

A New Method to Develop Highly Specific Models for Regulatory DNA Regions

Kornelie Frech, Kerstin Quandt and Thomas Werner

GSF - National Research Center for Environment and Health, Institute of Mammalian Genetics, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany

Abstract. We present a new modular concept to construct organizational models for transcriptional regulatory DNA units. The method requires a training set of at least 10 sequences and a simple initial model (e.g. two characteristic transcription factor binding sites). The final model is generated by computer analysis directly from the sequences. 20 Lentivirus long terminal repeats (LTRs) and an initial model consisting of only two elements (TATA box and polyA signal) resulted in a final model of 10 elements which recognized all of the more than 100 available Lentivirus LTRs while rejecting all other known LTR types. Database searches with this Lentivirus LTR model demonstrated the very high specificity of our method.

1 Introduction

The timely appearance of proteins at correct locations is crucial for the functionality of an organism. This is mainly controlled at the level of transcriptional regulation which involves the interaction of proteins with their cognate DNA binding sites. These DNA elements are organized into regulatory units for individual genes to create promoters, enhancers or silencers (including scaffold attachment regions, SARs and locus control regions, LCRs). The specificity of a regulatory unit is usually determined by the spatial organization of binding sites for common and cell- or tissue-specific transcription factors (TF-sites). These sites are separated by non conserved “spacer”-DNA precluding global alignment strategies for identification of regulatory units.

We already developed methods for the identification of individual TF-sites (Frech *et al.*, 1993, Quandt *et al.*, 1995b, Wolfertstetter *et al.*, 1996), allowing identification of individual elements of regulatory units. Recently, we have shown that the quality scores assigned by one of these methods (ConsInspector, Frech *et al.*, 1993) correlate to some extent with biological functionality (Quandt *et al.*, 1995a). However, transcriptional regulation is not an inherent property of individual binding sites but is determined by the context of binding sites. We also developed a method to assess the context by distance correlation of different TF-sites which allowed us to identify basic organizational features of yeast promoters (Quandt *et al.*, 1996a, Quandt *et al.*, 1996b).

There are other methods which utilize the uneven distribution of TF-sites in promoter and non-promoter sequences for recognition of polymerase II promoters

(PROMOTER SCAN (Prestridge, 1995) and FunsiteP (Kondrakhin *et al.*, 1995)). Model-building approaches like GeneLang (Dong and Searls, 1994), GeneParser (Snyder and Stormo, 1995), or GRAIL (Uberbacher and Mural, 1991) focus mainly on gene prediction. Only GRAIL includes few basic features of the promoter regions in the last release.

Here, we present a novel method (ModelGenerator) based on the concept of a strictly modular design of regulatory units which generates a model by combining several types of individual elements with the overall spatial organization of the regulatory unit. Another program designated ModelInspector is able to scan sequences of unlimited length for matches to the generated models. ModelInspector showed an extraordinary high specificity for the models tested.

2 ModelGenerator Algorithm

2.1 Pre-analysis Steps

ModelGenerator requires a set of at least 10 sequences containing a single type of regulatory unit and a very simple initial model (e.g. two characteristic elements and their sequential order). Further potential elements can be identified by analysis of individual sequences by the program CoreSearch (Wolfertstetter *et al.*, 1996). Elements found consistently in the majority of sequences by comparative analyses of the whole sequence set can then be selected for the model.

2.2 Generation and Recognition of Models

ModelGenerator defines two types of individual elements. *Determining elements* are part of the basic organizational framework of the regulatory unit (e.g. TATA box and/or initiator for TATA box containing promoters). They have to be present in almost all of the sequences and can be detected with a tolerable number of false positive matches. *Non-determining elements* are not *per se* clearly associated with the regulatory unit (i.e. occur frequently throughout the sequence like low energy hairpins) and are distinguished solely by association with other elements.

First, all potential determining elements are located either by a consensus search (ConsInspector, Frech *et al.*, 1993 or MatInspector, Quandt *et al.*, 1995b) or by other search routines included in ModelGenerator (e.g. secondary structure elements, direct repeats). Potential matches are rated according to their score assigned by the search program. For each sequence of the training set, ModelGenerator then identifies all possible combinations of determining elements which reach the user-defined thresholds and occur in the same sequential order (D-sets). At this point, more than one potential D-set is possible in a single sequence. The *basic set* is build of all sequences with a single D-set (at least 6 sequences are required). Distance histograms for adjacent elements are then generated from the basic set. A single D-set within sequences with more than one potential D-set can then be selected by the maximum

element score and the maximum distance score based on the distance histograms. These sequences are added to the basic set. The sequence between two determining elements is designated a *region*. The determining elements and the regions define the basic framework of the regulatory unit model.

