Nucleic Acids Research, 1993, Vol. 21, No. 7 ***1655–1664***

Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids

Kornelie Frech, Günter Herrmann[1] and Thomas Werner*
Institut für Säugetiergenetik and [1]Institut für Medizinische Informatik und Systemforschung, GSF-Forschungszentrum für Umwelt und Gesundheit mbH, Ingolstädter Landstraße 1, D-8042 Neuherberg, Germany

Received July 22, 1992; Accepted March 5, 1993

ABSTRACT

We present a method to determine the location and extent of protein binding regions in nucleic acids by computer-assisted analysis of sequence data. The program ConsIndex establishes a library of consensus descriptions based on sequence sets containing known regulatory elements. These defined consensus descriptions are used by the program ConsInspector to predict binding sites in new sequences. We show the programs to correctly determine the significant regions involved in transcriptional control of seven sequence elements. The internal profile of relative variability of individual nucleotide positions within these regions paralleled experimental profiles of biological significance. Consensus descriptions are determined by employing an anchored alignment scheme, the results of which are then evaluated by a novel method which is superior to cluster algorithms. The alignment procedure is able to include several closely related sequences without biasing the consensus description. Moreover, the algorithm detects additional elements on the basis of a moderate distance correlation and is capable of discriminating between real binding sites and false positive matches. The software is well suited to cope with the frequent phenomenon of optional elements present in a subset of functionally similar sequences, while taking maximal advantage of the existing sequence data base. Since it requires only a minimum of seven sequences for a single element, it is applicable to a wide range of binding sites.

INTRODUCTION

Computer assisted analysis of nucleotide and amino acid sequences is a powerful tool to reveal similarities between nucleic acid or protein elements. Multiple alignments allow determination of phylogenetic relationships between homologous sequences of different species or subspecies (1, 2, 3). Though most of these alignments rely on the similarity of relatively long sequence stretches, a meaningful analysis can be carried out with shorter regions chosen properly (4). Similar or statistical methods have been applied to study transcriptional control sequences and other short regulatory structures, identifying highly conserved core sequences (i.e. TATAA, 5, 6, 7, 8). However, none of these methods was able to yield information about the actual length of a regulatory sequence. While consensus sequences were derived for many of the regulatory DNA elements (e.g. transcription factor binding sites, splicing/translation signals) definitions of the consensus limits remained arbitrary. Attempts to utilize the information content of individual consensus positions restricted the determination of the consensus extension to sequence regions aligned without gaps (9, 10). Other methods based on a suboptimal search for the highest conserved k-word (11, 12) or on expectation maximization (13, 14) yield excellent results. However, these algorithms neither include antibiasing of closely related sequences nor are they able to reject sequences necessitating exclusion of such sequences by the user. Though this are very elegant methods, they still cannot deal with closely related sequences and cannot reject false positive matches. Neural networks, though promising in principle, can only be applied to sequence elements for which numerous examples are known, a rare situation for most of the known consensi (15,16).

Biological significance as deduced by molecular analysis usually resides in a distinct region which includes a highly conserved consensus core and extends beyond the core in one or both directions. Since this is an intrinsic feature of the DNA, appropriate sequence analysis should allow extraction of such information.

We combined a conventional alignment scheme with a new evaluation method based on comparison of information contents in order to predict the extension and orientation of biologically relevant regions. This method is capable of antibiasing. In contrast to all published applications of the information content of nucleic acid sequences, our algorithm is capable of distinguishing between real binding sites and false positives. To test the reliability of this method, we focused on well characterized binding sites for transcription factors, for which sufficient data on sequence requirements and functionality were available.

* To whom correspondence should be addressed

SYSTEM AND METHODS

Database searches and sequence retrievals were carried out on a VAXstation 3200 (Digital Equipment Corporation). Sequences were taken from Genbank 68 and EMBL 27 and are referred to by their accession numbers. Program development was conducted either on a VAXstation 3200 or on an Alliant FX 2800 supermini-computer running UNIX (Berkeley 4.3) operating system.

Multiple alignments

A search string which is part of the best conserved core sequence of a binding site (e. g. GCAAT for the C/EBP site) is used as guideline for the multiple alignments. The alignments are anchored at the first nucleotide of this string from which the alignment proceeds in both directions. Alignments were obtained based on the method described by Sellers (17). In a preliminary step, sequences are aligned to each other to determine a distance matrix used for the construction of a minimum spanning tree on which the final alignment is based. In essence, the method begins with an alignment of the two most similar sequences in the set and then serially adds the next most similar sequences or sequence subsets. In contrast to virtually all commonly used methods, we do not discard any information upon aligning subsets. Instead of converting one subset to a consensus, we prefer to keep the full nucleotide distribution matrices of all subsets in the alignment, although this sacrifices speed. The distance of two subset consensus sequences at a position i is defined as

$$[1] \qquad D(i) = \sum_{a \epsilon B} \sum_{b \epsilon B} P_1(a,i) * S(a,b) * P_2(b,i)$$

$B = \{A,C,G,T,gap\}$

P_1, P_2 = nucleotide distribution matrices representing the two subset consensi
S(a,b) = matrix of single base distances ('scoring matrix')

We used a distance of 0 for identical nucleotides, a distance of 0.7 for A/G and T/C pairs and a distance of 1.0 for all other combinations. Gaps were assigned a distance of 0 to another gap and a 1.0 to all nucleotides.

The reduced distance for A/G and T/C matches was introduced after analysis of the Keytool (IntelliGenetics Suite database) NA sites which revealed a 4-fold bias for A/G and T/C substitution as compared to other substitutions (A/C, G/C, G/T). The reduced distance for A/G and C/T matches did not alter the general alignment but improved details resulting in higher consensus index scores as compared to the unitary matrix.

This algorithm covers the distance between individual sequences, between a subconsensus and an individual sequence, and between two subconsensi. Gaps are preserved in all successive alignments according to Feng & Doolittle (18), and nucleotide scoring by the distance matrix is averaged at each position across the entire set of sequences included in the alignment.

Antibiasing for closely related sequences

Closely related sequences should not be weighted as independent sequences in order to avoid biasing of the consensus. The program allows for optional weighting of closely related subsets in the alignment during determination of the minimum spanning tree. Weighting is carried out according to equation [2] which results in weight contributions between 0.5 and 1.0 for each branch.

$$[2] \qquad SW = 1 - 0.5 * e^{-(d/WF)}$$

SW = sequence weight
d = calculated distance from dual alignment
WF = weight factor (user defined)

The effective weight for each sequence is the product of the weight contributions from all branches leading to that sequence. Thus, a subcluster of 2 or more identical sequences will result in a total sequence weight comparable to a single independent sequence in the final alignment.

