Review

Finding and Decrypting of Promoters Contributes to the Elucidation of Gene Function

Thomas Werner[1,2,*]

[1]Genomatix Software GmbH, Landsberger Strasse 6, D-80339 München, Germany
[2]Institute for Experimental Genetics, GSF National Research Center for Environment and Health, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany

Edited by E. Wingender; received and accepted 2 January 2002; published 13 March, 2002

ABSTRACT: The combination of full-scale genomic sequencing with high throughput expression analysis provides a new and largely unexploited basis for in silico functional genomics. Recent break through developments in locating and analyzing promoters now allow extending functional genomics in silico far beyond identification of protein sequences into the complex regulatory structures and mechanisms of the genome. However, only first examples of this new type of approach are emerging at present and intensive further developments of bioinformatics tools will be required before such analysis can become large-scale routine in genomic sequence analysis. Nevertheless, the door to a new dimension of functional analysis of the genomic sequence is open. Finally, only the tight integration of the enormous amount of knowledge gained from proteins sequence analysis with the complementary information about gene regulation will afford us with a more complete picture of the networks than constitute life.

KEYWORDS: genome annotation, gene regulation, promoter prediction, promoter modeling, functional context, transcription factor binding sites, promoter modules, RANTES, chemokine, coexpressed genes

INTRODUCTION

The number of genes so far identified or estimated from the human genomic nucleotide sequence turned out to be somewhere between 30,000 and 40,000 genes. Given the enormous size of the human genome (about 3 billion basepairs) this is both surprising as well as a bit disappointing. Only 2% to 3% of the human genome are now believed to be coding for proteins. As a consequence most of the genomic sequence and the functions hidden within will remain in the dark, even after all gene-encoded proteins will have been identified. About 40% of the genome can be accounted for by repetitive sequences, which are not void of function as well [3]. Nevertheless, even putting them aside, more than half of the genomic sequence remains an unknown territory.

[*] E-mail: werner@genomatix.de

Electronic publication can be found in *In Silico Biol.* **2**, 020313< http://www.bioinfo.de/isb/2002/02/0023/>, 13 March 2002

Proteins are the most visible and prominent proponents of life, however even a complete inventory of all proteins would never ever be able to give rise to a complex multicellular organisms. Appearance of individual proteins has to be tightly controlled in time and space, whole sets of proteins need to be expressed in a well-orchestrated manner during embryonal development [1]. The regulation of gene transcription and posttranscriptional control mechanisms is probably the most crucial part of life for any organisms in this respect.

LARGE SCALE IN SILICO ANNOTATION

The availability of large continuous regions of human genomic sequences in concert with new developments in bioinformatics allows a more systematic approach to tackle regulatory functions of gene expression directly on sequence level. This type of approach is completely independent of the nature and function of the proteins encoded by the respective genes and may even yield clues to functional assignment of some of the proteins still labeled unknown in the human genome annotation.

PROMOTER FINDING

One of the most crucial steps in the quest for genomic gene regulation is locating promoters of human genes in the genomic sequence. In contrast to the yeast system this is a difficult task as human promoters may be tens of kilobases upstream of the coding region due to the existence of non-coding leading exons [16]. This step takes prevalence over identification of other regulatory sequences (e.g. enhancers) because ultimately nothing can influence the transcription of a gene unless it has an effect on the promoter. Therefore, promoters act as central processing units (CPUs) of gene transcription, integrating all signals influencing gene transcription on the molecular level.

Finding promoters in mammalian genomic sequences is complicated by the diverse nature of promoters. First of all, promoters do not carry any clearly outstanding sequence signature by which they could be pulled from the genomic sequence [18]. Therefore, all methods to locate promoters have to rely on combinations of more subtle features, none of which is really promoter specific. Of course this translated directly into specificity problems and almost all existing methods are haunted by huge amounts of false positives [6]. Traditionally this problem was circumvented by an approach that I would like to call “sheltered environment” which helps to escape from the perils of low specificity. If only short stretches of DNA are being analyzed (1 kb to 10 kb or 15 kb were quite popular ranges) the number of false positives always remains acceptable and it was possible to focus on sensitivity rather than to worry about specificity. Quite naturally, this is not really de novo promoter finding as the information where the gene is has to be used in order to select the few KB most likely containing the promoter. In this manner several methods have been developed and were tested in a comparison [6,13,22].

