Functional Analysis of Human Gene Promoters

CORINA SAMOILA[1]*, ANDREI ANGHEL[1], MARILENA MOTOC[1], ALFA XENIA LUPEA[2], LIVIU TAMAS[1]

[1] V”.. Babe°” University of Medicine and Pharmacy, Biochemistry Department, 2 P-ta Eftimie Murgu, 300041, Timisoara, Romania

[2] Politehnica University of Timi°oara, Faculty of Industrial Chemistry and Environmental Engineering, 2 P-þa Victoriei, 300006, Timisoara, Romania

The complexity of gene control processes and the perspectives opened by understanding how to control a target gene are several important reasons for researchers to try to find out more information about gene control mechanisms and the molecular networks responsible for that. The promoter is a central part of the regulatory network, being responsible for the transcription initiation process. The functional analysis of human promoters comprises combinations of methods that connect the special organizational features of the human promoters with the aspects of gene control by investigating in vivo and in vitro interactions. Combining data from different approaches, the functional analysis of the promoter represents the key to understand the gene control process, the mechanisms and molecular networks involved. Consequently, these helps to define the therapeutic strategies to be adopted in the case of a particular disease to control the expression of the genes associated with that disease.

Keywords: gene control promoter, transcription factors, computational, functional analysis

The molecular events contributing to gene control process are very complex as a result of multiple possibilities to control gene expression at different stages [1, 2] by action of complex molecular networks comprising a variety of controlling elements [3-5]. Usually situated upstream of the transcription start site of a gene but often overlapping with the first exon of a gene, the gene promoter is a DNA cis-controlling sequence that initiate the gene transcription process. The promoter region contains short DNA sequences (5-25 bp) named transcription factor binding sites (TFBSs) able to recognize and bind specifically controlling proteins (transcription factors or TFs) to form the transcription initiation complex [6].

Organizational features of the human promoter determine functionality of the promoter in the context of a specific signaling. These organizational features are represented by the modular structure of the promoter with individual modules organized in a specific manner [7, 8]. The promoter modules are composed of multiple transcription binding sites arranged within a defined pattern [9]; it determines a promoter response to the specific signaling [8, 10] depending on the type of cell or tissue. The functionality of the promoter is provided by the hierarchical organization of TFBs in promoter modules [7] that permits the interactions between controlling proteins, TFBs and the components of the basal transcription system in a specific manner.

In this review, we will outline some strategies and methods used for the functional analysis of human promoters based on a combination of advanced computational prediction of promoter and TFBs with in vitro and in vivo experimental techniques to understand the contribution of a particular TFBs, module or entire region to gene expression process.

Experimental part

Bioinformatics Analysis

The computational approach of promoter analysis has known a rapid advance in the last years permitting identification of the promoter region and TFBSs, to construct models based on different combinations of regulatory elements or to compare the genomic sequences with the aim to find the controlling regions.

Identification of promoters

The mammalian promoters show a large diversity as a consequence of a large number of possible TFBs combinations [12, 13] or as a consequence of the regulatory elements positioning not only upstream of the gene but within large regions [13]. That diversity of human promoters associated with the possibility to have multiple different start sites of the mammalian genes [14, 15] make difficult to identify the promoter. Computational approaches in human promoter identification have made easier the task of human promoter prediction, identifying 13-54% of promoters from genomic sequence [11], although the number of false positive rates is estimated at approximately one per kilobase [16] for most computational methods.

Alignment of the full-length cDNA sequences to the human genome sequence is frequently used to obtain promoter sequence, to predict the transcription start sites and to identify the sequences immediately upstream of the transcription initiation site or overlapping with it. The computational methods developed to locate a promoter in the genomic sequence are based on different algorithms such as those using oligomers composition (PromoterInspector [18]) or the density of the TBSs in the promoter (TSSG, TSSW [19], Proscan [20]). The consensus promoter predictor (CONPRO) developed by Liu et al [21] permits aligning of gene transcripts (expressed sequence tags [ESTs] or mRNA) with genomic sequence and, using a combination of five computational methods for promoter prediction, confers a higher confidence for promoter prediction [21]. The sensitivity and specificity of the computational methods is enhanced by combining them with CpG island information [22], CpG islands being clusters of CpG dinucleotides residing in a non-random fashion within approximately 50 % of mammalian gene promoters, each promoter of this type being associated with one or more CpG islands [23].