The alignment constructed in this way already contains all information necessary to delimit the biologically important region. However, retrieval of this information requires additional evaluation of the alignment results (see below) which does not influence the alignment in any way.

Evaluation of alignment results

The composition of individual positions of the final consensus is transformed into the consensus index (C_i). This is based on the entropy definition given by Shannon (19), which was adapted by Schneider et al. (10) for nucleotide sequences. We made an additional modification to allow the inclusion of gaps and incorporated [3] into our programs:

$$[3] \quad C_i = 100/\log 5 * \left(\sum_{b \epsilon B} p(i,b) * \log p(i,b) + \log 5 \right) \qquad 0 <= C_i <= 100$$

C_i = consensus index at position i of the multiple alignment
p(i, b) relative frequency of element b at position i of the multiple alignment

$B = \{A, C, G, T, gap\}$

The sequence weight factors defined in [2] are incorporated in [3] by:

$$p(i,b) = \left(\sum_{s \epsilon N} f(s,i,b) * SW(s) \right) / \sum_{s \epsilon N} SW(s)$$

$N = \{all\ sequences\}$
f(s,i,b): absolute frequency of element b in sequence s at position i
SW(s): sequence weight of sequence s

Formula [3] will yield the value of 100 for a totally conserved nucleotide and 0 for equal distribution of all 4 nucleotides or a gap at a single consensus position. Thus the conservation of individual nucleotide positions in the consensus is correlated with the numerical value of the consensus index. C_i values are averaged over a window of 3 nucleotides in order to simulate the sterical influence of nearest neighbors.

Delimitation of the significant region

The extension of a significant region is determined by a threshold operation based on the mean of the averaged consensus indices. The actual threshold is identical to the mean value excluding the scoring of the search string. A user-defined number of consecutive positions n (default 2) is allowed to score below the threshold and will be included in the significant region only if followed by at least by $2n+1$ positions with a C_i above threshold. In all other cases, the last position scoring above threshold will mark the last nucleotide of the significant region.

Usually, there is no difference in prediction of the significant region for use of either the mean or the threshold as cutoff criterion.

The consensus description includes nucleotide distribution and consensus index of each position and extent of significant regions. All consensus descriptions are collected in a library which is used by ConsInspector to identify binding sites.

Determination of the alignment validity

Each sequence and 30 randomly shuffled versions of the same sequence (search string fixed) are individually aligned to a consensus of all other sequences. The validity is calculated according to formula [4] which is designed to emphasize the important, i.e. best conserved positions (protein contact points) over less important positions (no direct protein contact). The formula was based on results of crystal structure analysis and mutational analysis of DNA/protein interactions. Proteins are most sensitive to changes at single nucleotide positions at the contact sites whereas they obviously tolerate variations at positions with which they are not in direct base contact. Thus, the contact sites are characterized by an extraordinary high degree of nucleotide conservation. Neither the Sellers alignment nor the information content appear to be sensitive enough to single nucleotide variations at contact sites to allow discrimination of potentially significant and insignificant matches. Formula [4] increases the penalty of a sequence variation at these positions up to fivefold as compared to the conventional information content, while it is relatively tolerant to variations at less conserved positions. Therefore, we feel that [4] represents a simple implementation of the 'discrete contact position' principle almost all proteins seem to follow. Furthermore, since each sequence is compared to random sequences of the same nucleotide composition (shuffled original sequence) the influence of a nucleotide bias (e.g. high G/C content) is neutralized.

This simple formula [4] is sufficient to allow discrimination of more than 90% of all random sequences from proven sites, despite conservation of the core in the random sequences. A sequence is included only if the calculated ΔC for the real sequence is at least 1 standard deviation above mean ΔC for the shuffled sequences. In all other cases, the sequence will be rejected.

$$[4] \qquad \Delta C = 1/n^* \sum_{i \epsilon S} \left(C_i^* \sum_{b \epsilon B} \mathrm{fcons}(i,b)^* \mathrm{fseq}(i,b) \right)$$

C_i = consensus index at position i
fcons(i,b) = relative frequency of element b at position i of the consensus
fseq(i,b) = frequency of element b at position i of the test sequence (0 or 1)
n = length of significant region(s) predicted from analysis of all sequences
S = {positions | positions within significant region(s)}
B = {A, C, G, T, gap}

We chose one standard deviation above mean as criterion for inclusion of a sequence in the consensus, since random sequences (as described above) rarely reach this threshold. Additional regions of significance can be included in the evaluation of alignment validity the program. Thus, it can favor sequences containing such additional regions and create a consensus preferably composed of sequences containing the additional element. This efficiently filters the initial set of sequences with regard to the presence of further elements. However, since this sort of filter may impair analysis of single elements it is an optional feature.

The 'twilight zone' and alignment stability

Dealing with alignments of short sequences with relatively low scores always calls for very cautious judgement of the results. If a sequence is accepted or rejected on a relatively low score this may be due to the choice of random numbers used in the calculations. There is a 'twilight zone' where no clear decision is possible if the sequence scores close to one standard deviation above mean of random sequences. In order to avoid arbitrary decisions in these cases, the program is able to perform a more extensive analysis. In brief, this consists of a ConsIndex analysis repeated 30 times with all sequences accepted for the consensus in the initial analysis. A twilight analysis can be carried out for each rejected sequence individually by forced inclusion of this sequence in the initial alignment. If a particular sequence of the consensus (including the initially rejected sequence) is rejected in more than 10 of the 30 test runs it is to be considered as a false positive and should be removed from the consensus. Otherwise it may be a weak but real match and should remain in the consensus. We did not include this procedure as the default rejection scheme since it is very time-consuming and confirms the results of the standard method in most cases.

