BIOINFORMATICS

Vol. 14 no. 3 1998
Pages 244–251

Transcription regulatory region analysis using signal detection and fuzzy clustering

L.Pickert, I.Reuter, F.Klawonn[1] and E.Wingender

Gesellschaft für Biotechnologische Forschung mbH, Mascheroder Weg 1, D-38124 Braunschweig, Germany and [1]Fachbereich Elektrotechnik und Informatik, Fachhochschule Ostfriesland, Constantiaplatz 4, D-26723 Emden, Germany

Received on October 1, 1997; revised on December 1, 1997; accepted on December 9, 1997

Abstract

Motivation: *Presently available programs for the recognition of potential transcription factor binding sites in genomic sequences generally yield a huge amount of output. These output lists have to be filtered to obtain biologically significant elements, which is highly laborious work to be done manually.*

Results: *We developed a strategy for systematic verification and improvement of the underlying profiles, and for their contextual analysis by a fuzzy clustering approach using non-redundant libraries of search profiles as a prerequisite.*

Availability: *The tools mentioned in the paper are available upon request.*

Contact: *ewi@gbf.de*

Introduction

To our present knowledge, control of gene expression occurs largely, if not mainly, at the level of transcriptional regulation. This is achieved by a large functional class of proteins, the transcription factors, which bind with relaxed specificity to short genomic sequence elements comprising 5–25 bp (for a review, see McKnight and Yamamoto, 1992; Wingender, 1993). The recognition patterns of many transcription factors have been defined either as consensus sequence strings using the 15-letter IUPAC code, collections of which are available in the literature (Locker and Buzard, 1990; Faisst and Meyer, 1992), or in databases (Wingender *et al.*, 1996a, 1997). Although the use of these consensus strings is quite widespread and popular, it is much more appropriate to apply nucleotide distribution matrices (positional weight matrices, 'profiles') to describe and identify transcription factor binding sites (Bucher, 1990; Chen *et al.*, 1995; Quandt *et al.*, 1995). A large collection of such matrices is part of the TRANSFAC database (Wingender *et al.*, 1997) and is used by the program MatInspector (Quandt *et al.*, 1995). To provide the user with optimized threshold values and to improve these profiles, additional attempts are required to characterize the recognition patterns available in terms of the false-positive and false-negative matches they produce. Further improvement can conceivably be achieved by including additional characteristics such as structural features in the search routines (Wingender *et al.*, 1996b).

However, even optimized recognition patterns and procedures will still most likely yield lots of false-positive and false-negative matches. An apparently perfect signal will be without biological significance if placed in the wrong genomic sequence context, and even a highly aberrant element may gain function if placed optimally in relation to other elements. This is taken into consideration by algorithms like that applied by PromoterScan (Prestridge, 1995). However, this program does not yet consider the quality of the matches and the possibility of flexible assignment of matches to more than just one cluster of potential binding sites. For these reasons, we have developed a program that detects clusters of predicted transcription factor binding sites by two-dimensional fuzzy cluster analysis, taking the DNA position as the first dimension and the quality of the matches as the second dimension. However, this methodology requires non-redundant input lists of potential transcription factor binding sites, the prediction of which has to be as reliable as possible. As has been shown previously (Quandt *et al.*, 1995; Frech *et al.*, 1997a,b), a matrix approach is generally superior to sequence comparison approaches using, for example, IUPAC consensus strings. This paper will therefore describe an approach for the definition of individual thresholds for each profile contained in the matrix library used by the program MatInspector. Moreover, a general matrix comparison approach is described to establish a non-redundant library as a prerequisite for the fuzzy clustering analysis of potential transcription factor binding sites. Applying this method to a series of experimentally characterized regulatory regions reveals that it is suitable to filter potential promoters and enhancers out of the background noise of potential single sites.

Methods and algorithm

Matrix comparison

To compare two nucleotide distribution matrices, the weights within each column were normalized, the absolute difference of corresponding cells in corresponding columns was

calculated and summed up for each column. In order to obtain the smallest difference for both matrices, they were shifted along each other, starting and ending with an overlap of at least four positions. Thus, there are three different cases to handle: left overhang of the smaller matrix, total overlap, and right overhang of the smaller matrix. The empty cells in the overhanging regions were filled with 0.25 since they do not contribute to the specificity of that matrix. The distance value, *dist*, of two matrices *MAT1* and *MAT2* was then calculated according to equation (1):

$$dist = \frac{\sum_{i=1}^{w} \sum_{B} |f_{B,MAT1}(i) - f_{B,MAT2}(i)|}{w} \quad (1)$$

where i is the column index, w is the width of both matrices (i.e. the sum of the widths of both matrices minus their overlap) and f_B is the normalized frequency of nucleotide B in either matrix 1 or 2.

