A small automaton for word recognition in DNA sequences and its application to consensus analysis of regulatory elements in DNA regions controlling gene expression.

Christophe Lefèvre & Joh-E Ikeda
Ikeda Genosphere Project/ ERATO
Tohkai University, School of Medicine
Isehara, Kanagawa 259-11. Japan
E-mail: clefevre@niguts.nig.junet

Abstract

A method for pattern analysis of DNA sequence data is considered. A space economical automaton for word recognition was presented elsewhere together with an algorithm for its compilation in linear time. An algorithm for the localization of words including imperfect matches (motif search) was developed. A program was implemented on the Macintosh and used extensively for the representation of the word composition of DNA data. We explore different sets of regulatory sequences to illustrate the performance of this method. In mammalian DNA, this analysis reveals "consensus motifs" corresponding to functional (or putative) cis-acting elements mediating the regulation of gene expression.

Introduction

The rapidly growing body of available DNA sequence information and the undertaking of large sequencing projects have created challenging interests for developing computational approaches to assist the molecular biologist in the identification of significant components in DNA. Consensus patterns defining genetic control elements are examples of such components. In eukaryotes, genetic regulatory elements, also called transcriptional or cis-acting elements, are short DNA motifs forming specific recognition sites for the binding of factors that regulate transcription over variable distances. Therefore, the molecular recognition of the DNA molecule by protein factors is an important aspect of the mechanisms governing the processing of biological information in living organisms. Moreover, the presence of transcriptional elements in the DNA indicates the existence of a regulatory code consisting primarily of the repertoire of functional regulatory sites. The characterization of this code is a very active domain of current molecular biology. As DNA sequences are being unraveled at an increasing rates, the analysis of this genetic information is an attractive and promising approach towards the deciphering of this regulatory code.

DNA can be looked upon as a molecular inscription in genetic language. Nucleic acid sequences can be viewed as words over the alphabet of nucleotides and become objects for linguistic analysis (Brendel 1984, 1986). In this context, the problem of word recognition in DNA is a challenging approach to a better elucidation of the molecular mechanisms ruling the expression of genetic information.

A method for the analysis of the word composition of DNA sequence data on a microcomputer is considered. A space economical automaton for word recognition, 20 to 30% smaller than the smallest automaton recognizing all sub-strings in a string (Blumer 1985), was presented together with an algorithm for its compilation in linear time (Lefèvre 1993, 1). This data structure allows the implementation of fast, on-line, string search functions. An algorithm for the localization of imperfect matches (motif search) was developed. The method is used for the representation of the word composition of DNA data. In regulatory DNA, this analysis reveals "consensus motifs" that correspond to functional cis-acting elements mediating the regulation of gene expression through the recognition of local DNA structure by specific transcription factors. For example, analysis of eukaryote genes responsive to heat shock treatment and mammalian growth hormone promoters identified the main elements of transcriptional activity (Lefèvre 1993, 2). Here, other examples are considered to illustrate the performance of the program. They include: mammalian oncoviruses, insulin and prolactin regulatory regions.

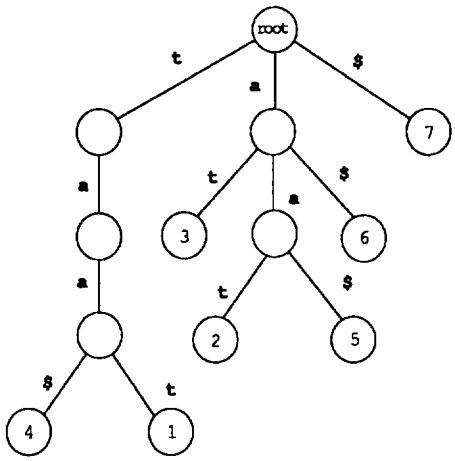
Method

The program was implemented on the Macintosh using object oriented C programming. All computations presented in this paper were made on a macintosh IIfx.

