A Profile HMM for Recognition of Hormone Response Elements

Maria Stepanova[1], Feng Lin[2], and Valerie C.-L. Lin[3]

[1] Bioinformatics Research Centre, Nanyang Technological University,
50 Nanyang Drive, Singapore 637553
[2] School of Computer Engineering, Nanyang Technological University,
Block N4, Nanyang Avenue, Singapore 639798
[3] School of Biological Sciences, Nanyang Technological University,
60 Nanyang Drive, Singapore 637551
{mari0004, asflin, cllin}@ntu.edu.sg

Abstract. Steroid hormones are necessary for most vital functions of vertebrate organisms, and act within cells via interaction with their receptor molecules. Steroid hormone receptors are transcription factors. Identification of Hormone response elements (HREs) on DNA is essential for understanding the mechanism of gene regulation by steroid hormones. In this work we present a systematic approach for recognition of steroid HREs within promoters of vertebrate genomes, based on extensive experimental dataset and specifically reconstructed Profile Hidden Markov Model of putative HREs. The model can be trained for further prediction of HREs in promoters of hormone responsive genes, and therefore, investigation of direct targets for androgen, progesterone and glucocorticoid hormones. Additional documentation and supplementary data, as well as the web-based program developed for steroid HRE prediction are available at http://birc.ntu.edu.sg/~pmaria.

1 Introduction

A large number of ontogenetic and physiological processes within different organisms - from fungi to human - are regulated by a small group of steroid hormones. It can be hardly to over-evaluate the significance of steroid hormones for the life cycle during the whole period of development of an individual. Steroid hormones play a central role in the regulation of all aspects of female reproductive activity leading to the establishment and maintenance of pregnancy [1]. Also steroid hormones are essential for male fertility [2], some of them have been implicated in the cardiovascular [3], immune [4], and central nervous systems [5], as well as in bone function [6].

Steroid hormone family includes estrogen, progesterone, androgens, glucocorticoids, and mineralocorticoids, which are synthesized of cholesterol and secreted by endocrine cells [7]. The steroid hormone receptors (HRs) are intracellular transcription factors that exist in inactive apoprotein forms either in the cytoplasm or nucleus [8]. Connection of a hormone results in allosteric change of conformation of the receptor (this process is known as "activation of a receptor") that

J.C. Rajapakse, L. Wong, and R. Acharya (Eds.): PRIB 2006, LNBI 4146, pp. 15–22, 2006.

raises affinity of the receptor to DNA; it allows a receptor to bind to specific parts (hormone response elements, or HREs) of DNA molecule inside a nucleus and to adjust transcription of cis-linked genes. In addition to regulating transcription, steroid hormones occasionally regulate gene expression by affecting mRNA stability and translational efficiency [7].

Consensus steroid Hormone Response Elements contains symmetric imperfect repeats; namely, direct repeats, palindromic, and inverted palindromic repeats, of hexameric half-site sequence 5'-AGAACA-3'. These half-sites are usually divided by 3bp-long spacer [9] (except for Estrogen Response Element (ERE) which has some other distinctive features and is not included in this work [10]). In natural promoters, HREs display a great diversity in nucleotide sequence, some of which may contribute to a degree of receptor specificity, whereas other nucleotide substitutions may be incidental. Mutational analysis allows estimating relative significance of every position within the response element. It is worth mentioning works by Dahlman-Wright et al. [9], Barbulescu et al. [11], Truss et al. [12] and a review by Evans [13], where specific structure of HREs in described in a series of experiments.

Activated HRs are usually considered as classic vertebrate transcription factors, and classic method of transcription factor binding sites (TFBSs) can be used for prediction of steroid HREs too. A review of possible approaches for the task of recognition of binding sites in general has recently been published by Wasserman and Sandelin [14]. Unfortunately, these methods are of very low specificity due to great diversity of TFBS.

A possible way to improve the accuracy of prediction is to take into account the specific structure of a particular TFBS, and reconstruction of the model with consideration of its specific features. Specific HRE-like patterns have lately become an object of interest of several research groups: works by Favorov et al. [15], Sandelin and Wasserman [16], Bono [17] mainly focus on specific HRE-like structures, and the work by Bajic et al. [10] describes a method and a tool for the steroid hormone estrogen. However, the performance of the proposed NHRE works is limited due to insufficient training sets, as well as the high level of false positives inherent for single nucleotide position frequency-based models.