After completion of this comparison, the smaller matrix was reversed and inverted, and underwent the same procedure. The lowest value obtained from all these comparisons was taken as the distance value for the two matrices.

To assess the 'biological similarity' of the transcription factors the two matrices refer to, we took advantage of a comprehensive transcription factor classification scheme have developed earlier (Wingender, 1997; http://transfac.gbf.de/TRANSFAC/cl/cl.html). This TF classification is, in fact, mainly a classification of DNA-binding domains (DBD), defining a hierarchy of superclasses, classes, families, subfamilies (optional), 'genera' and 'species' of transcription factors. Since the DNA-binding specificity is part of the definition of a transcription factor family, the family assignment of two factors can be used now as an already defined qualitative similarity measure. For this, we assigned a value of '+1' to matrix pairs, which belong to the same transcription factor or to transcription factors belonging to the same family. Those of distantly related factors (belonging to distinct families, but to the same class) were marked by a value of 0.5. If the factor of one matrix is as yet ill characterized, a value of 0 was assigned, whereas all remaining matrix pairs referring to obviously unrelated factors received '–1'. These values were averaged over all matrix pairs falling into matrix distance intervals of 0.02 (see above).

Compilation of exon sequences

A standard negative test set consisting of genomic mammalian exon ≥2 sequences was extracted from the EMBL database Release 48 (datafiles: hum1.dat, hum2.dat, mam.dat, rod.dat).

For computational purposes, only entries with at least 35 bp upstream and 35 bp downstream of the exon sequences were selected. This has been done because these sequences were also to be used for analyses with the program ConsInspector (Frech *et al.*, 1993). This program requires a sequence context of normally 70 bp. However, for MatInspector runs, only the exon sequences themselves were used. No redundancy check was applied. This resulted in a set of 20 397 (exon 2 = 4052) sequences with 4 223 453 (exon 2 = 956 468) bp and a mean length of 207 (exon 2 = 236) bases (without the adjacent 35 bp).

The test results showed that the sample of exon 2 sequences alone is sufficient to obtain reliable results, whereas the use of all exons ≥2 provides similar results, but is more time consuming.

Compilation of positive test sets

The SITE table of the TRANSFAC database provides binding sites of defined transcription factors, including a 'quality' measure reflecting the strength of the experimental evidence given in the literature for each of these sites. Table 1 lists the quality criteria we applied.

Matrices which were generated on the basis of TRANSFAC SITE entries are classified according to the least quality of sites included. The TRANSFAC MATRIX identifier (ID line in the flat files) reflects this value by appending it instead of the continuous numbering applied for matrices which were taken from the literature, e.g. the matrix for the vertebrate transcription factor CREB generated from sites with a quality not less than two has the identifier V$CREB_Q2.

Table 1. Criteria for experimental evidence of transcription factor binding sites

Quality	Criteria
1	Factor binding as for quality 2 plus functional evidence
2	Binding of highly purified or recombinant factor
3	Factor binding altered by a specific antibody
4	Factor binding competition with a well-characterized binding sequence
5	*Bona fide* element
6	Not classified

Therefore, the sequences from which the 'q' matrices were generated are available in the database. Similarly, most genomic sequences underlying published matrices are already part of the TRANSFAC database. In contrast, sequences selected for binding to a given transcription factor by *in vitro* selection and amplification have not been included in the database, but have been published in most cases. They are used only in those cases where not enough genomic binding sites have been characterized yet.

Moreover, all matrices were controlled and, in some cases, had to be corrected for suboptimal alignment or simple counting errors. However, not all published matrices come with the underlying sequence set, in which cases the authors have to be contacted. In total, out of the 259 matrices under

consideration, 62 matrices were to be corrected and 24 are still missing the sequence data.