Now that close to 3 billion bps of human genomic sequence are available such restrictions render any methods obsolete for large-scale analysis irrespective of their value for analysis of small sequences. In order to cope with a true genomic scale (not to mention any “post-genomic” analysis) unrestricted analysis of the whole raw genomic sequence for promoters became mandatory. It is only in this way that additional knowledge reaching beyond gene discovery methods can be gained.

We have developed and published an entirely new concept to locate promoters in genomic sequences [15]. We have initially applied this method (PromoterInspector) to human chromosome 22 in order to carefully evaluate the performance by comparison with the known annotation of the chromosome [16]. In the meantime we have completed analysis of the whole human genome draft based on the golden path sequence and came up with predictions of about 20,000 promoter regions. This collection constitutes the

Genomatix Promoter Resource (GPR) and is marketed by Genomatix. According to our estimates of sensitivity and specificity outlined below we calculated the number of genes in the human genome to be in the range of 35,000 to 40,000. This is remarkably well within the range found by any gene counts or gene conservation approaches, given that we did not take any known data about genes of gene counts into consideration.

The sensitivity of PromoterInspector was initially reported to be 43% [15] but has been confirmed by a much larger set of known promoters to be 50% by now. Specificity was also found to be 43% in the initial publication. Again enlarging the test set by more than one order of magnitude indicated a specificity of at least 85% suggesting some undocumented promoters in the small initial test set. Specificity as well as sensitivity were confirmed by observations of many scientists applying PromoterInspector to various sequences of their own (PromoterInspector is accessible for academic scientists at http://www.genomatix.de). However, neither our own tests nor the various observations confirming our data really proof those numbers. Unless we are sure of the annotation of the human genome (missing no genes!) all values for specificity or sensitivity must be seen as best estimates not as the ultimate truth.

FUNCTIONAL CONTEXT OF GENES

Although PromoterInspector affords us with the largest collection of human promoter regions available so far (at least with that specificity), this does not yield any clues about functional features of those promoters or genes. The most important step after locating promoters is to understand their functional anatomy. The most important elements involved in promoter functions are binding sites for effector proteins, called transcription factors (TFs). Only by interaction of a suitable set of such TFs with the promoter sequence the activator complex and subsequently the transcriptional initation complex can be formed [21]. However, a simple map of promoter elements, which can be easily obtained by analysis of promoter sequences for transcription factor binding sites, is not sufficient to understand promoter functions [e.g. 14]. This is the reason why I prefer the term functional anatomy because the context of TF binding sites are often more important than the binding sites themselves [e.g. 20]. For example a perfect binding site located outside the appropriate context will bind its cognate protein but may not elicit any biological function in transcription. On the other hand, many promoters contain relatively weak binding sites that may be unable to bind their cognate protein on their own but such sites have been shown to be functional in many cases [e.g. 2].

It was realized already several years ago that promoters seem to be composed of modular units conveying special functionality to the promoter, e.g. rendering a promoter responsive to a specific signaling pathway or induction in a cell or tissue specific manner [7]. Such functionally defined modules were first defined as regions with no further specification, later on presence of selected sets of TF binding sites were found to be associated with promoter modules [17], and the most stringent definition adds specific orientation and distances between those TF binding sites to the concept [10,12]. The smallest possible promoter module on sequence level is a combination of just two TF binding sites, which has been termed “composite element” by Kel *et al.* [11]. Molecular promoter modules (those including specific internal organization of the TF binding sites involved) can overlap physically with each other and may be located on different strands of the DNA in different promoters. These features illustrate why detection of promoter modules by in silico methods is such a difficult task. Alignment procedures cannot account for the extraordinary flexibility off molecular promoter modules, which is why more sophisticated modeling approaches are required [8].

The functional context of a gene is a composition of the context (e.g. binding partners) of the encoded protein and the regulation of its expression in time and space [14]. For example, proteins acting together in a pathway are often coexpressed at least under specific circumstances (e.g. immune functions). Part of this functional context can be derived from analysis of promoter sequences, as specific modules are often associated with coexpression of groups of genes [19].

The example I would like to use is RANTES, a well-studied chemokine where most of the results of the promoter analysis can be verified by experimental evidence. I will give a short summary of the study because it is published in more detail in Fessele *et al*. [5].