This framework also provides the basis to identify additional elements, for example poorly conserved TF-sites, low energy secondary structure elements, or elements present only in a subset of the sequences. Analysis of individual sequences for these non-determining elements is restricted to the regions as defined above. Elements within one region are selected if present in at least a significant subset of all sequences. Thus, all non-determining elements compatible with the basic framework are defined.

Finally, a complete set of distance histograms is calculated for all elements. The complex regulatory unit model consists of descriptions of all individual elements, their mean scores, their relative frequencies, their sequential order, and the distance distributions observed in the training set.

Another program designated ModelInspector utilizes these models to scan sequences for new regulatory units matching the model. First all matches for individual determining elements which score above the thresholds defined by the model are located in the new sequence. These individual elements are combined to match the organization of the D-set characteristic of the model. If the sum of scores of determining elements exceeds a given threshold non-determining elements are identified and the sum of scores of all elements is calculated. This total element score and the score for element distances are then used to evaluate the fit of complex elements to the model. An extensive description of the algorithm and the programs ModelGenerator and ModelInspector will be presented elsewhere (Frech *et al.*, in preparation).

3 Results

3.1 Lentivirus LTR Model

We applied our methods to define a model for Lentivirus LTRs (the currently most prominent Lentiviruses are the human immunodeficiency viruses). More than 100 individual Lentivirus LTRs are present in the Human Retroviruses and AIDS database (Myers *et al.*, 1994) with a wide range of global sequence similarities (from 37% to almost 100%, determined by the GCG program Gap).

We used a training set of 20 different Lentivirus LTRs ranging from 350 bp to 855 bp in length. To analyze the TATA boxes present in the training set, we established a consensus profile with the program ConsInd (Frech *et al.*, 1993) and compared this profile with the TATA box profile of C-type LTRs. None of the C-type TATA boxes reached the cutoff score of the Lentivirus TATA box consensus (1.81) while all were above the cutoff score of the C-type TATA box consensus (1.48). Only three Lentivirus TATA boxes scored above C-type cutoff and one Lentivirus TATA box did not reach the Lentivirus cutoff score (Table 1). This

indicated that the derived TATA box consensus profile is very specific but not sufficient to identify the LTRs.

Our initial LTR model consisting only of the TATA box and the PolyA signal allows definition of three initial regions (upstream of TATA box, between TATA box and polyA signal, and downstream of polyA signal). Analysis of these regions with ConsInspector (Frech *et al.*, 1993), MatInspector (Quandt *et al.*, 1995b), CoreSearch (Wolfertstetter *et al.*, 1996), and ModelGenerator revealed eight additional elements which are part of the organizational framework of Lentivirus LTRs. We were able to define six elements as common to all Lentivirus LTRs (including the TATA box and the polyA signal) while four elements were found to be specific for primate Lentivirus LTRs (Fig. 1). A more detailed description of the Lentivirus LTR model is given in Frech *et al.* (1996).

3.2 Specificity of Lentivirus LTR Model

ModelInspector correctly identified all available Lentivirus LTRs (more than 100) from the Human Retroviruses and AIDS database. Thus, the model is sufficiently complete to identify Lentivirus LTRs though it may not contain all elements important for these LTRs.

Another crucial point is whether this model is also specific for Lentivirus LTRs. We analyzed 36 LTRs from C-type, B-type, D-type, Spuma, and HTLV-BLV retroviruses with ModelInspector. ModelInspector rejected all 36 LTRs of other type although the overall sequence similarity of MoMuLV LTR (C-type) to Lentivirus LTRs is within 1% of that among distant Lentivirus LTRs.

3.3 Database Search

We scanned the primate section of GenBank Release 92.0 (47,659,902 nucleotides in 53,102 sequences) with our Lentivirus LTR model. The search thresholds were set below the weakest Lentivirus LTR match of the training sequences. With these parameter settings ModelInspector detected 100 new candidate LTRs on both strands of the primate sequences of GenBank. Even if all of these new matches would be regarded as false positive matches this would be a false positive rate of about one per million nucleotides (exact: 1 per 953,200 nucleotides). These results demonstrate the discriminate power of our approach and some of these matches may represent new hitherto unknown Lentivirus LTRs.

