Design and Development of a FPGA-based Cascade Markov Model for Recognition of Steroid Hormone Response Elements

Maria Stepanova
Bioinformatics Research Centre, Nanyang Technological University, 50 Nanyang Drive, Singapore 637553
mari0004@ntu.edu.sg

Feng Lin
School of Computer Engineering, Nanyang Technological University, Block N4, Nanyang Avenue, Singapore 639798
asflin@ntu.edu.sg

Valerie Lin
School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551
CLLin@ntu.edu.sg

Abstract

Steroid hormones are necessary for vital functions of most of vertebrates. Hormone molecules act within cells via interaction with their receptor proteins which are transcription factors. Identification of Hormone Response Elements (HREs) in DNA is essential for understanding the mechanism of hormone-mediated gene expression regulation. We present a systematic approach for recognition of HREs within promoters of vertebrate genes. The proposed approach is based on an experimentally validated dataset and a specifically reconstructed cascade Markov model for HRE recognition with reference to its complex composition. The approach provides a reliable accuracy for HRE prediction, and may be extended to other sequence motifs with repeated or multi-component structure. The developed FPGA implementation of the Markov model is tested using a Virtex-4 board. The model is trained for prediction of HREs in promoters of hormone-responsive genes, and for further study on direct targets for androgen, progesterone and glucocorticoid hormones.

1. Introduction

Hormone response elements are binding sites for transcription factors belonging to the family of steroid hormone receptors. In general, recognition of transcription factor binding sites (TFBSs) through experimental research is a slow and tedious task. Through the advent of computational biology, many statistical models and algorithms have been developed to examine regions of the DNA that might harbor TFBSs in a quick and efficient manner, but current computational methods for prediction of TFBSs in DNA are still not as reliable as experimental approaches.

The major challenge for *in silico* recognition of TFBSs is their weak conservation which results in low selectivity of statistical models. In particular, the diverse patterns of the TFBSs make it difficult for naïve statistical approaches to specifically distinguish functional binding sites and neutral DNA. As it has been shown in many experiments [1–3], the true predictions are often accompanied by large numbers of false positives, thus making the results of computational prediction nearly meaningless. An essential way to improve TFBS prediction accuracy for a particular transcription factor (or a family of those) is to enhance the model selectivity by employing additional features that are specifically innate to the pattern of interest, though at a cost of losing generality of modeling.

The general idea of HRE acting as a DNA dimer has been known for years, but only thorough mutational analysis might help to explore the tiny composition of a functional response element. In literature, an exhaustive mutation analysis of HRE was reported by Nelson et al. [4]. The authors performed selection assays for detection of response elements to three steroid hormones with both high binding affinity and specificity, but in their work, no flexibility of the right half-site of the HRE dimer was allowed. Although this assumption had a base, as the right half-site had previously been reported to be well conserved [5], this half-site still admitted some single nucleotide substitutions in non-contact points [6]. In fact, the functional HRE sequences display both conservation and diversity, and the latter makes it a challenging task

978-1-4244-1694-3/08/$25.00 ©2008 IEEE

to establish a computational model for reliable HRE prediction.

We have previously reported our findings concerning preferences of hormone receptors towards their target DNA sequences [7], mainly based on Position Weight Matrix (PWM) prediction method. Statistic characteristics at each nucleotide position of the HRE sequence were studied similar to those from mutational analysis experiments. While the position frequency distributions are undoubtedly important for prediction of HREs and allow constructing an easily interpretable HRE motif profile, we believe that statistic features between the nucleotides, or nucleotide transition patterns, may provide a new dimension in modeling of DNA sequence motifs.

Hence, in this paper, we study a method for predicting HRE using the profile Markov model which is specifically designed according to the dimeric structure of the HRE consensus. We propose a cascade Markov model with specific state transition matrix for each constituent of the complex HRE structure, namely, the two half-sites and flanking regions around them, and implement the model on FPGA as a parallel architecture

2. Construction of HRE Training Database

Accuracy of a statistic model largely depends on construction of the training datasets. One can easily achieve very high sensitivity and specificity of motif prediction with just a few sequences used for training and testing, but this result tends to be meaningless because the relative variance will also be very high.