The RANTES promoter is only about 300 bp in length and contains six distinct binding regions each of which contains several potential TF binding sites. Nevertheless, this small region is sufficient to direct differential RANTES expression in a variety of cell/tissue types. We were able to dissect the RANTES promoter into five distinct but overlapping submodels, based on factor binding to the RANTES promoter experimentally verified in five different cell lines (Figure 1) [4]. The advantage of these in silico models was that they can be used to explore the nucleotide sequence databases for other promoters with a similar modular setup (all analyses were done with the GEMS Launcher software package from Genomatix, Munich). Examination of the mammalian sections of the EMBL nucleotide database with all five models resulted in less than 60 matches in total, proving an extraordinary selectivity of the models [5]. We then proceeded to check each individual match for known functional relationship with RANTES or chemokine function in general and found more than 60% of all matches to belong to the functional context of chemokines (Figure 2). This is quite remarkable as no other information but the promoter models was used in the search.

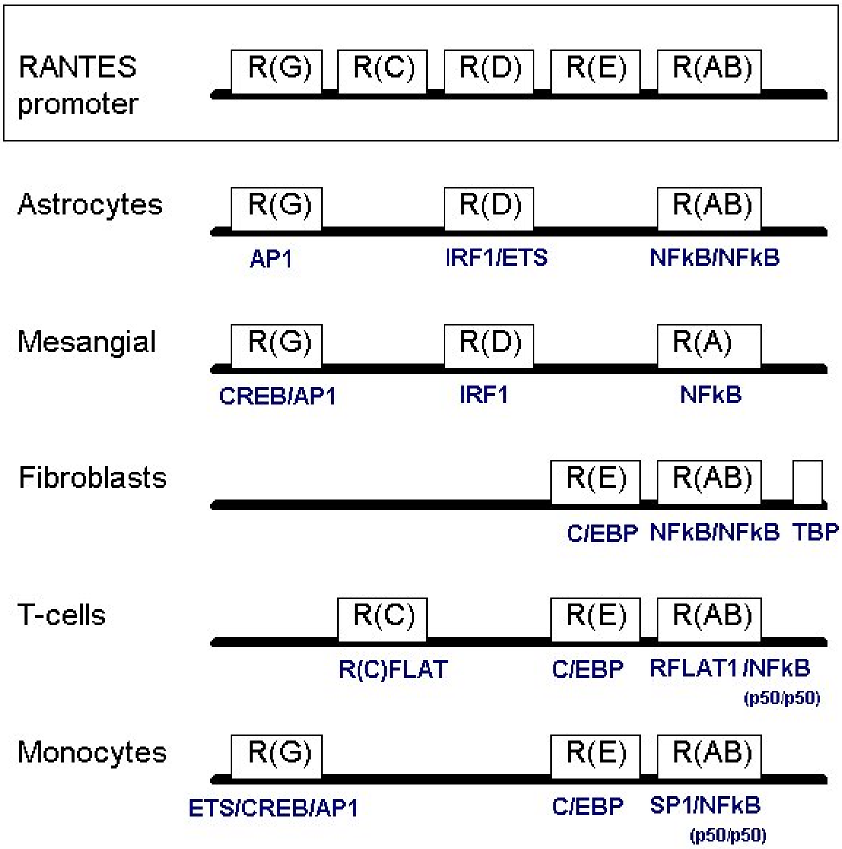

Fig. 1. The box shows the general setup of the RANTES promoter consisting of 5 different binding regions. Below the box the five submodels relevant in the respective cell types are shown. Below the binding regions individual TF binding sites are indicated that were used in the in silico promoter models. Note that different factor for the same binding region are used in different models.

