic sense to mean similarity arising from common ancestry only.

+ For example, internal IVS endonuclease sites might act as a negative regulatory control complementing the presumptive positive control for the sRNA in mRNA splicing (15).

REFERENCES

1. Smith, T. F., Waterman, M. S., and Sadler, J. R. (1982) Nuc. Acids Res. (this vol.).
2. Grantham, R., et al. (1981) Nuc. Acids Res. 9, r. 43.
3. Proudfoot, N. J., et al. (1980) Science 209, 1329.
4. Efstratizdis, A., et al. (1980) Cell 21, 653.
5. Dickerson, R. E. (1980) Nature.
6. Dayhoff, M. O. (1978) Atlas of Protein Sequence and Structure, Vol. 5, NBR Silver Spring, MD.
7. Smith, T. F., Waterman, M. S. and Fitch, W. M. (1981) J. Mol. Evol. 18, 38.
8. Waterman, M. S., Smith, T. F. and Beyer, W. A. (1976) Adv. Math. 20, 367.
9. Needleman, S. B. and Wunsch, C. D. (1970) J. Mol. Biol. 48, 443.
10. Proudfoot, N. J. and Brownlee, G. G. (1976) Nature 263, 211.
11. Benoist, C. et al. (1980) Nuc. Acids, Res. 8, 127.
12. Breathneck, R. et al. (1978) PNAS (USA) 75, 4853.
13. Mount, S. (1982) PNAS (USA), in press.
14. Hentschel, C. C. and Birnstiel, M. L. (1981) Cell 25, 301.
15. Lerner, M. R., J. A. Boyle, S. L. Wulin, J. A. Steitz (1980) Nature 283, 220.
16. Darnell, J. E. (1982) Nature 297, 365.
17. Smith, T. F. and M. Waterman (1981) J. Mol. Biol. 141, 195.
18. Taylor, J. H. and Watanabe, S. (1981) ICN-UCLA.
19. Schmid, C. W. and Jelinek, W. R. (1982) Science 216, 1065.
20. Moreau, J. et al. (1982) Nature 295, 260.
21. McKnight, S. L. and Kingsbury, R. (1982) Science 217, 316.
22. Dickerson, R. E. and Drew, H. R. (1981) J. Mol. Biol. 149, 761.
23. Dickerson, R. E. et al. (1982) Science 216, 475.
24. Patel, D. J., Pardi, A. and Itakura, K. (1982) Science 216, 581.
25. Waterman, M. S. and Smith, T. F. (1978) Math. Biosciences.
26. Sauer, R. T. et al. (1982) Nature 298, 447.
27. Pabo, C. O. and Lewis, M. (1982) Nature 298, 443.
28. Pabo, C. O. et al. (1982) Nature 298, 441.
29. Ohlendorf, D. H. et al. (1982) Nature 298, 718.
30. Smith, T. F. and Sadler, J. R. (1971) J. Mol. Bio. 59, 273.