There are a few public databases of TFBSs or *cis*-regulatory modules, among them are TRANSFAC [1], JASPAR [2], PReMod [3]. However, much of the time there is no collaboration between these databases, resulting in duplicated data that needs to be sifted through in order to avoid errors. By scanning through these public databases, we have found only approximately 50 HREs. In particular, publicly available version of TRANSFAC contains GRE matrix calculated for 38 binding sites; JASPAR has ARE weight matrix with 24 sequences used. Genomatix TFBS database used for MatInspector tool [4] is capable of for predicting of response elements of the glucocorticoid, androgen and progesterone receptors (ARE, GRE and PRE), but this database was constructed based on only one experimental result [5] with a strong limitation for HRE selection process. That is, a reliable and exhaustive (but non-redundant) dataset of HREs for training and testing purposes is a must for further development of any HRE prediction model; otherwise, any possible experiments will be inconclusive.

For this purpose, a set of DNA sequences of experimentally verified HREs was collected from literature. At first, we collected only progesterone response elements, but this hormone looked like to be the least investigated among all steroid hormones, so we considered a canonical assertion that glucocorticoid, androgen and progesterone receptors tended to share the same response elements on DNA. The collected set contained 661 HRE sequences extracted from more than 200 literature sources. The amount of progesterone response elements in the collection was 66, in addition to 377 glucocorticoid HREs and 218 androgen HREs. These sequences are housed in our database Tiger HRE DB [6].

3. Building a Cascade Markov Model for HRE Prediction

A Markov chain is an extension of a finite state machine with the Markov property [1]. The Markov model can be used to represent the probability of transition between adjacent nucleotides in a DNA sequence. However, a naïve Markov model does not associate the transition probabilities with the position of the nucleotides in the sequence. It is thus necessary to define a more sophisticated model to incorporate the targeted complex sequence pattern in computation of the transition probabilities, which leads to our solution for the HRE prediction. With reference to the suggested HRE structure [2], we designed a cascade model in which the two HRE half-sites, the internal spacer, and the two flanking regions were considered as five components of the HRE model. Each component model had its own transition parameters, while the top-level model computed the transition probabilities between them. Transition probabilities for the component HRE models and probabilities for inter-model transitions were trained using maximum likelihood estimation.

The defined positions of half-sites are highly position-specific due to the presence of contact points typical for the HRE. The binding site for HRE consists of three domains: two half-sites and a spacer. In our model, we also considered flanking regions since they were reported to be involved into the process of binding [3]. Therefore, the number of domains to be included in the model is five. Each of these domains is expected to have its own properties (i.e. internal transition probabilities), so the Markov models for the corresponding domains have to be examined and trained separately. We created a Markov model for

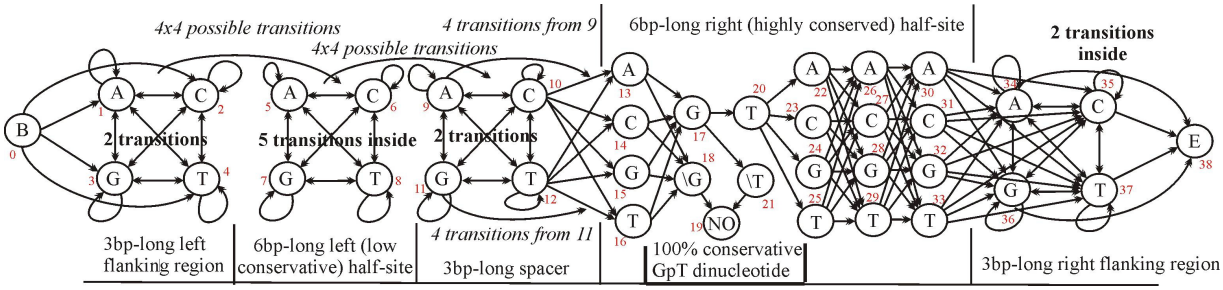


Fig. 1. A five-stage cascade Markov model for HRE recognition

representation of steroid HRE motif as shown in Fig. 1.