Transformation of MatInspector score into cluster data points

To facilitate visualization of potential transcription factor binding sites and their score, we transform the scores calculated by MatInspector (matrix similarity; see Quandt *et al.*, 1995) into an inverse quality function T_{inv} according to equation (2). As a result, high-scoring matches are close to the DNA position axis to which they are projected.

$$T_{inv}: [0,X_{max}] \times [Y_{min},Y_{max}] \rightarrow [0,U_{max}] \times [0,V_{max}] \qquad (1)$$

$$(x,y) \rightarrow \left(x, \frac{V_{max}}{(Y_{max} - Y_{min})} (Y_{max} - y) \right) \qquad (2)$$

where X_{max} is the sequence length in bp, Y_{max} is the estimated upper limit for quality of false positives, Y_{min} is the estimated lower limit for quality of true positives, V_{max} is the maximal range of quality in cluster space, U_{max} is the sequence length in cluster space, and $U_{max} = X_{max}$ and $Y_{min} < Y_{max}$.

The cluster algorithms are described elsewhere in greater detail (Pickert *et al.*, 1997).

Application of TFC to the SV40 genome

The complete simian virus 40 (SV40) genome (EMBL accession number V01380; 5243 bp) was analysed with MatInspector using default parameters. Although most of the known SV40 elements are also part of the training sets used to construct the relevant matrices, they represent only one out of 5–108 underlying sequences for each matrix. Thus, the search for individual elements is not significantly biased by the training set used.

All suggested hits with a matrix similarity value of <0.9 were removed from the output list (first filter). The data of the selected potential TF binding sites were transformed as described above, arbitrarily choosing a V_{max} value of 1000 since experience has shown that it is reasonable for fuzzy clustering to adopt a similar scale for both the *y*- and the *x*-axis. Subsequently, false hot spots are eliminated using a shrink filter which defines those matrices that recognize identical sites according to the matrix comparison (see above; second filter). The allowed maximal cluster size was 500 bp.

Results and discussion

General strategy

Search routines for potential transcription factor binding sites, such as PatSearch (Wingender *et al.*, 1996b) or MatInspector (Quandt *et al.*, 1995), normally produce a considerable output which is difficult to evaluate manually. To clear this kind of result list, we developed the following strategy:

1. to reduce the redundancy of potential matches due to very similar search patterns;
2. to improve the reliability of the individual search patterns;
3. to clear the output list of suggested binding sites which lack a significant context.

Comparison of transcription factor binding profiles

To reduce redundancy in search patterns, especially in the matrices from TRANSFAC, all matrices from the TRANSFAC MATRIX table were compared with each other. The distance values obtained ranged between 0.105 and 1.347.

Subsequently, we correlated all pairwise matrix similarities with the biological relationship of the transcription factor they refer to (see above). Although the resulting values intentionally reflect a qualitatively rather than quantitatively exact measure, they can be used to estimate the correlation between both the computed matrix distance and the similarity of the factors they refer to. This biological similarity assessment, on the other hand, is not sufficient for filtering redundant matrices since there are examples where distinct matrices are known to describe the interaction of one factor with several subpopulations of binding sites (e.g. Oct-1).

Counting the density of related or unrelated factors within certain distance value intervals revealed those ranges where matrices can be taken as closely related without further proof (range A, <0.22; Figure 1), where most of them are related (range B, between 0.22 and 0.48), where most of them are unrelated (range C, between 0.48 and 0.64) or the value above which they are unrelated anyway (not shown in Figure 1; >0.64). For the matrix pairs in ranges B and C, their relatedness must be proven individually (Figure 1). Among the 33 411 matrix pairs of the 259 individual matrices, only 25 (0.07%) fall in range A, 325 (0.97%) in range B and 1813 (5.4%) in range C. This means that the redundancy in the whole matrix library is very limited. As a result of these studies, any new matrix entered into the database can now easily be evaluated with regard to pre-existing profiles that may recognize the same pattern.

