HMMotif: predicting sequence motifs governing constitutive exon splicing

Jing Xing[1,2], Lei Huang[3], Jianhua Lu[1], Xiaobo Zhou[4], Hongwei Li[5*] and Jiawen Bian[5*]
1. Institute of Statistics, Hubei University of Economics, Wuhan, 430205, CHINA
2. Institute of Geophysics & Geomatics, China University of Geosciences, Wuhan, 430074, CHINA
3. Department of Information Science, School of Mathematical Sciences and LMAM, Peking University, Beijing, 100871, CHINA
4. Department of Diagnostic Radiology, Center for Bioinformatics & Systems Biology, Wake Forest University-School of Medicine, Winston-Salem, NC 27103, USA
5. School of Mathematics and Physics, China University of Geosciences, Wuhan, 430074, CHINA
*hwli@cug.edu.cn, jwbian@cug.edu.cn

Abstract
A computationally effective algorithm HMMotif is proposed to detect two kinds of splicing regulatory factor, which are ESEs and ESSs. HMMotif can make full use of the contextual information from the vicinity of 5' and 3' splice sites on the DNA to predict the hidden state of ESEs and ESSs simultaneously. A constitutive exons database is used for the extracting the EST of the concern area from UCSC database. Experiments on this data verified the effectiveness by comparing HMMotif with the traditional threshold based method.

Keywords: Splicing regulatory elements, ESE, ESS, Hidden Markov Model.

Introduction
One of the fundamental steps in the transfer of genetic information from DNA to protein is the splicing of RNA transcripts. In this process, relatively small exons (~100 nt) are selected from among generally much larger introns (thousands of nucleotides) and are joined to form mature mRNA. Splicing relies on the correct identification of exons and must be exactly recognized within pre-mRNAs despite being extremely short compared with intronic regions.[1] In addition to splice site signals at the exonic 5' and 3' ends, accurate discrimination of exons and introns requires additional auxiliary elements.[2-4] Sequences around the splice junctions-the 5' and 3' splice sites (5'ss and 3'ss) are also important for splice site recognition.[5] The sequence or structure context in the vicinity of the 5'ss and 3'ss motifs is known to play an important role in splice site recognition.[6]

It is now established that sequences within exon bodies have a prominent role in promoting exon definition and inclusion in mature transcripts. The best understood the so-called exonic splicing enhancers (ESEs) represent exonic elements. ESE sequences, which enhance splicing at nearby sites,[7] are an important component of its context. ESEs represent binding sites for SR proteins which are thought to have a role in the initial steps of spliceosome assembly.[8-10] Sequences that act as exonic splicing silencers (ESSs) have also been described and studied[11-13] but are less well characterized compared with ESEs. In many instances, ESSs have been shown to bind negative regulators belonging to the heterogeneous nuclear ribonucleoprotein (hnRNP) family.[14-15] The function of ESEs and ESSs appears to be especially important for the regulation of alternative splicing events but these sequences probably also play a relevant role in the definition of constitutive exons.

Recent research on the searching of ESEs and ESSs is mainly based on comparing the frequencies between the potential area and the background area such as the vicinity of exons and introns or the area between exons and pseudoexons area.[1,10,16-17] At such conditions, a threshold or cutoff value is needed for the discrimination of motifs from the reference sequence, however, a good threshold is hard to obtain for avoiding the missing of any true motif.[5]

Hidden Markov model (HMM) is widely used in many fields such as speech and handwriting recognition, text classification as well as DNA and protein classification.[18] It is observed that both ESEs and ESSs have a large frequency differential between exon area and intron area while the proportion of ESEs in exons is much higher than that in introns and inversely for ESSs.[10] A new algorithm HMMotif is developed for motif finding of ESEs and ESSs by incorporating the frequency differention of the two kinds of motifs above with HMM. The frequency differentiation for ESEs is considered to be between the vicinity of 5' and 3' end of exons and the vicinity of 5' and 3' end of introns.[5] With the introducing of HMM, hard threshold for motifs inference can be avoided and more probability power is also expected to gain from the contextual information of the sequence concerned.