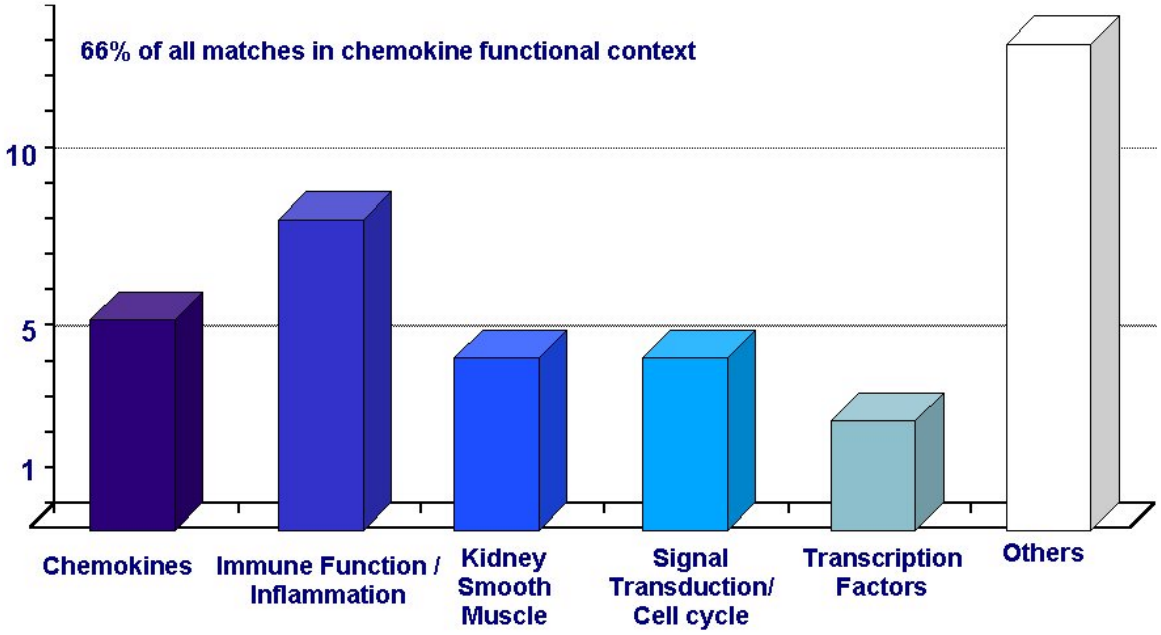

Fig. 2. Distribution of genes with promoter anatomy similar to RANTES within the functional context of RANTES. The blue bars indicate the number of matches found in the mammalian sections of the EMBL database (release 66). More intensive coloring indicates closer functional relationship. The white bar indicates matches where no functional relationship was known between RANTES and the genes found.

These results indicate that careful analysis of promoter sequences for their internal functional anatomy is a very powerful approach towards elucidation of the genomic functional context (in terms of other genes) of the genes in question. This also allows inferring on at least some functional features of an unknown protein by analysis of other genes found by functional promoter context similar to the RANTES example.

However, in most cases the detailed experimental dissection of the promoter will not be available as in the RANTES example. Fortunately, there are also in silico methods that allow dissecting at least part of the functional anatomy of a promoter in the absence of experimental promoter analysis. In principle there are two strategies possible. Both are based on the recent successes of systematic high throughput analyses. The first approach takes advantage of the parallel genome projects currently under way for several organisms as well as the existing nucleotide sequence databases.

In many cases it is possible to obtain promoter sequences for a particular genes from two or more species (orthologous genes). A comparative promoter analysis can reveal the modular backbone of TF binding sites that was evolutionary conserved in the promoter despite missing similarity on nucleotide sequence level [9]. Such analysis can be carried out with GEMS Launcher almost fully automatically and the methodology was shown to produce highly specific promoter models.

A more time-consuming strategy is possible based on expression array data, which is rewarding the extra efforts by revealing a much more detailed functional structure of the promoter. Genes found to be co-expressed with the gene of interest can be selected from expression array data and their promoters can be derived from GPR (promoters for about half of the genes will be more than sufficient for the purpose). These promoter sequences can then be analyzed the same way as the orthologous promoters to reveal promoter modules [detailed in19]. The difference here is that the data set from one expression array will reveal only the module responsible for the effect observed under the specific experimental conditions used (e.g. stimulation of genes by a hormone). This may be a module containing only two or three TF binding sites. Repeating this type of analysis with another set of genes coexpressed with the gene of interest and different condition (e.g. heat shock) will reveal another module associated with the other conditions. This way an extensively fine-structured anatomy of the promoter can be obtained that will provide at least the same resolution we obtained for RANTES by direct promoter analysis.

REFERENCES

[1] Arnone, M.I. and Davidson, E.H. (1997). The hardwiring of development: organization and function of genomic regulatory systems. *Development,* **124**, 1851–1864.

[2] Colgan, J. and Manley, J.L. (1995). Cooperation between core promoter elements influences transcriptional activity in vivo. *Proc. Natl. Acad. Sci. USA* **92**, 1955–1959.

