Published in final edited form as:

Methods Mol Biol. 2017 ; 1552: 177–184. doi:10.1007/978-1-4939-6753-7_13.

Finding RNA-protein interaction sites using HMMs

Tao Wang¹, Jonghyun Yun², Yang Xie¹, and Guanghua Xiao^{1,*}

¹University of Texas Southwestern Medical Center, Quantitative Biomedical Research Center, Department of Clinical Sciences, 5323 Harry Hines Blvd, Dallas, TX 75390, United States

²University of Texas at Arlington, Department of Mathematics, 1225 West Mitchell St, Arlington, TX 76019-0408, United States

Abstract

RNA-binding proteins play important roles in the various stages of RNA maturation through binding to its target RNAs. Cross-linking immunoprecipitation coupled with high-throughput sequencing (CLIP-Seq) has made it possible to identify the targeting sites of RNA-binding proteins in various cell culture systems and tissue types on a genome-wide scale. Several Hidden Markov model-based (HMM) approaches have been suggested to identify protein-RNA binding sites from CLIP-Seq datasets. In this chapter, we describe how HMM can be applied to analyze CLIP-Seq datasets, including the bioinformatics preprocessing steps to extract count information from the sequencing data before HMM and the downstream analysis steps following peak-calling.

Keywords

Hidden Markov Models; RNA-binding proteins; interaction sites

1. Introduction

RNA-binding proteins (RBPs) participate in RNA translation, splicing and editing events, and play essential roles in mRNA maturation and downstream regulation of cellular events [1]. Thus, accurate identification of RBP binding targets is important to a systematic understanding of transcription, translation and other biological processes within cells.

A technique known as cross-linking immunoprecipitation coupled with high-throughput sequencing (CLIP-Seq) has been developed to study genome-wide RNA-protein interactions [2–4]. CLIP-Seq experiments are further subdivided into 3 types: 1) high-throughput

*Corresponding Author, Guanghua.Xiao@UTSouthwestern.edu.

¹)Some of the sequencing reads could be mapped across splicing junctions. However, usually less than 5% of all the CLIP reads are mapped across splice junctions. The reads mapped across splice junctions represent different binding events from those reads mapped within these introns, and therefore need to be handled separately. One possible solution is to map reads to genome and then map again the unaligned reads to transcriptome. Then the CLIP clusters identified from the genome and transcriptome can be analyzed at the same time.

²)Nowadays, sequencing data can reach exceptionally large sizes (many >50M reads). Therefore, the CLIP clusters that form from overlapping tags could be huge in number, and the length of the clusters can also be very long. These data present a lot of computational challenges. Therefore, it is advisable to place a limit on the minimum number of tags for a cluster and focus analysis on the larger CLIP clusters.

sequencing of RNA isolated by cross-linking immunoprecipitation (HITS-CLIP) [2,5,6]; 2) photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation (PAR-CLIP) [7–9]; and 3) individual-nucleotide resolution CLIP (iCLIP). HITS-CLIP utilizes UV cross-linking of proteins with RNA and introduces mutations in the sequencing data. More specifically, the mutations are induced on the cDNAs generated in the reverse transcription step from the RNA fragments, when the reverse transcription enzyme incorporates an incorrect nucleotide at the site of the cross-linked nucleotide due to attachment of the remaining residues of the covalently bound protein. However, the type of mutation is not well-defined and may vary for different proteins [5,10]. PAR-CLIP utilizes photoreactive ribonucleoside analogs for incorporation into RNA, and some of the analogs that are cross-linked by UV treatment at a later step will result in specific nucleotide substitution events. For example, 4-thiouridine treatment will induce T->C mutations at crosslinking sites [9]. The iCLIP experiment, which involves a self-circularization step to capture the truncation of cDNA reads, is less commonly used than HITS-CLIP and PAR-CLIP [11]. In iCLIP experiments, cDNA counts, rather than total tag count and mutant tag count, are used to infer protein binding sites.