Material and Methods
Problem formulation and HMMotif model specification: For ESEs and ESSs, we take 100 nt downstream of 5' and upstream of 3' for the sequence in exons. For sequence in introns, we also take 100 nt upstream of 5' and downstream of 3'. For the sequence in exons, the whole exons sequence is taken if the length of the sequence is less than 100 nt. All the sequence are taken out and then joined to form a new sequence for ESEs and ESSs searching respectively.

HMMotif is used to predict ESEs and ESSs simutaneously. For each of them, the length of each possible motif is denoted as l, the number of all possible motifs on the considered genome is noted as L. We consider three states for each l -length possible motif: (1) ESE, (2) normal

expressed l-length sequence (not a motif), (3) ESS. For simplicity, we denote the three states as {aa, ab, bb}. Aim is to predict the hidden state for each l-length sequence. The hidden state I and observation for each position are noted as:

$\boldsymbol{S} = \{s_t\}(t = 1, 2, \cdots, L) \in \{v_i\}(i = 1, 2, \cdots, I)$

$\boldsymbol{O} = \{\boldsymbol{o}_t\}(t = 1, 2, \cdots, L)$ respectively.

where $\{v_i\}_{i=1}^{I}$ are all states considered. The underlying regulatory characters are taken as the hidden states. For simplicity, we note state {aa} as state 1, state {ab} as state 2 and state {bb} as state 3 in the following initial state distribution and state transition matrix.

Initial state distribution:

$\boldsymbol{\pi} = \{\pi_1, \pi_2, \pi_3\}$, $\pi_i = \mathrm{P}(s_1 = v_i \mid t = 1)$

State transition matrix:

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = (\boldsymbol{A}_1^T, \boldsymbol{A}_2^T, \boldsymbol{A}_3^T) \ ,$$

$a_{ij} = \mathrm{P}(s_{t+1} = v_j \mid s_t = v_i)$

Emission probability distribution:

$\boldsymbol{B} = \{b_{s_t}(\boldsymbol{o}_t)\}$

The observation to be considered for each l-length sequence includes the frequency of the l-length sequence at the vicinity of two ends of exons and the frequency at the vicinity of two ends of introns (both for ESEs and ESSs). $b_{v_i}(\boldsymbol{o}_t)$ is calculated as a conditional probability, given the hidden state:

$$b_{s_t}(\boldsymbol{o}_t) = \mathrm{P}(\{q^t, r^t\} \mid s_t = v_i) = \begin{pmatrix} q^t + r^t \\ q^t \end{pmatrix} u_i^{q^t} (1 - u_i)^{r^t} \quad (1)$$

where q^t and r^t are the frequencies of the l-length t-th sequence on the exon part and intron part respectively. $\{u_i\}_{i=1}^{I}$ is the binomial distribution parameter for each l-length sequence on the sequence.

Prior distribution of HMM: Initial distribution of π is taken as Dirichlet distribution with hyper-parameter $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3)$, $\boldsymbol{u} = (u_1, u_2, u_3)$ is taken conjugately according to a beta distribution with hyper-parameter $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ as follows:

$$\mathrm{P}(\boldsymbol{\pi} \mid \boldsymbol{\delta}) = Dirichlet(\boldsymbol{\pi} \mid \boldsymbol{\delta}) \quad (2)$$

$$\mathrm{P}(u_k \mid \alpha_k, \beta_k) = Beta(u_k \mid \alpha_k, \beta_k) \quad (3)$$