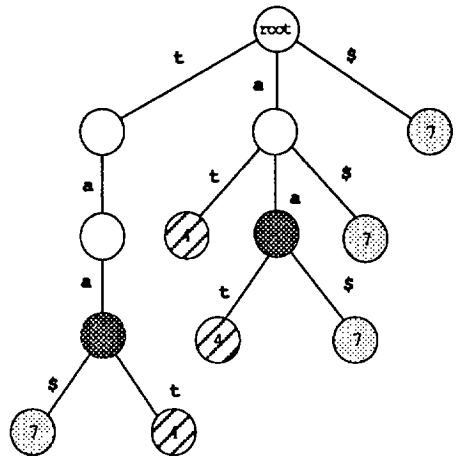
When doing extensive sub-string search in a sequence or main string of text, it is sometimes more efficient to preprocess the sequence in order to build an auxiliary index table (Weiner 1973). Such data structures known as position or suffix trees have been extensively developed and used in a wide variety of string matching problems (Aho 1974), including diverse applications to speed up biological sequence analysis (Clift 1986; Altschul 1990). In order to provide a high level of interactivity for the inspection of the sequence data, our approach was

Figure 1

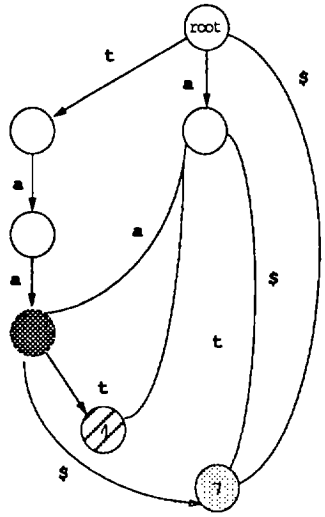
Construction of a compact lexical tree for word recognition in the sequence :"taataa$".



A) Position tree



B) End label Position tree



C) Position End-set tree

In A the classic position tree is represented. The leaves of the tree are labelled with the starting positions of the word composing the path from the root to the corresponding leaf.In B the leaves of the position tree are labelled with the ending positions rather than the starting positions of the corresponding words.Several branches become equivalent upon end position labelling, as illustrated in the figure. Equivalent nodes can be grouped in their end-set equivalence class to produce the compact position end-set tree as represented in C.

developed around the use of the original and space economical position end-set tree. This data structure, compiled during a preprocessing step, is briefly presented in figure 1 and, is described in details elsewhere together with an algorithm for its construction in linear time (Lefèvre 1993, 1). When searching for a sub-string in a sequence, the main advantage of the position tree is that every word is uniquely represented by a single path, even if the sub-string is encountered many times in the sequence. This word search facility can be annotated with word counts and used for the efficient enumeration of words matching a motif or pattern. Because transcriptional consensus elements rarely contain insertions or deletions we consider only misspells during motif search. We described elsewhere in detail a recursive algorithm which allows for the implementation of user defined pattern specification (Lefèvre 1993, 2). motifs are taken as a distribution of mismatches along word length. The user may impose restrictions upon this distribution, deciding that the n^{th} mismatch is not allowed before a match, with n-1 mismatches, of a defined minimal length. To produce a representation of the word composition in one or several concatenated DNA sequences, all the words composing the sequence(s) are simply matched against each other while counting the number of hits. During this phase, the listing of all the words in the data is also optimized through the use of the position tree. We distinguish between two types of word properties: 1) the counts of words of a preset length (k-tuple occurrence) 2) the length of words occurring above some preset value. Plotting these attributes as a function of word position along the sequence produces either of two types of profiles: 1) a k-tuple occurrence histogram or 2) a landscape (Stormo 1990). Peaks in the histogram shows the positions of k-tuple occurring with high frequency. Peaks in the landscape indicate unusually long words satisfying a minimal occurrence. This processing displays a global view of the properties of the words where peaks reveal the position of words presenting unusual features. Depending upon the nature of the DNA data under consideration, these words may pinpoint to biologically significant components. In search for putative control elements in regulatory DNA, sequence data assembly can follow two rationales. Taking into account the mechanism of evolution, the first principle postulates that regulatory sites are more sensitive to mutations than the surrounding spacing sequence, so that they tend to be conserved among evolutionally related species. This suggests that "consensus motifs" should be preferentially found in the regulatory regions of equivalent genes across related species. Hence, motifs presenting unusual properties in a concatenated sequence of related regulatory regions may pinpoint to transcriptional elements. The second rationale considers the molecular mechanisms involved in the specific recognition of DNA by transcription factors. The postulate is that the primary sequence constitutes the main determinant of the local feature of the DNA structure which is truly involved in the interaction with transcription factors. Therefore, in regulatory regions of "unrelated" genes sharing some aspect of gene expression, consensus motifs would be expected to point to cis-acting elements mediating the common transcriptional response. A main feature of the program is its high level of interactivity during the inspection of the resulting word graphs. In this phase the use of the position tree supports word search in real time. Zooming is available for exploration of the graphs in details around a region of interest. Interesting words can be selected and the position of their relatives can be displayed almost instantly. An option is also available for the display of the alignment of related words.