The analysis of CLIP-Seq data mainly consists of three steps: the bioinformatics analysis that extracts count information from the raw sequencing data, the statistical analysis that carries out peak calling, and the downstream analysis that relates the significant interaction sites to their biological meanings. While this chapter centers on application of HMM in the peak calling step, we will also briefly touch upon the other two steps in analysis of CLIP-Seq data.

2. Sequencing data preprocessing

CLIP-Seq analysis usually starts with SAM/BAM files, which contain the mapped reads from high-throughput sequencing data. Since CLIP-Seq experiments involve PCR amplification from cDNA libraries with limited complexities, removal of PCR duplicates is an essential step for CLIP-Seq experiments. The iCLIP technique explicitly introduces a random barcode, thus making it relatively easy to define PCR artifacts from the sequencing data. But the duplicate removal step for HITS-CLIP and PAR-CLIP is not trivial. One popular approach that has been adopted in many studies [12–14] is to define reads that have exactly the same mapping coordinates as PCR duplicates. Unfortunately there is no single best answer to this question, and the scenario is even more complicated for paired-end sequencing reads.

For HITS-CLIP and PAR-CLIP, unique tags overlapping by at least one nucleotide are grouped together to form CLIP clusters after PCR duplicates are collapsed, and those not overlapping with any other tags are discarded. Then the clusters are divided into bins of small length (for example, 5 bp) and the total tag count in each bin is calculated. If mutant information is utilized in the statistical analysis, mutant tag counts will also be calculated for each bin.

However, for iCLIP datasets, the preprocessing step differs substantially. After removing PCR duplicates, the first nucleotide upstream of each mapped cDNA, defined as the “cross-

link nucleotide,” is counted and the counts are binned. Since the cross-link nucleotides exist in a sparse, discontinuous manner around RBP binding sites, it might be a good idea to expand each cross-link nucleotide by a few base pairs to smooth the tag data, as is done in dCLIP [14].

3. Hidden Markov Model (HMM)

The bins coming from one single CLIP cluster form an observational sequence. There are many CLIP clusters identified from the sequencing data, which form multiple observational sequences. The HMM needs to analyze all the observational sequences simultaneously. Our introduction to HMM is subdivided into four parts.

1) Transcript abundance normalization

Currently, most CLIP-seq studies do not conduct transcript abundance measurements [6,15] and, accordingly, most current CLIP-seq analysis tools, such as PARalyzer [16], also do not consider transcript abundance. However, taking background transcript abundance into account will be very helpful for more accurately identifying RBP binding sites, since the size of each CLIP cluster is affected by both the transcript level and the binding strength of the RBP.

2). Model specification

We assume that for the t -th bin of each region d , the spatially associated tag count $n_t^{(d)} > 0$ is governed by a Markov latent variable $q_t^{(d)}$, which spans two exclusive states as follows:

$$q_t^{(d)} = \begin{cases} 1 & \text{if unenriched;} \\ 2 & \text{if enriched.} \end{cases}$$

Markov chains in different clusters are assumed to be independent, but the chains share identical forms of transition and emission probabilities. A graphical representation of the HMM is shown in Figure 1.

2.1) Transition probabilities—An initial state of Markov chains are described by an initial distribution $\xi(q_1^{(d)})$. The Markov transition matrix Λ describes stochastic transition behaviors of $q_{t+1}^{(d)}$ relying on the previous state $q_t^{(d)}$.

$$\Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix},$$

where Λ_{rv} denotes the transition probability from state r to v . That is,

$$\Lambda_{rv} \equiv \Pr(q_{t+1}^{(d)} = v | q_t^{(d)} = r).$$

2.2) Emission probabilities—A mixture of negative binomial (NB) distributions have been commonly used to fit the tag count in next-generation sequencing data due to over-dispersion. The NB density on $n_t^{(d)} - 1 | q_t^{(d)}$ is assumed. The parametrization of the NB (a, b) involves the shape parameter $a > 0$ and the inverse scale parameter $b > 0$. Assuming a to be identical for all states, the conditional density of $n_t^{(d)}$ given $q_t^{(d)}$ is