In this work we present a systematic approach for recognition of HRE within promoters of vertebrate genomes, based on extensive experimental data collected from literature and a classic method commonly used for profile modeling - Profile Hidden Markov Model [18]. The model can be used for prediction of HREs for further investigation of androgen, progesterone and glucocorticoid primary target genes.

2 Methods

2.1 Data Collection

Seven hundred of experimentally verified binding sites for Androgen, Glucocorticoid and Progesterone nuclear receptors were collected from the biomedical literature. For a binding site to be accepted into the collection a convincing experimental evidence was required - at least validated for binding in vitro, and demonstrated to mediate a response through plasmid transfection assays. Further requirement was a positive identification of the interacting steroid hormone receptor and an experimentally based identification of the binding site positions.

A binding site was not included into the collection if correspondent literature source contained ambiguous or insufficient information. In particular, if experimental data showed only location of protected region, but the position of binding site was predicted by sequence analysis on basis of comparison with known ARE/PRE/GRE consensus; or if binding site was predicted by only transfection assay (or other indirect method), without showing immediate receptor-DNA interaction. To avoid over-fitting of the model we included a particular HRE into the database only once even if a particular binding site was mentioned twice or more as verified by different experimental methods, and correspondent primer had been retrieved from one source.

Reported bound sequence was included with three flanking nucleotides in both directions. Positions of two half-sites of the response element were recorded if this information was given; if not - the internal structure of the response element was determined based on pairwise alignment of the sequence with known consensus binding site.

All retrieved binding sites were joined into Tiger HRE database. Every entry of this database is characterized by i. response element nucleotide sequence (if known, positions of two half-sites to which a receptor bind as a dimer were indicated); ii. a steroid hormone for which receptor binding was detected (if the same binding site was reported to bind to two or three steroid hormone receptors in the same literature source, it corresponded to several entries); iii. corresponding hormone-regulated gene (if existed and mentioned); iv. species from whose genomic DNA (used in the experiment) with the response element was retrieved; v. relative position from transcription start site (if this response element was retrieved from promoter or enhancer region or first exon of any hormone-regulated gene); vi. experimental method of binding detection; vii. reference. After implementation of proposed algorithm for HRE recognition, each entry from the database was supplemented with corresponding probability value for each HRE sequence.

Final version of database was implemented as a table within MySQL database system.

2.2 Hidden Markov Model Algorithm for HRE Recognition

The proposed profile HMM is depicted in Fig.1. It represents per se a composition of 5 independent HMMs for each constituent part of the HRE pattern - two flanking regions, two half-sites for dimer binding, and a spacer separating them. Each of these constituent domains is expected to have its own properties (i.e. internal transition probabilities), so has to be examined and trained separately in; transition probabilities between two consecutive ones also must be evaluated.

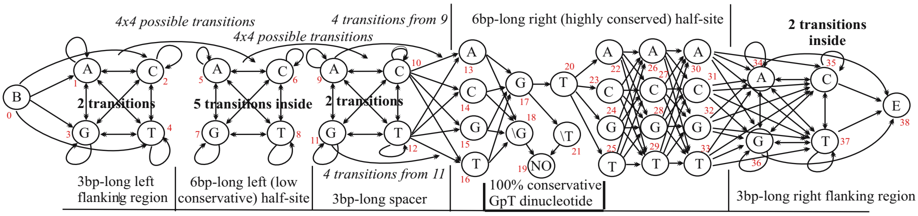

Fig. 1. Hidden Markov Model for HRE recognition

As the right half-site is found to demonstrate conservation close to a rate of 100%, a more specific topology of state transitions is defined. And also, as a dinucleotide GpT was shown to be a characteristic feature of almost all functional HREs (as shown in all works mention in the Introduction section), it is made a necessary component of an input sequence by the profile HMM. In this way, if a path leads to state 19, the model emits "NO" and the probability of the sequence is set to 0. However, there are some differences in lengths of training sequences (not all of them are denoted with flanking regions in corresponding literature). Hence, normalization procedure for probability value is used - logarithm of probability is divided by sequence length. Also prior distribution is used from position frequency matrices. Alignments of experimentally verified HREs from Tiger HRE DB were used for the Maximum Likelihood (ML) estimation of transition probabilities with the profile HMM. Probability value received with use of this method is further denoted as HMMS (Hidden Markov Model Similarity) and calculated as a product of transition probabilities come across when aligning the sequence and the reconstructed HMM. Received values for parameters of the HMM are given in the Supplemental Info section. Then, moving a 21bp-long window down the given sequence (being scanned for HREs), recognition procedure is performed for longer sub-parts of DNA.