The model structure satisfies the described conditions: different domains are considered independently, and transition probabilities between them are determined. The transition probabilities for the first component model, which corresponds to states 1 through 4, are determined using 3bp flanking regions at the beginning of the HRE, the probabilities for the second model (states 5-8) correspond to the left half-site, etc. The fact that the right half-site is highly conservative is reflected in the structure of the fourth component model, which corresponds to the states 13-31. The transition probabilities in this part of the model are position-dependent, unlike the probabilities in the other parts of the sequence. Based upon the collected PFMs, where GpT dinucleotide in the right half-site is observed in 99% cases, we decide that the presence of GpT should be declared as an essential feature. If this dinucleotide is absent at the specified position in the sequence being tested, which corresponds to states 18 and 21 in the model, then the sequence cannot be an HRE (and referred to transition to the state 19).

Given the length of each constituent part of HRE, namely the flanking regions are of length 3 and so on, we define the number of possible transitions inside each of the Markov models. In particular, there are two transitions allowed in the flanking regions models, so that three states are visited and the 3bp-long sequence is returned.

There exist certain differences in lengths of training sequences as not all of them have flanking regions annotated in the literature. Hence, normalization procedure for resulting Markov probability value is used – logarithm of probability is divided by the sequence length. Also prior Markov distribution, in case of the starting position of the sequence different from B (i.e. if no left flanking region annotated), is taken from position frequency matrices; otherwise this distribution is considered as uniform.

After the sequence is processed by the model, its Markov probability is stored into a database and is subject to a threshold for decision making. The model is designed to be used as a part of a multiple-feature prediction framework where the putative HREs are first selected by a simple method, such as PWM, with high sensitivity. However, if the model is intended for screening of a long DNA region, an optimization based on a systolic array or a similar technology should be used.

4. Fine-Grained Parallelization with FPGA

In the software implementation, we often limit the prediction accuracy to avoid the prohibitively long execution time. Based on the Field-Programmable Gate Arrays (FPGA), we developed a hardware-accelerated solution to exploit the modeling capability of the proposed HRE Markov model. The entire model consists of consecutive component models, but in fact these models can be processed in parallel if additional input prepocessing is involved. We can expect the acceleration benefiting from fine-grained parallelization, with optimization of logic interconnections which is a significant advantage of the FPGA technology in comparison with other application-specific approaches to hardware design.

4.1 Schematic Design of the Cascade Markov Model

We use the ADM-XRC-4 PCI board with the Xilinx Virtex-4 chip which contains 135,168 logic cells and 5,184Kbit of embedded RAM. Initially the RTL (register-transfer level) description in Verilog HDL was simulated by creating test benches to validate the system. Then, after the synthesis engine had mapped the design to a netlist, the netlist was translated to a gate level description where simulation was repeated to validate the synthesis results. Finally, the design was laid out in the FPGA at which point propagation delays were added and the simulation ran again with these values back-annotated onto the netlist.

The configured FPGA chip is then used as a co-

processor for the computing system. The C++ application reads the input DNA sequences, converts the letters of DNA alphabet into binary numbers (two bits per nucleotide positions, and 21 positions per input), and sends the 42bit-long input vector to the FPGA board. It also obtains the output Markov probability from the board, and proceeds it to the decision making scheme for HRE recognition. Transition probabilities were obtained using the maximum likelihood approach, and stored in the memory as 32 bit unsigned fixed point numbers.