Strategy for verification of search patterns

In the TRANSFAC database on transcription factors and their binding sites, the MATRIX table provides a library of 259 nucleotide distribution matrices (Release 3.1) (Wingender *et al.*, 1997). They may have been derived from experimentally proven genomic binding sites, either pre-compiled in the literature or by ourselves, from random selection studies or from consensus descriptions generated with ConsIndex (Frech *et al.*, 1993). From this collection, 220 matrices were selected as a library for the MatInspector tool (Quandt *et al.*, 1995). The selection was made by defining a random

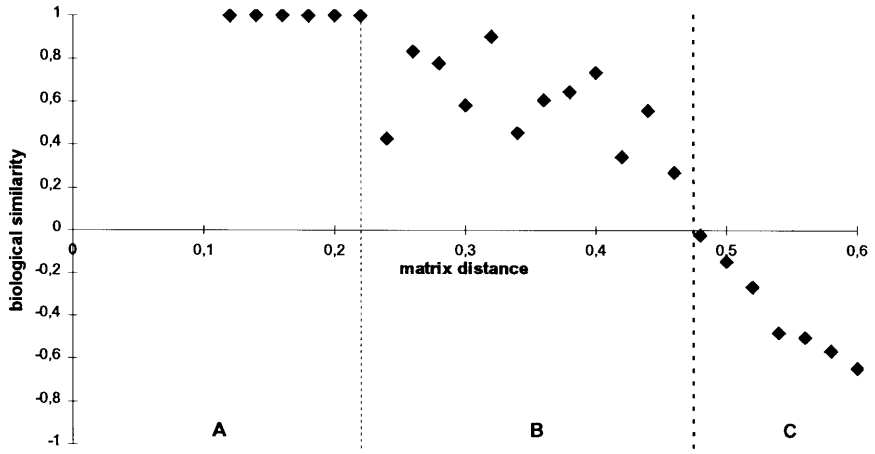

Fig. 1. Matrix comparison results. The expected biological similarity according to the transcription factor classification averaged over intervals of 0.02 of calculated matrix distance shows three zones that can be distinguished. (A) All matrices in the interval are closely related. (B) The matrices are mostly related. (C) The matrices in this interval are mostly not related.

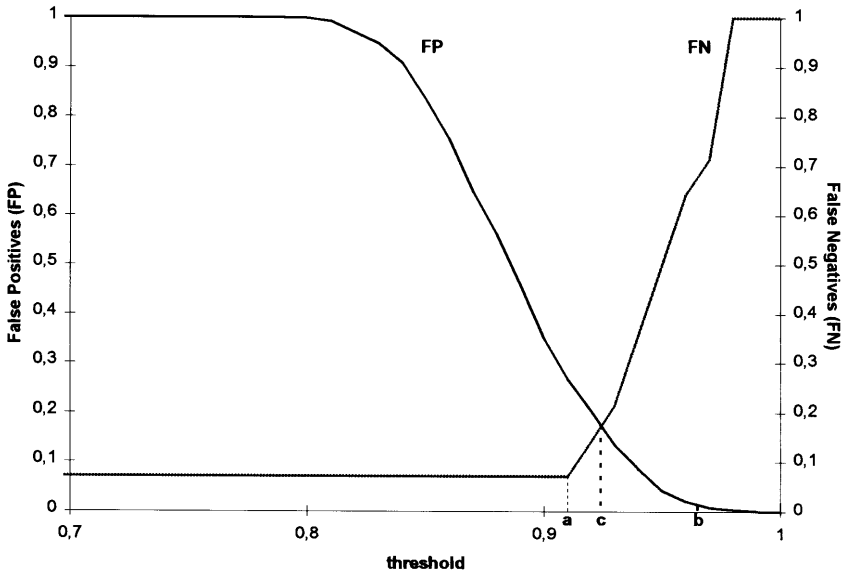

Fig. 2. False positives and false negatives of TRANSFAC matrix V$AP1_Q2 (the core similarity threshold was set to 0.85 in both cases). From these curves, we derive three thresholds recommended to the user: (a) to minimize false negatives; (b) to reduce false positives to 1%; (c) to minimize both error rates.

expectation value which reflects the number of expected matches per kilobase of a random sequence and setting a cut-off of 5.0 at standard default thresholds. To verify the use of these matrices as well as of the residual matrices which still could be of use under different parameters, it is necessary to determine the selectivity, sensitivity and positive prediction value for each of them. This means to investigate the number of false-positive (FP) and false-negative (FN) matches which each of these profiles produces, and thus to provide the basis for recommending optimal threshold values for each matrix.