Results and Discussion

The eukaryote promoter database (Bucher 1991) is composed of updated references to well characterized transcription control regions stored in the EMBL database library. Most promoter regions were assembled using this handy referencing.

Consensus in mammalian oncovirus promoter regions.

Word graphs produced from the analysis of 14 mammalian oncoviruses promoters are represented were computed. In figure 2-A and B, only perfectly conserved words (no mismatch) are considered. With a total sequence length of 6330 nucleotides, the graphs are computed in about three seconds. In the landscape (figure 2-A) peaks indicate the position of extended perfect repeats with a minimum of 10 occurrences. The longest such repeat is 15 nucleotides. Its sequence, "caaacaggatatctg", occurs 10 times in the data and generates the 10 major peaks in the landscape. This DNA sequence corresponds to a transcriptional element recognized by the enhancer binding protein (Johnson 1987). A small region of the landscape has been enlarged underneath to show the details around one of the peaks in the murine leukemia virus promoter (REENVXA). The peak is due to a doublet corresponding to the overlapping sequences "caaacaggatatctg" and "aggatatctgtgga" 15 and 14 nucleotides in length respectively. Both are sub-strings of the enhancer core transcriptional element. A shorter repeat of 12 nucleotides with the sequence "tctcgcttctgt" indicates a second consensus. It is noteworthy that this sequence, when present, always occurs at a fixed position about 60 nucleotides upstream from the transcriptional start site. To our knowledge this sequence has not been shown to be a functional transcriptional element but our analysis suggests its possible importance. Figure 2-B shows a histogram obtained for words 14 nucleotides in length. Here, peaks indicate words occurring at high frequency. Not surprisingly, the most repeated words correspond to sequences related to the conserved enhancer core element described above. Landscapes and histogram are two different ways to look at the word composition of sequence data, but word length and word occurrence are