2.3 Accuracy Estimation

For assessment of accuracy of our predictions by profile HMM, we used cross-validation approach for sensitivity assessment, that is, 70% of collected dataset used for training vs. 30% for testing; and we generated 10 random 'DNA' sequences, each being 50Mbp long, with all 'nucleotides' equally frequent and all positions independent, for estimation of occurrence of signals (random estimation, or re-value) using prediction level on a random basis.

2.4 Web-Based Tool for HMM Prediction of HREs

The publicly available version of the program allows users to input the sequence in FASTA, GenBank, EMBL, IG, GCG or plain format by either pasting it into an input box or by reading it from a text file. Also user can select accuracy level with use of provided table of sensitivity and specificity correspondence. Allowable length of submitted sequence is up to 5kb, and of course it should not be shorter than pattern length of 21 bp = 3 + 6 + 3 + 6 + 3 bp (two-half-sites separated by a 3bp spacer in consensus, together with two f3bp-long lanking regions). The resulting output will include: relative position of found match within submitted sequence; direct/complimentary DNA strand (if option of inclusion of complimentary strand is selected before the search started); actual nucleotide composition of found HRE; novel/known HRE (known means presented in the training set); HHM-based probability. For further investigation, the user can submit the sequence to other web-base tools for recognition (reviewed above) to estimate presence of other binding sites in the surrounding area and predict functionality of a potential regulatory complex. For the user to perform analysis of the promoter region of the gene of interest, it is necessary to extract promoter region from any public database of promoters (for example, BEARR [19]) and submit a sequence to the form.

3 Results

3.1 Database of Hormone Response Elements

The benchmark dataset for training and testing of the model was collected from 174 different biomedical literature sources (in the final version to date of paper submission, it is 712 hormone response elements included into the database). Such a collection has no analogs in the current public and commercial databases of TFBS profiles considering hormone response elements. While a few of the regulatory elements are derived from genes in insects and birds, most of the sites are mammalian - with 89% of all sites from human or rodent genes.

It is also worth mentioning that most collections do not filter out confirmed binding sites from recognized ones: when a DNA region was found to exercise promoter activity, regions similar to HRE consensus are sought in the long promoter sequence by computational methods. Our aim was to collect sites with binding affinity, whatever their structure is, so in the current dataset only experimentally confirmed binding sites were included into the collection

3.2 Accuracy of Prediction

The Hidden Markov Model provides a versatile method for sequence transition pattern recognition. A specifically designed HMM with its states, emission letters and transition probabilities can best characterize the transition patterns in the nucleotide sequence of interest. We designed and implemented a profile HMM, taking into account specific structure of HRE sequences being recognized.

In the current work HMM approach allowed to achieve 88% of sensitivity with re-value of 1:1217bp (threshold of normalized probability 0.33) and a level of prediction 1:6.4kb with 63% of true positives (threshold 0.36). Its sensitivity and re-values were evaluated as described in previous sections. Considering the trade-off between sensitivity versus specificity, we selected threshold of 0.343 with sensitivity of 79% and specificity of 1 prediction per 3.9kb for future analysis of hormone responsive genes.

In the web-based version of the model, the accuracy level is a user-defined parameter. If in the query sequence, HRE patterns are not reported by the system, the user may increase the sensitivity (by decreasing the threshold) and repeat the analysis. Conversely, the user can reduce the sensitivity level if the detected ERE patterns seem to be false positive predictions. Reduction in sensitivity should decrease the number of potential false positives.

