Regulatory DNA sequences: predictability of their function

Edgar Wingender, Thomas Heinemeyer, David Lincoln

Gesellschaft für Biotechnologische Forschung mbH,
Mascheroder Weg 1, D-3300 Braunschweig, Germany

Summary

Functional analysis of DNA sequences has to consider potential regulatory regions. Prediction of promoters and enhancers merely from sequence information has to be based on an amount of experimental data that is statistically reliable. For this purpose, a database for transcription regulating sequences and proteins has been established. Due to the high degree of degeneracy of transcription factor recognition sequences, an appropriate score system has been developed that weighs the importance of the individual bases within a potential element thus enhancing the probability of correct prediction. Furthermore, accumulation of potential sites may indicate promoter/enhancer regions.

Introduction

When large amounts of genomic sequences are analyzed it is of obvious interest to recognize open reading frames and to predict the possible function(s) of the proteins they appear to code for and, by systematic comparison, to increase our knowledge of the physics and chemistry of their structure and function. Enhanced biological understanding, however, will arise mainly from the knowledge of the regulatory mechanisms that a gene might be subject to. The expression pattern of proteins obviously has to be controlled to achieve and to maintain specific cellular activities, and this control operates at all levels between the (silent) information laid down in the DNA sequence and the function of a protein or the activity of an enzyme. At present however, it is thought that most regulatory mechanisms are exerted at the transcriptional

level. For efficient transcription, several general transcription factors are required (Fig. 1), one of them (TFIID) specifically interacts with the TATA box present in most but not all eukaryotic promoters (Fire et al., 1984).

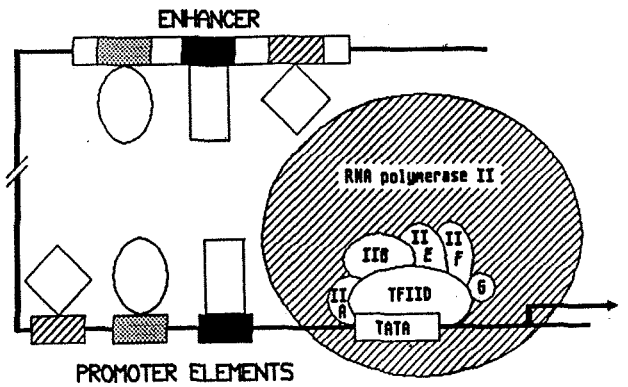

Fig. 1: A transcription initiation complex for eukaryotic RNA polymerase II consists of several general transcription factors (TFIID-TFIIG; see Sumimoto et al., 1990). Its assembly and activity to start RNA synthesis (arrow) is controlled by DNA binding factors that interact with specific short sequences in promoter and enhancer regions.

Additionally, several upstream promoter elements may interact with transcription enhancing proteins, and the same DNA-protein interactions may also occur within enhancer sequences far upstream or downstream of a coding region or even within introns (for review, see Wingender, 1988 & 1990). Some of these sequences and the factors binding to them are responsible for the stage-specific regulation of genes, whose products may also control the further development of an embryonic organism, as has been shown for the class of homeotic genes (Levine & Hoey, 1988). Other factors and their recognition sequences govern cell/tissue-specific gene expression. Among them are most of the proteins containing a so-called POU-domain such as the OCT2 factor(s) which is essential for the immunoglobulin gene expression. Finally, there are target sequences for DNA-binding proteins which are involved in the signal transduction cascades conferring the successful communication between cells and

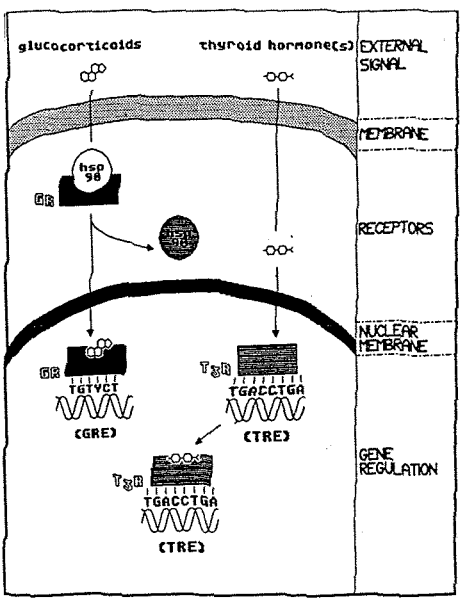

Fig. 2: Signal transduction by short hydrophobic molecules may be governed by their direct binding to a chromosomally located receptor as in the case of thyroid hormone. Glucocorticoids, however, bind to a cytoplasmic receptor form which is subsequently enabled to translocate into the nucleus and to bind to the appropriate target sequence.

