Search algorithm for pattern match analysis of nucleic acid sequences

Robert Harr[1], Mikael Häggström[2] and Petter Gustafsson[1]

[1]Department of Microbiology and [2]Department of Computer Science, University of Umeå, S-901 87 Umeå, Sweden

Received 10 November 1982; Revised and Accepted 14 February 1983

ABSTRACT

A new type of search algorithm to find biological information inherited in nucleic acid sequences was developed. The algorithm is of pattern match type and is based on the fact that genetic information often is a function of a predictable statistical occurence of the four bases within parts of the sequence. The search algorithm compares the known statistical pattern of bases in e.g. a promoter, with an unknown sequence and calculates the statistical significande of the match at all positions in the unknown sequence. The program was tested on 54 published prokaryotic promoters. 44 or 49 could be found with 1 or 4 false answers, respectively. The program was also used on plasmid pBR322. All promoters functioning in an in vitro transcription system were found (tet, anti-tet, p4, bla and ori) except the so called p5 promoter. A search for donor and acceptor sites was performed in a human HLA genomic sequence that contains six introns. Five of the possible six donor and acceptor sites were found.

INTRODUCTION

Computer programs for DNA sequences have become a necessary tool in molecular biology during the last years. A number of papers describing computer programs have been published (for a review, see ref.1).

Two principally different search algorithms to find regions of homology within or between DNA sequences exists. The first algorithm compares two sequences base per base and the operator defines minimum length of homology and minimum number of hits within the homology. A simple type of this algorithm is used by Staden (2,3) while Korn **et al** (4) have presented a more sophisticated type that allows looping out of nucleotides. This type of search algorithm has also been combined with a graphic presentation of the result (5). The second type of search algorithm, the dot matrix method, has been presented by Maizel **et al** (6). The program places a dot in a 2 dimensional graph at each position between the two sequences where a homology exists of the length that was defined by the operator.

Both types of search algorithm have the disadvantage that they can only

search a DNA sequence on the basis of hit or non-hit. It has become clear during the last years that the degree of allowed variation of the four bases, A,T,C and G within structures like promoters, ribosom binding sites or splice points, differs from position to position in the structure (7,8,9). A progressively growing number of such structures have been sequenced and the significance of the statistical variation of their base pattern is increasing. Recently, an article was published (10) that shows that the "perceptron" algorithm (11) can be used to generate a weighting function which afterwards can be applied to separate two classes of sequences. In the article it was used to create a base pattern of the ribosome binding region in E. coli and to deduce if ribosome binding sites could be found in unknown sequences.

However, it is also important to simplify the search for known patterns in an unknown sequence. To make it possible to search for structures involving a statistically significant pattern e.q. a promoter, we developed a novel type of pattern match algorithm for computers. Using the developed program we could find 49 of the 54 promoters published by Siebenlist et al (7) with only 4 false answers. The program was also able to find all but one promoter in the plasmid pBR322 DNA sequence (12), both weak and strong promoters in the E. coli L10 and L11 operons and to find 5 out of 6 splice points located within the DNA sequence of a human HLA antigen (13).

Description of the program

The program is written in PASCAL and consists of around 700 statements. The data analysis was done on a DEC VAX 11/750 (VMS) connected to a DEC VT-100 video terminal. A copy of the program is available on request.

The program consists of two parts:

i) A statistical data table for the pattern to be found

ii) A calculation of the statistical significance of each match

When the program is initiated, the operator is asked to define sequence, type of search, data tables to be used and region within the sequence to be searched (Fig. 1). A number of data tables have been included in the program but the operator can also design his own. As the last step, the operator sets the level of significance and the program searches the sequence for significant patterns.

We have found that a significance level of 0.002 when using equation (1) and 0.0005 when using equation (2) is appropriate to use when a search for prokaryotic promoters is performed.

The time consumption for the program is marginal. Even applications on small computers are possible.

```
RUN PR1

TYPE NAME OF SEQUENCE FILE
E.G.     SV40.DAT   <CR>

PBR322N

DEFINE TYPE OF SEARCH

STANDARD PROMOTERS  6+6 BASES          TYPE  1<CR>
STANDARD SPLICE POINTS                 TYPE  2<CR>
ARBITRARY PATTERNS                     TYPE  3<CR>
1

 DEFINE STATISTICS

 ALL         6+6          YOU TYPE  1<CR>
 BK,ALFA     6+6          YOU TYPE  2<CR>
 BK,LAMBDA   6+6          YOU TYPE  3<CR>
 RIBOSOM     6+6          YOU TYPE  4<CR>
 ANABOL      6+6          YOU TYPE  5<CR>
 ST          6+6          YOU TYPE  6<CR>
 YOUR OWN MTRL NR         YOU TYPE  7<CR>
 SEARCH                   YOU TYPE  8<CR>
 YOUR OWN MTRL FILE       YOU TYPE  %<CR>
 MTRL ASSIGNED STAT       YOU TYPE  A<CR>
 HELP                     YOU TYPE  ?<CR>
 SAME                     YOU TYPE  S<CR>

1

   FIRST,LAST SEQ  NO:
   INTEGERS WITH SPACE IN BETWEEN
1 5000

GIVE LIMIT FOR SEQ. SEARCH
0.002

       T T G A C A                T A T A A T
       ===========                ===========
    10 T T G A C A   15        33 T T T A A T   38   a = 0.102273   # = 0.91
  4143 T T G A A T 4148      4166 T T C A A T 4171   a = 0.006085   # = 0.80
  4197 T T G T C T 4202      4220 T T G A A T 4225   a = 0.002466   # = 0.79
GIVE LIMIT FOR SEQ. SEARCH
```

