Analyzing ChIP-seq Data based on Multiple Knowledge Sources for Histone Modification

Dafeng Chen
School of Information Science, Nanjing Audit University, Nanjing 211815, China
Email: windking@nau.edu.cn
Deyu Zhou
School of Computer Science and Engineering, Southeast University, Nanjing 210029, China
Email: zhoudeyu@gmail.com
Yuliang Zhuang
School of Information Science, Nanjing Audit University, Nanjing 211815, China
Email: Zhuangyl@nau.edu.cn

***Abstract*—ChIP-seq is able to capture the genomic profiles for histone modification by combining chromatin immunoprecipitation (ChIP) with next generation sequencing. However, enriched regions generated from peak finding algorithms are evaluated only based on the limited knowledge acquired from manually examining the relevant biological literature. This paper proposes a novel framework of incorporating multiple knowledge sources, consisting of information extracted from biological literature, Gene Ontology, and microarray data, in order to precisely analyze ChIP-seq data for histone modification. The information is combined in a unified probabilistic model to rerank the enriched regions generated from peak finding algorithms. Through filtering the reranked enriched regions using some predefined threshold, more reliable and precise results could be generated. The combination of the multiple knowledge sources with the peaking finding algorithm produces a new paradigm for ChIP-seq data analysis.**

***Index Terms*—ChIP-seq, histone modification, reranking, information extraction**

I. INTRODUCTION

Histones, acting as spools around which DNA binds, is the chief protein components of chromatin. Histones are subject to lots of posttranslational modifications, such as lysine acetylation, lysine and arginine methylation, serine and threonine phosphorylation, and lysine ubiquitination and sumoylation [1]. Histone modifications may alter the electrostatic charge of the histone resulting in a structural change in histones or their binding to DNA. Histone modifications may be the binding sites for protein recognition modules which recognize acetylated lysines or methylated lysine, respectively. Overall, histone modifications affect chromosome function in may ways. Thus, posttranslational modifications of histones create a mechanism for the regulation of a variety of normal and disease-related processes.

ChIP-seq [2], which combines chromatin immunoprecipitation (ChIP) with next generation sequencing, is able to capture the genomic profiles for histone modification and transcription factor (TF). It is characterized by high resolution, cost effectiveness and no complication.A large amount of data have recently been generated using the ChIP-Seq technique, therefore calling for new analysis algorithms.

To discover the exact locations of TF binding sites from ChIP-seq data, a number of algorithms, such as CisGenome [3], MACS [4], PeakSeq [5], QuEST [6], sPP [7], Useq [8] and SISSRs [9], have been proposed. TF binding is mainly governed by sequence specificity. Therefore TF binding sites are typically correlated with very localized ChIP-seq signals in the genome. On the contrary, many modification marks consist of broad domains, which are believed to stabilize the chromatin state. Moreover, the signals for histone modifications, histone variants and histone-modifying enzymes are usually diffuse and lack of well-defined peaks, spanning from several nucleosomes to large domains encompassing multiple genes. As such, peak-finding algorithms employed to find TF binding sites with strong local enrichment are unsuitable for discovering these generally weak signals from DNA modification marks.

To the best of our knowledge, only few methods, e.g. ChIPDiff [10] and SICER [11], have been published focusing on analyzing ChIP-seq data specifically for histone modification. ChIPDiff attempts to identify differential histone modification sites by computationally comparing two ChIP-seq libraries generated from different cell types. Instead of partitioning the genome into bins and computing the fold-change of the number of ChIP fragments in each bin, ChIPDiff modeled the cor relation as a hidden Markov model (HMM) where transmission probabilities were automatically trained in an unsupervised manner. By inferring the states of histone modification changes using the trained HMM parameters, the correlation between consecutive bins is taken into account. Nevertheless, ChIPDiff fails to compare more than two ChIP-seq libraries. Instead of

doi:10.4304/jsw.7.12.2755-2762

comparing two ChIP-seq libraries, SICER partition the genome into non-overlapping windows with fixed size. Islands (potential ChIP-enriched domains) are identified as clusters of eligible windows separated by gaps of a size less than a predetermined threshold. Then, a clustering method is employed to score each island.