The Markov model implemented on FPGA consists of seven main units: five units for the component Markov models for the two half-sites of HRE, two flanking regions, and the spacer between the half-sites; the memory unit, which stores the transition probabilities for each component model; and the merging unit, which receives the resulting Markov probabilities from the five component model units, and returns the overall Markov probability for the input. The input to the FPGA-implemented Markov model is

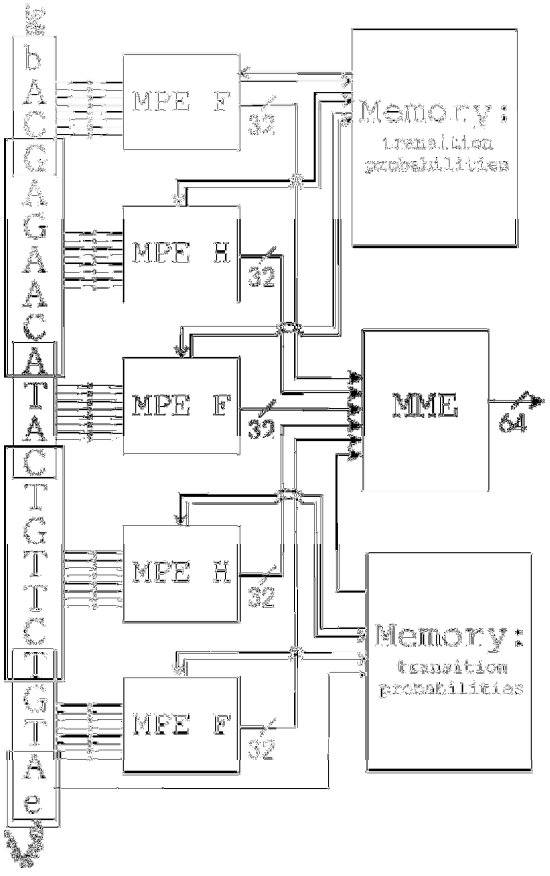


Fig. 2. The schematic diagram of FPGA implementation of the HRE Markov model

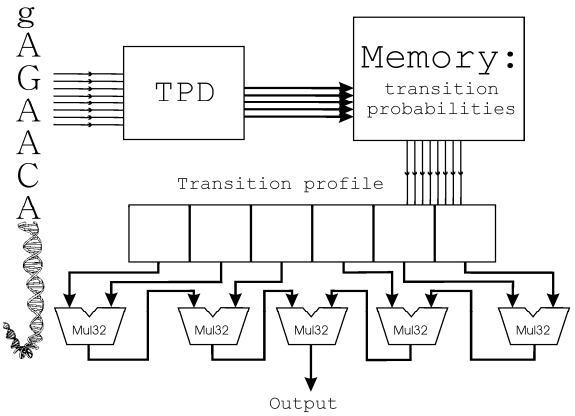


Fig. 3. The dataflow scheme of the Markov processing element MPE H for the cascade Markov model. The element processes a 6bp-long half-site of an HRE. The MPE F element has similar structure, though uses only two multiplications for the 3bp-long flanking regions. The half-site sequence is in upper case, while the letter in lower case is from the preceding element.

a DNA sequence. It is then split into five partially overlapping subsequences, each of which is processed by a component Markov model, as shown in Fig. 2.

By involvement of the overlapping areas instead of a consequence of component Markov models, we can make these models to operate in parallel. The overlapping area for the splitting of the input sequence into subsequences is for the Markov property of the model since the current state depends on the previous state only. When the first-order Markov model is considered for HRE modeling, one previous nucleotide is merged together with the following component constituent as its beginning state. For the higher-order Markov model, it would be necessary and sufficient to only increase the size of overlapping for the sequence splitting.

5. Logic Interconnection in the Markov Model

Each of the five component HRE Markov models was implemented in a form of Markov processing element (MPE). Two types of processing elements were proposed for the two main types of model atomic constituents of HRE pattern, in particular, for the 3bp-long neighboring regions (MPE F, or Markov processing element for flanking areas and the spacer), and for the 6bp-long HRE half-sites (MPE H, or Markov processing element for the half-site, shown on the Fig. 3). All Markov processing elements have their

own memory storage units for corresponding transition probabilities (that is, the physical memory is distributed, as it allowed us to use less area resources in comparison with a common memory unit for the entire Markov model), and are connected to the Markov merging element (MME) which is in turn connected to the output of the Markov model FPGA module.