Table 1. TATA box ratings (in C_i - scores)

	TATA position ^a	C-type TATA	lentiviral TATA
C-type LTRs			
akvltr	331	4.48	-1.69
baevltr	390	5.36	0.30
galv	341	1.61	-0.50
rmcfltr	342	5.05	-0.81
ssv	331	4.27	-1.93
erv3ltr	462	2.83	1.38
agmerltr2	383	4.08	0.20
re3ltr	469	2.75	1.33
hsrtr1	425	3.27	0.28
amulvltr	342	4.42	0.15
fcvenvc	318	3.11	-0.24
fcvgmb	312	2.67	-0.42
hsersp1b	527	3.41	-0.38
homltr	352	4.08	-0.44
fcvfltr	532	2.85	-0.46
catrd114a	371	4.07	1.27
hsersp3	383	3.34	0.37
musergl1	243	3.86	0.49
Lentivirus LTRs			
sivltr	466	-0.77	4.30
hivlai	425	1.21	5.87
hiv2ben	525	-1.96	5.12
resivsmm	485	-1.66	4.61
sivcps	198	1.79	4.62
sivsyk	428	1.12	2.71
sivmne	488	-1.91	5.63
sivcpz	445	1.20	5.99
hivmvp	432	0.99	6.34
sivagm90	475	-1.09	4.18
sivsabl	450	-1.09	5.35
sivmm251	476	-1.71	5.04
vlvcg	233	-1.17	4.40
bimpevpp	358	-0.93	0.51
olvvcg	243	1.97	3.92
ceavltr	304	-1.66	4.02
eia5ltr	183	0.35	4.46
fivltr	208	1.56	3.77
olvtransac	224	0.75	3.69
pumaltr	162	2.18	5.09