After discovering enriched regions using a peak finding algorithm, validation of the results is typically performed based on some limited knowledge acquired from biomedical literature, such as experimentally validated genes relating with the histone modification. It is also possible to validate the correctness of the discovered enriched regions through QPCR (real-time Quantitative Polymerase Chain Reaction detecting system) experiments; but this is too costly and labor intensive and is therefore seldom adopted in practice. Thus, the prevailing approach of validating the discovered enriched regions is the former method which uses limited knowledge acquired from biomedical literature. However, it suffers from the following drawbacks:

- Amount of knowledge for validation. Most knowledge for validation are obtained by hand-curated the relevant experimental results described in biomedical literature, which is laborious, time consuming, and error-prone. Moreover, it has been demonstrated that biomedical literature is growing at a double-exponential pace, it thus becomes extremely hard for biologists to be updated with the most up to-date knowledge from biomedical literature.
- Source of knowledge for validation. Existing approaches mainly use knowledge extracted from biomedical literature for validation. It is worth to exploit knowledge from other sources, such as results from microarray data analysis, or knowledge inferred from Gene Ontology.
- Handling of contradictory knowledge. It is possible that the results discovered by peak finding algorithms are contradictory to the knowledge obtained from biomedical literature. There lack of effective methods in handling such a situation.

This paper explores an efficient way to improve the precision of genomic-wide chromatin modification profiles. A framework of incorporating information extraction into a probabilistic model for reranking discovered enriched regions (candidate histone modification sites) is comprehensively investigated. To improve the histone modification sites discovery results, the external knowledge sources, such as information extracted from biomedical literature, microarray data, and Gene Ontology, are employed to re-score enriched regions. The rationale be hind this is that biomedical literature, microarray data, and Gene Ontology are reliable resources for describing the gene expression level in some specific cell lines, while the histone modifications are major epigenetic factors regulating gene expression. Therefore, there is some casual relationship between histone modifications and the knowledge sources which can be used to improve the accuracy of discovered histone modification sites.

II. Related Work

This section presents the existing work in two areas, information extraction for genes regulated by histone modification, and reranking based on multiple knowledge sources.

A. Information Extraction for Genes Regulated by Histone Modification

Large amount of experimental and computational biomedical data, specifically in the areas of genomics and proteomics have been generated along with new discoveries, which are accompanied by an exponential increase in the number of biomedical publications describing these discoveries. In the meantime, there has been great interest with scientific communities in literature mining tools to sort through this abundance of literature and find the nuggets of information such as protein-protein interactions, gene regulation and so on, which are most relevant and useful for specific analysis tasks.

To mine information from the biomedical literature, two steps are crucial. One is named entity recognition (NER) which recognizes names of biomedical entities, such as gene, proteins, cells and diseases. The other is information extraction. In general, current approaches for biomedical information extraction can be divided into three categories, computational linguistics-based methods, rule-based methods and machine learning and statistical methods.

Corinna [12] developed an approach for identifying histone modifications in biomedical literature with Conditional Random Fields (CRFs) and for resolving the recognized histone modification term variants by term standardization.

Many systems [13–17], examples including EDGAR [18], BioRAT [19], GeneWays [20] etc.,have been developed to extract protein-protein interaction from text. To the best of our knowledge, there are no existing approaches focusing on mining the gene information regulated by histone modification.

B. Reranking based on Multiple Knowledge Sources

Recently, reranking algorithms have been quite popular for data mining and natural language processing. The idea behind reranking is that some information which is crucial for generating ranking scores is not incorporated in the ranking algorithm used. Therefore, there is a need for a reranking algorithm to rerank results by incorporating these information.

For example, documents can be represented in the vector space model used in information retrieval. In traditional information retrieval, given a query q, retrieved documents are presented in a decreasing order of the ranking scores with respect to the content information. In addition to content, documents are interconnected to each other through an explicit or latent link. Thus, many recent methods take into account link-

based information. However, one of the issues is that those ranking algorithms typically treat the content and link information separately, and each document is assigned a score independent of other documents for the same query. Reranking algorithm leverage the interconnection between documents/entities to improve the ranking of retrieved results [21].