Figure 1. Example of user communication of the program. A number of predefined tables exists. Statistical table ALL will cause the program to use the table presented in Fig. 2. Table BK,LAMBDA was calculated from lambda promoters, RIBOSOM was calculated from ribosomal operon promoters and ANABOL was calculated from promoters for biosynthetic operons (7). In the experiment shown, all promoters with a stastical value above 0.002 are listed. The calculation was done according to equation 2. The statistical value calculated from the equations are presented in the tables to the right of the found promoter sequences, were a is the value calculated from equation 2 and x is the value from equation 3.

i) A statistical data table: For each biologically significant DNA sequence that involves a consensus sequence e.g. a promoter of transcription it is possible to create a data table showing the occurrence of each of the four bases at each position. The significance of such a table increases of course when the number of sequenced structures increases. In Figs. 2 and 3

		T	T	G	A	C	A		T	A	T	A	A	T
	#													
T	#	85	87	13	17	9	31		89	9	50	17	7	100
A	#	6	11	0	61	17	52		0	89	24	65	65	0
G	#	4	0	81	2	7	11		7	2	7	15	7	0
C	#	6	2	6	20	67	6		4	0	19	4	20	0

Figure 2. The statistical values used in the search for a standard promoter. Data were taken from Siebenlist et al (7). The -35 region is shown to the left and the -10 region to the right. The distance between the regions was set to 15 to 19 base pairs. The figures used in the tables are the occurence of each base at every position in the promoter region. The consensus sequence is given above the table.

such data is shown. Fig. 2 is calculated from the 54 promoters published by Siebenlist et al (7). To increase the significance new tables can be constructed were promoters with similar function have been extracted and put together. In Fig. 3 the statistical variation of both the donor and acceptor sites in eukaryotic mRNA according to Breathnach and Chambon (9) is shown. We have found it convinient to use a data table of 9 bases for a donor site search and a combination of 5 and 12 bases for a search for possible acceptor sites. Also a table around the splice junction in splice RNA (or cDNA) can be constructed. The program allows the construction of any type of data table

DONOR 9 BASES TABLE

		C	A	A	G	G	T	A	A	G	T
	#										
T	#	4	13	10	0	100	7	10	3	62	
A	#	42	59	10	0	0	56	68	10	16	
G	#	12	13	76	100	0	33	13	86	10	
C	#	41	14	4	0	0	4	9	1	12	

ACCEPTOR 5 BASES TABLE

		C	A	G	G	T G
	#					
T	#	21	0	0	11	37
A	#	4	100	0	22	13
G	#	1	0	100	47	32
C	#	74	0	0	20	18

ACCEPTOR 12 BASES TABLE

		T	T	T	C	T	T	X	C	A	G	G	T G
	#												
T	#	44	52	46	43	50	55	25	21	0	0	11	37
A	#	14	7	15	8	5	7	24	4	100	0	22	13
G	#	6	9	11	6	8	6	22	1	0	100	47	32
C	#	36	32	28	43	37	32	29	74	0	0	20	18

Figure 3. Tables used in the search for donor and acceptor sites in eukaryotic genes. The data were taken from Breathnach and Chambon (9). The consensus sequence is given above the tables.

according to the operators wish. It also allows the combination of any number of tables with any distance in between the tables.

ii) Calculation of the statistical significance of each match: When the search is performed the consensus sequence is moved along the test sequence, one position at a time. At each position, the test sequence is compared with the consensus sequence and the statistical value (see Figs. 2 and 3) for each of the bases at each position in the test sequence is taken from the data table. The calculation of statistical value of the whole match can now be performed according to one of the following equations.

$$\frac{n_1 \times n_2 \times \ldots \times n_{m-1} \times n_m}{a_1 \times a_2 \times \ldots \times a_{m-1} \times a_m} \qquad (1)$$

$$\frac{n_1 \times n_2 \times \ldots \times n_{m-1} \times n_m}{a_1 \times a_2 \times \ldots \times a_{m-1} \times a_m} \times f(d) \qquad (2)$$

$$\left[\frac{n_1}{a_1} + \frac{n_2}{a_2} + \ldots + \frac{n_{m-1}}{a_{m-1}} + \frac{n_m}{a_m}\right] \times \frac{1}{m} \qquad (3)$$

were n_i is the score for base at position i taken from the data table, a_i is the score for the most frequent base at position i, m is the number of bases in the match sequence and f(d) is a factor that is used in procaryotic promoter search where the distance between the -35 and -10 region also is of importance for the evaluation of the match.

All three equations give a value of 1 for a perfect match.

Equations 1 and 2 calculate the significance in a proper statistical way. However, as is the case with the table in Fig. 2, position 12, where so far only T has been found, the calculation according to (1) and (2) gives a value of 0 for base A, G, and C in position 12 thereby biasing the result. For such cases equation 3 is advantageous to use.

Equation (2) has been used to weigh the importance of the distance between the -10 and -35 boxes in a prokaryotic promoter according to Mandecki and Reznikoff (14) and Stefano and Gralla (15). For the different distances between the boxes we have accordingly used the following values for f(d): f(17) = 1 ; f(16) and f(18) = 0.15 ; f(15) and f(19) = 0.02.