where $\boldsymbol{\delta} = (100, 1000, 100)$ by assuming that most l-length sequence will be at normal state. We also set α = (600, 500, 400) and β = (400, 500, 600) by assuming that the probability of state {aa} occurring at the stated position is some larger than that it is not occurring and is vice versa for state {bb}. For the initial distribution of state transition matrix, the initial distribution of $\boldsymbol{A}_i$ is as follows:

$$\mathbf{P}(\boldsymbol{A}_i \mid \boldsymbol{\gamma}_i) = Dirichlet(\boldsymbol{A}_i \mid \boldsymbol{\gamma}_i) \quad (4)$$

where $\boldsymbol{\gamma}_1 = (1000, 100, 100)$, $\boldsymbol{\gamma}_2 = (100, 1000, 100)$ and $\boldsymbol{\gamma}_3 = (100, 100, 1000)$. Since the sum of elements in $\boldsymbol{A}_i$ should be equal to probability 1, a normalization for $\{\boldsymbol{A}_i\}_{i=1}^{I}$ is performed after each iteration of HMMotif.

It is noted that the number level of the hyper-parameters above looks some large. Actually, the ratios of the random numbers produced by the corresponding hyper-parameters will be closer to the ratios of the numbers in the hyper-parameters with the increase of the number level of hyper-parameters.

Estimation of HMM parameters: For simplicity, the model parameters of HMM are $\lambda \overset{\Delta}{=} (\boldsymbol{\pi}, \boldsymbol{u}, \boldsymbol{A})$ and one learns the unknown HMM by using EM algorithm and computing the maximum likelihood estimation when the observed data are incomplete. [18] The aim is to find the model parameter λ maximizing the observation probability i.e. $L(\boldsymbol{o}, \lambda) \square \mathrm{P}(\boldsymbol{o} \mid \lambda)$ or log $\mathrm{P}(\boldsymbol{o} \mid \lambda)$ where the later one is usually used when the length of the observation is large. We use a special case of EM algorithm, Baum-Welch algorithm[19] to learn the unknown parameters. For the training of HMM, the following auxiliary function $Q(\lambda, \bar{\lambda})$ is used as the objective function for the optimization of the HMM parameters.

$$Q(\lambda, \bar{\lambda}) = \Sigma_{\boldsymbol{s} \in \boldsymbol{S}} \log \mathrm{P}(\boldsymbol{O}, \boldsymbol{S} \mid \bar{\lambda}) \mathrm{P}(\boldsymbol{S} \mid \boldsymbol{O}, \lambda) \quad (5)$$

It is proved that maximizing the following auxiliary function $Q(\lambda, \bar{\lambda})$ can lead to the increase of the likelihood $\mathrm{P}(\boldsymbol{O} \mid \lambda)$, i.e. $\max_{\bar{\lambda}} Q(\lambda, \bar{\lambda}) \rightarrow \mathrm{P}(\boldsymbol{O} \mid \bar{\lambda}) > \mathrm{P}(\boldsymbol{O} \mid \lambda)$. [19]

Given model parameter set λ , $\mathrm{P}(\boldsymbol{O}, \boldsymbol{S} \mid \lambda)$ can be calculated as:

$$\mathrm{P}(\boldsymbol{O}, \boldsymbol{S} \mid \lambda) = \pi_{s_0} \prod_{t=1}^{L} a_{s_{t-1} s_t} b_{s_t}(\boldsymbol{o}_t) \quad (6)$$

Replacing term $\mathrm{P}(\boldsymbol{O}, \boldsymbol{S} \mid \lambda)$ in (5) with (6), (5) can be rewritten as:

$$\begin{aligned} Q(\lambda, \bar{\lambda}) &= \sum_{\boldsymbol{s} \in \boldsymbol{S}} \log \bar{\pi}_{s_0} \mathrm{P}(\boldsymbol{O}, \boldsymbol{S} \mid \lambda) \\ &+ \sum_{\boldsymbol{s} \in \boldsymbol{S}} \sum_{t=1}^{L} \log \bar{a}_{s_{t-1} s_t} \mathrm{P}(\boldsymbol{O}, \boldsymbol{S} \mid \lambda) \\ &+ \sum_{\boldsymbol{s} \in \boldsymbol{S}} \sum_{t=1}^{L} \log \bar{b}_{s_t}(\boldsymbol{o}_t) \mathrm{P}(\boldsymbol{O}, \boldsymbol{S} \mid \lambda) \end{aligned} \quad (7)$$