organs. Among them, the relatively simple mechanisms by which small hydrophobic molecules such as steroid or thyroid hormones seem to act (Fig. 2) contrast with the complex and subtle pathways by which surface receptor-binding agents trigger expression of specific genes via adenylate cyclase, phospholipases, ion channels (Fig. 3) or tyrosine phosphorylation events.

It would therefore be of great value to be able to recognize eukaryotic promoters which per se do not display a "promoter consensus sequence", to identify the elements that constitute these promoters, to conclude and to verify the transcription factors that may bind to them and from this, to predict the involvement of the respective gene into the regulatory networks of the organism. Moreover, aberrant regulatory sequences may cause pathological disorders (Superti-Furga et al., 1988). Thus, their analysis is also of clinical relevance as has been shown in some examples.

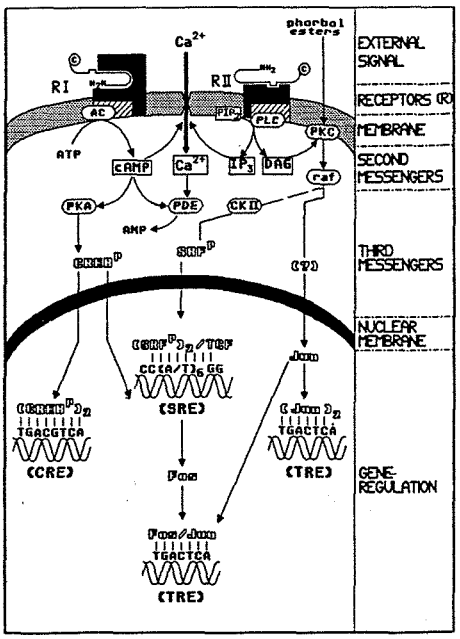

Fig. 3: Some signal transduction pathways and their chromosomal targets. An external ligand such as a peptide hormone binds to a receptor that may be coupled, e.g., either to adenylate cyclase (AC) or to phosphlipase C (PLC). cAMP as the intracellular second messenger may lead to the activation of a cAMP-response element binding protein, CREB, that binds to a distinct DNA sequence. Metabolites of PLC activity result in an activation cascade finally inducing transcription factors such as SRF (serum response factor), c-Fos and c-Jun. PKA: protein kinase A; PDE: phosphodiesterase; PKC, protein kinase C; CK II, casein kinase II.

Establishment of a database

To be able to develop appropriate prediction algorithms, however, one has to rely on a considerable amount of experimental information. This means, that the statistics must be based on a large number of known and characterized binding sites for each transcription factor to draw, e.g., reliable consensus sites. Once this material is available we would also be able to compare new protein binding sites that have been identified using one of the wide-spread footprinting techniques with the whole library of known sites which may suggest which factor might be involved in the particular interaction. Moreover, we could screen new sequences to see if they contain homologues to one of the known sites.