Each MPE has a transition probability detection (TPD) unit for preprocessing of the input sequence (Fig. 3). For each input DNA sequence, the TPD unit returns the memory indexes, which are then used for extraction of the corresponding transition probability values from memory. Thus, the transition profile of the input sequence is generated and processed. In case of higher order models, more preceding elements should be involved, and the transition profile becomes (in general, exponentially) larger.

For calculations of Markov probabilities, the unsigned fixed point notation is used. Here, inside of MPEs, we use 32 bits to represent a numerical value of probability with 32 fractional bits, while for the merging element we extend it to 64 bits. The reason is that after a series of tests, 32 fractional bits have been found to cause notable underestimation of resulting Markov probabilities for most of HRE inputs. Indeed, for the length of the input sequence of 21 bp, the precision of 64 bits used for fraction part is enough, though for longer inputs it may be reasonable to involve logarithmic transformations so as to replace multiplications by additions.

Arithmetic operations used here are standard procedures: 36bit×36bit multiplication is implemented as a finite state machine which regulates the sequence of pairwise 18bit×18bit multiplications accomplished by the two dedicated hardware multipliers. Multiplications were not replaced by logarithmic additions in order not to overload the limited number of logic gates. When 64 bits are used instead of 32 for fractional number representation, four embedded multipliers are involved, instead of two, into Markov probability calculations. It is necessary to note, though, that the architecture of frequently used arithmetic operations is always a trade-off between limited resources and corresponding latency.

The FPGA clock frequency was set to 100 MHz

Summary of the ensued implementation of the cascade Markov model is as follows:

```
Logic elements: 21,396 of 135,168 (16%)
RAM:            160Kbit of 5184Kbit (3%)
I/O pins:       120 of 960 (13%)
DSP slices:     42 of 96 (44%).
```

6. Analysis of Computational Performance and Complexity

The complexity of the Markov model training is $O(\omega^2*L*n)$ where L is the length of the pattern of interest, ω is the size of the alphabet, and n is the length of the sequence to process. The computational complexity is higher than that of conventional position-specific models because the amount of possible transition patterns is square to the cardinality of the alphabet. However, for the developed five-stage model of HREs, the complexity is lower than it might be in average, because we consider the GpT dinucleotide to be essential for binding as it is the most important contact point for protein-DNA interaction; therefore, we can eliminate the candidate sequences which do not have this pattern in the required position. It allows us to gain the average speed-up as much as 16 times only with the software implementation.

The FPGA speed-up for the multi-stage Markov model is examined using 1Mb of randomly generated sequences of the DNA alphabet. The list of the testing sequences is preprocessed by the encoding application, and then the sequences are submitted into the board, one at a time, after a handshake signal for the completion of calculations for the previous input is received. In this case, we do not benefit from the possible re-distribution of the data flow, though it is possible to decrease the latency even more if the MPEs modules start processing the next input while the MME module calculates the current output. For testing purposes, we also did not include the selection procedure based on high conservation of the GpT dinucleotide in the right HRE half-site.

The results of test on computational performance of different implementation of the models can be summarized as follows:

- with the IBM 4-way server (4 CPUs each of 3.17GHz, 3.25GB RAM, Win 2003 Server), it takes about 60 seconds to process 1 Megabyte of DNA text by the five-stage Markov model;
- by the FPGA-based hardware acceleration, it allows to screen 1Mb of DNA within 8 seconds of runtime, thus allowing up to 8X speed-up.

That is, the achieved speed-up value is due to not only the parallelization itself, but also involves the advantage of application-specific logic-interconnection design benefiting from the use of FPGA. The scalability of the design was also measured using complete genomic sequences of human and mouse that are nearly 3 billions of base pairs-long each, and a similar acceleration of nearly an order of magnitude was achieved.