Reranking approaches in the natural language processing domain attempt to improve upon an existing probabilistic parser by reranking the output of the parser. Reranking has benefited applications such as name-entity extraction [22], semantic parsing [23] and semantic labeling [24]. Most reranking approaches are based on discriminative models while base parers are mostly based on generative models. The reason behind is that generative probability models such as hidden Markov models (HMMs) or hidden vector state (HVS) models provide a principled way of treating missing information and dealing with variable length sentences. On the other hand, discriminative methods such as support vector machines (SVMs) enable us to construct flexible decision boundaries and often result in performance superior to that of generative models. The combination of generative and discriminative models could leverage the advantages of both approaches.

III. Proposed Framework

The overall process of the proposed framework is shown in Figure 1 which takes the form of the three main processes. Firstly, millions of short reads generated from the deep sequencing platform are mapped to reference genome. After peak finding, enriched regions are discovered. Secondly, information extraction based on a statistical model aims to extract information about genes which are regulated by histone modification. Information about the environment for these regulations will also be extracted. The extracted information will be combined with the external knowledge sources such as gene ontology and results mined from microarray data to form inputs to a probabilistic model, which is then employed for re-ranking the discovered enriched regions.

A. Information Extraction based the Conditional HVS model

In order to extract genes regulated by histone modification, they need to be first identified through named entity recognition. After that, the genes regulated by histone modification can be extracted through relation extraction. For the first step, CRFs or SVMs can be employed to recognize genes regulated by histone modifications. For the second step, we are particularly interested in relation extraction from biomedical literature based on the Hidden Vector State (HVS) model. The HVS model was originally proposed in [25] and has been successfully applied in biomedical domain for protein-protein interactions extraction [26, 27].

Given a model and an observed word sequence W =(W_1 … W_T), semantic parsing can be viewed as a pattern recognition problem and the most likely semantic representation can be found through statistical decoding. If assuming that the hidden data take the form of a semantic parse tree C then the model should be a push-down automata which can generate the pair <W,C> through some canonical sequence of moves D = (d_1 … d_T). That is

$$P(W,C)=\prod_{t=1}^{T}P(d_t|d_{t-1}\cdots d_1) \qquad (1)$$

When considering a constrained form of automata where the stack is finite depth and <W,C> is built by repeatedly popping 0 to n labels off the stack, pushing exactly one new label onto the stack and then generating the next word, it defines the HVS model in which conventional grammar rules are replaced by three probability tables. Given a word sequence W, concept vector sequence C and a sequence of stack pop operations N, the joint probability of P (W,C,N) can be decomposed as

$$P(W,\boldsymbol{C},N)=\prod_{t=1}^{T}P(n_t|\boldsymbol{c}_{t-1})P(c_t[1]|c_t[2\cdots D_t])P(w_t|\boldsymbol{c}_t) \qquad (2)$$

where C_t, the vector state at word position t, is a vector of D_t semantic concept labels (tags), i.e. C_t=[C_t[1], C_t

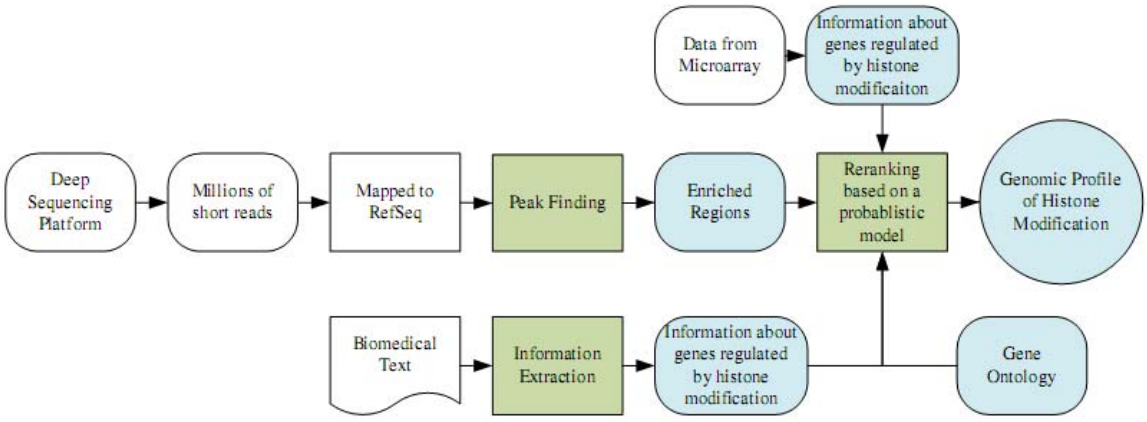

Fig. 1 The framework of incorporating multiple knowledge resources for analyzing ChIP-seq Data.