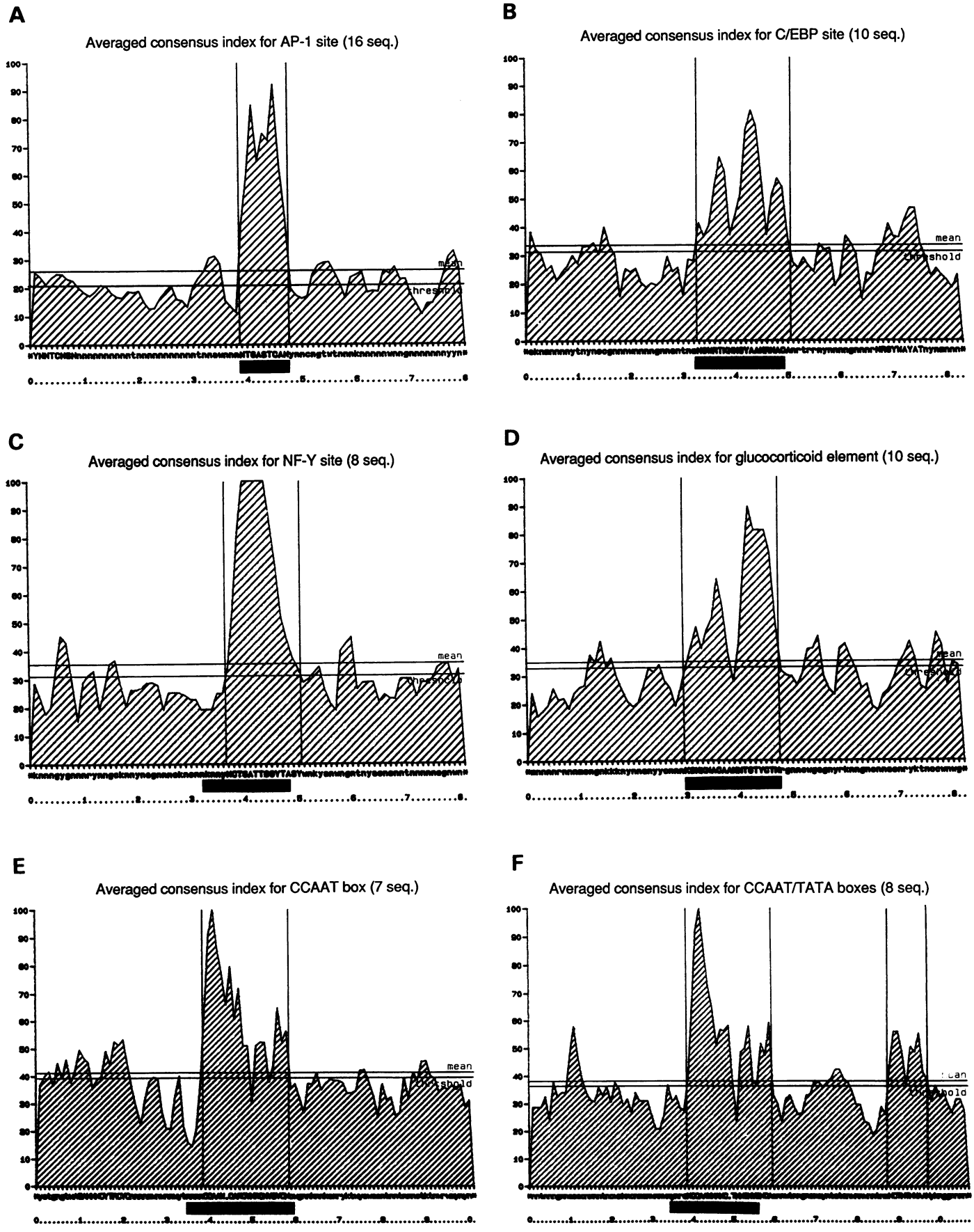
Alternatively, the twilight analysis can be used to examine the stability of the ConsIndex alignments. For this purpose a random sequence (containing the core string) is forced into the initial alignment. The rejection scheme and the alignment were found not to be disturbed by one totally unrelated sequence; all originally matching sequences are still accepted and the random sequence is rejected in at least 90% of all test runs.

Program output

The program ConsInspector generates an extensive results file listing all parameters used in the analysis, the accepted sequences along with their scoring, the initial consensus alignment used in the analysis, and the final alignment of the tested sequences to this consensus. If the consensus is updated with the new sequence, the new consensus alignment is given along with a minimum distance list and a semigraphical representation of the computed profiles for the averaged consensus index (consensus scores and consensus index are optional additions). The profiles are converted to an HP-GL file which can be plotted on any HP-GL compatible plotter (see fig. 1 and 2 for plot examples and Appendix I for an example of a complete output).

Availability and compatibility of the program

The program ConsInspector is written in C and is available for Vax-Computers running VMS (at least 5.3) and all UNIX-based computers. Sequence input is accepted from sequence files containing individual or multiple sequences (Genbank and EMBL) in either the IG (IntelliGenetics) or the GCG (Genetic Computer Group, Inc.) format. The program including a consensus library is available from the authors either on standard 0.5 inch magnetic tape (1600 bpi, 6250 bpi, UNIX only), an Exabyte video 8 or on a TK50 cassette (VMS). Users will receive the program at no charge and are asked to send their request along with their preferred storage medium and a self-addressed prepaid return-container. Updates of the program and the consensus library will be available on a subscription basis for which a nominal fee will be charged.


A
Averaged consensus index for AP-1 site (16 seq.)
mean
threshold
B
Averaged consensus index for C/EBP site (10 seq.)
mean
threshold
C
Averaged consensus index for NF-Y site (8 seq.)
mean
threshold
D
Averaged consensus index for glucocorticoid element (10 seq.)
mean
threshold
E
Averaged consensus index for CCAAT box (7 seq.)
mean
threshold
F
Averaged consensus index for CCAAT/TATA boxes (8 seq.)

RESULTS

Though successful alignment of core regions for generation of a consensus does not require gaps in the alignment, this is not true for alignments of the complete binding sites. Protein binding sites on DNA can accommodate nucleotide insertions or deletions without loss of function, as shown for the AP-1 binding site (20) or the glucocorticoid element (21), necessitating inclusion of gaps in the alignment method. Furthermore, DNA/protein interactions are influenced by nearest neighbor contacts either directly or via the DNA conformation (22). In order to account for this phenomenon, we averaged the consensus indices over a window of three positions which proved to be the best range.

The consensus index allows definition of biologically important regions

According to current knowledge, transcriptional control regions are composed of functionally important short regions (mostly protein binding sites) and spacer regions apparently for proper positioning of the binding sites. This is also true for several complex binding sites that are composed of individual elements. These basic elements are present in prokaryotic as well as in eukaryotic sequences and many of them are known and compiled in transcription factor databases (23, 24, 25). Thus analysis of individual binding sites should yield valuable information on DNA-function.

Mutations and even short deletions or insertions in spacer regions can be tolerated in many cases without loss of transcription control function and therefore these regions are believed to be biologically less significant. On the other hand, similar changes inside binding regions often abolish their function and interfere with protein binding, indicating high biological significance of these nucleotide stretches. The purpose of our method is to distinguish functionally significant regions from surrounding spacers which we show for 7 DNA binding elements. For alignments we chose sequences with proven functionality and avoided 'bona fide' sites identified solely by similarity to a short consensus core.