Figure 2
Mammalian Oncovirus promoter regions

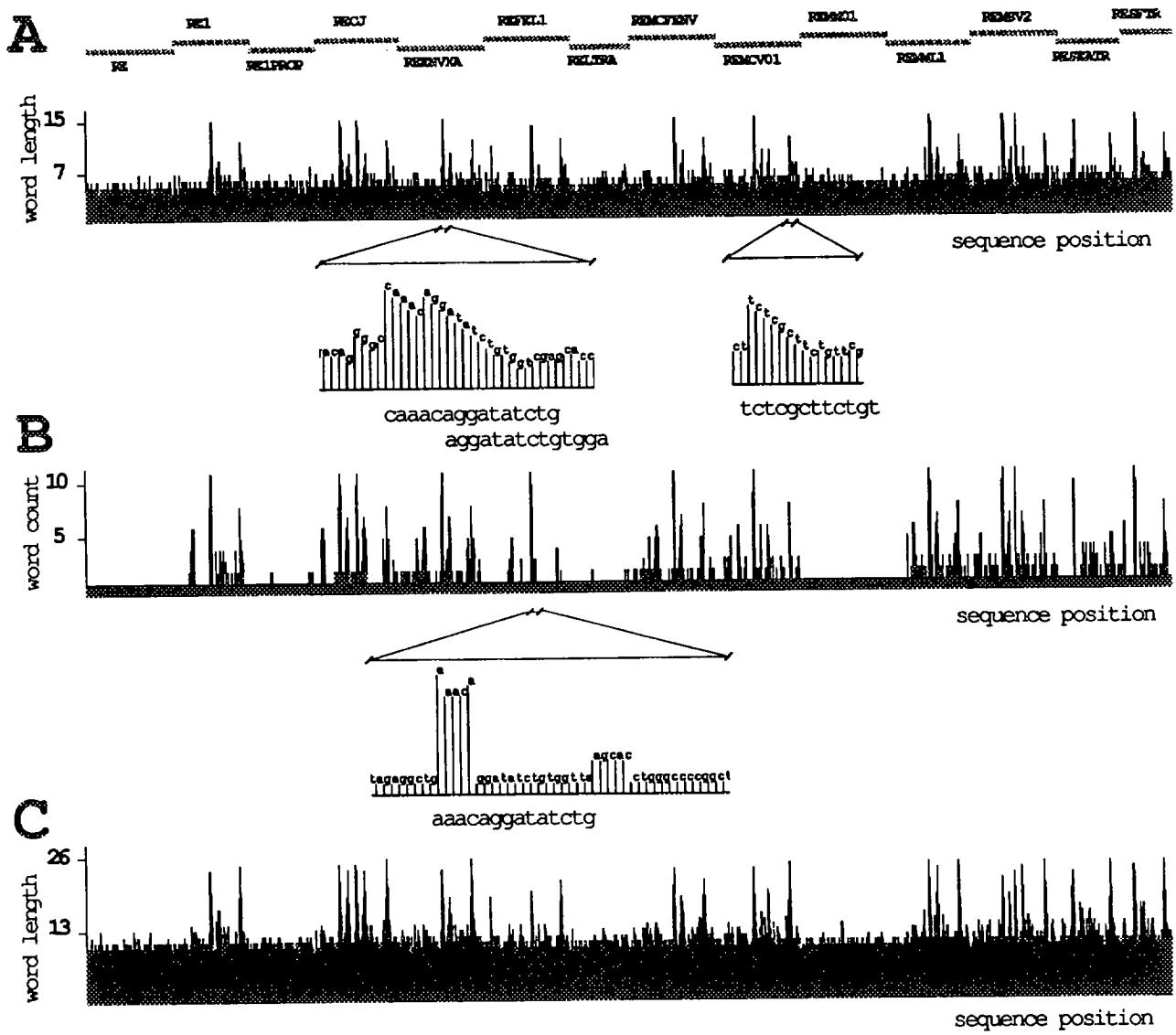


Figure 2: Consensus analysis in the promoter regions of mammalian oncoviruses. The sequences are labeled with the sequence name of their respective files in the EMBL gene database library. RE : Bovine Leukemia Virus, RE1 : Simian Sarcoma Virus, RE1PROP : Human T-cell Leukemia Virus, RECJ : AKV-Murine Leukemia Virus, REENVXA : Murine Leukemia Virus NZB, REFEL1 : Gardner-Arnstein Feline Leukemia Virus, RELTRA : Human T-cell Leukemia Virus 2, REMCFENV : Mink Cell Focus Forming Murine Leukemia Virus, REMCV01 : Friend Mink Cell Focus Inducing Virus, REMM01 : Mouse Mammary Tumor Virus, REMML1 : Murine leukemia virus, REMSV2 : Moloney Mouse Sarcoma Virus, RESEATR & RESFTR: Gibbon Ape Leukemia Viruses. Promoter regions were identified from the Eukaryotic Promoter Database. Total sequence length is 6330 nucleotides. The computation of graphs presented in figures 2-a & 2-b, where no mismatching is involved, takes about three seconds on a macintosh IIfx. **A)** Landscape showing the extent of perfectly conserved words with a minimum of 10 occurrences in function of their respective positions in the concatenated data. Two major consensus regions have been enlarged to show the detailed sequence (see text). **B)** Histogram showing the counts of perfectly conserved 14 nucleotides words. A major peak in REFEL1 has been enlarged. This peak correspond to the enhancer core transcriptional element. **C)** Landscape showing the extent of motif containing up to 3 mismatches, where the n^{th} mismatch is not allowed before 2n letters, with a minimum of 10 occurrences. Computation took 36 seconds. A three peak pattern can be clearly seen in some sequences (RE1, REENVXA, REMCFENV, REMCV01, REMML1).