Equation (2) was used to search random sequences, each consisting of 10.000 bases, for promoter like patterns. The number of patterns found in 10.000 bases with a significance value above 0.002 never exceeded 1. 3 to 5 patterns with a value above 0.0005 were regularly found depending on which random sequence was used.

RESULTS

a) Search for "standard" promoters

The search algorithm was tested on the 54 published promoters taken from Siebenlist et al (7) that was used to construct the data tables. All promoters were entered as one long sequence (3240 bases), where each promoter comprised 60 bases as can be seen in Fig. 4. Using equation 1 and a significance level of 0.002 the program finds 43 promoters with only one "false" answer (column 1). However, one anomality is found. Interestingly, the program does not find the lacP115 promoter but a more significant pattern 12 bases upstreams which is the lac-wild type promoter. Using equation 2 which also weighs the importance of the distance between the -35 and -10 boxes, according to Mendecki and Reznikoff (14) and Stefano and Gralla (15), and a significance level of 0.0002, 4 more promoters are found with 2 "false" answers (column 2). Still the lac-wild type promoter is pointed out as a more significant pattern than the lacP115 promoter mutant (column 2). The second "false" answer is the galP2 promoter which correctly is found as a more significant pattern 12 bases from the weaker galP1 promoter listed in the table of Siebenlist et al (7). The result shows that the discriminatory effect of the search increases when the distance between the two boxes is taken into account.

To investigate what happens if the significance level is decreased drastically we performed a search using equation 1 and a significan-ce level of 0.0002. The result can be seen in Fig 5 were the new promoters found are shown. Now the program finds 50 of the 54 but as can be seen the noise level has increased drastically and 8 more "false" promoters are opened. The "false" promoters (underlined sequences in Fig. 5) that now are detected very often overlap the promoter found at a higher level of significance. In the cases of lamCIN, T7 C, T5 26 and G4 B the -10 region is relocated 1 or 2 bases from the first one found. A minor number of promoters are new structures located at a different position than the first one found. It was difficult to find promoters araBAD, bioA, lacIq and araC. These promoters are hidden in the background noise. Some of those four promoters are only active in the presence of actabolite activator protein.

The following promoters gave a high value of significance (0.1 or higher) using eq. (1): T7 A2, 0.1125;PhiX D,0.1009;G4 D,0.2656;lacUV5, 0.1625;tet,0,1023;str,0.48;rrnDXE2,0.722;rrnA1,0.2833;rrnA2,0.1615. These 9 promoters can be considered as "strong" promoters in the sense that they show a high resemblence to the consensus sequence. Two promoters gave a value