7. Experimental Results

To evaluate the sensitivity of the developed cascade model, we used the experimental setup with different proportions of the collected dataset of HREs used for training and testing. The typical values of the Markov probability for functional HREs were found to be around 0.35. The accuracy of HRE prediction by the five-stage Markov model is summed up in Table 1, and if Fig. 4, the ROC curve is given.

Table 1. Accuracy of HRE prediction by the five-stage cascade Markov model.

β	Sensitivity, %			RE, bp^{-1}
	70/30	50/50	100/100	
0.30	95.3	93.9	97.3	192
0.31	94.9	93.2	96.5	293
0.32	94.2	92.2	95.9	458
0.33	89.3	85.0	93.1	737
0.34	84.8	75.5	88.0	1325
0.35	73.0	66.3	79.2	2942
0.36	63.5	61.8	66.6	4254
0.37	60.0	58.3	61.0	6734
0.38	51.3	46.8	55.9	15342
0.39	41.1	37.0	42.2	35876
0.40	33.4	28.6	33.8	723462

Notation: β - the Markov probability threshold, RE – random expectation value (1 prediction per RE base pairs of randomly generated nucleotide sequence).

In Table 1, the results of three series of independent accuracy tests are given. First, the tests were performed with 70% of the HRE collection used for training and the rest for testing. Second, the training/testing split was into half and half (50/50 column), and the sequences for training sample were randomly selected from the database without replacement. Finally, the entire HRE dataset was used for training and testing; it is noted in the table by the 100/100 column. For the cascade five-stage Markov model, we achieved the average sensitivity of 85% with random expectation of 1:1325bp, and a level of prediction rate of 1:6.5kb with 60% of correctly predicted HREs. The AUC value for the model is 0.924, which is lower that of the position weight matrix methods (reported 0.953 and 0.941, respectively [1]), but still in the range of good prediction rates for the problems of TFBS recognition.

The five-stage Markov model provides a versatile method for modeling the transition pattern of the HRE sequence. However, it is reasonable to believe (and this strategy is already widely used in the design of ensemble models [2]) that if two pattern recognition methods consider different properties of the object under recognition, their combination may outperform each single method. Here we use a nucleotide frequency model based on position weight matrix for enhancing the Markov model by establishing a position-transition ensemble for HRE recognition.

In particular, for combined HRE prediction, we designed a unanimous voting system. A HRE is predicted only if the single- and dinucleotide PWMs support the HRE predictions by the cascade Markov model. Using this scheme, we eliminate a large amount of false positives. The combined position-transition prediction accuracy was tracked in comparison with each of the methods involved.

Figure 4 shows the ROC curves for different possible combinations of the statistic methods of HRE prediction. The ROC curve for the single nucleotide PWM is labeled with PWM1. The curve for the dinucleotide PWM is labeled with PWM2. The curve for the combination of position weight matrix methods is labeled with "PWM1+PWM2". Label MM corresponds to the Markov model described in this paper. Finally, the label "Tiger" denotes the ROC curve for the combination of all the three prediction methods implemented as a unanimous voting scheme.

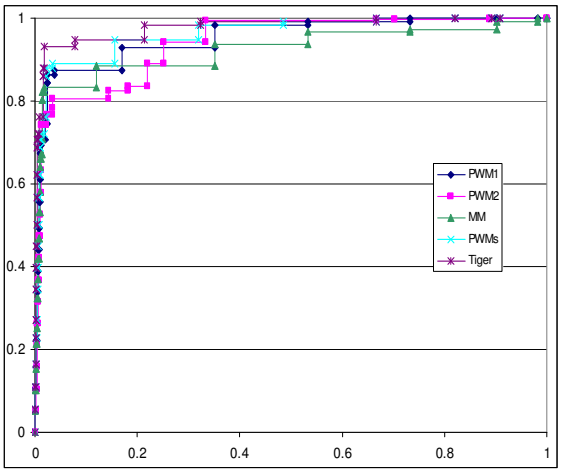


Fig. 4. The ROC curves for the combined prediction of HREs in comparison with the single- and dinucleotide PWM methods and the five-stage cascade Markov model alone.