Such a data collection has been published several years ago (Wingender, 1988) and has been continuously updated in the meantime. This compilation essentially consists of five tables. The main listing gives the genes whose promoters have been investigated, the cellular source of the interacting proteins,

```
ID   HS$6-16_1          STANDARD;  DNA;  14 BP.
XX
AC   R00001;
XX
DT   29-JUL-1990   (DATA ENTRY)
XX
DE   6-16.
XX
OS   HUMAN (HOMO SAPIENS, MAN, HOMME, MENSCH).
OC   EUCARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMALIA;
OC   EUTHERIA; PRIMATES.
XX
RN   [1]  (aa)
RA   Dale T.C., Ali Imam A.M., Kerr I.M., Stark G.R.;
RT
RL   Proc. Natl. Acad. Sci. USA 86:1203-1207(1989).
RN   [2]  (a)
RA   Porter A.C.G., Chernajowski Y., Dale T.C., Gilbert C.S.,
RA   Stark G.R., Kerr I.M.;
RT
RL   EMBO J. 7:85-92(1988).
RN   [3]  (ba)
RA   Dale T.C., Rosen J.M., Guille M.J., Lewin A.R., Porter A.G.C.,
RA   Kerr I.M., Stark G.R.;
RT
RL   EMBO J. 8:831-839(1989).
XX
CC   EWI.
CC   Data edited (30-JUL-1990) by Thomas Heinemeyer.
XX
DR   CELLINE       human/HeLA+IFN; Bristol 88+; HFF+.
DR   METHODS       3, 4a, 4b, 4d, 1f.
XX
KW   Protein interacting regions.
XX
FH   Key        From     To       Binding factor
FH
FT   TRANSFAC   -127     -89       E factor.
SQ   SEQUENCE     14 BP; 7 A; 1 C; 4 G; 2 T;
     gGGAAAaTGA AACT
```

Fig. 4: The first entry of the transcription factor database. Shown is one protein recognition sites of the human gene 6-16, the three references that have described it, the cell lines in which it has been studied, the methods (in a coded form), the position upstream of the transcription start site, the binding factor and the sequence interacting with this protein.

the position of the binding, a code for the method that has been used for identification of this site, the sequence or sequence motif that has been proposed by the authors to mediate the protein binding, the designation of the binding factor and the references. The cells as well as the methods are decoded in two separate lists. In another table, the transcription factors are listed together with their source, the genes they have been shown to recognize, their molecular weights and some of their structural features. In this column, it is shown whether a protein is of zinc finger type and, if so, how many fingers of CCHH- or C_4-type it has, or whether it is a helix-loop-helix protein, whether it possesses a leucine zipper, which could make it prone to dimerization, or whether it has a glutamine-,

proline- or serine-threonine-rich domain that all have been shown to occur in transcription-activating domains of transcription factors. Finally, it is shown whether there are known synonyms for this protein or whether there are related proteins in other species.

We are now transferring these data into a computer-readable format to establish a real relational database. One entry of the first table is shown in Fig. 4; it gives the location of the "E factor"-binding to the promoter of the interon-β-induced gene 6-16.

The updated version presently contains approximately 1500 protein binding sites in the regulatory regions of 340 genes. These are covered by approximately 480 transcription factors. However, the latter figure is somewhat problematic since it includes some known and certainly also some undiscovered synonyms, but on the other hand, many transcription factors have recently been shown to consist of whole families of related proteins. There are particularly few examples of factors from the plant kingdom. The information content of the database has been collected from some 890 references.

Problems in the prediction of recognition sites

In this section, some examples of transcription factor recognition sites will be described, and the specific problems we are faced with when we try to develop methods to predict these sites simply from sequence information will be discussed.

1st problem: there are more recognition motifs than are actually used.

When we screen the EMBL data base for several transcription factor consensus sites, e.g. for the factors AP-2, AP-3, AP-4, CREB/ATF, NF-1, OTF or Sp1, a large number of hits are found (Tab. 1). The recognition motifs for these eukaryotic transcription factors are not only present in eukaryotic DNA such as the human genome (labelled "HS"), but also in a

Tab. 1: Transcription factor consensus sites found in the EMBL database

factor	recognition sequence	total hits	EC	HS	HS/EC
AP-1	TGASTCA				
AP-2	*CCSC(AG)GGC* GCC(CT)GSGG	9507	259	2001	7.7
AP-3	TGTGGWWW WWWCCACA	11833	511	2597	5.1
AP-4	(CT)CAGCTG(CT)GG *CC(AG)CAGCTG(AG)*	500	30	158	5.3
CREB	TGACGTCA	524	37	63	1.7
NF-1	(CT)GG(AC)N_{5-6}(GT)CCA	10561	545	2151	3.9
OTF (H2B)	ATGCAAAT ATTTGCAT	2260	70	354	5.1
(SV40)	ATGCAAAG CTTTGCAT	1607	32	349	10.9
Sp1	GGGCGG CCGCCC	20176	926	4476	4.8