[3] Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., Ainscough, R., Almeida, J.P., Babbage, A., Bagguley, C., Bailey, J., Barlow, K., Bates, K.N., Beasley, O., Bird, C.P., Blakey, S., Bridgeman, A.M., Buck, D., Burgess, J., Burrill, W.D., O'Brien, K.P. *et al*. (1999). The DNA sequence of human chromosome 22. *Nature* **402**, 489–495.

[4] Fessele, S., Boehlk, S., Mojaat, A., Miyamoto, N.G., Werner, T., Nelson, E.L., Schlondorff, D. and Nelson, P.J. (2001). Molecular and in silico characterization of a promoter module and C/EBP element that mediate LPS-induced RANTES/CCL5 expression in monocytic cells. *FASEB J.* **15**, 577–579.

[5] Fessele, S., Maier, H., Zischek, C., Nelson, P.J. and Werner, T. (2002). Regulatory context is a critical part of gene function. *Trends Genet.* **18**, 60–63.

[6] Fickett, J.W. and Hatzigeorgiou, A.G. (1997). Eukaryotic promoter recognition. *Genome Res.* **7**, 861–878.

[7] Firulli, A.B. and Olson, E.N. (1997). Modular regulation of muscle gene transcription: A mechanism for muscle cell diversity. *Trends Genet.* **13**, 364–369.

[8] Frech, K., Danescu-Mayer, J. and Werner, T. (1997). A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. *J. Mol. Biol.* **270**, 674–687.

[9] Frech, K., Quandt, K. and Werner, T. , (1998). Muscle actin genes: A first step towards computational classification of tissue specific promoters. *In Silico Biol.* **1,** 0005.

[10] Kel, A., Kel-Margoulis, O., Babenko, V. and Wingender, E., (1999), Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. *J. Mol. Biol.* **288,** 353–376.

[11] Kel, O.V., Romaschenko, A.G., Kel, A.E., Wingender, E. and Kolchanov, N.A. (1995), A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Res.* **23**, 4097–4103.

[12] Klingenhoff, A., Frech, K., Quandt, K. and Werner, T. (1999), Functional promoter modules can be defected by formal models independent of overall nucleoside sequence similarity. *Bioinformatics* **15**, 180–186.

[13] Ohler, U., Harbeck, S., Niemann, H., Nöth, E. and Reese, M.G. (1999), Interpolated Markov chains for eukaryotic promoter recognition. *Bioinformatics* **15**, 362–369.

[14] Rothenberg, E.V. and Ward, S.B. (1996). A dynamic assembly of diverse transcription factors integrates activation and cell-type information for interleukin 2 gene regulation. *Proc. Natl. Acad. Sci. USA* **93**, 9358–9365.

[15] Scherf, M., Klingenhoff, A. and Werner, T. (2000), Highly Specific Localization of Promoter Regions in Large Genomic Sequences by PromoterInspector — A Novel Context Analysis Approach. *J. Mol. Biol.* **297**, 599–606.

[16] Scherf, M., Klingenhoff, A., Frech, K., Quandt, K., Schneider, R., Grote, K., Frisch, M., Gailus-Durner, V., Seidel, A., Brack-Werner, R. and Werner, T. (2001), First pass annotation of promoters on human chromosome 22. *Genome Res.* **11**, 333–340.

[17] Wasserman, W.W. and Fickett, J.W.,(1998), Identification of regulatory regions which confer muscle-pecific gene expression. *J. Mol. Biol.* **278**, 167–181.

[18] Werner, T., (1999), Identification and characterization of promoters in eukaryotic DNA sequences. *Mamm. Genome* **10,** 168–175.

[19] Werner, T. (2001), Cluster analysis and promoter modelling as bioinformatics tools for the identification of target genes from expression array data. *Pharmacogenomics* **2**, 25–36.

[20] Yamauchi, M., Ogata, Y., Kim, R.H., Li, J.J., Freedman, L.P. and Sodek, J., (1996), AP-1 regulation of the rat bone sialoprotein gene transcription is mediated through a TPA response element within a glucocorticoid response unit in the gene promoter. *Matrix Biol.* **15**, 119–130.

[21] Zawel, L. and Reinberg, D. (1995), Common themes in assembly and function of eukaryotic transcription complexes. *Annu. Rev. Biochem.* **64**, 533–561.

[22] Zhang, M.Q. (1998), Identification of human gene core promoters in silico. *Genome Res.* **8**, 319–326.