^awith respect to LTR start position

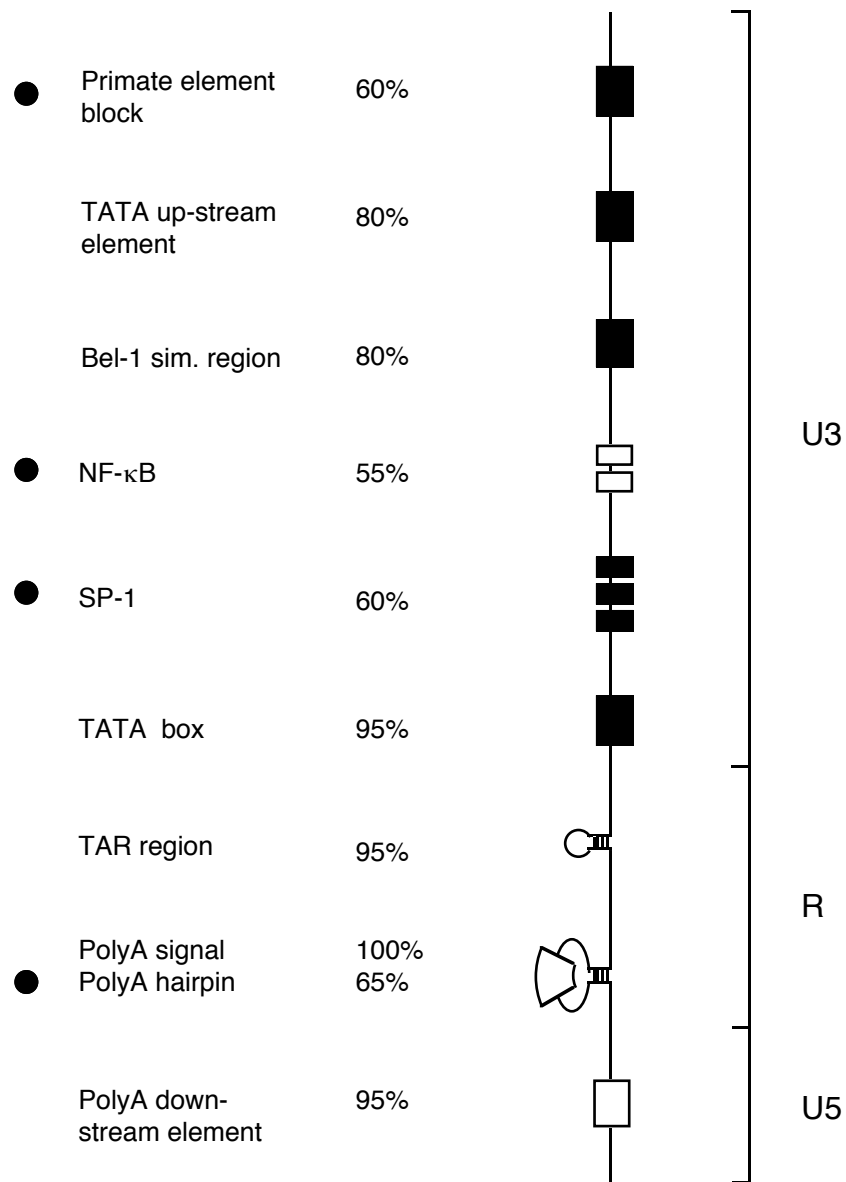


Fig. 1. Graphical representation of Lentivirus LTR model.

Consensus elements identified by ConsInspector are shown as black boxes, elements found by MatInspector are shown as white boxes, and hairpins are indicated by a simple hairpin symbol regardless of their true structure. Also given is the frequency of the respective elements (in % of training sequences containing the element). Primate specific elements are marked by a black dot.

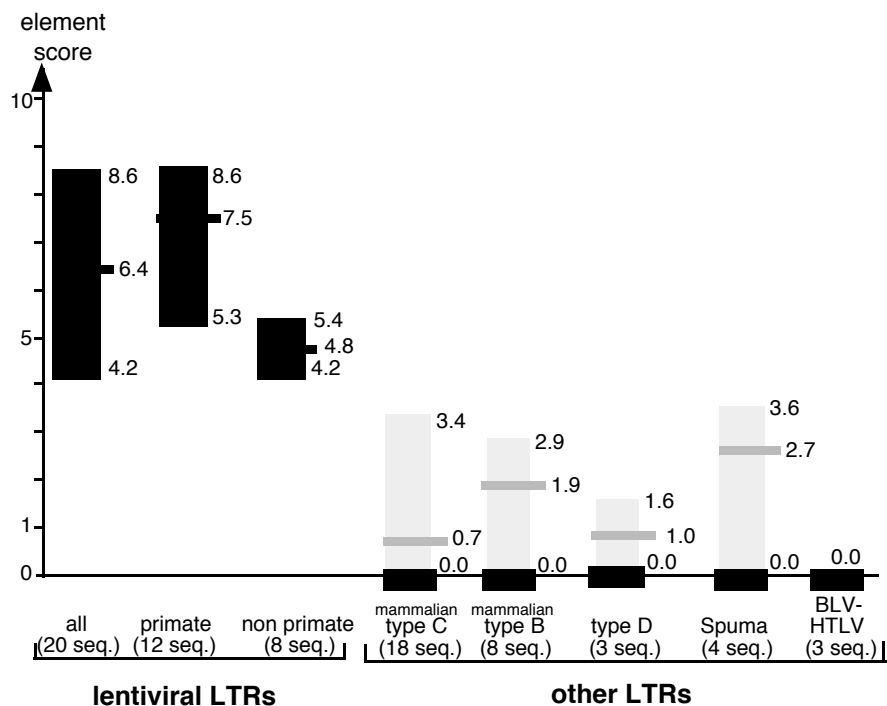


Fig. 2. Specificity of Lentivirus LTR model (from Frech *et al.*, 1996)

Total element scores determined by ModelInspector with the Lentivirus LTR model for different LTR types are shown. For each set of sequences the minimum, maximum, and mean score is given. Element scores marked by black bars are reached with default parameter settings (element score threshold for determining elements: 50%, element score threshold for all elements: 50%, distance score threshold: 30%). Element scores indicated by grey bars were obtained with lower thresholds.

4 Discussion

We described a new algorithm to model transcriptional regulatory regions implemented into a program package consisting of the ModelGenerator program to develop the model and the ModelInspector program capable of searching whole databases for matches to predefined models. ModelInspector has no limit for sequence length allowing the analysis of sequences of several million nucleotides in length such as complete yeast chromosomes. A set of at least 10 sequences and a simple initial model are required as input for ModelGenerator. Most of the model is generated by sequence analysis allowing the development of complex models revealing unknown elements which are part of the organizational framework of the final model. We showed this to work for the regulatory regions of Lentiviruses and

demonstrated the specificity of the model by analyzing LTRs of all other types with the Lentivirus LTR model.