HS/EC (total number of sequences): 4.7

prokaryotic genome like that of Escherichia coli ("EC" hits). These and a number of human (eukaryotic) hits are likely to be "false positives". In most cases, the ratio of human to E.coli hits resembles that of the total number of sequences of both organisms (4.7) deposited in the EMBL library. One exception is the CREB/ATF site which is even over represented in E.coli, presumably due to its similarity with the prokaryotic cAMP-response element (Lin & Green, 1989).

We have looked at the Sp1 site in greater detail. This transcription factor belongs to the class of the DNA-binding zinc finger proteins and contains serine/threonine and glutamine-rich stretches as transcription-activating domains (Kadonaga et al., 1987). Its binding to a number of promoters has been demonstrated, among them are viral and cellular house-keeping as well as subtly regulated genes.

How can the "specificity" of identification of Sp1 sites be enhanced? We tried to combine the Sp1 consensus with a TATA box sequence (Tab. 2). The range of possible distances between both elements has to be relatively large, since the Sp1 sites known so far are up to 500 base pairs from the transcription start site. Applying this more restricted search strategy, we obtain a considerable lower number of hits. However, this did not improve the specificity. This is due to the degenerate, short TATA sequence (TATAWW) we have used in combination with the Sp1 element over a large distance range allowed. Moreover, since TATA boxes may be even more degenerate than the sequence we have used or even may be absent, we will loose some of the actual Sp1 sites ("false negatives").

Tab. 2: Combinations of Sp1 elements

recognition sequence	total hits	EC	HS	HS/EC
GGGCGG	9821	551	2180	4.0
CCGCCC	10355	375	2296	6.1
Σ:	20176	926	4476	4.8
$GGGCGGN_{20-500}TATAWW$	1344	70	215	3.1
$CCGCCCN_{20-500}TATAWW$		66	256	3.9
Σ:		136	471	3.5
$GGGCGGN_{1-100}GGGCGG$	1351	25	400	16.0
$GGGCGGN_{1-100}CCGCCC$	746	14	183	13.1
$CCGCCCN_{1-100}GGGCGG$	781	10	179	17.9
$CCGCCCN_{1-100}CCGCCC$	1386	18	409	22.7
Sp1-Sp1-TATA	105	2	27	13.5
Sp1 (degen.)	2445	196	526	2.7
$Sp1(deg)N_{10-100}Sp1$	99	2	36	18.0

Sp1 (degenerated): (GT)(GT)GGCG(GT)(GA)(GA)(CT)

However, a property of Sp1 sites that could be advantageously applied is that they are frequently clustered. By looking for such clustered Sp1 motifs in all possible orientations, a strong improvement in the specificity is achieved (Tab. 2).

Thus, clustering of elements may be a valuable criterion for reliable identification of regulatory regions.

Similarly, a huge number of hits with low specificity for eukaryotes has been found when the hexanucleotide TGTYCT or the inverse element was applied as search pattern (Tab. 3). This sequence is the second half-site of the glucocorticoid regulatory element (GRE) which is sufficient for binding of the glucocorticoid receptor. Here also the ratio of human to E.coli sequences could be greatly increased by either combination of GRE second half-sites. Not too surprisingly, the full GRE, though highly degenerate in its first half, lead to only one single hit in the known E. coli sequences (Tab. 3).