3.3 Analysis of Steroid Hormone Primary Target Genes

In this study, we estimated our model using the reported progesterone responsive genes [20]. Although a particular gene might be hormone-regulated by any of indirect pathways, primary target genes are supposed to contain HRE in their regulatory area. For a list of 380 human PR-regulated genes we selected their promoters (areas [-3000; +500] relative to annotated transcription start sites) from NCBI Genbank database (build 35.1), and scanned them using the strategy described above with optimal values of thresholds for recognition. A set of all human genes was used as a potential control of 'noise' level.

The average number of the found PREs in promoter area for 380 PR-responsive genes from the list is 1.06 while for total set of human genes this value is 0.62 HREs per promoter.

Another negative control is through implementation of the ERE recognition within promoters of PR-responsive genes, because progesterone primary target genes are considered not to be estrogen-regulated. We used database of EREs [10] for exactly the same PWM training and testing procedure and selected thresholds for recognition to keep the same sensitivity value as for PRE prediction - 79%. The average number of EREs is 0.66 per promoter of PR-responsive genes.

The highest frequencies of PREs were found in promoter areas of human CMAH gene (encoding for cytidine monophosphate-N-acetylneuraminic acid hydroxylase) and for AOX1 (aldehyde oxidase 1) - 6 and 5 per promoter region respectively. Also there were 7 genes with 4 predicted PREs (1.8% of total 380), 34 - with 3 found matches (8.9% of 380), 62 with 2 (16.3%) and 118 with only one promoter-located PRE being predicted (31.1% of total 380 reported PR-responsive human genes).

The highest probability of being steroid hormone primary target gene was found for human MMP1 gene encoding for matrix metalloproteinase 1 (interstitial collagenase). Its promoter contains three predicted HREs and two of them are adjacent (which have been previously reported to have very high chance to be functional [21]). Steroid hormone progesterone was previously reported to reduce level of human MMP1 gene expression significantly [22]. The second significant PR-responsive gene NGRF was also reported to be progesterone-regulated [23].

3.4 Proposal for Modeling of Secondary Response

It is well-known that transcription regulatory mechanisms, being rather complicated themselves, when considered from secondary response point of view, become even more intricate. However, with more experimental information becoming available, it is very suggestive to look further and investigate induced effects of the first level of regulation.

In the current list of PR-regulated genes there are at least 8 genes whose product proteins are involved in transcriptional regulation. Among them there is one gene FOSL1 which has been proved to be a primary target. However, even this information can provide important information.

The Fos gene family consists of 4 members: FOS, FOSB, FOSL1, and FOSL2. These genes encode leucine zipper proteins that can dimerize with proteins of the JUN family, thereby forming the transcription factor complex AP-1. As such, the FOS proteins have been implicated as regulators of cell proliferation, differentiation, and transformation (i.e. the processes in which progesterone regulation is extremely important) information. For example, IL-8 gene is also known to be progesterone regulated. However, FOS transcription factor has been recently reported to be involved in regulation of IL-8 gene [24]. Therefore, it is at least reasonable to look at the putative pathway of regulation: progesterone → human FOSL1 gene → Fos transcription factor → regulation of IL-8.

For conclusion, we present a novel program for identification of a class of steroid hormone response elements (HREs) in genomic DNA, including HREs for androgen, glucocorticoid and progesterone. The detection algorithm uses Profile Hidden Markov Model representation of the sequence of interest, and takes into account its specific structure. After series of independent tests on several large datasets, we estimated appropriate combination of sensitivity and specificity as 79% and specificity of 1 prediction per 3.9kb. Users can further investigate selected regions around the identified HRE patterns for transcription factor binding sites based on publicly available TFBS databases, estimate promoter sequences to be hormonally-regulated, and therefore, predict steroid hormone primary target genes.