$$p(n_t^{(d)} = n | a, b_v, q_t^{(d)} = v) = \binom{n+a-2}{a} \left(\frac{b_v}{1+b_v} \right)^a \left(\frac{1}{1+b_v} \right)^{n-1}, n > 0,$$

where $b_1 > b_2$. Here, $\mu_v \equiv [n_t - 1 | q_t^{(d)} = v] = a \cdot b_v^{-1}$, and $\text{Var}[n_t - 1 | q_t^{(d)} = v] = \mu_v + \mu_v^2 \cdot a^{-1}$, where a^{-1} is the dispersion parameter for the NB distribution. According to the definition of the HMM states, it is expected that $\mu_1 < \mu_2$, and so the ordering on b_v 's is imposed.

2.3) Likelihood—Let $\phi = (\pi, \Lambda, a, b_1, b_2)$, $\mathbf{n}_d = (n_1^{(d)}, \dots, n_{T_d}^{(d)})$, and $\mathbf{q}_d = (q_1^{(d)}, \dots, q_{T_d}^{(d)})$. The likelihood of ϕ is given as

$$\prod_d p_\phi(\mathbf{n}_d) = \prod_d \sum_{\mathbf{q}_d} p_\phi(\mathbf{q}_d, \mathbf{n}_d),$$

where the summation is taken over all possible configurations of $\mathbf{q}_d$. The sequential structure of HMMs facilitates the computation of the joint density.

$$p_\phi(\mathbf{q}_d, \mathbf{n}_d) = \pi(q_1^{(d)}) \prod_{t=1}^{T_d} \Lambda(q_{t+1}^{(d)} | q_t^{(d)}) p_\phi(n_t^{(d)} | q_t^{(d)}),$$

where we use $\Lambda(v|r)$ to denote Λ_{rv} .

3) Model inferences

3.1) Baum-Welch expectation maximization algorithm—The Baum-Welch expectation maximization algorithm [17] is used to estimate parameters in the HMM. The iterative algorithm repeats two steps (E-step and M-step) to update a current estimate ϕ^c until a stopping criterion is met. For simplicity, we drop the cluster index d , and observation sequences are denoted by $n_{1:t} = (n_1, \dots, n_t)$.

Forward-backward recursions are essential in implementing the E-step. In the E-step, we evaluate an expectation

$$G(\phi | \phi^c) \equiv \mathbb{E}_{\phi^c} \left[\sum_d \log p_\phi(\mathbf{q}_d, \mathbf{n}_d) \right],$$

where the expectation is taken with respect to the conditional density $\prod_d p_{\phi^c}(\mathbf{q}_d | \mathbf{n}_d)$. The filtering distributions $\pi_t(q_t) \equiv p_{\phi^c}(q_t | n_{1:t}, q_0)$ can be evaluated recursively as

$$\pi_t(q_t) = \begin{cases} \frac{1}{M_t} \psi_t(q_t) \xi^c(q_t) & \text{if } t=1; \\ \frac{1}{M_t} \psi_t(q_t) \sum_{r \in Q} \Lambda^c(q_t|r) \pi_{t-1}(r) & \text{if } t>1, \end{cases}$$

where $\psi_t(q_t) \equiv p_{\phi^c}(n_t|q_t)$, M_t is the normalizing constant reconciled by $\sum_{q_t \in Q} \pi_t(q_t) = 1$, and $Q = \{1, 2\}$ is the state space. The backward probability is computed as

$$\kappa_{t,T}(q_t) = \begin{cases} M_{t+1}^{-1} \sum_{v \in Q} \kappa_{t+1,T}(v) \psi_{t+1}(v) \Lambda^c(v|q_t) & \text{if } t < T; \\ 1 & \text{if } t = T. \end{cases}$$

Then, the smoothing distribution $\pi_{t|T}(q_t) \equiv p_{\phi^c}(q_t|n_{1:T})$ can be computed recursively as

$$\pi_{t|T}(q_t) = \frac{1}{M} \kappa_{t,T}(q_t) \pi_t(q_t),$$

where $M = \sum_{r \in Q} \Lambda^c(q_{T+1}|r) \pi_T(r)$.