Tab. 3: Glucocorticoid regulatory elements

sequence	total hits	EC	HS	HS/EC
TGTYCT	23513	653	5277	8.1
AGRACA	31505	747	6529	8.7
$TGTYCTN_{10-200}TGTYCT$	2503	41	620	15.1
$TGTYCTN_{10-200}AGRACA$	2800	33	745	22.6
$AGRACAN_{10-200}AGRACA$	4560	46	952	20.7
$(TGA)GT(AT)CAN_{2-3}TGTYCT$	68	1	25	25

2nd problem: Most factors recognize degenerate sites

Unlike restriction enzymes, transcription factors do not bind to an unambiguously defined DNA sequence. E.g., the Sp1 consensus site has been redefined with increasing experimental material. Thus, the original GGGCGG has been extended and weakened to (G/T)GGGCGGRRY (Jones et al., 1986) and then to (G/T)(G/T)GGCG(G/T)(G/A)(G/A)(C/T) (Briggs et al., 1986). Moreover, this as well as other elements can occur in either orientation which means that Sp1 can bind to at least 64

different sequences. This raises the question, what features really define actual Sp1 sites and help to differentiate them from unfunctional sites?

One possible solution to this problem is simply to extend the sequence under consideration and to analyze nucleotide preferences in the environment of the known sites. In a first approach, we aligned 41 known Sp1 sites which were extended by 10 more nucleotides on either side and determined the frequency of each nucleotide in every position (Tab. 4). These frequencies were now used as "scores" to evaluate any particular sequence for its Sp1 binding character simply by calculating the sum of frequencies of those bases found in it. Multiplication of the frequencies with their binary logarithm would transform these values to a parameter indicating the information content of these sequences in bits (Schneider et al., 1986). Comparison of the sequence information values would lead to identical results.

On average, a score of 12.90 ± 0.15 can be assigned to the known Sp1 sites. If the scores for 100 arbitarily selected potential human Sp1 sequences were calculated, a significantly lower value of 11.65 was obtained. Even more reduced is the score for 100 E.coli sites with Sp1 homology, for which a value of 10.99 was determined. Thus, these scores might serve as a first criterion to identify a potential factor binding site.

Tab. 4: Base frequencies around Sp1 sites

A:	5	4	6	7	5	8	5	8	11	2	3	0	0	0	2	0	9	7	2	3	6	6	7	4	4	7
C:	9	14	19	10	14	9	9	7	11	3	0	0	0	40	0	0	2	1	28	14	8	7	6	5	16	8
G:	21	17	11	18	17	21	19	14	11	26	37	41	41	1	39	40	27	31	4	10	12	19	21	28	18	18
T:	6	6	5	6	5	3	8	12	8	10	1	0	0	0	0	1	3	2	7	14	15	9	7	4	3	8
	g	g	c	g	g	g	g	g	a c g	g	G	G	G	C	G	G	g	g	c	c t	t	g	g	g	g	g

In summary, the development of appropriate score systems may be valuable in the optimal alignment as well as in the identification of potential recognition sites.

3rd Problem: Some sequences interact with more than one factor

If we consider the possibility of predicting from sequence information the regulation of a gene, the situation is complicated by the fact that some sequences interact with different transcription factors. Among them are the CCAAT-boxes, which interact with CP1 and CP2 (or HAP proteins in yeast), with C/EBP, CTF/NF-1 or CDP (Wingender, 1990). At least the recognition motifs of C/EBP and CTF/NF-1 have now been defined more clearly (Ryden & Beemon, 1989; Meisternst et al., 1988, Dean et al., 1989, Cordingley & Hager, 1988):

C/EBP: T(TG)NN G(TC)AA(TG)

CTF/NF-1: (CT)GGAN$_{5-6}$(T/G)C C AA T

But in order to discriminate between the binding sites of these factors unambigously, more experimental material is required. In other cases, a certain element interacts with functionally different but related factors that govern cell-specific transcriptional enhancement or repression, respectively. This is the case for the octamer-binding factors (OTF or OCT). Here, the evaluation of the corresponding DNA sequence information would have to be based on the detailed knowledge of the cell-specific factor expression pattern.

4th Problem: One sequence element interacts with several factors of different function

This problem extends the complications in functional interpretation of DNA sequences described before. E. g., the high-affinity binding site for AP-2, CCGCCCGCC, also contains a Sp1 site (bold face). The factors CREB/ATF and AP-1 recognize related motifs: TGACGTCA and TGA(CG)TCA, respectively. It has been shown, that both factors (in variable heterodimeric

compositions) may mutually interact with their binding sites (Ivashkiv et al., 1990). Similar to the AP-1 element is the estrogen-regulatory element (ERE) as it appears, e.g., in the promoter of the Xenopus laevis vitellogenin gene: TGACTG.