Analysis of false positives

Generally, published reports about transcription regulatory sequence elements do not list false positives, i.e. sequences which match a given consensus but do not mediate the

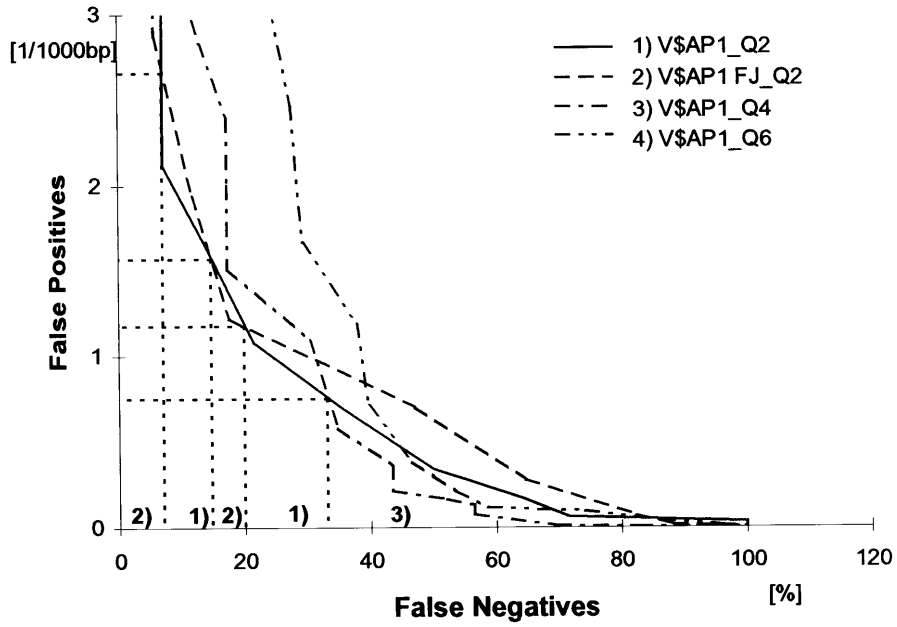

Fig. 3. False positives (FP) versus false negatives (FN) for different AP-1 matrices from TRANSFAC. The indicated matrices will be used in the outlined intervals as follows [Matrix: (FN | FP)]: 2 (0–7.1 | >2.66); 1 (>7.1–14.7 | >1.57–2.66); 2 (>14.7–20 | >1.175–1.57); 1 (>20–33.1 | >0.75–1.175); 3 (>33.1–100 | <0.75–0).

suspected function. Instead, negative test sets are frequently compiled from computer-generated random sequences, but this method does not take account of special features of real-life sequences (GC richness, etc.). Since regulatory regions have been found in nearly all parts of a genome, including intronic sequences, in far upstream or downstream regions and even within repetitive elements, we decided to extract all exon ≥2 sequences from the EMBL data library to assess the amount of FP produced by a given search pattern of a transcription regulatory signal. We omitted exon 1 sequences since it is known that they may also contain functional transcription factor binding sites. Although we cannot rule out that some of the matches found in exon ≥2 sequences would gain function when artificially placed in a promoter context, these sequences most likely do not exert any regulatory effect in the natural genomic context. The major problem of using exon sequences as a negative test set, however, may arise from some codon bias which could cause some over- or underrepresentation of certain nucleotide patterns. Since this will concern only a limited subset of patterns, if any, this issue will be neglected here, but will be dealt with in a separate systematic evaluation.

When we examined the number of FP using a series of nucleotide distribution matrices, we reproducibly found that restricting the analysis to exon 2 sequences gives nearly the same results as using the complete library. Therefore, most of the following investigations were carried out with the exon 2 library.

Combining FP and FN criteria

Combining both FP and FN values raises the question of how to normalize them to comparable scales. FN range between 0.0 (all true sites found) and 1.0 (no true positives identified). In contrast, FP values are normally referred to all potential match positions, i.e. to any position a sequence exhibits, and therefore are much smaller.

We therefore exploited the specific MatInspector algorithm which pre-selects for matches with a highly conserved core sequence and used this as the 100% rate. In this case, both FP and FN can be plotted using the same scale (Figure 2). Now, three thresholds can be recommended to the user: one which reduces the false positives to a pre-defined value (e.g. 1%), another one which minimizes false negatives (in the example shown, one AP-1 binding site out of 14 sequences in the positive test set does not contain the core sequence), and a third which represents a compromise between both requirements.