two interdependent features of the statistical properties of words. However, in practice, histograms of word counts are valuable only if a significant number of sequences are available. This is rarely the case for DNA sequence data. When few sequences are available, the landscape is an alternative particularly relevant to the analysis of evolutionally related sequences. Because evolution operates through discrete and random variations of the DNA sequence due to mutation events, regions where variations are forbidden for the maintenance of vital processes tend to be preserved over extended length. Extended length of conserved regions is precisely what the landscapes represents. Another advantage of the landscape approach is that one needs only to fix a target occurrence for the consensus. Generally, this value relates to the number of sequences where one is seeking a consensus. This parameter is easier to estimate than the target word length for an histogram. However, as illustrated in figure 2 both approaches can be used in synergy for a more complete analysis. Figure 2-C shows the landscape obtained when allowing for up to three mismatches. If we restrict the mismatch distribution such that the n^{th} mismatch can only occur after 2n nucleotides, this landscape is computed in 36 seconds. All the peaks in this landscape refer to 3 sub-strings. Beside the enhancer core and the "tctcgcttctgt" signal characterized above, a new consensus best represented by the sequence "ggccaagaacagatggtccccaga" is characterized. It is noteworthy that this sequence always occurs about 50 nucleotides downstream of the enhancer core. These consensus lead to a typical three peak pattern in a subset of sequences (RE1, REENVXA, REMCFENV, REMCV01, REMML1). In two sequences: RECJ and REMSV2, a 5 peak pattern is seen. This is due to the presence of two copies of the enhancer core element, with their respective accompanying copies of the third consensus 50 nucleotides downstream. In this data set, note that out of 14 mammalian viral oncovirus promoters, 4 sequences show no consensus at all in any of the graphs presented in figure 2. These correspond to the mouse mammary tumor virus, a bovine and two human leukemia virus strains. The set with consensus includes murine (5 strains), feline, primate (Gibbon) leukemia viruses besides simian and Murine sarcoma viruses. Based on this consensus analysis the latest viruses fall in one set regarding their transcription control region. An important feature of our program is that not all sequences need to carry a consensus element.

Consensus in the insulin promoter.

Figure 3-A shows the landscape obtained with 8 mammalian insulin promoter sequences. Two large peaks above the human and chimpanzee sequences correspond to a highly polymorphic regions composed of many copies of a short sequence arranged in tandem repeats (Bell 1982). This arrangement generates the large peaks which tend to hide other shorter consensus signal. However, when the user is confronted with such an observation, the program provides interactive functions to facilitate analysis. For example, it is possible for the user to zoom in on the

Figure 3
Consensus in the insulin promoter

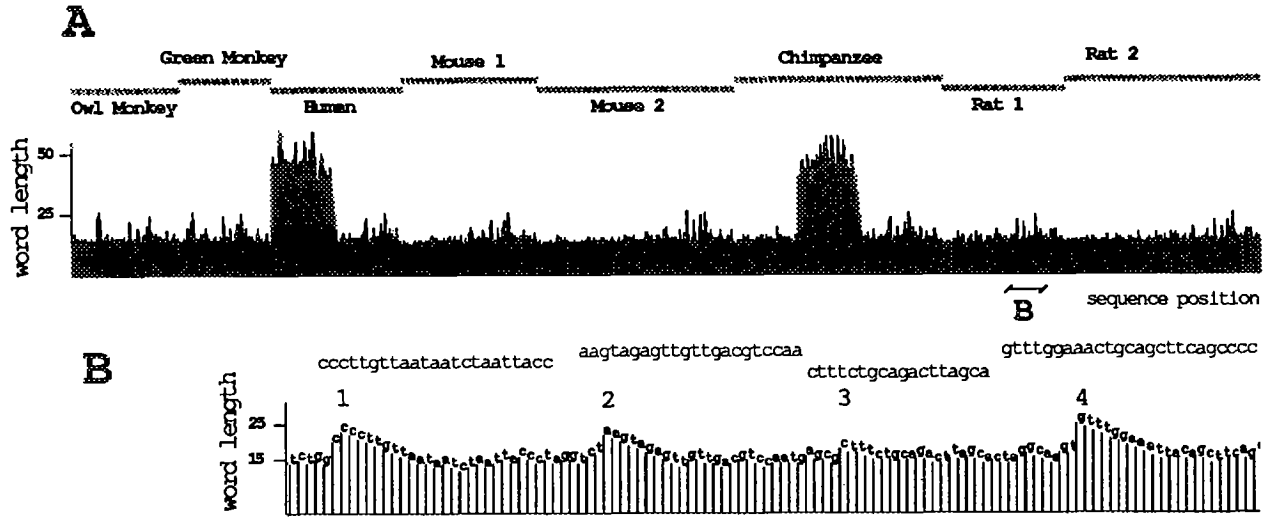


Figure 3: Consensus in the insulin promoter. EMBL library sequence names: Owl monkey: ATINS, Green monkey: CEPPINS, Human: HSINSV, Mouse gene1: MMINSIG, Mouse gene2: MMINSIIG, Chimpanzee: PTPPINS, Rat gene1: RNINSI, Rat gene2: RNINSII. Only the upstream proximal 5' flanking regions is considered. Total sequence length 5828 nucleotides. **A**) Landscape allowing for up to 5 mismatches (the n^{th} mismatch is not permitted before 2n nucleotides) and displaying motifs occurring at least 8 times. Computation takes 97 seconds. **B**) Enlargement in a short region of the rat gene1 promoter.