Definition of transcriptional control elements

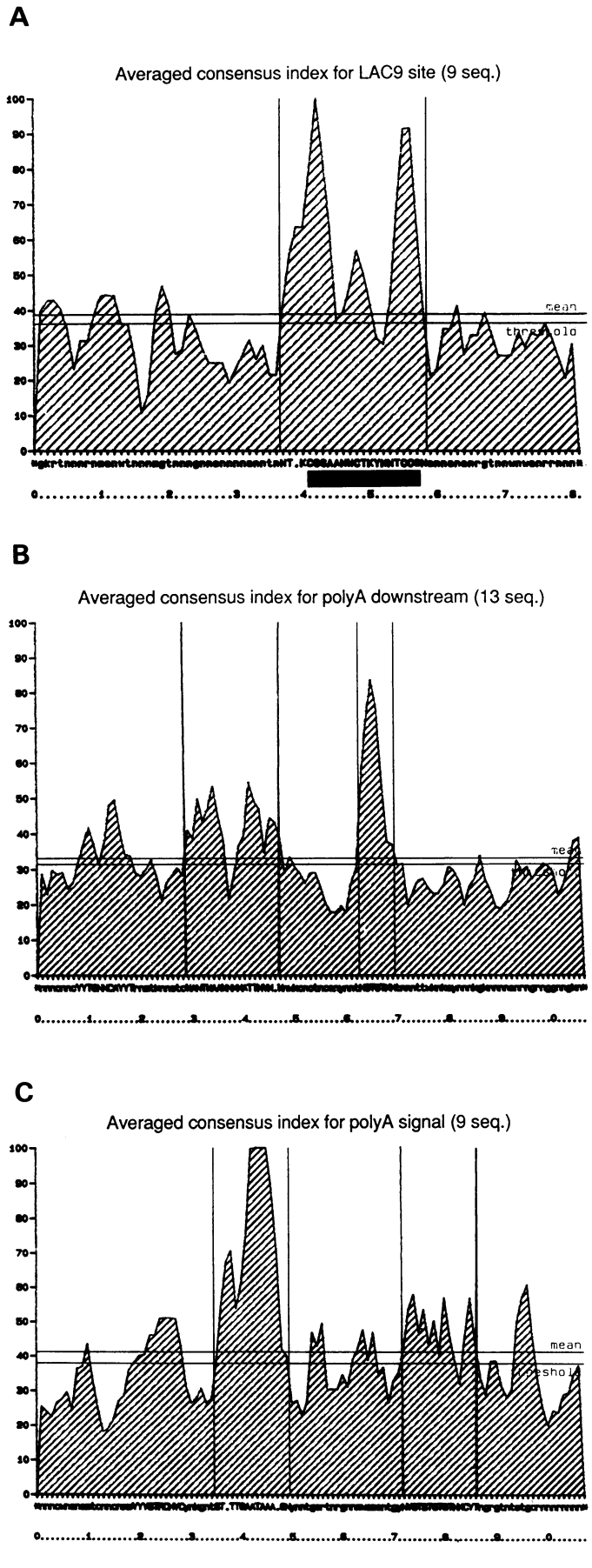
In the following we have delimited several transcriptional control elements and compared the results with data from molecular analysis. 'X's refer to nucleotide positions of the consensus located outside the highly conserved core region, regardless of the IUPAC code assigned in the consensus sequence. This allows easy comparison of the extent of the regions with those determined experimentally.

Consensus sequences for binding sites of transcription/enhancer factors AP-1, C/EBP, NF-Y, the glucocorticoid responsive elements (GRE) and the CCAAT box in retroviral long terminal repeats (LTR) were determined. All of the calculated regions were in agreement with experimental evidence as is evident from Fig. 1.

The AP-1 region XTGASTCAX (Fig. 1a) corresponds with the well-known AP-1 consensus described in the literature (20, 22, 26, 27). The region protected in footprint experiments for C/EBP (CAAT-Enhancer Binding Protein) binding (28) was determined to be X_5TKNNGYAAKX$_4$ which agrees with our prediction X_5KNNGYAAKX$_5$ (Fig 1b). NF-Y is another CCAAT box binding factor which is distinct from factors like C/EBP (29). We found a significant region X_4ATTGGX$_5$ including the invariant ATTGG box (CCAAT in the opposite strand, Fig. 1c). Though this differs from the footprint (X_9ATTGGX$_3$), it matches the binding analysis performed by these authors (single transversion mutations). The region within which a single transversion decreased NF-Y binding by 70% or more relative to the wildtype sequence was X_5ATTGGX$_5$ in concurrence with our prediction.

While 'bona fide' glucocorticoid responsive elements (GRE) are widespread in the database, only few sequences have been shown to be functional. Jantzen et al. (29) compiled the GGTACAN$_3$TGTTCT consensus from 10 functional sequences and have shown that approximately the same region is protected in footprint experiments (common region in footprints: X_{11}TGTTCTX$_5$), which is similar to the footprint X_{10}AGTGCTX determined by Strömstedt et al. (30) for the osteocalcin GRE. ConsIndex determined the region X_{11}TGTYCTX using the same sequence set (29) including the osteocalcin GRE (Fig. 1 d). Moreover, X_{11}TGTCCTX is almost identical to the region shown to be covered by the glucocorticoid receptor dimer in the crystal structure, thus confirming biological importance of this region (21). While the core sequence of the retroviral CCAAT boxes is only CCAAW, a much longer stretch of DNA is protected in footprint experiments (X_4CCAATX$_{15}$; 31). The

Figure 1. Analysis of single component regulatory elements. The averaged consensus index is plotted over the entire aligned region. The consensus sequence according to the definition of Cavener (5) is printed below the consensus index profile. The IUPAC code for nucleotide ambiguity is used. The consensus determined to be significant is given in upper case and marked by two vertical lines. The overall average of the consensus index and the threshold used to define the consensus region are represented by horizontal lines. The extension of protection against DNAse I determined in footprint experiments are shown as grey bars below the consensus sequence. Sequences used for the footprint experiments are in the set used for consensus prediction. All sequences are identified by a short code and their unique accession numbers. (**a**) 16 AP-1 binding sites. Sequences aligned: BPV-1 X02346,HPV-6 X00203, HUMMET2AA M15244, SV40XX V01380, PCGR K02989, HPV-18 X05015, HPV-8 M12737, HPV-33 A M12732, HPV-16 A K02718, HPV-33 B M12732, HPV-16 B K02718, HPV-33 C M12732, RBFCG K02708, HUMCN2A M16567, HPV-11 M14119, HPV-31 J04353. The footprint shown was determined by Quinn et al. (38). (**b**) 10 C/EBP binding sites. Sequences aligned: SV40 V01380, HSV tk V00463, AMV K00388, FSV gag J02194, AMV A M24159, FSV gag2 J02194, AMV B M24159, RAV-0 K03527, RERSV6 V01197, RSV J02024. The footprint was determined by Ryden & Beemon (28). (**c**) 8 NF-Y binding sites. Sequences aligned: MSV J02263, MUSMHKBA M11847, E-BETA K00123, HUMTKA M13643, XLHSP70 X01102, SUSH2B1G1 X04681, E-ALPHA M17389, SUSH2B1G2 X04681. The footprint was determined by Dorn et al (29). (**d**) 10 Glucocorticoid response element (GRE). Sequences aligned: HUMMET2a X00504, RATTOG5 A X05145, RATTAT M15863, RATTOG5 B X05145, HUMGH1 J00148, MMTV A K01045, MSV A V01185, MSV B V01185, MMTV B K01044, HUMBGP J03858. The footprint was determined by Strömstedt et al (31). (**e**) 7 retroviral CCAAT boxes. Sequences aligned: S71LTR J03061, RE3 K02016, ERV3 K02017, ERV1 K02919, AGMER M11390, AKV J01998, HSERSPC1 X06275. The footprint was determined by Tsukiyama et al. (32). (**f**) 8 retroviral CCAAT boxes, alignment elongated by 5 nucleotides to the right. This alignment identifies an additional consensus which includes the TATAA box region. Sequences aligned: HL1 K02722, RE3 K02016, ERV3 K02017, ERV1 K02919, AGMER M11390, AKV J01998, HSERSPC1 X06275, HSRTR1 M16779. The footprint is identical to Fig. 1 d).