Because of the somewhat arbitrary choice of the core similarity applied for MatInspector runs, the characteristics shown in Figure 2 would also have to be recorded in dependence on this parameter for each profile. As an alternative, we plotted

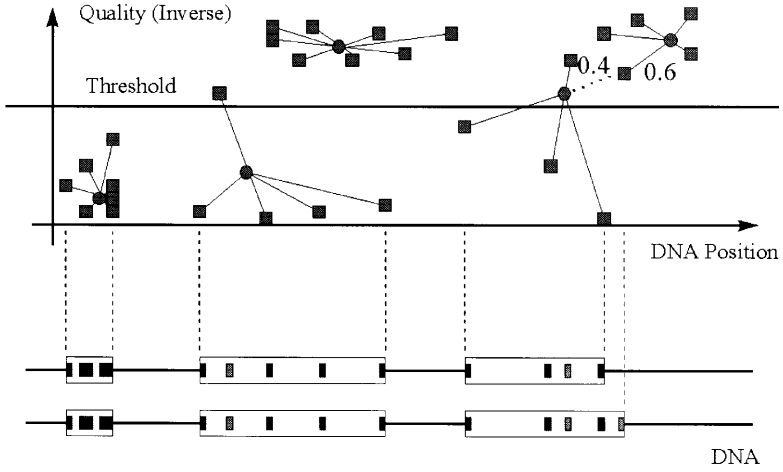

Fig. 4. Fuzzy clustering approach for (potential) transcription factor binding sites. An output list of, for example, a MatInspector sequence analysis run is transformed to data points of the coordinates' position and inverse quality (derived from the MatInspector scores). TFC introduces 'prototypes', points around which clusters are then assembled. The precise position of prototypes is optimized for minimal sum of distances, the number of prototypes is increased stepwise until all clusters fall within a user-defined maximal cluster width. Throughout the analysis, the degrees of belongingness of each data point to all clusters is calculated and stored (see example top right). Data points of those clusters which have at least one member below (i.e. better than) the threshold are projected onto the positional axis.

FP against FN for all related matrices (Figure 3). From these curves, the user may obtain the number of FP to be expected when entering the portion of FN he may be willing to tolerate, or vice versa. In those cases where more than one profile is available for a certain transcription factor, the program will automatically choose the most suitable pattern.

Identification of clusters of potential transcription factor binding sites

Functional transcription factor binding sites generally appear in clusters which represent, for example, promoters or enhancers. In contrast, isolated sequence elements normally are not biologically significant even when they match a consensus pattern perfectly. Therefore, searching for clustered potential transcription factor binding sites may help to filter insignificant items from the output lists of MatInspector or other sequence scanning routines.

For this purpose, we applied fuzzy cluster algorithms to a two-dimensional cluster analysis for the following reasons.

1. The two-dimensional analysis allows one to consider the scoring of the potential sites in addition to the position along the DNA; this is necessary since (a) the bulk of low-scoring matches can thus be eliminated semi-interactively, and (b) individual data points within a 'cloud' of suggested low-scoring matches may nevertheless be assigned to nearby clusters without being lost by setting an *a priori* threshold.
2. Fuzzy clustering calculates for each data point a defined degree of belongingness to every suggested cluster. This enables one to assign data points to different clusters flexibly, according to the parameters chosen. This relieves the necessity to artificially assign data points which are located just between two or more clusters to one of them. Moreover, when it has been decided not to consider a cluster for further analysis, individual members can nevertheless be 'saved' by reassigning them to another one. Also, a fuzzy approach is best suited to deal with vague data such as scored binding sites.

The scores of MatInspector output (or of another sequence scanning routine) are first loaded into a relational database. In a next step, two filters can be employed.

1. Setting a scoring threshold allows one to get rid of obvious false-positive matches; to avoid time-consuming new analysis runs when the proceeding analysis suggests repeating it with a different threshold, it is advisable to apply threshold filtering on this level of analysis rather than adjusting it for the search routine employed (e.g. MatInspector).
2. A shrink matrix allows one to filter matches produced by highly similar search profiles; to keep them as independent data points would mean to generate clusters (or hot spots) artificially. The basis of the shrink matrix is provided by the matrix comparison approach described above.