		TTGACA	<--15-19-->	TATAAT	--> START	STATISTICAL SIGNIFICANCE LEVEL (Eq. 1) ∂ = 0.002	(Eq. 2) ∂ = 0.0002
T7 A3	GTGAAACAAAACGG	TTGACA	ACATGAAGTAAACACGG	TACGAT	GTACCAC A TGAAACGAC	+	+
T7 A1	TATCAAAAAGAGTA	TTGACT	TAAAGTCTAACCTATAG	GATACT	TACAGCC A TCGAGAGGG	+	+
lam PR	TAACACCGTGCGTG	TTGACT	ATTTTACCTCTGGCGGT	GATAAT	GGTTGC A TGTACTAAGG	+	+
lam PRM	AACACGCACGGTGT	TAGATA	TTTATCCCTTGCGGTGA	TAGATT	TAACGT A TGAGCACAAA		+
lam PO	TACCTCTGCCGAAG	TTGAGT	ATTTTTGCTGTATTTGT	CATAAT	GACTCCT G TTGATAGAT	+	+
lam PL	TATCTCTGGCGGTG	TTGACA	TAAATACCACTGGCGGT	GATACT	GAGCAC A TCAGCAGGAC	+	+
lam C17	GGTGTATGCATTTA	TTGCA	TACATTCAATCAATTGT	TATAAT	TGTTAT C TAAGGAAATA	+	+
lam CIN	TAGATAACAATTGA	TTGAAT	GTATGCAAATAAATGCA	TACACT	ATAGGT G TGGTTTAATT	+	+
lam PR	TTAACGGCATGATA	TTGACT	TATTGAATAAAATTGGG	TAAAT	TGACTCA A CGATGGGTT	++	++
lam 434PR	AAGAAAAACTGTAT	TTGACA	AACAAGATACATTGTAT	GAAAAT	ACAAGAA A GTTTGTTGA	+	+
T7 C	CATTGATAAGCAAC	TTGACG	CAATGTTAATGGGCTGA	TAGTCT	TATCTT A CAGGTCATCT	+	+
T5 26	CTTAAAAATTTCAG	TTGCTT	AATCCTACAATTCTTGA	TATAAT	ATTCTC A TAGTTTGAAA	++	+
T5 25	CATAAAAAATTTAT	TTGCTT	TCAGGAAAATTTTTCTG	TATAAT	AGATTC A TAAATTTGAG	+	+
PHiX A	AATAACCGTCAGGA	TTGACA	CCCTCCCAATTGTATGT	TTTCAT	GCCTCC A AATCTTGGAG	+	+
fd X	TCTTAATCTTTTTG	ATGCAA	TTCGCTTTGCTTCTGAC	TATAAT	AGACAG G GTAAAGACCT	+	+
fd II	ACAAAACATTAACG	TTTACA	ATTTAAATATTTGCTTA	TACAAT	CATCCT G TTTTTGGGGC	+	+
fd III	TTAAGAAATTCACC	TCGAAA	GCAAGCTGATAAACCGA	TACAAT	TAAAGGC T CCTTTTGGA	+	+
fd V	TATTAACGTAGATT	TTTCCT	CCCAACGTCCTGACTGG	TATAAT	GAGCCA G TTCTTAAAAT	++	++
SV 40	GAATGCAATTGTTG	TTGTTA	ACTTGTTTATTGCAGCT	TATAAT	GGTTACA A ATAAAGCAA	+	+
T7 A2	ACGAAAAACAGGTA	TTGACA	ACATGAAGTAACATGCAG	TAAGAT	ACAAATC G CTAGGTAA	+	+
PhiX D	TAGAGATTCTCTTG	TTGACA	TTTTAAAAGAGCGTGGAT	TACTAT	CTGAGTC C GATGCTGT	+	+
PhiX B	GCCAGTTAAATAGC	TTGCAA	AATACGTGGCCTTATGGT	TACAGT	ATGCCC A TCGCAGTTC	+	+
G4 A	GCTCCAAATAATGC	TTGACT	AATACTCAATCACCACTC	TAATAT	GCCTCCC A TCAAACGG	+	+
G4 B	GGCAAATAAATAGC	TTGCAA	AACACGTGGCCTTATGGT	TACTCT	ATGCCC A TCGCAGTCC	+	+
G4 D	TAAACAATCAATGC	TTGACA	TACTGAAAGAACGTGGCC	TATTAT	CCACATC G TCAACTGA	+	+
fd II'	TTTGAATCTTTGCC	TACTCA	TTACTCCGGCATTGCATT	TAAAAT	ATAT G AGGGTTCTAAA		
fd IV	TGATAAATTCACTA	TTGACT	CTTCTCAGCGTCTTAATC	TAAGCT	ATCGCT A TGTTTTCAA	+	+
M13-RNA	CTAAAGACTTTTTC	ATGAGG	AAGTTTCCATTAAACGGG	TAAAAT	ACGTAAT G CCACTACG		
fd VIII	GATACAAATCTCCG	TTGTAC	TTTGTTTCGCGCTTGG	TATAAT	CGCTGGG G GTCAAAGATG	+	+
lac UV5	TAGGCACCCCAGGC	TTTACA	CTTTATGCTTCCGGCTCG	TATAAT	GTGTGG A ATTGTGAGC	+	++
gal P2	CACTAATTTATTCC	ATGTCA	CACTTTTCGCATCTTTGT	TATGCT	ATGGTT A TTTCATACC		+
ara BAD	TTAGCGGATCCTAC	CTGACG	CTTTTTATCGCAACTCTC	TACTGT	TTCTCCAT A CCCGTTT		
bio A	GGGCCTTCCAAAAC	GTGTTT	TTTGTTGTTAATTCGGTG	TAGACT	TGTAA A CCTAAATCTT		
lac P115	CTTTACACTTTATG	CTTCCG	GCTCGTATGTTGTGTGG	TATTGT	GAGCGG A TAACAATTTC	+	+
lac IQ	GACACCATCGAATG	GTGCAA	AACCTTTCGCGGTATGG	CATGAT	AGCGCCC G GAAGAGAGT		
trp E.c.	TCTGAAATGAGCTG	TTGACA	ATTAATCATCGAACTAG	TTAACT	AGTACGC A AGTTCACGT	+	+
trp S.m.	GTGCAAAAAGAGGG	TTGACT	TTGCCTTCGCGAACCAG	TTAACT	AGTACAC A AGTTCACGG	+	+
trp S.t.	TACTGAAATAGGTG	TTGACA	TTATTCCATCGAACTAG	TTAACT	AGTACGA A AGTTCACAT	+	+
bio B	TTGTCATAATCGAC	TTGTAA	ACCAAATTGAAAAGATT	TAGGTT	TACAAGTC T ACACCGAT		+
bio P98	TTGTTAATTCGGTG	TAGACT	TGTAAACCTAAATCTTT	TAAATT	TGGTTT A CAAGTCGATT	++	++
ara C	TTCTGCCGTGATTA	TAGACA	CTTTTGTTACGCGTTTT	TGTCAT	GGCTTT G GTCCCGCTTT		
tet	AAGAATTCTCATGT	TTGACA	GCTTATCATCGATAAGC	TTTAAT	GCGGTA G TTTATCACAG	+	+
str	TCGTTGTATATTTC	TTGACA	CCTTTTCGGCATCGCCC	TAAAAT	TCGGC G TCCTCATATTG	+	+
spc	CCGTTTATTTTTTC	TACCCA	TATCCTTGAAGCGGTGT	TATAAT	GCCGC G CCCTCGATATG	+	+
rpo B	TGTAAACTAATGCC	TTTACG	TGGGCGGTGATTTTGTC	TACAAT	CTTACC C CCACGTATAA	+	+
L11	CGGCGATTTAATCG	TTGCAC	AAGGCGTGAGATTGGAA	TACAAT	TTCGC G CCTTTTGTTTT	+	+
tyr tRNA	TCTCAACGTAACAC	TTTACA	GCGGCGCGTCATTTGA	TATGAT	GCGCCCC G CTTCCCGATA	+	+
rrn D1	GATCAAAAAAATAC	TTGTGC	AAAAAATTGGGATCCC	TATAAT	GCGCCTCC G TTGAGACGA	+	+
rrn X1	ATGCATTTTTCCGC	TTGTCT	TCCTGAGCCGACTCCC	TATAAT	GCGCCTCC A TCGACACGG	+	+
rrn DXE2	CCTGAAATTCAGGG	TTGACT	CTGAAAGAGGAAAGCG	TAATAT	ACGCCAC C TCGCGACAGT	+	+
rrn E1	CTGCAATTTTTCTA	TTGCGG	CCTGCGGAGAACTCCC	TATAAT	GCGCCTCC A TCGACACGG	+	+
rrn A1	TTTTAAATTTCCTC	TTGTCA	GGCCGGAATAACTCCC	TATAAT	GCGCCACC A CTGACACGG	+	+
rrn A2	GCAAAAATAAATGC	TTGACT	CTGTAGCGGGAAGGCG	TATTAT	GCACACC C CGCGCCGCTG	+	+
gal P1	AATTTATTCCATGTCACACT	TTTCGCATCTTTGTTATGCTATGGT	TATTTC A TACCATAA				+