** email: biochim@umft.ro; Tel.: 0256 204250 /413*

FBSs prediction

The next step after promoter identification is prediction of TFBSs based on the frequency of analyzed TFBS in the genome. PWM method (position weight matrix) will create a matrix pattern based on the probability of the base to appear in the specified position of the consensus sequence [24]. The MatInspector program allows searching of sequences for matches with the consensus matrix description; the matrix similarity is calculated following an algorithm described by Quandt et al. [25] and higher scores indicate higher probability of the sequence analyzed to match with the consensus matrix. The applicability of methods depends of the size matrices libraries. The largest library available for public [26] is MatInspector library containing today 634 matrices; also, the program Match [27] allows searching from a TRANSFAC 6.0 public database containing 336 matrices. The limitation of PWM methods is that they permit identification of TFBS but not all the predicted TFBS are functional. To reduce numbers of false positive results, the computational predictions were adapted to the idea that the organizational features of the functional promoters are conserved in two directions: vertically (inter-species) and horizontally (intra-species) [28].

Analysis of expression profile of co-controlled genes

Analysis of expression profile of controlled genes provides data about controlling elements from a cluster of genes with similar expression patterns. They belong to the same cluster, contributing to a specific biological pathway and may be controlled at the transcriptional level [28]. Analysis of a cluster of genes has the aim to discover similar controlling regions. Then the computational methods are used to generate possible models for these shared regulatory regions and sequence databases are scanned to identify new genes comprising similar regions [29]. The strategy has applied well for yeast [30], but for mammalian promoters the analysis is more complex [28].

Comparison between species (comparative genomics or phylogenetic footprinting)

Frequently used in analysis of mammalian promoters, comparison between species is applied to find out evolutionary conserved regions which are assumed to be functional regions. This comparative analysis is carried out using bioinformatics tools to align DNA sequences for similarities and homologous regions. Several algorithms are based on local alignment (BLAST [34] or PIPMaker [32]), others are based on global alignment (Clustal W [33]). The FootPrinter program [34] identifies short conserved motifs 10 bp in length within 2 or more aligned DNA sequences from related species being more specific and informative than that of PIPMaker in identifying small homologous regions [32]. Using multiple computational approaches seems to be the best method to find TFBs with a higer probability to be functional. For example, Wasserman et al. have identified all three known major muscle specific TFs (SRF, MEF2 and MYF) based on a cross-species analysis between human and rodent skeletal-muscle specific genes combined with different other approaches [35]. The comparative genomics combined with TRANSFAC database searching and statistics was also applied successfully to identify the AP1, RUNX2 and CREB as PTH responsive transcription factors by Qiu et al [36].

In vitro analysis of human promoters

The experimental strategies used in functional promoter analysis, based on in vitro and in vivo methods, have two major purposes: to understand the interactions of the protein complexes with the promoter sequences (using electromobility shift assay [EMSA], chromatin immunoprecipitation assay [ChIP] or cromatin immunoprecipitation assay coupled with microarray [ChIP-chip assays]) and to understand how these interactions influence the expression gene (reporter assay, RT-PCR, microarrays). The aim of such experimental strategies is to understand how all these factors interact to form the transcription initiation complex, starting the transcription process to produce mRNA.

The analysis based on in vitro experiments includes the quantification of mRNA by real-time RT-PCR, EMSA and some of the reporter gene assays.