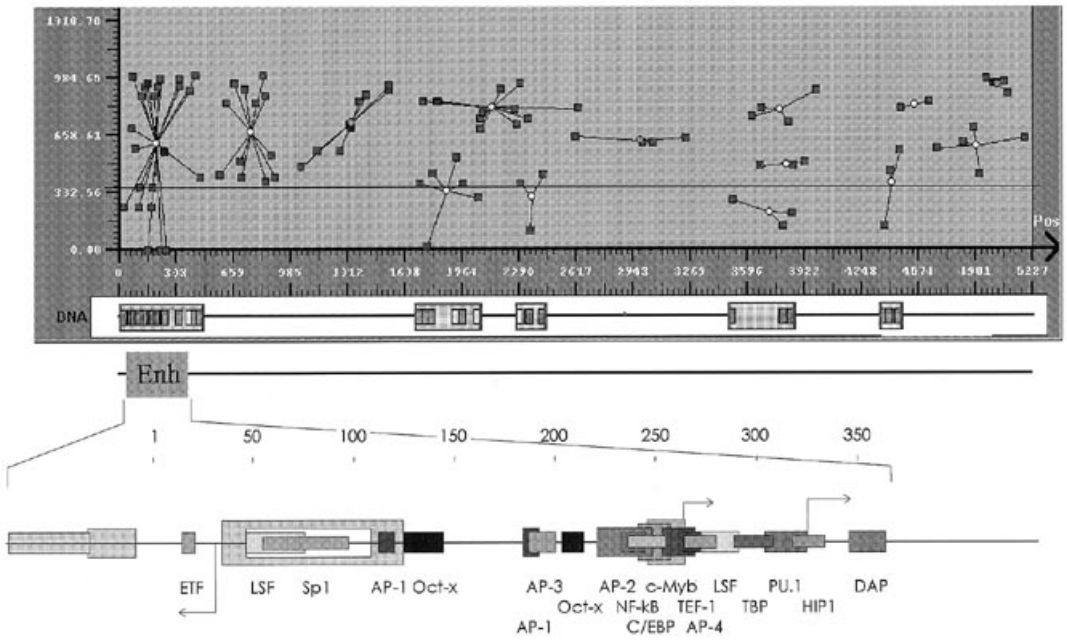

Fig. 5. Fuzzy cluster analysis of transcription factor binding sites in the SV40 genome. The suggested potential TF sites are suggested by MatInspector (Quandt *et al.*, 1995). The data set was pre-filtered for high-scoring matches (matrix similarity > 0.90) and subjected to clustering using the Gath and Geva algorithm. Note that the scores are transformed to an inverse scale in order to display high-scoring sites close to the DNA position axis. Clusters with at least one data point below the threshold indicated by the straight line have been accepted, data points belonging to these clusters are projected onto the DNA position axis. The lower part shows the experimentally verified pattern of TF binding sites of the SV40 enhancer as recorded in the TRANSFAC database, the arrows point to transcription start sites according to the feature table of EMBL/GenBank # J02400.

The scores are converted into a quality function. For the sake of better visualization, this quality is plotted inversely, thereby assembling the high-scoring matches near the DNA projection axis and the low-scoring ones in the upper part of the plot (Figure 4). It is important to scale the quality axis properly in relation to the positional axis since this defines the cluster space and thus significantly influences the partitioning of data points, and therefore the results of the analysis. It has been shown that the best results are obtained by setting the quality range 2–3 times higher than the expected size of transcription regulating regions (promoters, enhancers), normally around 300–500 bp (Pickert *et al.*, 1997).

We implemented two fuzzy clustering algorithms: fuzzy *c*-means (FCM) (Bezdek, 1981) and Gath and Geva (GG) (Gath and Geva, 1989). FCM is very fast and searches for spherical clusters of comparable size. GG extends the algorithm to different cluster shapes and sizes. The number of clusters is automatically increased stepwise until the size of all those clusters which match the adjusted quality criteria is within a pre-defined size, e.g. <500 bp. A cluster matches the quality criteria if at least one of its data points is better than a user-defined threshold. This rule considers the fact that highly degenerate elements become functional if placed in a proper context with other sites. While this principle is well known from many promoter examples, it is debatable whether a single high-scoring site is sufficient and represents just a first experimental approach. Each cluster is arranged around a prototype which, in a simplified view, can be considered as that point to which all data points of this cluster have minimal distance. In each step, new prototypes are introduced within those clusters which are larger than the pre-defined maximal size.