region determined by our program is $CCAAWX_{14}$ (Fig 1e). Though this is extremely asymmetric relative to the core sequence, ConsIndex positioned the consensus largely in agreement with the footprint. Extension of the alignment in the 3'-direction did not change the predicted CCAAT consensus significantly ($CCAAWX_{14}$ versus $XCCAAWX_{14}$) but defined a second conserved region XCTNTMWAAX which coincides with the TATAA box region; the sequence TATAA complies with the TMWAA consensus (Fig. 1f).

Since ConsIndex detected a TATA consensus region 35 nucleotides downstream of the CCAAT box region, the method seems capable of recognizing multiple component elements provided a moderate distance correlation exists. The analysis of other sequence elements connected by spacers proved this to be true in all cases tested (see below).

Multicomponent elements

The *Kluyveromyces lactis* protein LAC9 is homologous to GAL4 and binds to a 16 bp binding site upstream of the lactose-galactose operon with dyad symmetry. The important binding sites are located at the far ends of the sequence $CGGN_{11}CCG$. Halvorsen et al. (32) performed a detailed mutational analysis of this region in filter binding assays. We analyzed a set of 10 sequences given by these authors and found a significant region $X_3CGGN_{11}CCGX$ (Fig. 2a) which is in very good agreement with the methylation interference footprint data ($CGGN_{11}CCG$).

We also reanalyzed the retroviral CCAAT box region this time anchoring the alignment to the TATAA box. This also indicated a second region located 5' to the TATAA box including the consensus CCNMW (which covers the CCAAT sequence). The distance of 35 nucleotides between the first C of CCNMW and the first T of TATAA in the final consensus sequence was exactly the same as in the CCAAT box alignment (data not shown), although individual distances of CCAAT and TATA boxes in the aligned sequences vary between 30 and 60. This indicates that the overall alignment is not affected by different anchor points. Although a CCAAT box-like element was detected, the extent of the significant region was not predicted correctly. This

Figure 2. Analysis of multicomponent regulatory elements. See legend to Fig. 1 for details of the plots. The extension of protection in a methylation interference footprint experiment is shown as grey bar below the consensus sequence of the LAC9 binding site. All sequences are identified by a short code and their unique accession numbers. (**a**) 9 LAC9 binding site from Kluyveromyces lactis. Sequences aligned: YSKGALB M18108, YSKLAC12 A X06997, YSKLAC12 B X06997, YSKGAL A X07039, GAL10UASI X07039, YSKGAL B X07039, GAL10UASII X07039, YSKLAC4 X00430, GAL7UASI X07039. The methylation interference footprint was determined by Halvorsen et al. (33). (**b**) 13 polyA downstream signal sequences. Furthermore, the polyA signal sequence AATAAA and another upstream sequence are found to be significant in this alignment. Sequences aligned: HUMINS01 J00265, RATPSBPA3 V01259, RATACCYB J00691, CHKHBADA2 J00854, MUSHBBMIN V00722, HUMHBB J00179, RABHBB1A1 J00659, MUSHBBMAJ J00413, RATINSII J00748, CHKHBBCOM J00858, Dogins J00042, DUKHBAD X01831, HUMHB4A J00153. (**c**) 9 polyA signal sequences AATAAA. The region containing the downstream signal as well as an upstream element are predicted in this analysis in addition to the AATAAA sequence. Sequences aligned: GGCOLA2C X01614, MUSHBBMIN V00722, RATACCYB J00691, GOTHBAI J00043, HUMHBA4 J00153, MUSHBA J00410, RATINSII J00748, HUMINS01 J00265, DUKBAD X01831.

suggests that the extent of a region should only be determined from alignments around the anchored position.

The polyadenylation signal is another example of a two or multiple component element. The sequence AATAAA is not sufficient for efficient polyadenylation of most mRNA sequences (34). The authors derived a consensus sequence for a polyadenylation downstream signal YGTGTTYY located 20 to 30 nucleotides downstream of the polyA signal which was shown to be involved in protein binding (35). ConsIndex predicted an XGTGTKX consensus which is slightly shorter than the published consensus. Interestingly, the program identified two additional significant regions, and identified a consensus correlating to the approximate position of the polyA signal, despite the relatively weak distance correlation (20−30). Realignment with emphasis on finding of additional peaks revealed a better defined second region starting with WANTWAA which includes the AATAAA canonical signal (Fig. 2b). Therefore, we reanalyzed this set of sequences, focusing the alignment on the polyA signal this time. The program predicted a downstream consensus sequence including the published downstream signal located at the same distance as in the previous alignment (Fig. 2c). As in the case of the CCAAT/TATAA boxes a second conserved sequence was identified solely by sequence comparisons.