Figure 4. Promoters found using equation 1 and 2. The table was constructed from reference 7. The consensus sequence for the -35 box (TTGACA), -10 box (TATAAT) and the transcriptional start points are indicated above the sequences. Promoters found are indicated by + in the two right tables. ++ means two found promoters. Underlined sequences shows either a discrepancy between the positon of the promoter as defined in reference (7) and the position found using the computer program or the position of an extra promoter found by the computer. No + means that no promoter was found.

below 0.001. Those were:lam prm,0.0003;bio B,0.0003 and can thus be considered as "weak" promoters as discussed above. Also the four promoters that are difficult to detect must be considered as "weak".

	TTGACA	<--15-19-->	TATAAT	--> START
lam CIN	TAGATAACAATTGA	TTGAAT	GTATGCAAATAAATGCA	TACACT ATAGGT G TGGTTTAATT
T7 C	CATTGATAAGCAAC	TTGACG	CAATGTTAATGGGCTGA TAGICT	TATCTT A CAGGTCATCT
T5 26	CTTAAAAATTTCAG	TTGCTT	AATCCTACAATTCTTGA	TATAAT ATTCTC A TAGTTTGAAA
fd X	TCTTAATCTTTTTTG ATGCAA		TTCGCTTTGCTTCTGAC TATAAT	AGACAG G GTAAAGACCT
G4 B	GGCAAATAAATAGC	TTGCAA	AACACGTGGCCTTATGGT TACICT	ATGCCC A TCGCAGTCC
fd IV	TGATAAATTCACTA	TTGACT	CTTCTCAGCGTCTTAAATC	TAAGCT ATCGCT A TGTTTTCAA
lac UV5	TAGGCACCCCAGGC	TTTACA	CTTTATGCTTCCGGCTCG	TATAAT GTGTGG A ATTGTGAGC
bio B	TTGTCATAATCGAC	TTGTAA	ACCAAATTGAAAAGATT	TAGGTT TACAAGTC T ACACCGAT
gal P1	AATTTATTCCATGT CACACT	TTTCGCATCTTTGTTATGC	TATGGT	TATTTC A TACCATAA

Figure 5. Additional promoters found with equation 1 using a level of 0.0002. The additional promoters found are indicated as underlined structures. See Fig. 4 for further details.

b) Search for promoters in plasmid pBR322

Plasmid pBR322 has been sequenced (12) and its promoters have been mapped under the electron microscope within +/- 100 bp using an in vitro transcription system (16). In the in vitro transcription system, 6 promoters were found corresponding to the anti-tet (P1), tet (P2), bla (P3), ori (Pp), 110 base pair transcript (P4) and c-amp-promoters (P5). We performed a promoter search in the DNA sequence of pBR322 trying to find promoter like patterns corresponding to the promoters found with the in vitro transcription assay. The result is seen in Fig 6. All promoters except P5 (cAMP-promoter) could easily be found. At the same significance level also 4 other patterns similar to promoters were found. However, no ribosome binding sites could be found after those promoter patterns. The promoters in front of the oriRNA showed up as a double promoter. Promoter P5 could only be found when the significance level was decreased to a level where 4 more "false" promoters were found.

c) Search for promoters in the L10 and L11 operons of Escherichia coli The L10 and L11 operons are located together at position 89.5 min on the Escherichia coli genetic map. The L11 operon contains the genes rplK and rplA which codes for the ribosomal proteins, L11 and L1, respectively. The L10-operon contains four genes rplJ, rplL, rpoB and rpoC, with rplJ as the promoter proximal gene. The genes rplJ and rplL codes for the ribosomal proteins L10 and L7/L12, respectively. The two promoter distal genes, rpoB and rpoC, codes for the ß and ß' subunits of the RNA-polymerase (Fig 7). Both operons,