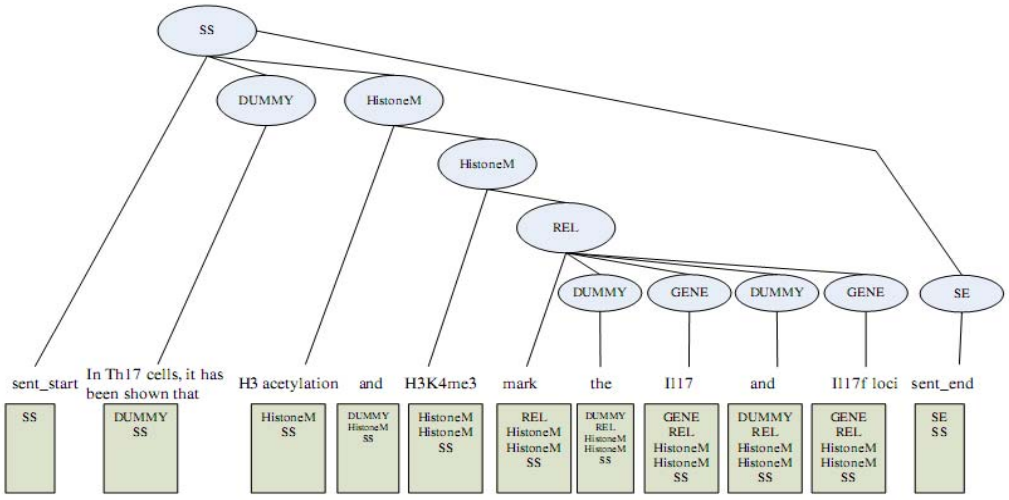

Fig. 2 Example of a parse tree and its vector state equivalent.

[2], .. C_t [D_t]] where $C_t[1]$ is the preterminal concept label and C_t [D_t] is the root concept label (SS in Fig. 2), nt is the vector stack shift operation at word position t and take values in the range 0, . . . , D_{t-1} and $C_t[1] = Cw_t$ is the new preterminal semantic tag assigned to word w_t at word position t.

An example parse tree is illustrated in Figure 2 which shows the sequence of HVS stack states corresponding to the given parse tree. State transitions are factored into separate stack pop and push operations constrained to give a tractable search space. The result is a model which is complex enough to capture hierarchical structure but which can be trained automatically from only lightly annotated data.

The HVS model computes a hierarchical parse tree for each word string W, and then extracts semantic concepts C from this tree. Each semantic concept consists of a name-value pair where the name is a dotted list of primitive semantic concept labels. For example, the top part of Figure 2 shows a typical semantic parse tree and the semantic concepts extracted from this parse would be in Equation 3

HistoneM = H3 acetylation
HistoneM.HistoneM = H3K4me3
HistoneM.HistoneM.REL.GENE = IL17
HistoneM.HistoneM.REL.GENE = IL17f (3)

The HVS model parameters are estimated using an EM algorithm and then used to compute parse trees at run-time using Viterbi decoding. In training, each word string W is marked with the set of semantic concepts C that it contains. For example, if the sentence shown in Figure 2 was in the training set, then it would be marked with the four semantic concepts given in equation 3. For each word wk of each training sentence W, EM training uses the forward-backward algorithm to compute the probability of the model being in stack state c when wk is processed. Without any constraints, the set of possible stack states would be intractably large. However, in the HVS model this problem can be avoided by pruning out all states which are inconsistent with the semantic concepts associated with W. The details of how this is done are given in [25].

The original HVS model takes a form of a generative model which makes it difficult to incorporate background knowledge or non-local features. We propose to represent the model as a conditionally trained graphical model similar to the CRFs. The HVS model can be viewed as a graphical model. Assuming the vector state stack depth is limited to be 4, that is, there are at most 4 semantic tags (states) relating to each word position. C_t is the vector state corresponding to the word W_t. S_t is the stack shift operation which consists of popping N_t semantic tags from the previous vector state C_{t-1} and pushing one pre-terminal semantic tag to the stack and thus producing C_t.