The consensus index profile correlates with biological data

As can be seen from figures 1 and 2, individual profiles of the regions selected for analysis differ significantly. We compared these profiles with data available on the importance of individual positions within these regions. Fig. 3 shows superposition of profiles of biological significance (gray areas) derived from

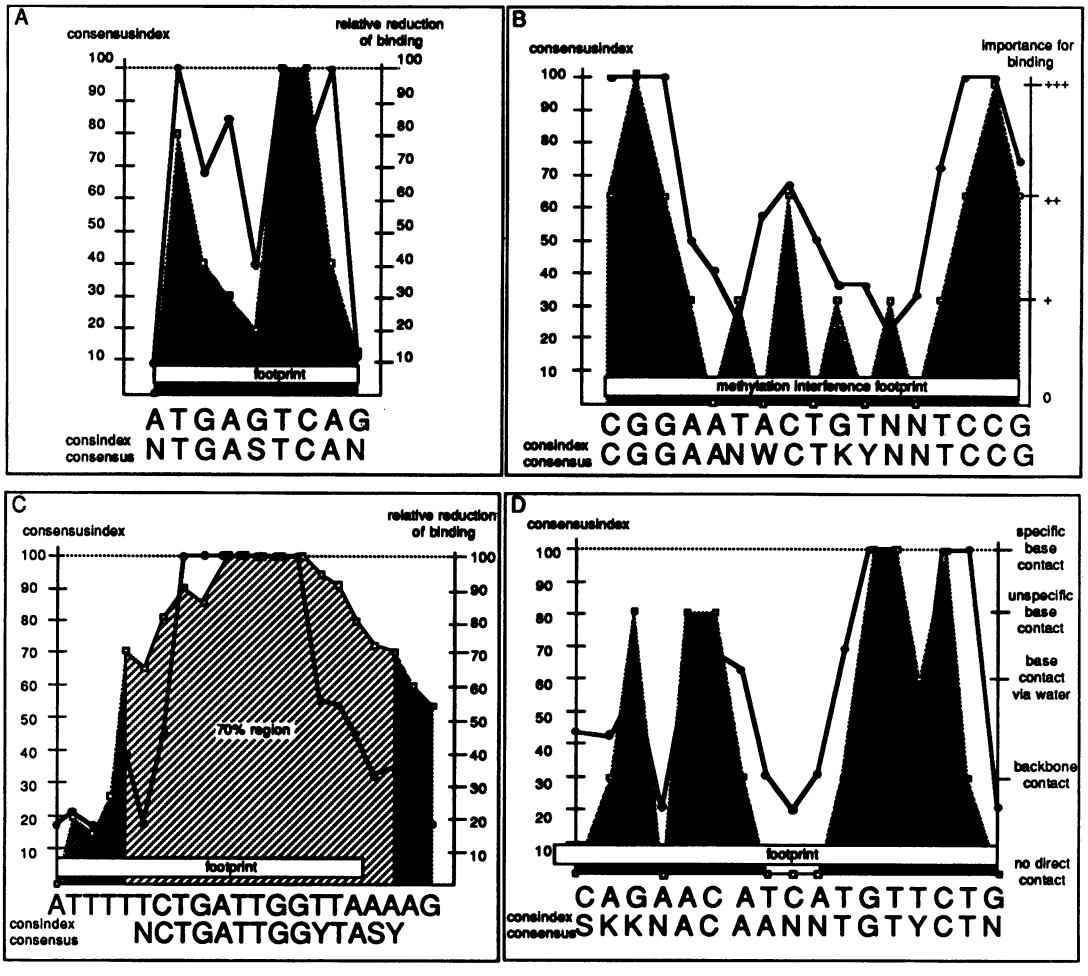


Figure 3. Comparison of consensus index profiles derived from the alignments shown in Fig. 1 and Fig. 2 (black line) with experimentally determined significance profiles for four consensus sequences (gray areas). Below the profiles the sequence reported for the experimental verification is given in the upper line and the significant consensus determined by ConsIndex in the lower line. White bars superimposed at the bottom of the profiles indicate protection by DNAse I or methylation interference in footprint experiments as detailed in Fig. 1 and Fig. 2. A. The AP-1 site. With exception of the fourth nucleotide ConsIndex predicted the relative significance of individual consensus positions parallel to the mutational analysis performed by Risse et al. (27). B. The LAC9 binding site. The experimental profile determined by Halvorsen et al. (33) and the consensus index profile are congruent. Note the correct prediction for nucleotide 8 (central C) as intermediate significant. C. The NF-Y binding site. The footprint is not congruent either with the consensus index profile or the transversion mutation results given by Dorn et al. (29). The region predicted by ConsIndex is a nearly perfect match to the experimental 70% reduction range. D. The glucocorticoid responsive element (GRE). The predicted consensus index profile reflects exactly the contact sites determined by crystallographic analysis of the DNA-protein complex. Note that the contacts made in the left half of the site are not sequence specific as determined by Luisi et al. (21) while the contacts in the right half are sequence specific.

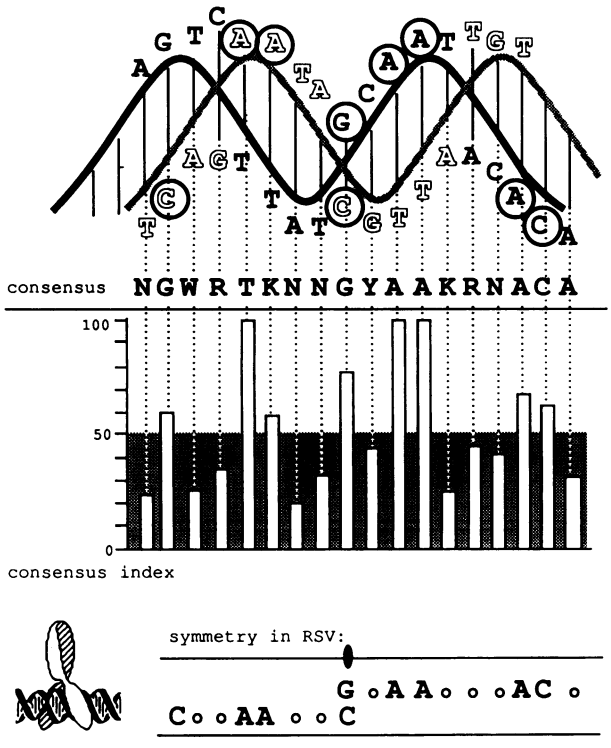


Figure 4. Predicted binding model for C/EBP dimer derived from ConsIndex profile. The consensus region of the RSV binding site is given superimposed onto a simplified helix model. Below the corresponding consensus sequence, consensus index values are shown as bars with the area below 50 indicated as shaded background. In the lower part the dyad symmetry element defined by the nucleotides with a consensus index above 50 is given along with an idealized binding model.

mutational studies or crystallography with the respective consensus index (black lines). The AP-1 site was found to be asymmetric with respect to binding of AP-1 (27). With the exception of the fourth nucleotide (A) the consensus profile parallels the biological profile (Fig 3a). The same is true for the LAC9 binding site (Fig. 3b). Even the relative importance of the central single nucleotide (C) demonstrated in vitro is precisely reflected in the consensus index profile. For the NF-Y binding site the consensus index agrees principally with the binding profile (Fig. 3c), marking the significant binding region (70% binding reduction) with the exception of a single nucleotide at the 5′ boundary. Interestingly the binding assays as well as ConsIndex showed the 5′ nucleotides (ATTTT) to be rather unimportant for NF-Y binding, despite their protection in the footprint assay.