Reporter gene assay

There are several possibilities to verify that a promoter region is involved in the control of gene expression. Reporter assays require a construct plasmid containing the fragment of interest from promoter sequence ligated upstream of the reporter gene in a specific vector.The regulatory activity is analysed by assessing the activity of the reporter gene under different conditions. The construct plasmid can be introduced in a given type of cell by transient or stable transfection procedures followed by measurement of activity of the reporter gene. Additional analyses involve successive deletion or/and mutagenesis of the promoter and analysis by reporter assays to define the specific sequences of promoter region responsible for the gene activation or gene repression. The method was applied successfully to characterize the elements required for basal transcription of many human genes. Ugliati et al. [37] have identified the fragment with higher activity in the analyzed type of cell from the two fragments of human *CR2* (complement receptor type 2) proximal promoter sequence cloned upstream of a luciferase reporter gene, indicating that all elements required for basal transcription of the gene are localized in the -315/+75 bp fragment. By further truncation of that promoter fragment and luciferase assays they have identified the start transcription site for basal transcription and the sequences which mediate that basal transcription of CR2 gene.

Real-time reverse transcription PCR (or quantitative, qRT-PCR)

The method permits the quantification of the amount of messenger RNA (mRNA) transcribed from a gene as a measure of gene expression in chosen conditions of time and cell type. The real-time reverse transcription PCR requires isolation of mRNA from a cell sample, production of cDNA by reverse transcription and some steps from a real-time PCR amplification of the cDNA that permit quantification of the DNA amount obtained after each round of amplification. There is a large number of protocol variants for the real-time using fluorescence detection which differ by enzymes, primers, probes and amplicon combinations, but the method is simple to use and capable of high-quantification [38] .

Electromobility shift assay (EMSA)

EMSA is applied to identify the sequence-specific DNA-binding proteins (such as transcription factors) in vitro. In the promoter analysis EMSA is usually combined with mutagenesis in order to find important binding sequences within the promoter. EMSA requires incubation of the protein from a crude nuclear or whole cell extract with a labeled DNA fragment which contains the specific protein binding site. To investigate the binding specificity, competition assays of unlabelled specific and non-specific competitors are used in excess, to out-compete specific interactions [39, 40]. In a recent study, in order to investigate the role of XBP1 (a key transcription factor in the endoplasmatic reticulum (ER) stress response patway) in neuronal cells, Chihiro et al. [41] have analyzed the WFS1 (an ER stress response-related gene) which was up-regulated by XBP1. Using a WFS1 promoter assay they found as yet-unidentified ERSE-like motif (highly similar with ER stress response element -ERSE) that is critical for the regulation of WFS1 by XBP1. Analyzing the binding of XBP1 to this sequence by EMSA, the results have shown that there is not a directly binding between them, other transcription factors being probably bound to the ERSE sequence of WFS1.

Results and discussions

In vivo analysis of human promoters

DNase I hypersensitivity assays

The methods are used to detect the modifications of chromatin structure happened in vivo, permitting the binding of the TFs to TFBSs within a gene promoter to regulate gene transcription. Thus, DH assay is a useful technique for indicating the existence of transcription factors bound *in vivo* to regions within a gene promoter [42]. The classical method requires a nuclei isolation step from tissues or cells, followed by treatment with DNase I. The DNA is purified from treated nuclei and digested with a restriction enzyme. Restriction fragments are separated by agarose gel electrophoresis and blotted onto nitrocellulose membranes. The open regions of chromatin representing DNase I hypersensitive (DH) sites are detected by hybridization with a radiolabelled probe. In the case when such DH sites are present, the autoradiograph shows additional bands to those resulting from the digestion with restriction enzyme. The major limitation of the classical DNase I hypersensitivity method results from manipulation of DNA solution; also, the DNA molecules exceeding ~ 20 Kb cannot be analyzed using a conventional agarose gel electrophoresis [43]. To overcome these limitations, Matthew E. Pipkin and Mathias G. Lichtenheld [43] have proposed a *Mega-DNase I hypersensitivity Analysis (MDHA)* method which uses the DNase I treated nuclei embedded in low-melting point agarose for DNA purification and replacing the conventional agarose gel electrophoresis with field inversion gel electrophoresis (FIGE) they can facilitate separation of DNA molecules in size 1-10^3 Kb.