The update of model parameters π_i and a_{ij} with constraints $\sum_{i=1}^{N}\pi_i = 1$ and $\sum_{j=1}^{N} a_{ij} = 1$ can be obtained by maximizing the first and second term of (7) with respect to $\boldsymbol{\pi}_i$ and a_{ij} respectively as follows:

$$\bar{\pi}_i = \frac{\mathrm{P}(\boldsymbol{O}, s_0 = i \mid \lambda)}{\mathrm{P}(\boldsymbol{O} \mid \lambda)} \tag{8}$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T} \mathrm{P}(\boldsymbol{O}, s_{t-1} = i, s_t = j \mid \lambda)}{\sum_{t=1}^{T} \mathrm{P}(\boldsymbol{O}, s_{t-1} = i \mid \lambda)} \tag{9}$$

The update for $\{u_i\}_{i=1}^{I}$ can be obtained by maximizing the third term with respect to $\{u_i\}_{i=1}^{I}$, however, the close-form expression is not available due to the complicated structure of the observation term $f(\{q_i^t, r_i^t\}_{i=1}^{L_t})$. Newton iteration was used with respect to the first and second derivation of the third term as follows:

$$u_i^{new} = u_i^{old} - \frac{\sum_{t=1}^{T} \frac{\partial f}{\partial u_i} * \frac{\mathrm{P}(\boldsymbol{O}, s_t = i \mid \lambda)}{f}}{\frac{\partial f}{\partial u_i}\left\{\sum_{t=1}^{T} \frac{\partial f}{\partial u_i} * \frac{\mathrm{P}(\boldsymbol{O}, s_t = i \mid \lambda)}{f}\right\}} \tag{10}$$

The forward and backward algorithm[19] is used to update π_i and α_{ij}. In the implementation of Baum-Welch, the update of u_i is used for the update of the emission probability. Finally, Viterbi algorithm[19] was used to infer the hidden states of each l-length sequence in study.

Results

Dataset: For ESEs and ESSs detection, we use the constitutive exon database and their flanking sequence for ESEs and ESSs detection were used. The constitutive exon ESTs are obtained from http://hexevent.mmg.uci.edu/cgi-bin/ HEXEvent/HEXEventWEB.cgi and then the EST was used to obtain the vicinity sequence from hg19 at http:// genome.ucsc.edu/cgi-bin/hgGateway.

Zhang and Chasin[10] observed that both ESEs have a sharp decreasing from the vicinity of exon inside to the vicinity of exon outside while ESSs have a sharp increasing from the vicinity of exon inside to the vicinity of exon outside observed it. So the constitutive exon database and its flanking intron area can be used by HMMotif to detect ESEs and ESSs simultaneously. The flanking length for the 5' and 3' end are taken as 100nt. The length of ESEs and ESSs are both taken as 6.[5]

Statistical performance: For the extracting of motifs from the result of HMMotif, the 6-length sequence was taken as potential ESEs having more than half of the observations having hidden state ESE and it is similar for the detection of ESSs. We compared the potential ESEs and ESSs extracted from HMMotif with that of Threshold Setting (TS) method.[5]

Two thresholds are considered here to illustrate the effectiveness of HMMotif, which are 1.5 and 2 representing the ratio between the vicinity of exon and intron for ESE and between the vicinity of intron and exon for ESS respectively. The number of ESEs detected by HMMotif and TS methods, as well as their overlaps is presented in table 1. The number of ESSs detected by HMMotif and the TS methods, as well as their overlaps is given in table 2.