Given a word sequence W, concept vector sequence C and a sequence of stack pop operations N, the conditional HVS model takes the form

$$P_\Theta(\boldsymbol{C}, N|W) = \frac{1}{Z_w}\exp\Big(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(\boldsymbol{c_{t-1}}, n_t, W, t) + \sum_{t=1}^{T}\sum_{k}\mu_k g_k(c_t[1], c_t[2\cdots D_t], W, t) + \sum_{t=1}^{T}\sum_{k}\nu_k h_k(\boldsymbol{c_t}, W, t)\Big) \quad (4)$$

where $\Theta = <\lambda_1, \lambda_2, ...; \mu_1, \mu_2, ...; \nu_1, \nu_2, ...>$ is the parameter vector of the conditional HVS model. f_k, g_k, h_k are arbitrary feature functions over their respective arguments, and λ_k, μ_k, ν_k are the corresponding learned weights for each feature function.

Inference for the conditional HVS models can be performed efficiently with dynamic programming. Parameter estimation can be performed with standard optimization procedures such as iterative scaling, conjugate gradient descent, or limited memory quasi-Newton method(L-BFGS).

B. Reranking based on a Probabilistic Model

To rerank the enriched regions generated from a peak finding algorithm, we need to first select some essential

features based on the multiple knowledge sources. Suppose the enriched region R and its related gene G, its information extracted from text I_T , results mined from microarray data I_M, and information inferred from Gene Ontology I_O are defined as follows:

- Information extracted from Text IT , for the pair < Histone Modification, G>, is defined as the probabilistic score that is generated from the conditional HVS model.
- Results mined from Microarray, I_M is defined as the expression level results obtained from microarray data for G.
- Information inferred from Gene Ontology I_O describes the trust level of inference that this gene is regulated by the histone modification. I_O is defined as the score of inference based on gene ontology.

Overall, it can be observed that the higher the value of I_T , I_M, and I_O, the strong confidence of the correctness of the enriched region.

We use the above parameters I_T , IM and I_O to calculate Score, the overall score of the enriched region R. Based on these scores, enriched regions generated from peak finding are reranked. It should be noted that up to this point, the relationship between Score and the above parameters is not apparent and it could be linear or non-linear. We thus investigate several ways to describe this relationship by constructing three models including a log-linear regression model, neural networks, and support vector machines.

1. Log-linear Regression Model

For the log-linear regression model, Score is defined as

$$\log Score = \beta_t I_T + \beta_m I_M + \beta_o I_O + \beta_0, \qquad (5)$$

which is a combination of the above three defined parameters. To estimate the coefficients β = (βt, βm, βo, β_0), the method of least squares is applied and the coefficients β are selected to minimize the residual sum of squares

$$RSS(\beta) = \sum_{i=1}^{M} (\log Score' - \beta_t I_{Ti} - \beta_m I_{Mi} - \beta_o I_{Oi} - \beta_0)^2 \qquad (6)$$

where M is the number of training data and log Score′ is the true value of Score.

2. Neural Networks

The central idea of neural networks is to extract linear combinations of the inputs as derived features, and then model the target as a nonlinear function of these features.The model based on neural networks has the form

$$Score = \sum_{m=1}^{M} g_m(\omega_m^T X) \qquad (7)$$

where X = (I_T , I_M, I_O) and ωm,m = 1, 2, . . . ,M is unit 3-vectors of unknown parameters.

3. Support Vector Machines

Support vector machines produce nonelinear boundaries by constructing a linear boundary in a large, transformed version of the feature space.

The model based on support vector machines has the form:

$$Score = h(X)^T \beta + \beta_0 \qquad (8)$$

where h_m(X),m = 1, . . . ,M are basis functions and X = (I_T , I_M, I_O).

IV. Experimental Results

The proposed framework of analyzing ChIP-seq data based on multiple knowledge sources for histone modification are evaluated in two parts, information extraction and re-ranking based on multiple knowledge sources.

The information extraction system works as follows.At the beginning, abstracts are retrieved from MED-LINE and split into sentences. Gene names, other biological terms are then identified based on a pre-constructed biological term dictionary. And histone modifications are identified using a classification model. After that, each sentence is parsed by the semantic parser employing the conditional HVS model. Finally, information about genes related to histone modification is extracted from the tagged sentences using a set of manually-defined simple rules. An example of the procedure is illustrated in Figure 3.