Chromatin Immunoprecipitation (ChIP) assay

The DNA-protein interactions can be studied by Chromatin Immunoprecipitation (ChIP) assay which allows the detection of the specific transcription factors bound to the specific sites from gene promoter in the cells. Treating the cells with cross-linking reagent as formaldehyde**,** the proteins are cross-linked to DNA in the living cells. The cells are lysed to isolate nuclei and the nuclei are sonicated to shear the genome randomly [44]. Using a factor-specific antibody, the protein-DNA complex is immunoprecipitated and isolated followed by isolation and purification of the DNA target. The DNA fragment is cloned in a plasmid vector and sequenced by using vector-specific restriction enzymes [45] or the DNA fragment can be marked with fluorophores and hybridized to a genomic microarray [46]. Several types of genomic microarrays are associated with ChIP [47-50] either using selected or randomized promoters, human genomic fragments with high CpG content or using a continuous genomic sequence to find transcription factor binding sites [46].

Conclusions

Taking account of the large number of possible combinations for promoter organization, the utility as well as the limitations of using bioinformatics methods are evident. The computational approaches of promoter and TFBS prediction have a large number of false positive results. Not always an identified TFBS by this type of method proved to be functional; due to a specific context created by specific interactions between TFs and TFBSs, a particular TFBS can become functional or, contrary, can become non-functional. To overcome these limitations, the computational analysis is accompanied by experimental in vivo and in vitro methods that validate the findings of bioinformatics analysis. Each experimental method, of course, has its own advantages and limitations. The main disadvantage of in vitro methods is that the cells in culture do not represent cells within a normal physiological environment, despite the large applicability of these methods. The reporter gene assays are applied in most of the promoter analyses, being successfully used, for example, to demonstrate that the IL-6 interleukin serum level variations that appear in systemic–onset juvenile chronic arthritis (S-JCA) results from a difference in the control of IL-6 interleukin expression due to a C/G polymorphism in the promoter of that gene [51]. Despite the fact that gene expression can be easily measured, the reporter assays are measuring reporter activity rather than expression of a given gene, but is not that easy to culture and analyze primary cells of individuals with known promoter genotypes. Comparative with reporter assays, the RT-PCR method permits the quantification of mRNA corresponding to the protein of interest, but the quantified level of mRNA may not reflect the real level of protein produced by the cells [52] because many regulatory events of protein synthesis proceed at the post-transcriptional level. Besides RT-PCR, various other methods are available to study the level of gene expression such as: northern blots [53], S1 nuclease protection [54] and serial analysis of gene expression (SAGE) [55]. From the same category of methods, the array–based technologies such as cDNA or oligonucleotide arrays [56, 57] were developed rapidly in the last several years to analyze a large variety of sample in the same time in various conditions. Thus, the method can provide information about gene expressions in different types of cells or about the clusters of gene that shows similar expression patterns. Although the parallel analysis of gene expression provides useful information about biological pathways, it does not provide details about mechanisms of molecular interactions. The DNA-protein and protein-protein interactions are analyzed by in vivo and in vitro methods such as EMSA, DNase I hypersensitivity, in vivo and in vitro footprinting and ChIP which can offer more details about the molecular mechanisms. As an example, the ChIP-based target gene method was successfully used to study the role of the transcription factors in immune response [45], but the observed limitations of the ChIP technique resides from that not all interactions detected by this method have a functional effect and the interpretation of the results requires additional assays [45]. Also, in a ChIP assay, the location of binding site should be approximately known to amplify the DNA sequence-binding protein in order to sequence it [45]. The DNase I hypersensitivity assay offers more details about the sites where the TFs are bound in vivo, but it does not show which proteins are bound there. The major limitation of in vivo footprinting assays is that they don't provide clear results if a protein is bound to the specific sequence analyzing only a fraction of cells [58]. However, using different strategies by combining the computational and experimental methods is recommended to have more confidence in the results; the interpretation of the results requires attention because analysis not always results in a functional effect. Moreover, combining the information provided by various methods in promoter analysis will contribute to the complex study of the molecular mechanisms responsible for a particular response and to the development of therapeutic pathways to influence the expression of candidate genes for a particular disease.