landscape and take a close look at the sequence. Any word generating a signal can then be selected and the position of its relatives can be displayed almost instantly. When doing this, one can realize that the large peaks relate only to the regions in the human and chimp sequence. This excludes them as insulin promoter consensus. In figure 3-B, the landscape is enlarged over a short region of the rat gene1 where 4 mammalian consensus peaks are found. They pinpoint to 4 subsequences of 23, 22, 18 and 25 nucleotides from left to right, respectively. Two of these sequences (labeled 1 and 3 in the figure) correspond to characterized transcriptional elements (Ohlsson 1986). In the case of peak 3, the signal is short and its significance could be overlooked. On the other hand, peak 4 suggest a putative transcriptional element.

Consensus in the prolactin promoter.

Two landscapes computed from 5 prolactin promoter sequences are presented in figure 4 when considering up to 5 (figure 4-A) or 2 (figure 4-B) mismatches. Consensus is found in the mammalian species but no signal is present in the salmon sequence. This sequence includes only 160 base pairs upstream from the transcriptional start site. However consensus is found in this area in the mammalian sequences. Similarly, in a study of the growth hormone promoter, including 6 mammals and two fish, consensus was found only among mammalian regulatory sequences (Lefèvre 1993, 1). In the rat (Sprague-Dawley) sequence extending 2 kilobases upstream from the transcriptional start site, peaks are seen in the distal regions. These peaks do not correspond to a prolactin consensus but, in stead, to regions containing short "ca"