The conditional maximization method is used to implement the M step. The initial distribution is updated by $\xi(v) \propto \sum_d \pi_{1|T_d}^{(d)}(v)$. The transition probabilities are updated on the constrained space $(\sum_r \Lambda_{rv} = 1)$ as

$$\Lambda_{rv} = \frac{\sum_d \sum_{t=1}^{T_d} P_{\phi^c}(v, r | \mathbf{n}_d)}{\sum_{v \in Q} \text{numerator}},$$

where the numerator can be expressed as

$$P_{\phi^c}(v, r | \mathbf{n}_d) = \begin{cases} \pi_{1|T_d}^{(d)}(v) & \text{if } t=1; \\ (M^{(d)} \cdot M_t^{(d)})^{-1} \kappa_{t,T_d}^{(d)}(v) \psi_t^{(d)}(v) \Lambda^c(v|r) \pi_{t-1}^{(d)}(r) & \text{if } 1 < t \leq T_d. \end{cases}$$

The inverse scale parameter b_v can be updated by

$$b_v = \frac{a^c \sum_d \sum_t \pi_{t|T_d}^{(d)}(v)}{\sum_t (n_t^{(d)} - 1) \pi_{t|T_d}^{(d)}(v)}.$$

Updates for a need to be done using numerical optimization methods such as the Newton-Raphson.

3.2) Viterbi algorithm—After the convergence of the EM is observed, the Viterbi algorithm [18] is implemented along with the EM estimate $\hat{\phi}$ to obtain the most likely sequence of states that generates tag counts.

The algorithm recursively finds $q_t^{(d)}$ that maximizes

$$\omega(q_t^{(d)}=v)=\psi_t^{(d)} \cdot \max_{q_{t-1}^{(d)}} \omega(q_{t-1}^{(d)}) \cdot \Lambda(v|q_{t-1}^{(d)}),$$

where $\omega(q_1^{(d)}=v)=\xi(v) \psi_1^{(d)}(v)$.

From the *maximum a posteriori* path, the set of the enriched bins $E=\{(t, d) | q_t^{(d)}=2\}$ is obtained.

4) Mixture models

Mutation counts can be used to pinpoint binding sites among enriched bins. Let n_i and x_i denote tag and mutation counts for bins in E . We assume a mixture of binomial distributions with a latent variable s_i , where $s_i=1$ if the bin contains binding sites; 0 otherwise.

$$x_i | n_i \sim \begin{cases} \text{Bin}(n_i, p_1) & s_i=1; \\ \text{Bin}(n_i, p_0) & s_i=0, \end{cases}$$

where $p_1 > p_0$ are mutation probabilities. The EM algorithm can be easily implemented to estimate parameters $\theta = (p_0, p_1, \text{ and } w=p(s_i=1))$ in the mixture model. Then, a user-defined cut-off for the posterior probability $p_\theta(s_i=1 | \text{observations})$ is applied to identify RBP binding sites over bins in E .

4. Downstream analysis

1) For PAR-CLIP and HITS-CLIP data, after identification of significantly enriched RBP binding sites, it is interesting to identify single bases with high ratios of characteristic mutations within these sites. These bases pinpoint the location of RBP-RNA interaction at almost single-nucleotide resolution. However, one difficulty with this analysis is that the mutation ratios are usually very low. This is especially true in some HITS-CLIP datasets, where the highest observation mutation ratios are only around 1–2%. In addition, SNPs that exist in the cell culture or tissue could obscure this analysis. Such false positives can be distinguished by conducting background sequencing experiments as control or by comparing them to SNP databases in the public domain.

2) One of the downstream analyses that can be performed here is to relate the significant RBP binding sites to genes or annotation features (CDS, intron, 5' UTR, 3' UTR, etc.). When relating RBP binding sites to genes, some researchers prefer to mask tRNAs, rRNAs and other repetitive sequences from this analysis, since such binding sites usually represent uninteresting or artifact binding events. Most CLIP-Seq experiments do not distinguish

nucleus RNAs from cytoplasmic RNAs. Since libraries made in this way could contain cDNAs converted from nascent mRNAs, it is possible that a significant portion of CLIP reads will be mapped to reference gene introns. In some studies, intron-locating peaks even constitute the majority of the RBP binding targets.