Bibliography

1.STRACHAN, T., READ, A.P., Human molecular genetics**.** 2nd edition. New York: Wiley, 1999

2.CAREY, M., SMALE, S.T., NY: Cold Spring Harbor Laboratory Press, 2000

3.WERNER, T. Mamm. Genome **10**, 1999, p.168

4.FRISCH, M., FRECH, K., KLINGENHOFF, A., CARTHARIUS, K., LIEBICH, I., WERNER, T., Genome Res. **12**, 2002, p.349

5.ROBERTSON, K. D., Oncogene 21, 2002, p.5361

6.ZAWEL, L. AND REINBERG, D., Annu. Rev. Biochem. **64,** 1992, p. 533

7.KLINGENHOFF, A., FRECH, K., QUANDT, K., AND WERNER, T. Bioinformatics, **15**, 1999, p. 180

8.KLINGENHOFF, A., FRECH, K., AND WERNER, T. In Silico Biol. **2**, 2002, p. S17–S26

9.WERNER, T., FESSELE, S., MAIER, H ., NELSON, P. J. FASEB J. **17**, 2003, p. 1228

10.FIRULLI, A. B. AND OLSON, E. N., Trends Genet. **13,** p. 364

11.QIU P., Biochemical and Biophysical Research Communications **309** 2003, p. 495

12.LANDER, E.S., LINTON L.M., BIRREN B., NUSBAUM C., ZODY M.C, BALDWIN J., Nature, **409**, 2001, p. 860

13.HAYASHIZAKI Y., CARNINCI P., PLoS Genet, 2006, 2(4): e63

14.CARNICI P., KASUKAWA T., KATAYAMA S., GOUGH J., FRITH M.C., Science **309**, 2005, p.1559 -63

15.FICKETT, J.W., HATZIGEORGIOU, A.G., Genome Res. **7**, 1997, p861

16.KNUDSEN, S., Bioinformatics, **15,** 1999, p.356

17.SCHERF, M., KLINGENHOFF, A., WERNER, T**.,** J. Mol.Biol. **297,** 2000, p.599

18.SOLOVYEV, V., SALAMOV, A., ISMB, **5**, 1997, p. 294

19.PRESTRIDGE, D.S., J. Mol. Biol, **249**, 1995, p. 923

20.LIU R., STATES D., Genome Res., **12**, 2002, p.462

21.IOSHIKHES I.P., ZHANG M.Q., Nat. Genet., **26** , 2000, p. 61

22.ANTEQUERA, F., BIRD, A.**,** Proc. Natl Acad. Sci. USA, **90**, 1993, p. 11995

23.STORMO, G.D., Bioinformatics, **16**, 2002, p. 16

24.KERSTIN QUANDT, KORNELIE FRECH, HOLGER KARAS, EDGAR WINGENDER , THOMAS, WERNER**,** Nucleic Acids Res., **23**, 1995, p. 4878

25.CARTHARIUS K., FRENCH K., GROTE K., KLOCKE B., HALTMEIER M., KLINGENTOFF A., Bioinformatics, **21** , 2005, p. 2933

26.KEL A.E., GÖBLING E., REUTER I., CHEREMUSHKIN E., KEL-MARGOULIS O.V., WINGENDER E., Nucleic Acids Res., **31**, 2003, p. 3576

27.DÖHR S., KLINGENHOFF A., MAIER H., HRABÉ DE ANGELIS M., WERNER T. AND R. SCHNEIDER., Nucleic Acids Res, **33**, 2003, p.864

28.WERNER, T., Brief Bioinform., **1**, 2000, p 372

29.PILPEL, Y., SUDARSANAM, P., CHURCH, G.M., Nat.Genet. **29**, 2001, p.153

30.SCHWARTZ, S., ZHANG, Z., FRAZER, K.A., SMIT, A., RIEMER, C., BOUCK, J., GIBBS, R., HARDISON, R. AND MILLER, W., Genome Res., **10**, 2000, p. 577

31.ALTSCHUL, S.F., GISH ,W., MILLER, W., MYERS, E.W., LIPMAN, D.J**.,** J. Mol. Biol., **215**, 2000, p. 403

32.BLANCHETTE, M., TOMPA, M. Nucleic Acids Res., **31,** 2003, p. 3840

33.THOMPSON, J.D., HIGGINS, D.G., GIBSON, T.J., Nucleic Acids Res., **22**, 1994, p. 4673