The difference between the highly significant binding site, spacer and the less significant second binding site proven for the GRE by crystallography is also reflected in the consensus index profile (Fig. 3d). It should be noted that the base contacts in the first half scoring lower in the consensus index are unspecific contacts, while the base contacts in the second half are sequence specific (21). The profiles shown suggest a quantitative relation between the consensus profile and biological significance. However, more profiles have to be verified against biological data to prove this in general.

Many occurrences of AATAAA or CCAAT can be found which are most likely not functional (i.e. one gene may contain several AATAAA sequences but only one is biologically active in polyadenylation). The program ConsInspector analyzes a sequence in comparison to a consensus predefined by ConsIndex and determines whether a particular sequence contains a known consensus or only the respective core sequence by chance. ConsInspector takes advantage of a library of consensus sequences that ideally are deduced from sequences with known biological function. The program is fast and allows extension of predefined consensi.

DISCUSSION

We have developed a program that is capable of determining biologically important regions by a new procedure employing the consensus index differences in the evaluation of multiple alignments. This method is based on detecting nucleotide sequence conservation as observed in numerous DNA binding elements. However, it does not detect conserved secondary structure (hairpins) if the primary sequence is not conserved. The results of the program for 7 independent sets of sequences agree with experimental data derived from DNAse I protection footprints, binding studies with mutagenized DNA probes, and crystallographic determination of protein–DNA contacts. It appears that the extreme upstream and downstream positions are not reliable in terms of consensus extension and may in some cases depend on the particular alignment. Besides these 'fuzzy' termini, the extent of the consensus was stable in different alignments and after addition of several acceptable sequences. We expanded the CCAAT box alignment to 15 sequences by adding the corresponding region of various LTRs and predicted exactly the same CCAAT box region as in the shown alignment based on only 7 sequences ($CCAAWX_{14}$, data not shown). We tested up to 200 random shuffles for the rejection scheme and found 30 shuffles to be sufficient. Furthermore, the consensus index profile predicted the relative importance of individual positions within the consensus for all data available (4 sets of sequences). The program ConsInspector allows fast comparison of individual sequences with a library of consensus descriptions and is able to update the consensus descriptions. The output of this program is designed to give a maximum of information in one comprehensive figure, employing representations well-known from conventional alignment programs.

From our study it became clear that the minimal requirements of the program ConsIndex in the final alignment is a set of at least seven independent sequences aligned over a range of at least three to four times the size of the predicted significant region. This ensures that neither the alignment nor the predicted region varies considerably after adding an additional sequence or expanding the length of the alignment by 10% to 50%, which was confirmed by corresponding variations of the analyses. All results presented here were obtained with a fixed set of parameters for the alignments (see appendix) although slight variations of the parameters were found to improve individual alignments. The approach with one set of parameters is a safeguard against artificial matches produced by adjusting the parameters to individual sets of sequences.

Detection of biological significance by the consensus index difference approach

ConsIndex and ConsInspector rely on well known basic procedures. However, the novel concept to employ the consensus index in the validity check of the alignments without interfering

with the alignment procedure is one of the most important features underlying the unprecedented predictive power of the programs. The described consensus index difference method [4] is able to distinguish between real sequences (with proven biological functionality) and random sequences (containing a perfect core string, e.g. GCAAT for the C/EBP site) roughly in a 90:10 ratio with less than 5% false negatives. In contrast, conventional cluster algorithms employed in almost all multiple alignment approaches separate poorly in a 60:40 ratio including about equal amounts of false negatives and false positives. Thus cluster analysis cannot discriminate consensus members from random sequences.

The method by Hertz et al. (12) is able to tolerate a few random sequences but does not reject them. Their method of matrix comparison allows distinction of real matches from random matches but is very likely to fail to identify false positives which do include the best conserved core string. We analyzed the CCAAT-box example and determined the differences between a weak but real member, a false positive (with core string) and a random sequence (no core string.) Determination of the optimized matrix scores according to the method described in (12) resulted in a difference between the false positive and the real members that is almost the same as that between two real members, rendering discrimination between false positive and real member impossible. In detail the results were as follows:

Difference between real matches: 2.43 (ΔC_i), 3.2 (matrix score, according to 12), real versus false positive 13.08 (ΔC_i), 2.78 (matrix score), real versus random 28.64 (ΔC_i), 12.64 (matrix score). Thus our program separated the false positive match as well as the matrix score separated the random sequence (13.08/12.64) while the matrix score method clearly failed to separate the false positive match.

The consensus index profile together with this unique rejection scheme provide powerful tools for nucleic acid sequence analysis, yielding detailed information about biological importance not available from any established software.

As already discussed for the glucocorticoid receptor the consensus index profile of a binding site can yield information about likely protein contact positions and thus allows limited predictions of properties of the binding protein(s). This concept is illustrated in Fig. 4 for the C/EBP binding site. C/EBP sites bear only limited resemblance at the nucleotide level (36). This is reflected in the consensus containing many IUPAC ambiguities. However, comparison of the consensus profile with a very simple DNA helix model superimposed on the RSV C/EBP site in Fig 4 reveals well-conserved key positions in the binding site. The nucleotides with a consensus index above 50 define an element with dyad symmetry suggesting binding by either a protein containing two identical binding regions or by a dimer of the binding protein(s). Fig. 4 would suggest that a dimer could bind in the major groove to make contact to identical nucleotides located at very similar positions relative to the DNA helix. In fact the dimer binding model is supported by experimental data as well as by analysis with molecular modelling (36). However, the ConsIndex analysis allows this interpretation solely from the binding sites without any knowledge about the protein sequence and its structural features which are prerequisite for molecular modelling.

Due to experimental methodology biological analysis often reveals families of closely related sequences. Utilization of those sequences in multiple alignments results in biasing of the consensus towards the consensus of the related cluster of sequences. Since preselection of the sequences in order to eliminate this bias is tedious and necessitates omission of valuable data, the weighting scheme in our alignment procedure prevents this type of consensus biasing and allows inclusion of almost all experimental data. All methods based on a matrix scoring scheme without alignment preclude antibiasing since an alignment is a prerequisite for antibiasing. The matrix score of the binding site also is lower in unaligned sequences as compared to aligned sequences as shown by Hertz et al. (12).

Other approaches (11, 12, 13, 14) have the advantage of being able to define a binding site from scratch. However, they must compromise in specificity to do so, i.e. they cannot discriminate false positives. In contrast, our method requires higher quality input but offers a higher quality of discrimination enabling detection of false positives. Only the Cardon & Stormo algorithm is able to detect variable spacers, and the process is tedious, unless *a priori* information about the second element and/or spacer is available. Our method locates these additional elements automatically without any *a priori* information.