3) Another popular way to analyze CLIP-Seq peaks is discover sequence motifs or predict secondary structures around significant RBP binding sites. Since the mutant tag counts give much more precise information regarding RBP binding locations, it is advisable to extract RNA sequences in a short window (for example, 50bp) surrounding bases with high mutation ratios in significant clusters for analysis purpose. To discover sequence motifs, MEME or Homer can be used. If the dataset is PAR-CLIP, it could be interesting to investigate whether T->C mutations occur on the T bases within the significant motifs more often than the genome wide background. To predict secondary structures, software like RNAfold can be used. It is possible the RBPs recognize binding targets through secondary structures, rather than through the primary sequence.

References

1. Keene JD. RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet.* 2007; 8:533–543. [PubMed: 17572691]
2. Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature.* 2008; 456:464–469. [PubMed: 18978773]
3. Ule J, Jensen K, Mele A, Darnell RB. CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods.* 2005; 37:376–386. [PubMed: 16314267]
4. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, et al. CLIP identifies Nova-regulated RNA networks in the brain. *Science.* 2003; 302:1212–1215. [PubMed: 14615540]
5. Zhang C, Darnell RB. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat Biotechnol.* 2011; 29:607–614. [PubMed: 21633556]
6. Chi SW, Zang JB, Mele A, Darnell RB. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature.* 2009; 460:479–486. [PubMed: 19536157]
7. Hafner M, Lianoglou S, Tuschl T, Betel D. Genome-wide identification of miRNA targets by PAR-CLIP. *Methods.* 2012; 58:94–105. [PubMed: 22926237]
8. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, et al. PAR-CLIP—a method to identify transcriptome-wide the binding sites of RNA binding proteins. *J Vis Exp.* 2010
9. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell.* 2010; 141:129–141. [PubMed: 20371350]
10. Granneman S, Kudla G, Petfalski E, Tollervey D. Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc Natl Acad Sci U S A.* 2009; 106:9613–9618. [PubMed: 19482942]
11. Konig J, Zarnack K, Rot G, Curk T, Kayikci M, et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol.* 2010; 17:909–915. [PubMed: 20601959]
12. Ince-Dunn G, Okano HJ, Jensen KB, Park WY, Zhong R, et al. Neuronal Elav-like (Hu) proteins regulate RNA splicing and abundance to control glutamate levels and neuronal excitability. *Neuron.* 2012; 75:1067–1080. [PubMed: 22998874]
13. Boudreau RL, Jiang P, Gilmore BL, Spengler RM, Tirabassi R, et al. Transcriptome-wide discovery of microRNA binding sites in human brain. *Neuron.* 2014; 81:294–305. [PubMed: 24389009]
14. Wang T, Xie Y, Xiao G. dCLIP: a computational approach for comparative CLIP-seq analyses. *Genome Biol.* 2014; 15:R11. [PubMed: 24398258]

15. Neumann M, Bentmann E, Dormann D, Jawaid A, DeJesus-Hernandez M, et al. FET proteins TAF15 and EWS are selective markers that distinguish FTLD with FUS pathology from amyotrophic lateral sclerosis with FUS mutations. *Brain*. 2011; 134:2595–2609. [PubMed: 21856723]
16. Corcoran DL, Georgiev S, Mukherjee N, Gottwein E, Skalsky RL, et al. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol*. 2011; 12:R79. [PubMed: 21851591]
17. Welch LR. Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter*. 2003:1–14.
18. Viterbi AJ. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*. 1967; 13:260–269.

Figure 1.

Graphical representation of the HMM for cluster d : $\{n_t^{(d)}\}$ is the observable process (tag count), and $\{q_t^{(d)}\}$ is the hidden Markov chain. Independent Markov chains with identical transition rules are equipped for different clusters.