Table 1
Comparison of HMMotif with TS methods for ESEs

Motifs	HMMotif	TS_1.5	TS_2	Overlap
ESE	478	921	—	319(66.74%)
	478	—	440	268(56.07%)

It is observed that HMMotif can detect most of the ESE and ESS that TS can detect. With the increasing of ratio, HMMotif is still efficient in detecting the high-ratio motifs, which verifies the effectiveness of HMMotif. It can also be seen that the overlap proportion between HMMotif and TS method decreases with the increase of the ratio for ESS. It is not surprising as less motifs will be detected with the increase of the threshold. However, more fake motifs will be detected with the decrease of the threshold although the overlap between HMMotif and TS increases. So what threshold should be taken for the TS method is still a problem.

Table 2
Comparison of HMMotif with TS methods for ESSs

Motifs	HMMotif	TS_1.5	TS_2	Overlap
ESS	700	1185	—	580(82.86%)
	700	—	625	387(55.29%)

For the training of the parameters of HMM, one fourth of the data was used and considered as training, the initial value of three kinds of parameters i.e. π, u and transition probability A are generated according to the distribution. The trained parameters are given in table 3. It is observed that the trained values varied not too much compared with the initial values showing that the initial distribution of the parameters is very close to the true distribution of the parameters and the distribution of ESEs and ESSs obtained by HMMotif is consistent with the setting of the parameters.

Conclusion

A new tool HMMotif for splicing regulatory elements prediction of ESEs and ESSs has been proposed. HMMotif can make full use of the contextual information among each potential motif as each pair of coterminous l-length sequences share (l-1) -length sequence. It is observed that HMMotif can detect majority of the ESEs and ESSs simultaneously that threshold based method can detect and can be used for different length of motifs detection. Since HMMotif is based on the optimization of whole posterior probability, more probability power can be gained from the contextual information. Threshold can be avoided in the comparison between the exon flanking area and its reference area.

Table 3
Comparison of the parameters of HMMotif before and after training on one fourth of the data considered.

		Parameter
π	**initial**	(0.199900 0.611900 0.188300)
	trained	(0.234500 0.721312 0.126734)
u	**initial**	(0.663300 0.503313 0.333348)
	trained	(0.521870 0.354901 0.191122)
A	**initial**	$\begin{pmatrix} 0.836900 & 0.089300 & 0.073800 \\ 0.073800 & 0.886600 & 0.095400 \\ 0.014800 & 0.086800 & 0.898400 \end{pmatrix}$
	trained	$\begin{pmatrix} 0.812513 & 0.179130 & 0.008359 \\ 0.177593 & 0.707575 & 0.114834 \\ 0.014563 & 0.231420 & 0.754019 \end{pmatrix}$

Availability and Requirements

Tool home page: https://sites.google.com/site/hmmmotif/config/pagetemplates/hmm_motif1
Operating system: 64-bit Linux
Programming language: C
Other requirements: Linux Ubuntu 3.0.0 or higher
License: GNU GPL
Any restrictions to use by non-academics: License needed.

Acknowledgement

We would like to acknowledge the members of Translational Biosystems Lab of Xiaobo Zhou at Wake Forest University-School of Medicine and Dr. Jing Su and Dr. Chenglin Liu for their help with programming. This work was supported in part by the National Natural Science Foundation of China under Grants 61071188(Li), 61302138(Bian), 11126274 (Bian) and 61202460(Xing), the Fundamental Research Funds for the Central Universities, China University of Geosciences(Wuhan), under Grant CUGL140422(Bian) and NIH R01LM010185-03(Zhou), NIH U01HL111560-01(Zhou), NIH 1R01DE022676-01 (Zhou), U01 CA166886-01 (Zhou).