References

1. Conneely OM (2001) Perspective: Female Steroid Hormone Action. Endocrinology. 142(6):2194-2199
2. Eddy EM, Washburn TF, Bunch DO, Goulding EH, Gladen BC, Lubahn DB, and Korach KS (1996) Targeted disruption of the estrogen receptor gene in male mice causes alteration of spermatogenesis and infertility. Endocrinology. 137(11):4796-4805
3. Pelzer T, Shamim A, Wolfges S, Schumann M, and Neyses L (1997) Modulation of cardiac hypertrophy by estrogens. Adv Exp Med Biol. 432:83-89
4. Cutolo M, Sulli A, Capellino S, Villaggio B, Montagna P, Seriolo B, and Straub RH (2004) Sex hormones influence on the immune system: basic and clinical aspects in autoimmunity. Lupus. 13(9):635-638
5. Maggi A, Ciana P, Belcredito S, and Vegeto E (2004) Estrogens in the nervous system: mechanisms and nonreproductive functions. Annu Rev Physiol. 66:291-313
6. Kearns AE and Khosla S (2004) Potential anabolic effects of androgens on bone. Mayo Clin Proc. 79(4S):14-18
7. Tsai MJ and O'Malley BW (1994) Molecular mechanisms of action of steroid/thyroid receptor superfamily members. Annu Rev Biochem. 63:451-486

8. Alberts,B., Bray,D., Lewis,J., Raff,M., Roberts,K. and Watson,J. (1994) Intercellular signalling. Molecular Biology of the Cell. Garland Publishing, New York
9. Dahlman-Wright K, Siltala-Roos H, Carlstedt-Duke J, and Gustafsson JA (1990) Protein-protein interactions facilitate DNA binding by the glucocorticoid receptor DNA-binding domain. J Biol Chem. 265(23):14030-14035
10. Bajic VB, Tan SL, Chong A, Tang S, Strom A, Gustafsson JA, Lin CY, and Liu ET (2003) Dragon ERE Finder version 2: A tool for accurate detection and analysis of estrogen response elements in vertebrate genomes. Nucleic Acids Res. 31(13):3605-3607
11. Barbulescu K, Geserick C, Schuttke I, Schleuning WD, and Haendler B (2001) New androgen response elements in the murine pem promoter mediate selective transactivation. Mol Endocrinol. 15(10):1803-1816
12. Truss M, Chalepakis G, and Beato M (1990) Contacts between steroid hormone receptors and thymines in DNA: an interference method. Proc Natl Acad Sci USA. 87(18):7180-7184
13. Evans RM (1988) The steroid and thyroid hormone receptor superfamily. Science. 240(4854):889-895
14. Wasserman WW and Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet. 5(4):276-287
15. Favorov AV, Gelfand MS, Gerasimova AV, Ravcheev DA, Mironov AA, and Makeev VJ (2005) A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. Bioinformatics. 21(10):2240-2245
16. Sandelin A and Wasserman WW (2005) Prediction of nuclear hormone receptor response elements. Mol Endocrinol. 19(3):595-606
17. Bono HU (2005) SayaMatcher: Genome scale organization and systematic analysis of nuclear receptor response elements. Gene. 364:74-78
18. Eddy SR (1998) Profile hidden Markov models. Bioinformatics. 14(9):755-763
19. Vega VB, Bangarusamy DK, Miller LD, Liu ET, and Lin CY (2004) BEARR: Batch Extraction and Analysis of cis-Regulatory Regions. Nucleic Acids Res. 32(Web Server Issue):257-260
20. Leo JC, Wang SM, Guo CH, Aw SE, Zhao Y, Li JM, Hui KM, and Lin VC (2005) Gene regulation profile reveals consistent anticancer properties of progesterone in hormone-independent breast cancer cells transfected with progesterone receptor. Int J Cancer. 117(4):561-568
21. Tsai SY, Tsai MJ, and O'Malley BW (1989) Cooperative binding of steroid hormone receptors contributes to transcriptional synergism at target enhancer elements. Cell. 57(3):443-448
22. Lapp CA, Lohse JE, Lewis JB, Dickinson DP, Billman M, Hanes PJ, and Lapp DF (2003) The effects of progesterone on matrix metalloproteinases in cultured human gingival fibroblasts. J Periodontol. 74(3):277-288
23. Bjorling DE, Beckman M, Clayton MK, and Wang ZY (2002) Modulation of nerve growth factor in peripheral organs by estrogen and progesterone. Neuroscience. 110(1):155-167
24. Hoffmann E, Thiefes A, Buhrow D, Dittrich-Breiholz O, Schneider H, Resch K, and Kracht M (2005) MEK1-dependent delayed expression of Fos-related antigen-1 counteracts c-Fos and p65 NF-kappaB-mediated interleukin-8 transcription in response to cytokines or growth factors. J Biol Chem. 280(10):9706-9718