ModelGenerator and ModelInspector are quite different from the other approaches for definition of regulatory units in their general scope. Our approach focuses on the precise definition of highly specialized biological units rather than on global descriptions. For example, the promoter models used by FunSiteP (Kondrakhin *et al.*, 1995) account for the element composition of promoters but do not include detailed organizational features as the models generated by ModelGenerator. PROMOTER SCAN (Prestridge, 1995) will analyze pol II promoter candidates and does not differentiate between regulated or constitutive promoters. ModelInspector on the other hand will not detect all LTRs but is highly focused on the specific LTR subtype from which the model was derived. As demonstrated for the TATA box consensus profiles, a good deal of this high specificity is already inherited from the specificity of the individual element descriptions.

ModelGenerator develops the model from a set of sequences and is not restricted to promoters or LTRs. Any detectable organization of elements within a set of sequences can therefore provide the basis for model development (examples could be enhancers, scaffold attachment regions or even gene models). The ability to construct models for quite different genetic units is one of the main difference to other model based approaches like GeneParser (Snyder and Stormo, 1995), and GRAIL (Uberbacher and Mural, 1991) which are special-purpose gene prediction programs.

A clear advantage of all multi component models over detection of individual elements (e.g. TATA box) is the ability to tolerate missing individual elements without impairing recognition of the sequence by the model. The high specificity of our LTR models suggested that multi component ModelGenerator models might be suitable to distinguish between different promoter classes, e.g. identify promoters responding to a common pathway as a group. The general design of ModelGenerator also allows to use predefined models as elements for the construction of even more complex models. For example, the described LTR model could be used to develop a model for a complete provirus consisting of a 5'-LTR, a leader region, at least three reading frames for the common retroviral proteins (gag, pol, and env) and a 3'-LTR.

ModelGenerator integrates a variety of methods which allows extension of the basic model without further *a priori* knowledge or the necessity to alter the program. Thus, new information can be derived directly from the final model which may be a useful prospective analysis facilitating experimental design.

Acknowledgements

This work was supported in part by the BMBF Verbundprojekt GENUS 413-4001-01 IB 306 D (Förderschwerpunkt Bioinformatik) and by EU grant BI04-CT95-0226 (TRADAT).

References

- Dong, S., Searls, D.B.: Gene structure prediction by linguistic methods. *Genomics* **23** (1994) 540-551
- Frech, K., Herrmann, G., Werner, T.: Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. *Nucleic Acids Res.* **21** (1993) 1655-1664
- Frech, K., Brack-Werner, R., Werner, T.: Common modular structure of Lentivirus LTRs. *Virology* **224** (1996) 256-267
- Kondrakhin, Y.V., Kel, A.E., Kolchanov, N.A., Romashchenko, A.G., Milanesi, L.: Eukaryotic promoter recognition by binding sites for transcription factors. *Comp. Appl. Biosci.* **11** (1995) 477-488
- Myers, G., Wain-Hobson, S., Henderson, L.E., Korber, B., Jeang, K.-T., Pavlakis, G.N.: Human retroviruses and AIDS 1994. A compilation and analysis of nucleic acid and amino acid sequences. Database by Los Alamos National Laboratory (1994)
- Quandt, K., Frech, K., Herrmann, G., Werner, T.: A consensus match scoring system that is correlated with biological functionality. in *Bioinformatics: From Nucleic Acids and Proteins to Cell Metabolism* (Eds. D. Schomburg, U. Lessel) (1995a) 47-57
- Quandt, K., Frech, K., Karas, H., Wingender, E., Werner, T.: MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.* **23** (1995b) 4878-4884
- Quandt, K., Grote, K., Werner, T.: GenomeInspector: Basic software tools for analysis of spatial correlations between genomic structures within megabase sequences. *Genomics* **33** (1996a) 301-304
- Quandt, K., Grote, K., Werner, T.: GenomeInspector: A new approach to detect correlation patterns of elements on genomic sequences. *Comp. Appl. Biosci.* **12** (1996b) 405-413
- Prestridge, D.S.: Predicting Pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.* **249** (1995) 923-932
- Wolfertstetter, F., Frech, K., Herrmann, G., Werner, T.: Identification of functional elements in unaligned nucleic acids sequences by a novel tuple search algorithm. *Comp. Appl. Biosci.* **12** (1996) 71-80
- Snyder, E.E., Stormo, G.D.: Identification of protein coding regions in genomic DNA. *J. Mol. Biol.* **248** (1995) 1-18
- Uberbacher, E.C., Mural, R.J.: Locating Protein-Coding Regions in Human DNA Sequences by a Multiple Sensor Neural Network Approach. *Proc. Natl. Acad. Sci. USA* **88** (1991) 11261-11265