We applied the GG algorithm to analyse a MatInspector run for the whole SV40 genome. All hits have been passed through the two-filter procedures described in Methods and algorithm, and have been subjected to the fuzzy cluster analysis (Figure 5). In this case, we made use of the existing knowledge about the SV40 enhancer to adjust the threshold for rejecting most of the obviously irrelevant clusters. Among

the remaining clusters, the first one is located between positions 27 and 465, and comprises 25 suggested binding sites. Within the known enhancer region (approximately positions 1–350), 19 binding sites have been suggested, 13 of which have been proven experimentally. None of the known sites for which search profiles (matrices) are available have been omitted. However, 49 apparently false-positive matches (i.e. suggested binding sites for which no experimental evidence has hitherto been published) were removed from the MatInspector output list of this region.

We conclude that TFC clearly detected the SV40 enhancer region as the most dense cluster. Similar results have been obtained for the HBV enhancer I (data not shown). However, under the conditions applied, additional regions also revealed clustered potential transcription factor binding sites of much lower density. They do not appear to represent biologically meaningful regulatory regions. Lowering the cluster acceptance threshold drastically and changing the acceptance rule to more than just one high-scoring element per cluster, one could even sort out all other clusters except that representing the authentic enhancer. This clearly shows that more examples are required to adjust the rules under which TFC yields acceptable results, but that the tool itself has the potential to recognize regulatory genomic regions faithfully.

Acknowledgements

We would like to thank T.Werner and Y.Kondrakhin for helpful discussions, and K.H.Seifart for critically reading the manuscript. Part of this work was supported by EU grant BIO4-CT95-0226.

References

Bezdek,J.C. (1981) *Pattern Recognition with Fuzzy Objective Function Algorithm*. Plenum Press, New York.

Bucher,P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, **212**, 563–578.

Chen,Q.K., Hertz,G.Z. and Stormo,G.D. (1995) MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput. Applic. Biosci.*, **11**, 563–566.

Faisst,S. and Meyer,S. (1992) Compilation of vertebrate-encoded transcription factors. *Nucleic Acids Res.*, **20**, 3–26.

Frech,K., Herrmann,G. and Werner,T. (1993) Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. *Nucleic Acids Res.*, **21**, 1655–1664.

Frech,K., Quandt,K. and Werner,T. (1997a) Finding protein-binding sites in DNA sequences: the next generation? *Trends Biochem. Sci.*, **22**, 103–104.

Frech,K., Quandt,K. and Werner,T. (1997b) Software for the analysis of DNA sequence elements of transcription? *Comput. Applic. Biosci.* **13**, 89–97.

Gath,I. and Geva,A.B. (1989) Unsupervised optimal fuzzy clustering. *IEEE Trans. Pattern Analysis Mach. Intell.*, **11**, 773–781.

Locker,J. and Buzard,G. (1990) A dictionary of transcription control sequences. *DNA Seq.*, **1**, 3–11.

McKnight,S.L. and Yamamoto,K.R. (1992) *Transcriptional Regulation*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Pickert,L., Klawonn,F. and Wingender,E. (1997) Fuzzy cluster analysis for identification of gene regulating regions. In *Proceedings of the 7th IFSA World Congress*, Vol. IV, Prague, 56–61.

Prestridge,D.S. (1995) Predicting Pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.*, **249**, 923–932.

Quandt,K., Frech,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector—New fast and sensitive tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.

Wingender,E. (1993) *Gene Regulation in Eukaryotes*. VCH, Weinheim.

Wingender,E. (1997) Classification scheme of eukaryotic transcription factors. *Mol. Biol.*, (*Mosk).*, **31**, 483–497.

Wingender,E., Dietze,P., Karas,H. and Knüppel,R. (1996a) TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238–241.

Wingender,E., Karas,H. and Knüppel,R. (1996b) TRANSFAC database as a bridge between sequence data libraries and biological function. In Altman,R.B., Dunker,A.K., Hunter,L. and Klein,T.E. (eds), *Pacific Symposium on Biocomputing '97 (PSB'97)*. World Scientific, Singapore, pp. 477–485.

Wingender,E., Kel,A.E., Kel,O.V., Karas,H., Heinemeyer,T., Dietze,P., Knüppel,R., Romaschenko,A.G. and Kolchanov,N.A. (1997) TRANSFAC, TRRD and COMPEL: Towards a federated database system on transcriptional regulation. *Nucleic Acids Res.*, **25**, 265–268.