It has been demonstrated that such similarities may not be merely fortuitous but may have considerable physiological significance. Thus, the CREB binding to an AP-1 site may result in its repression (Lamph et al., 1990). The glucocorticoid receptor competes with CREB for binding to the promoter of the α-subunit of the human glycoprotein hormone (Akerblom et al., 1988; see Tab. 5). AP-1 (the Fos/Jun complex) binds to the (positive) vitamin D_3 regulatory element of the osteocalcin gene and has been suggested to be a negative regulator (Schüle et al., 1990). AP-2 and NF-1 bind to a single sequence of the growth hormone promoter in a mutually exclusive manner, but both stimulate the transcription (Courtois et al., 1990).

Tab. 5: Overlapping recognition sites

promoter:	sequence:	factors:
glycoprotein hormone α-su.	AGATCAAATTGACGTCA	CREB, GR
osteocalcin	GGTGACTCACCGGGTGAACGGG	AP-1, VDR
growth hormone	TGGCCTGCGGCCAG	NF-1, AP-2

Thus, assignment of potential factor binding sites may be ambiguous, either leading to "false positives" or to physiologically relevant ambivalence. Discrimination between both possibilities will be one of the challenges of the prediction programs to be developed.

References

Akerblom, I., Slater, E. P., Beato, M., Baxter, J. D., and Mellon, P. L. (1988) Science 241, 350-353

Briggs, M. R., Kadonaga, J. T., Bell, S. P., and Tjian, R. (1986) Science 234, 47-52

Cordingley, M. G., and Hager, G. L. (1988) Nucleic Acids Res. 16, 609-628

Courtois, S. J., Lafontaine, D. A., Lemaigre, F. P., Durviaux, S. M., and Rousseau, G. G. (1990) Nucleic Acids Res. 18, 57-64

Dean, D. C., Blakeley, M. S., Newby, R. F., Ghazal, P., Hennighausen, L., and Bourgeois, S. (1989) Mol. Cell. Biol. 9, 1498-1506

Fire, A., Samuels, M., and Sharp, P. A. (1984) J. Biol. Chem. 259, 2509

Ivashkiv, L. B., Liou, H. -C., Kara, C. J., Lamph, W. W., Verma, I. M., and Glimcher, L. H. (1990) Mol. Cell. Biol. 10, 1609-1621

Jones, K. A., Kadonaga, J. T., Luciw, P. A., and Tjian, R. (1986) Science 232, 755-759

Kadonaga, J. T., Carner, K. R., Masiarz, F. R., and Tjian, R. (1987) Cell 51, 1079-1090

Lamph, W. W., Dwarki, V. J., Ofir, R., Montminy, M., and Verma, I. M. (1990) Proc. Natl. Acad. Sci. USA 87, 4320-4324

Levine, M., and Hoey, T. (1988) Cell 55, 537

Lin, Y. -S., and Green, M. R. (1989) Nature 340, 656-659

Meisterernst, M., Gander, I., Rogge, L., and Winnacker, E. -L. (1988) Nucleic Acids Res. 16, 4419-4435

Ryden, T. A., and Beemon, K. (1989) Mol. Cell. Biol. 9, 1155-1164

Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986) J. Mol. Biol. 188, 415-431

Schüle, R., Umesono, K., Mangelsdorf, D. J., Bolado, J., Pike, J. W., and Evans, R. M. (1990) Cell 61, 497-504

Sumimoto, H., Ohkuma, Y., Yamamoto, T., Horikoshi, M., and Roeder, R. G. (1990) Proc. Natl. Acad. Sci. USA 87, 9158-9162

Superti-Furga, G., Barberis, A., Schaffner, G., and Busslinger, M. (1988) EMBO J. 11, 3099-3107

Wingender, E. (1988) Nucleic Acids Res. 16, 1879-1902
Wingender, E. (1990) CRC Crit. Rev. in Eukaryotic Gene Expression 1, 11-48