Figure 4

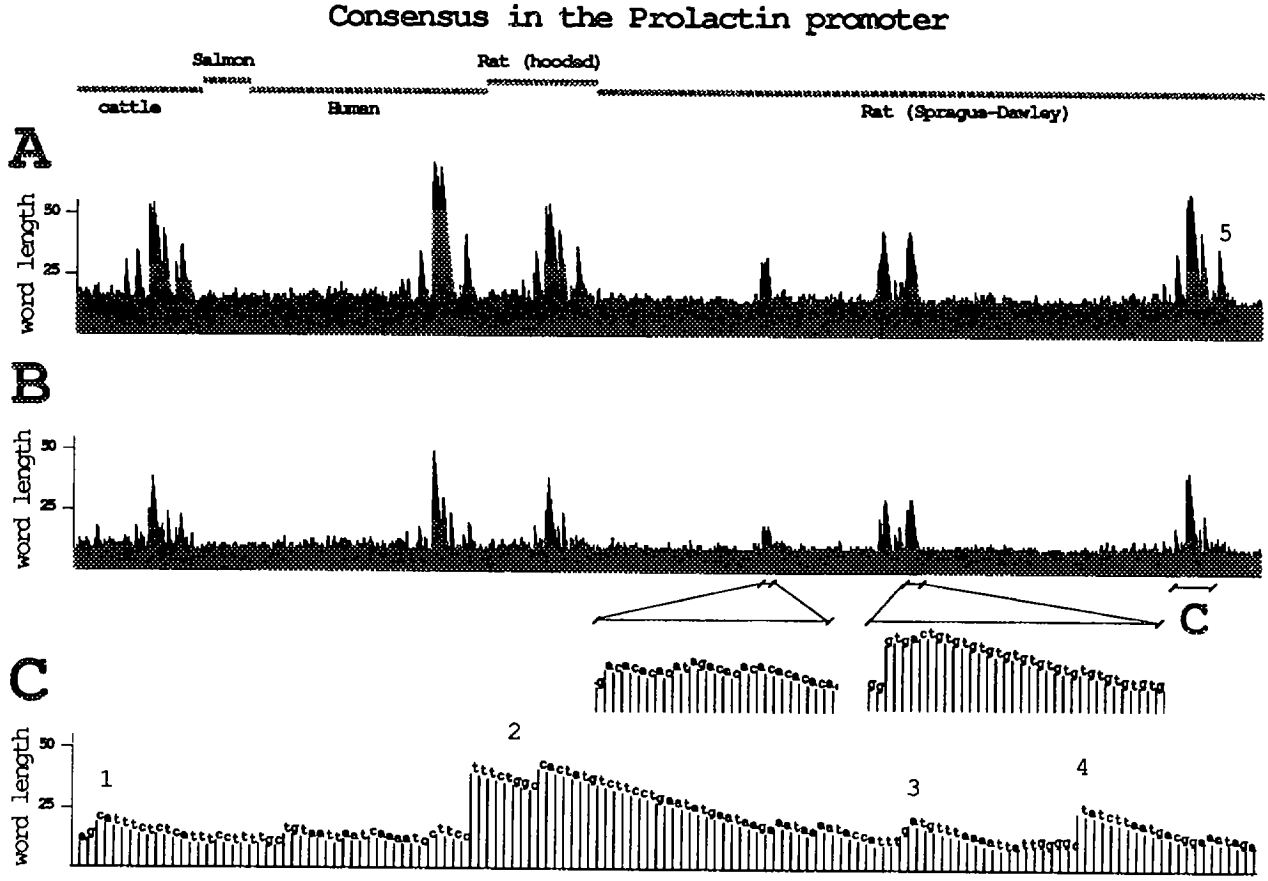


Figure 4: Consensus in the prolactin promoter. EMBL library sequence names: cattle: BOVPRLP1, Salmon: CYIPROLA, Human: HUMPRL1, Rat (Hooded): RATPRLHR, Rat (Sprague-Dawley): RATPRLSD. Only the upstream proximal 5' flanking regions is considered. Total sequence length 4794 nucleotides. A) Landscape allowing for up to 5 mismatches (the n^{th} mismatch is not permitted before 2n nucleotides) and displaying motifs occurring at least 4 times. Computation takes 87 seconds. B) Landscape allowing for up to 2 mismatches (the n^{th} mismatch is not permitted before 3n nucleotides) and displaying motifs occurring at least 4 times. Computation takes 17 seconds. Two short regions corresponding to a 'ca' and a 'tg' repeat are enlarged underneath. C) Enlargement of part of the landscape of figure 4-B in the consensus proximal prolactin promoter region of the rat (Sprague-Dawley) gene (see text for details).

and "tg" repeats. The landscape around these regions is enlarged underneath the figure to show sequence details. In the proximal region (200 base pairs), peaks correspond to mammalian consensus. This area is enlarged for the rat (Sprague-Dawley) sequence in figure 4-C. Consensus peaks are found in 4 areas (labeled 1 to 4 in the figure). Two regions (2 & 3) correspond to two previously characterized binding sites at -175 to -145 and -130 to -120 (Gutierrez-Hartman 1987). When more mismatches are permitted, a large consensus region includes peaks 2 and 3. Peak 1 corresponds to the -230 to -210 region. A binding site has been characterized with little details near this area at -210 to -190 (Gutierrez-Hartman 1987). This apparent discrepancy may be due to a low experimental resolution of the foot-printing experiment around this area. Peak 4 points to putative transcription element. In figure 4-A, under less stringent conditions, an extra peak is seen in the -65 to -25 region (labeled 5 in figure 4-A). Interestingly, this area includes part of the EGF and phorbol ester response elements (Elsholtz 1986) on the 5' border and the tataa box at the 3' border. In conclusion, this analysis of the prolactin promoter shows that consensus elements in the proximal upstream region of mammalian genes points to characterized and putative transcription elements.

conclusion

We have briefly discussed the use of an original automaton in the development of a consensus search program for the identification of unknown consensus "signals" in unaligned DNA. Other methods have been devised to search for patterns in DNA using concepts of expression search (Staden 1990), word enumeration (Waterman 1988), signal search (Bucher 1984) and matrix search (Stormo 1989, 1990; Hertz 1990; Lawrence 1990; Cardon 1992). We use an extensive word search approach. The position tree provides a flexible handle to the word structure of sequence data. Pattern recognition can be optimized with the implementation of user defined heuristics concerning pattern specification. Our approach proposes an efficient alternative when patterns can be defined more "stringently". This is useful for the study of highly related sequences from the genome of close mammalian species, where the most complex regulatory functions are found. As illustrated, in regulatory DNA, the method detects significant motifs. Transcriptional elements act in a position independent manner, and regulatory regions often include multiple elements of different types. Contrary to most previous consensus search methods, the program described here finds multiple consensus elements irrespective of their position. They also may be only present in a subset of sequences. The program presents to the user a global representation of word composition.