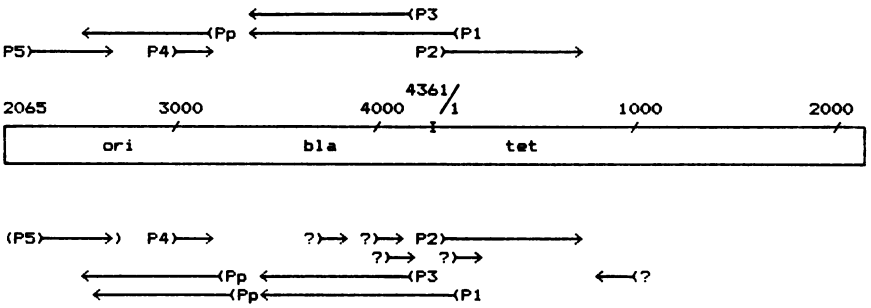
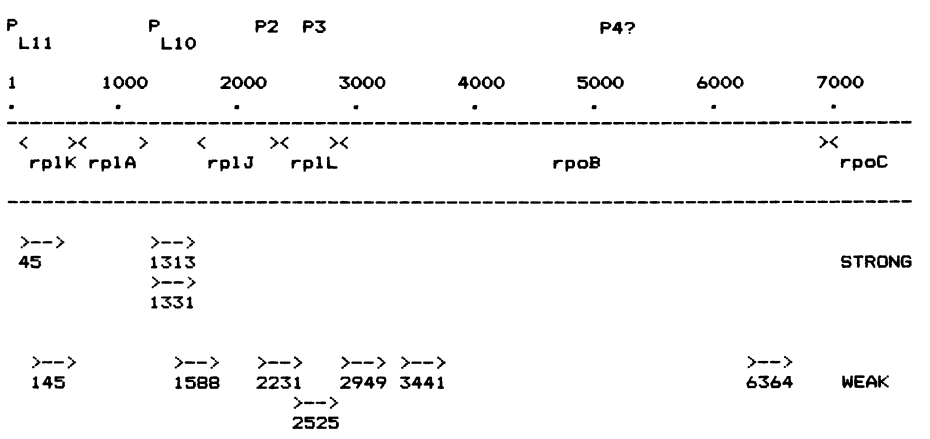


Figure 6. Promoter search in plasmid pBR322. The open box in the middle of the figure shows the position of the genes that have been located on the plasmid. Promoters found using an in vitro transcription system (16) are shown above the open box. Promoters found with the computer program are shown below the open box. ? denotes promoter-like patterns found with the computer program but that cannot be detected in vitro. Equation (1) and a level of 0.002 was used in the search.

except the last portion of rpoC have been sequenced as one continuous stretch of DNA and consists of 7604 basepairs (17,18,19). Two major promoter have been found in this region of the chromosome (20,31); one p_{L11}, is located just upstreams rplK and the other, p_{L10} (or p_{β}), in the 410 base pair intracistronic region between rplA and rplJ. Subclonings and gene fusions has been used to detect three weaker internal promoters in the L10 operon (21, 22,23). These promoters have sometimes been denoted P2, P3 and P4 (Fig. 7).

We searched the 7604 base pair sequence obtained from the EMBL database for promoter-like patterns. The result is also shown in Fig 7. Three significant promoter patterns can be found at positions 45, 1313 and 1331. The pattern at position 45 corresponds to p_{L11} and aligns well with the transcription start point at position 79 (17). The transcription start point for the p_{L10} promoter has also been mapped by in vitro transcription system on small DNA fragments to position 1348 (17) and the pattern found at position 1313 corresponds well with this initiation point. Interestingly enough, only 18 bases downstreams, at position 1331, a thirteen times more significant promoter pattern can be found. However, this pattern is seemingly not used in the in vitro transcription system.

Seven less significant promoter patterns are also found. From the compiled data of Linn and Scaife (21), Barry et al (22) and Ma et al (24), it can be deduced that the pattern at position 2231 is promoter P2 and the one found at position 2525 is the P3 promoter. The promoter patterns at position 3441 and 6364 can both be promoter P4 while a transcriptional activity has so



	T T G A C A			T A T A A T					
	==========			==========					
45	T T G C A C	50	68	T A C A A T	73	é =	0.003838	P(L11)
145	G T T A T T	150	168	T A T A A T	173	é =	0.000641	
1313	T T T A C G	1318	1336	T A C A A T	1341	é =	0.013319	P(L10)
1331	T T G T C T	1336	1354	T A T A A T	1359	é =	0.172222	
1588	G T G A C A	1593	1611	T T T T A T	1616	é =	0.001294	
2231	T T T A A A	2236	2253	T A A A C T	2258	é =	0.000267	P2
2525	C T G A A A	2530	2547	T A G A A T	2552	é =	0.000386	P3
2949	T T G T C A	2954	2971	T A T G G T	2976	é =	0.001089	
3441	T T G A C T	3446	3464	C A C T C T	3469	é =	0.000872	P4?
6364	C T G A A A	6369	6387	G A A A C T	6392	é =	0.000220	P4?

Figure 7. Promoter search in the L10 and L11 operon of Escherichia coli. The open box in the middle shows the position of the genes in the L10 and L11 operons. Above the open box, the positions for the in vivo promoters are shown. Below the box, the different strong and weak promoter patterns found with the search algorithm are shown. The table below indicates the nucleotide sequence and significance value of found promoter patterns. Equation (2) was used in the search.

far not been attributed to the remaining three patterns.

d) Search for splice points in a genomic clone of HLA124

Genomic DNA sequence coding for one human histoincompatibility immunoglobulin has been cloned and sequenced (13). The sequence that was taken from clone 12.4 is 4122 base pairs long. The gene contains six introns and the resulting protein was found to be composed of 338 amino acids. The junctions between exons and introns were defined in the same article. We used the compiled data for sequenced splice points presented by Brethnach and Chambon (9) in order to find the splice points using our search algorithm. A search for donor sites was first performed using the data table presented in Fig. 3 and a pattern of 9 bases assymetrically placed over the splice point. With a