34.HEID, C.A., STEVENS, J., LIVAK, K.J., WILLIAMS, P.M**.,** Genome Research, **6**, **1996**, p. 986

35.WALL, S.J., EDWARDS, D.R., Analytical Biochemistry, **300,** 2002, p. 269

36.COHEN, C.D., FRACH, K., SCHLONDOR, D., KRETZLER M., Kidney International, **61**, 2002a, p.133

37.ULGIAT,I D., PHAM, C., HOLERS, M.**,** The Journal of Immunology**,** **168**, 2002, p. 6279

38. BUSTIN, S. A, Journal of Molecular Endocrinology, **29**, 2002, p. 23

39.HENDRICKSON, W., Biotechniques , **3**, 1985, p. 198

40.REVZIN, A., Biotechniques, **7**, 1989, p. 346

41.KAKIUCHI, C., ISHIWATA M, , HAYASHI A, KATO T, Journal of Neurochemistry, **97**, 2006, p. 545

42.COCKERILL, P.N.**,** Methods in Molecular Biology. , Ed. Tymms, M.J., 2000, p. 28

43.PIPKIN, M. E., LICHTENHELD, G.M., Nucleic Acids Research, **34**, 2006, No. 4.

44.SHANG, Y., HU, X., DIRENZO, J., LAZAR, M.A., BROWN, M**.,** Cell., **103**, 2000, p. 843

45.WEINMANN, A. S., Immunolog , **4**, 2004, p. 381

46.WEINMANN, A. S., YAN, P. S., OBERLEY, M. J., HUANG, T. H.M, FARNHAM, P. J., Genes Dev., **16**, 2002, p. 235

47.REN, B.,CAM H., TAKAHASHI Y., VOLKERT T., TERRAGNI J., YOUNG R.A., DYNLACHT B.D., Genes Dev., **16**, 2002, p. 245

48.WEINMANN, A. S., BARTLEY, S. M., ZHANG, M. Q., ZHANG, T., FARNHAM, P. J. Mol. Cell. Biol , **2**1, 2001, p. 6820

49.MARTONE, R., EUSKIRCHEN, G., BERTONE, P., HARTMAN, S., ROYCE, T.E., LUSCOMBE, N.M., RINN, J.L., NELSON, F.K., MILLER, P., GERSTEIN, M., WEISSMAN, S., SNYDER, M., Proc. Natl Acad. Sci. USA , **100**, 2003, p.12247

50.HORAK, C.E., MAHAJAN, M.C., LUSCOMBE, N.M., GERSTEIN, M., WEISSMAN, S.M., SNYDER, M., Proc. Natl Acad. Sci. USA, 2002

51.FISHMAN, D., FAULDS, G., JEFFERY, R., MOHAMED-ALI V., YUDKIN, J.S., HUMPHRIES, S., WOO, P., J.Clin.Invest., **102**, 1998, p.1369

52.GYGI S.P., ROCHON Y., FRANZA B.R., AEBERSOLD R., Molecular and Cell Biology, **19**, 1999, p. 1720

53.ALWINE, J.C., KEMP, D.J., STARK, G.R. Proc. Natl Acad. Sci. USA , **74**, 1977, p. 5350

54.BERK, A.J., SHARP, P.A**.** Cell , **12**, 1977, p721–732

55.VELCULESCU, V.E., ZHANG, L., VOGELSTEIN, B., KINZLER, K.W., Science, **270**, 1995, p. 484

56.SCHENA, M., SHALON, D., DAVIS, R.W., BROWN, P.O., Science, **270**, 1995, p. 467

57.LOCKHART, D.J., DONG, H., BYRNE, M.C., FOLLETTIE, M.T., GALLO, M.V., CHEE, M.S., MITTMANN, M., WANG, C., KOBAYASHI, M., HORTON, H., BROWN, E.L., Nature Biotechnol., **14**, 1996, p. 1675

58. SUNG-HAE, L.K, VIEIRA, K., BUGERT, J., Nucleic Acids Research, **30**, 2002, 10

Manuscript received: 10.10.2007