For the combined predictions, such as "PWM1+PWM2" and "Tiger", the generation of ROC curves requires selection of a combination of two or three threshold values simultaneously in order to

balance sensitivity and specificity. To find an approximate solution, we change one parameter at a time by a small amount that is enough for only one sequence from the training data to be re-classified, and estimate the average specificity for each sensitivity value.

The combination of the three prediction methods eliminates a large number of the false positives while keeping the number of true positives. The sensitivity and random expectation values of the combined method are the functions of three variables and can be tuned in a three-dimensional space. For the problem of HRE recognition in gene promoters, we set the following values for recognition thresholds: for PWM1, the recognition threshold was 0.85; for PWM2, the recognition threshold was 0.73; for MM, the recognition threshold was 0.35. Using these values, we received a combination of sensitivity 73% and random estimation of 1 hit per 4.74kb. The corresponding AUC is 0.968, in comparison with the five-stage MM's AUC of 0.924, and PWM1+PWM2 AUC of 0.963.

8. Discussion and Conclusion

Several stochastic ways of modeling TFBS motif profiles exist. As TFBS sequences are typically short and degenerate, signal-to-noise ratio for their *de novo* prediction is usually low. Therefore, TFBSs are intrinsically hard to be accurately predicted by methods of statistic modeling. In this work, we presented a method for TFBS prediction based on a Markov model approach, and test it by the example of hormone response elements.

Unlike the PWM-based methods which consider only single or multiple nucleotide patterns of the HRE sequence consequently, in the proposed cascade Markov model, several component models can be designed to represent the actual constituents of the complex response element architecture. With this multi-stage approach, the HRE dimeric structure can be accurately described. When used in combination with the position-specific PWM methods, the multi-stage Markov model allows increasing the HRE prediction accuracy notably.

Based on the tests performed using an extensive HRE dataset, our findings are indeed promising. For comparison, the results of analysis of TRANSFAC-based prediction performance provided by Rahmann et al. [3] can be used. In that paper, the authors showed that a specificity level higher than 0.99 can be achieved for only 43 profiles (i.e. 7%) among 623 used for testing. Some profiles of high interest in practice like nuclear receptor binding sites were not included in that high-quality group. All other profiles resided below the specificity level of 99%, which was nearly disappointing because the corresponding prediction specificity was as low as 1 prediction per 0.1kb.

The proposed FPGA architecture of the cascade model allows for certain versatility in hardware design. In particular, the partially parallel structure can be redesigned for trade-off between latency and area. Furthermore, higher-order Markov models might be required for *in silico* prediction of sequence motifs, as it has been shown by Rajapakse and Ho for the case of genomic signals [4]. With reference to the described multi-stage model, the increase of the order of Markov model would result in a slightly modified memory management scheme and additional logic interconnections for distribution of the input to processing elements. For example, if we consider a second order model, with reference to Fig. 2, we only need to extend the borders for the operation coverage for the processing elements (vertical rectangles along the DNA sequence), and modify the size and mechanisms of control for the memory units.

To our knowledge, design and development of the cascade Markov model is a pioneering work. However, due to the model complexity, large genomic sequences may cause prohibitively long computing time. Thus, in order to make the approach useful in practice, we also studied the applicability of hardware-acceleration methods using FPGA technology. The conclusion is that the FPGA has been successful for enhancement of HRE prediction. The described hardware design refers to both advantages of FPGA – fine-grained parallelization and application-specific inter-connections. The achieved speed-up up to an order of magnitude allows us to conclude that the HRE modeling approach with the proposed hardware-acceleration architecture allows not to compromise best possible accuracy.

References

1. Freedman LP and Luisi BF (1993) On the mechanism of DNA binding by nuclear hormone receptors: a structural and functional perspective. *J Cell Biochem*. 51(2):140-150.

2. Smirnov AN (2002) Nuclear receptors: nomenclature, ligands, mechanisms of their effects on gene expression. *Biochemistry (Mosc)*. 67(9):957-977.