References

1. Senapathy P., Shapiro M. B. and Harris N. L., Splice junctions, branch point sites and exons: sequence statistics, identification and applications to genome project, *Methods Enzymol.*, **183**, 252-278 (**1990**)

2. Cartegni L., Chew S. L. and Krainer A. R., Listening to silence and understanding nonsense, exonic mutations that affect splicing, *Nat. Rev. Genet.*, **3**, 285-298 **(2002)**

3. Hiller M., Zhang Z., Backofen R. and Stamm S., Pre-mRNA secondary structures influence exon recognition, *PLoS Genet.*, **3**, e204 (**2007**)

4. Yeo G. W., Van Nostrand E., Holste D., Poggio T. and Burge C. B., Identification and analysis of alternative splicing events conserved in human and mouse, Proc. Natl. Acad. Sci., U. S. A., **102**, 2850-2855 (**2005**)

5. Fairbrother W. G., Yeh R. F., Sharp P. A. and Burge C. B., Predictive identification of exonic splicing enhancers in human genes, *Science*, **297**, 1007-1013 (**2002**)

6. Reed R. and Maniatis T., A role for exon sequences and splicesite proximity in splice site selection, *Cell*, **46**, 681-690 (**1986**)

7. Blencowe B. J., Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases, *Trends Biochem. Sci.*, **25**, 106-110 (**2000**)

8. Manley J. L. and Tacke R., SR proteins and splicing control, *Genes Dev.*, **10**, 1569-1579 (**1996**)

9. Mayeda A., Screaton G. R., Chandler S. D., Fu X. D. and Krainer A. R., Substrate specificities of SR proteins in constitutive splicing are determined by their RNA recognition motifs and composite pre-mRNA exonic elements, *Mol. Cell. Biol.*, **19**, 1853-1863 (**1999**)

10. Zhang X. H. and Chasin L. A., Computational definition of sequence motifs governing constitutive exon splicing, *Genes Dev.*, **18**, 1241-1250 (**2004**)

11. Staffa A., Acheson N. H. and Cochrane A., Novel exonic elements that modulate splicing of the human bronectin EDA exon, *J. Biol. Chem.*, **272**, 33394-33401 (**1997**)

12. Konig H., Ponta H. and Herrlich P., Coupling of signal transduction to alternative pre-mRNA splicing by a composite splice regulator, *EMBO J.*, **17**, 2904-2913 (**1998**)

13. Kan J. L. and Green M. R., Pre-mRNA splicing of IgM exons M1 and M2 is directed by a juxtaposed splicing enhancer and inhibitor, *Genes Dev.*, **13**, 462-471 (**1999**)

14. Chen C. D., Kobayashi R. and Helfman D. M., Binding of hnRNPH to an exonic splicing silencer is involved in the regulation of alternative splicing of the rat beta-tropomyosin gene, *Genes Dev.*, **13**, 593-606 (**1999**)

15. DelGatto-Konczak F., Olive M., Gesnel M. C. and Breathnach R., HnRNP A1 recruited to an exon in vivo can function as an exon splicing silencer, *Mol. Cell Biol.*, **19**, 251-260 **(1999)**

16. Goren A., Ram O., Amit M., Keren H., Lev-Maor G., Vig I., Pupko T. and Ast G., Comparative analysis identifies exonic splicing regulatory sequences-The complex definition of enhancers and silencers, *Mol. Cell*, **22**, 769-781 (**2006**)

17. Wang Z., Rolish M. E., Yeo G., Tung V., Mawson M. and Burge C. B., Systematic identification and analysis of exonic splicing silencers, *Cell*, **119**, 831-845 (**2004**)

18. Dempster A. P., Laird N. M. and Rubin D. B., Maximum likelihood from incomplete data via the EM algorithm, *J. Royal Stat. Soc.*, **B 39**, 1-38 (**1977**)

19. Rabiner L. R., A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE, **77**, 257-286 (**1989**).

(Received 09th January 2014, accepted 25th February 2014)