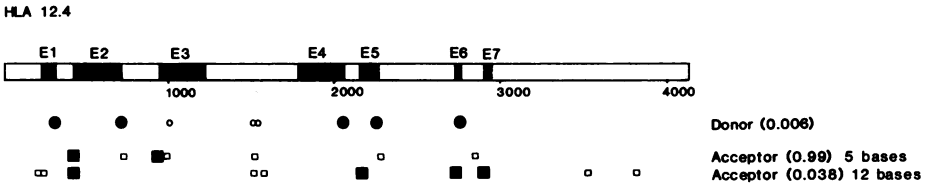


Figure 8. Search for donor and acceptor sites in HLA 12.4 sequence. Exons are marked E1 to E7. Filled circles and squares denotes correct donor and receptor sites, respectively. Open symbols denotes additional sites found. Equation (1) was used in the search.

discriminatory level of 0.006, 5 donor sites (IVS1, IVS2 ,IVS4 ,IVS5 and IVS6) were detected (Fig 8). Also 3 other donor like patterns were found. One was located very close to the acceptor site of IVS2 and the two others were located in the middle of IVS3. The donor site boundering IVS3 could only be detected when the discriminatory level was lowered to a point where the noise level had increased unacceptebly. The strategy was altered when acceptor sites were traced. First we used a data table of 12 bases (Fig. 3) and a low level of significance. Four acceptor sites were detected (IVS1, IVS4, IVS5 and IVS6)(Fig. 8). By shortening the data table to 5 bases but increasing the level of significance to a very high level (0.99) we were able to detect also the acceptor site of IVS2 (Fig. 8). A total of 11 "false" acceptor sites were also found. Two of these are located very close to the real acceptor sites (IVS2 and IVS6). Three others were found very close to the three "false" donor sites located in the middle of IVS3. Two "false" acceptor sites were found in close proximity to the donor sites of IVS2 and 5.

DISCUSSION

The search algorithm presented in this paper offers a new way to find patterns of bases inherited in genetic sequence. It requires a consensus sequence for a specific function e.g. promoters or splice points. The value of the search algorithm increases when more genetic material has been sequenced.

The program is flexible due to the fact that the pattern can be of any length and that any number of patterns can be linked together and that the distance between each pattern can be varied. The program in its present version has a small number of data tables built in but the operator has also the option to construct his own tables.

Two principally different equations to calculate the statistical significance of a pattern is presented, one of which multiplies the values for each position,the other is additive. All equations give a value of 1 for a perfect match, a feature which we find is both simple and logical to use. It is clear that the equation involving multiplication represents the "mathematically proper" way to perform the calculation. However, the equation has the disadvantage of biasing the result towards a predefined structure as is the case of the -10 region of a prokaryotic promoter where it is taken for granted that only T can exist in the last position. The equation involving addition does not have this disadvantage and may find its use in certain cases. It is however clear, that there is an absolute necessity to choose the right level of significance, irrespective of the equation used, before execution of the program.

In patterns like a promoter it can be shown (15,16) that the distance between the -10 and -35 regions is of importance for the function of a promoter. When this information was taken into account in the search it was possible to find more promoters with fewer false answers, thereby decreasing the noise level. In the articles mentioned above it was also shown that the strength of a promoter was greatly influenced by the distance between the two boxes. Optimum distance was found to be 17 bases. It can also be shown that the strength of a promoter most often increases when the sequence in the two boxes is altered towards the consensus promoter. It would be interesting to investigate if the algorithm presented in this paper also can be used to quantify detected structures.

In the pBR322 sequence, all promoters detected in an *in vitro* transcription system could be found but also four to eight other promoter like patterns out of which two were double promoters for the ori-RNA and the tet-gene, respectively. It would be interesting to perform an experiment to study the transcriptional activity of the detected patterns during *in vivo* conditions to see if the promoters found by the computer program may act as promoters *in vivo*. One way to improve the search algorithm would be to include patterns for also transcription initiation, ribosome binding and translation initiation. Such additions would strengthen the value of found promoter patterns. Of course, adding a pattern for ribosome binding would presumably mean that the ori-RNA promoter would never have been found.

All strong promoters found in vivo in the L10 and L11 operons could be easily found with our search algorithm. Interestingly, the algorithm found a 13 times more significant pattern very close to the tentative p_{110} promoter

picked out by Post et al (17). We think there are two possibilities for this anomality i) the right transcription start point was not picked up in the in vitro transcription system ii) if two promoter patterns overlaps each other the first is always choosen but the existance of two patterns increases the strength of the promoter. When the weak internal promotors were detected in the L10 operon 3 "false" patterns showed up, two of which were in the L10 operon and one in the intracistronic region between the two operons. Also here, an addition of a pattern for a translational startsignal would help in the evaulation of found patterns.

The HLA gene contains six introns and thus six donor and six acceptor sites. The search algorithm could find five of those but not the splice point for intron number three. Also three "false" donor and ten "false" acceptor patterns were detected. Two of the "false" donor sites and two "false" acceptor sites were found clustered in the middle of the third intron. It also became clear during the search that a combination of two acceptor site patterns, 5 and 12 base patterns, had to be used in order to find a maximum number of acceptor sites. Interestingely enough five "false" donor sites were found that matched the five base pattern perfectly which may interpreted to say that longer patterns in the sequence than five base pairs have to be involved to determine were an mRNA is to be spliced. However, using the longer pattern did not decrease the number of false acceptor sites. One of the acceptor sites found with the 5 base pattern did not show up when the 12 base pattern was used. Looking at the sequence around this site one can see that it is very GC-rich which does not follow the rules for the 12 base pattern. This indicates that the sequence around the 5 base box may have lower importance if the 5 base box is identical to the consensus. The rather high number of "false" donor and acceptor sites that were found is somewhat puzzling because the patterns used ought to be a very good discriminator againts the false structures. The result indicate that not yet discovered structures inherited in the sequences have an importance for the proper splice function.