The combined ability of ConsIndex and ConsInspector to include closely related sequences without bias, reject sequences during alignment and identify new elements on the basis of moderate distance correlation is unparalleled by any other program. Thus this software is a powerful tool for selection of candidate sequences for a certain regulatory element prior to experimental analysis of the sequences.

The basic construction of a consensus will require very powerful computers due to the excessive number of analyses necessitated by the iterative process of building, checking and rejections including the twilight analysis. However, comparison of a single sequence with a predefined consensus requires only a workstation and should also run on an appropriate PC.

The application of ConsIndex and ConsInspector is shown for protein binding sites in nucleic acids so far. In principle, the programs should be able to deal with any conserved pattern in nucleic or amino acid sequence alignments since the detection algorithm is independent of the alignment method. Thus, the best available method for protein alignment can be used. However, this is not yet established for protein alignments.

We are currently completing another program employing ConsIndex to locate and identify much more complex structures like retroviral LTRs based on principles previously outlined (37). Our new development already allows identification of unknown binding sites in unaligned sequences while retaining the unique ability to discriminate false positives. This requires no more input than the expectation maximization method (Wolfertstetter et al., in preparation).

Since this is the first computer assisted method explicitly allowing delimitation and discrimination of biologically important regions in sequence data, it is an extension of options available for sequence analysis. Especially in the light of the numerous genome sequencing projects, software of this kind is an important step towards rapid extraction of biological information from sequence data prior to selective molecular analysis.

ACKNOWLEDGEMENTS

We thank Dr Ruth Brack-Werner and Franz Wolfertstetter for critical reading of the manuscript and many helpful suggestions. We would especially like to thank Prof. Dr U.Ehling for his continuous support of this work. We thank both reviewers for their detailed comments which helped clarify the manuscript.

REFERENCES

1. Koonin, E. V. (1991) J. Gen. Virol. 72, 2197–2206.
2. Elena, S. F., Dopazo, J., Flores, R., Diener, T. O., and Moya, A. (1991) Proc. Natl. Acad. Sci. USA 88, 5631–5634.
3. Doolittle, R. F., Feng, D. F., McClure, M. A., and Johnson, M. S. (1990) Current Topics in Microbiology and Immunology 157, 1–18.
4. Werner, T., Brack-Werner, R., Leib-Mösch, C., Backhaus, H., Erfle, V., and Hehlmann, R. (1990) Virology 174, 225–238.
5. Cavener, D. R. (1987) Nucleic Acids Res. 15, 1353–1361.
6. Golemis, E. A., Speck, N. A., and Hopkins, N. (1990) J. Virol. 64, 534–542.
7. Galas, D. J., Eggert, M., and Waterman, M. S. (1985) J. Mol. Biol. 186, 117–128.
8. Mengeritsky, G., Smith, T. F. (1987) Comp. Appl. Biosci. 3, 223–227.
9. Goodrich, J. A., Schwartz, M. L., and McClure, W. R. (1990) Nucleic Acids Res. 18, 4993–5000.
10. Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986) J. Mol. Biol. 188, 415–431.
11. Stormo, G. D., Hartzell III, G. W. (1989) Proc. Natl. Acad. Sci. USA 86, 1183–1187.
12. Hertz, G. Z., Hartzell III, G. W., and Stormo, G. D. (1990) Comp. Appl. Biosci. 6, 81–92.
13. Cardon, L. R., Stormo, G. D. (1992) J. Mol. Biol. 223, 159–170.
14. Lawrence, C. E., Reilly, A. A. (1990) Proteins 7, 41–51.
15. O'Neill, M. C. (1991) Nucleic Acids Res. 19, 313–318.
16. Demeler, B., Zhou, G. W. (1991) Nucleic Acids Res. 19, 1593–1599.
17. Sellers, P. H. (1974) J. Appl. Math. 26, 787–793.
18. Feng, D.-F., Doolittle, R. F. (1987) J. Mol. Evol. 25, 351–360.
19. Shannon, C. E. (1948) The Bell System Technical Journal 27, 379–423.
20. Angel, P., Hattori, K., Smeal, T., and Karin, M. (1988) Cell 55, 875–885.
21. Luisi, B. F., Xu, W. X., Otwinowski, Z., Freedman, L. P., Yamamoto, K. R., and Sigler, P. B. (1991) Nature 352, 497–505.
22. Lee, W., Mitchell, P., and Tjian, R. (1987) Cell 49, 741–752.
23. Wingender, E. (1988) Nucleic Acids Res. 16, 1879–1902.
24. Ghosh, D. (1992) Nucleic Acids Res. 20, 2091–2093.
25. Bucher, P., Trifonov, E., N. (1986) Nucleic Acids Res. 14, 10009–10026.
26. Schüle, R., Umesono, K., Mangelsdorf, D. J., Bolado, J., Pike, J. W., and Evans, R. M. (1990) Cell 61, 497–504.
27. Risse, G., Jooss, K., Neuberg, M., Brüller, H.-J., and Müller, R. (1989) EMBO J. 8, 3825–3832.
28. Ryden, T. A., Beemon, K. (1989) Mol. Cell. Biol. 9, 1155–1164.
29. Dorn, A., Bollekens, J., Staub, A., Benoist, C., and Mathis, D. (1987) Cell 50, 863–872.
30. Jantzen, H.-M., Strähle, U., Gloss, B., Stewart, F., Schmid, W., Boshart, M., Miksicek, R., and Schütz, G. (1987) Cell 49, 29–38.
31. Strömstedt, P. E., Poellinger, L., Gustafsson, J. A., and Carlstedtduke, J. (1991) Mol. Cell. Biol. 11, 3379–3383.
32. Tsukiyama, T., Niwa, O., and Yokoro, K. (1989) Mol. Cell. Biol. 9, 4670–4676.
33. Halvorsen, Y-D., Nandabalan, K., and Dickson, R. (1991) Mol. Cell. Biol. 11, 1777–1784.
34. McLauchlan, J., Gaffney, D., Whitton, L. J., and Clements, J. B. (1985) Nucleic Acids Res. 13, 1347–1368.
35. Weiss, E. A., Gilmartin, G. M., and Nevis, J. R. (1991) EMBO J. 10, 215–129.
36. Lamb, P., Mcknight, S. L. (1991) Trends Biochem. Sci. 16, 417–422.
37. Brack-Werner, R., Barton, D. E., Werner, T., Foellmer, B. E., Leib-Mösch, C., Francke, U., Erfle, V., and Hehlmann, R. (1989) Genomics 4, 68–75.
38. Quinn, J. P., Farina, R., Gardner, K., Krutzsch, H., and Levens, D. (1989) Mol. Cell. Biol. 9, 4713–4721.