3. Nelson CC, Hendy SC, Shukin RJ, Cheng H, Bruchovsky N, et al. (1999) Determinants of DNA sequence specificity of the androgen, progesterone, and glucocorticoid receptors: evidence for differential steroid receptor response elements. *Mol Endocrinol*. 13(12): 2090-2107.

4. Lieberman BA, Bona BJ, Edwards DP, and Nordeen SK (1993) The constitution of a progesterone response element. *Mol Endocrinol*. 7(4):515-527.

5. Roche PJ, Hoare SA, and Parker MG (1992) A consensus DNA-binding site for the androgen receptor. *Mol Endocrinol*. 6(12):2229-2235.

6. Stepanova M, Lin F, and Lin VC (2006) In silico modeling of hormone response elements. *BMC Bioinformatics.* 7(4):S27.

7. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, et al. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*. 31(1):374-378.

8. Sandelin A, Alkema W, Engstrom P, Wasserman WW, and Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*. 32(Database issue):D91-D94.

9. Ferretti V, Poitras C, Bergeron D, Coulombe B, Robert F, and Blanchette M (2007) PReMod: a database of genome-wide mammalian cis-regulatory module predictions. *Nucleic Acids Res*. 35(D):122-126.

10. Cartharius K, Frech K, Grote K, Klocke B, Haltmeier M, et al. (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*. 21(13):2933-2942.

11. Stepanova M, Lin F, and Lin V (2006) Establishing a Statistic Model for Recognition of Steroid Hormone Response Elements. *Comput Biol Chem*. 30(5):339-347.

12. Eddy SR (1998) Profile hidden Markov models. *Bioinformatics*. 14(9):755-763.

13. Clote P (2000) Computational molecular biology: an introduction. John Wiley, New York.

14. Gusfield D (1997) Algorithms on Strings, Trees, and Sequences. Cambridge University Press.

15. Haelens A, Verrijdt G, Callewaert L, Christiaens V, Schauwaers K, et al. (2003) DNA recognition by the androgen receptor: evidence for an alternative DNA-dependent dimerization, and an active role of sequences flanking the response element on transactivation. *Biochem J*. 369(1):141-151.

16. Stepanova M, Lin F, and Lin VC (2006) Tiger HRE Finder - A Tool for Identification of Hormone Receptor Binding Sites in Query Sequences. *In Proc. of IMSCCS'06*. 1: 22-26.

17. Kuncheva LI and Whitaker CJ (2003) Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*. 51(2):181-207.

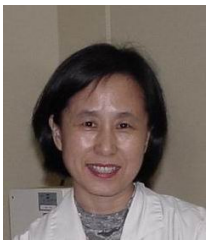
18. Rahmann S, Muller T, and Vingron M (2003) On the power of profiles for transcription factor binding site detection. *Stat Appl Genet Mol Biol*. 2(1): Article 7.

19. Rajapakse JC and Loi Sy Ho (2005) Markov Encoding for Detecting Signals in Genomic Sequences. *IEEE/ACM Trans Comput Biol Bioinf*. 2(2):131-142.

Authors

Maria Stepanova received her MSc degree in applied mathematics in Moscow Institute of Physics and Technology, Russia. She's currently a PhD student with the School of Computer Engineering, Nanyang Technological University (NTU), Singapore. Her research interests include bioinformatics, machine learning algorithms, mathematical statistics, and embedded systems for biomedical research.

Feng Lin received his PhD degree in computer engineering from Nanyang Technological University (NTU), Singapore. He is currently an associate professor with the School of Computer Engineering, Nanyang Technological University, Singapore. His research interests include bioinformatics, high-performance computing, visualization, and embedded systems for biomedical research.



Valerie Lin received her PhD degree from University of Reading, United Kingdom. She is currently an assistant professor with the School of Biological Sciences, Nanyang Technological University, Singapore. Her research interest is primarily focused on endocrine and paracrine regulation of mammary development and breast cancer.