As discussed above a DNA sequence like a promoter is composed of several patterns, some of which are involved in interaction with the surface of the RNA polymerase. To evaluate such an interaction to full extent it is not only necessary to take the pattern of bases into account but also the position of the bases in space. The first step towards such an evaluation was made in the article of Siebenlist et al (7). Our search algorithm can be modified according to this hypothesis and we believe that such a step would further

help the operator in the search biological significant patterns. A first step would be to present the result as a simplified 3-dimensional picture.

The published "perceptron" algorithm (10) produces a data table that looks similar to ours. However, the perceptron data table is produced in an entirely different way. The purpose of the perceptron is to separate two classes of sequences. It can be used to separate e.g. promoters from non-promoters. The result from the perceptron algorithm, however, looks similar to ours and can be used in an analogous way. We feel that our algorithm has two advantages over the perceptron algorithm. The first is the possibility to let the scientists knowledge direct the search towards predetermined structures by the construction of taylored data tables. The second advantage is that our algorithm permits the connection of any number of data tables with any distance between. The "perceptron" algorithm would produce a strange information table over prokaryotic promoters since it so far cannot take into account the variable distance between the -10 and -35 boxes.

The value of the presented algorithm will increase when more genetic sequences with inherited structures of biological significance will be found. We believe that such new sequence patterns may be composed of e.q. tRNA genes and enzyme active sites.

ACKNOWLEDGEMENTS

We wish to thank Tor Leif Lilja, the Department of Computer Science, University of Umeå, and Staffan Löf, the Swedish National Defence Research Institute, Umeå, for helpful advice. We also thank the staff at the Umeå Computer Central (UMDAC) Umeå, Sweden. We wish to thank Per Hagblom and Tomas Grundström for many valuable discussions. The work was supported by the Swedish Natural Science Research Council (Grant No 4629) and the Board for Technological Development (Dnr 81-3384).

REFERENCES

1. Applications of computers to research on nucleic acids, ed by D. Söll and R.J. Roberts (1982) Nucl. Acids Res. 10 vol 1
2. Staden, R. (1977) Nucl. Acids Res. 4, 4037-4051
3. Staden, R. (1978) Nucl. Acids Res. 5, 1013-1015
4. Korn, L. J., Queen, C. L., and Wegman, M. N. (1977) Proc. Natl. Acad. Sci. USA 74, 4401-4405
5. Harr, R., Hagblom, P., and Gustafsson, P. (1982) Nucl. Acids Res.
6. Maizel, J. V., and Lenk, R. P. (1981) Proc. Natl. Acad. Sci. USA 78, 7665-7669
7. Siebenlist, W., Simpson, R. B., and Gilbert, W. (1980) Cell 20, 269-281

8. Gold, L., Pribnow, D., Schneider, T., Shinedling, S., Swebilius Singer, B., and Stormo, G. (1981) Ann. Rev. of Microbiol. **35**, 365-404
9. Breathnach, R., and Chambon, P. (1981) Ann. Rev. of Biochem. **50** 349-383
10. Stormo, G. D., Schneider, T. D., Gold, L., and Ehrenfeucht, A. (1982) Nucl. Acids Res. **10**, 2997-3011 11. Minsky, M., and Papert, S. (1969) in Perceptrons, The MIT Press.
12. Sutcliffe, J. G. (1979) Cold Spring Harbor Symp. Quant. Biol. **43**, 77-90
13. Malissen, M., Malissen, B., and Jordan, B. R. (1982) Proc. Natl. Acad. Sci. USA **79**, 893-897
14. Mandecki, W., and Reznikoff, W. S. (1982) Nucl. Acids Res. **10**, 903-912
15. Stefano, J. E., and Gralla, J. D. (1982) Proc. Natl. Acad. Sci. USA **79**, 1069-1072
16. Stüber, D., and Bujard, M. (1981) Proc. Natl. Acad. Sci. USA **78** 167-171
17. Post, L.E., Strycharz, G.D., Nomura, M., Lewis, M., and Dennis, P.P. (1979) Proc. Natl. Acad. Sci. USA. **76**, 1697-1701
18. Delcuve, G., Downing, W., Lewis, M., and Dennis, P.P. (1980) Gene **11**, 367-373
19. Ovchinnikov. Y.A., Monastyrskava, G.S., Gubanov, V.V., Guryev, S.O., Chertov, O.Y., Modyanov, N.N., Grinkevich, V.A., Makarova, I.A., Marchenko, T.V., Polovnika, I.N., Lipkin, V.M., and Sverdlov, E.D. (1981) Eur. J. Biochem. **116**, 621-629
20. Yamamoto, M., and Nomura, M. (1978) Proc. Natl. Acad. Sci. USA **75**, 3891-3895
21. Linn, T., and Scaife, J. (1978) Nature **276**, 33-37
22. Barry, G., Squires, C.L., and Squires, C. (1979) Proc. Natl. Acad. Sci. USA **76**, 4922-4926
23. Newman, A.J., Linn, T.G., and Hayward, R.S. (1979) Mol. Gen. Genet. **169**, 195-204
24. Ma, J-C., Newman, A.J., and Hayward, R.S. (1981) Mol. Gen. Genet. **184**, 548-550