Because we wanted to offer the user maximum flexibility during the analysis of output data, the attractive real time word search provided by the use of automata originally motivated our approach. During this phase, efficient functions are available. The nature of a consensus and its occurring positions may be displayed instantly. This is the main feature of the program. Assessing the biological significance of consensus in DNA requires expertise and knowledge about the biological mechanism involved. The advantage of our approach is to provide a soft interface between the user and the sequence data. The pattern recognition abilities of the user can identify patterns at a higher level of organization. For example, in the mammalian oncovirus promoters, a three peak pattern was found. This kind of observation may lead to a better characterization of the rule governing specific interactions in the transcriptional machinery. Consensus motifs may appear due to different mechanisms. For example, short direct repeats in the prolactin and insulin promoters induce overlapping peaks devoid of meaning at the transcriptional level. During word enumeration, the filtration of overlapping consensus could be implemented at a computational cost. However, the user equipped with on-line routines may discard this consensus noise instantly.

Specific examples have been described to validate our approach in the analysis of a number of determinants of genetic regulation in DNA. The computations of word graphs for sequences in the 5 kilobase range were achieved in less than 2 minutes on the microcomputer.

The method could be improved in order to further explore the concept of lexical patterns or motifs in DNA sequences in connection with the molecular processes of recognition of mammalian DNA by transcription factors.

Acknowledgements. The authors would like to thank Dr Y. Nakano and K. Kojima for useful discussions. We also thank Dr Kim O'Hoy for her helpful comments on the manuscript.

References

Aho, A.V., Hopcroft, J.E. and Ullman, J.D.; (1974) The Design and Analysis of Computer Algorithm. Addison-Wesley, Reading, MA.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J.; (1990). A basic local alignment search tool. J. Mol. Biol., 215,403-410.

Bell G.I., Selby M.J., Rutter WJ.; (1982). Nature 295:31-35.

Blumer, A., Blumer J., Haussler D., Ehrenfeucht A., Chen M.T., Seiferas J.; (1985). Theoretical Computer Science 40, 31-55.

Brendel, V., Busse, H.G.; (1984). Nucleic Acids Res., 12, 2561-2568.

Brendel, V., Beckman, J.S. and Trifonov, E.N.; (1986). Journal of Biomolecular Structure & Dynamics, 4, 11-21.

Bucher, B.; (1991). EMBL Nucleotide Sequence Data Library Release 28, Postfach 10.2209, D-6900 Heidelberg.

Bucher, P., and Bryan, B.; (1984). Signal search analysis: a new method to localize and characterize functionally important DNA sequences. Nucleic Acids Res., 12, 287-305.
Cardon, L.,R. and Stormo G., D.; (1992). J. Mol. Biol., 223, 159-170.
Clift, B., Haussler D., McCornnell R., Schneider T.D. and Stormo G.D.; (1986) Nucleic Acids Res., 14, 141-161.
Ersholtz, H. P., Mangalam, H. J., Potter, E., Albert, V. R., Scott Supowit, A., Evans, R. M. and Rosenfeld, M., G.; (1986) Science 34, 1552-1557.
Gutierrez-Hartmann, A. et al. (1987) Proc. Natl. Acad. Sci. USA, 84, 5211-5215.
Hertz, G. Z., Hartzell, G. W. and Stormo G., D.; (1990). CABIOS, 6, 81-92.
Johnson, P. F. et al.; (1987) Genes & Development 1, 133-146.
Lawrence, C. E. and Reilly, A.; (1990). Proteins, 7, 41-51.
Lefèvre, C. and Ikeda,J.; 1 (1993) CABIOS, 9, Forthcoming.
Lefèvre, C. and Ikeda,J.; 2 (1993) CABIOS, 9, Forthcoming.
Nelson, C., Crenshaw, E. B., Franco, R., Lira, S.A., Albert, V. R., Evans, R. M. and Rosenfeld, M.G.; (1986). Nature, 322, 557-559.
Ohlsson, H. and Edlund, T.; (1986) Cell 45 , 35-44.
Staden, R.; (1990). In Method in Enzymology, Academic press., 183, 211-221.
Stormo G., D. and Hartzell, G., W. (1989). Identifying protein-binding sites from unaligned DNA fragments. Proc. Natl. Acad. Sci. USA., 86, 1183-1187.
Stormo, G.D.; (1990). In Method in Enzymology, Academic press., 183, 211-221.
Waterman, M.S.; (1988)., Waterman, M.S., Ed., CRC Press, Bocan Raton, Fla.,93-116.
Weiner, P.; (1973) IEEE 14th Ann. Symp. on Switching and Automata Theory 1-11.