To investigate the performance of the information extraction system, abstracts from PubMed and PubMedCentral are selected. Based on the search keyword “h3k4me3”, 211 abstracts are retrieved. In the similar way, 731 abstracts are retrieved from PubMed based on the search keyword “histone h3 lysine 4 methylation”. Abstracts about “h3k9ac” and “h3k27me3” are also retrieved in the similar way. These abstracts are split intosentences. The sentences with at least one gene or protein name and histone modification are kept and other sentences are filtered out. All the kept sentences are collected as the input for the information extraction system. After the parsing process described in Section 3, a list of histone modification-gene name pairs are generated. An example of the histone modification and gene name pair is given in Figure 3. To evaluate the precision of the extracted pair of histone modification and gene, some annotators qualified to PhD level worked on the abstracts and the extracted pair. Evaluation results show that the information extraction system achieved as high as 73.2% on precision, in which extracted pairs can be further annotated by some experienced researchers to ensure there correctness with little efforts.

To investigate the performance of the proposed framework, we worked on the ChIP-seq data for histone modification “H3K4Me3”, “H3K9Ac”,and “H3K27Me3” in three cell lines, EB, MK and HUVEC. The data were generate from Dr. Willem Ouwehand’s research group in the Department of Haematology of the University of Cambridge. The short reads are mapped to the reference genome using the Maq program. After mapping, the enriched regions are generated based on some peaking finding programs. Here, SISSRs [9] is employed. Part of

the enriched regions and their related genes for "H3K9Ac" in EB cell line are listed in Table 1.

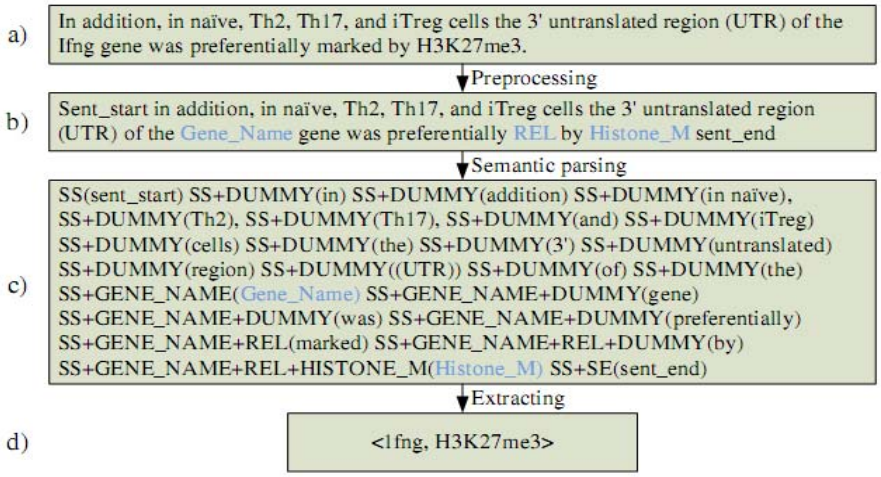

Fig. 3 An example of extracting genes related to histone modification.

TABLE 1
AN EXAMPLE OF ENRICHED REGIONS AND THEIR RELATED GENES GENERATED FROM THE PEAK FINDING ALGORITHM

Enriched Region	*Score of the region*	*Related genes id (gene name)*
chrX (18352600, 18354399)	71.0736603736	ENSG00000008086 (CDKL5)
chr15 (62903200, 62905399)	250.57069433	ENSG00000140451 (PIF1)
chr20 (41752800, 41753399)	18.7817973246	ENSG00000101057 (MYBL2)
chr1 (232808400, 232813799)	1256.27502925	ENSG00000168264 (IRF2BP2)

For the enriched regions ranked and selected by the peak finding algorithm, there are several possible changes of the score as shown in Table 2. As the purpose of analyzing ChIP-seq is to do some novel discovery, regions with type II or IV with final re-ranking medium score are paid more attention. In the following, two examples are given to illustrate how the regions with type II and IV are discovered. They also show the feasibility of our proposed framework.

For the regions with type II, a region at position from 87046200 to 87062199 on chromosome 16 is assigned a score of 1178 based on the number of short reads mapped to the regions. The ChIP-seq data are generated for H3K9Ac in EB cell line. The gene related to the region is ENSG00000179588 (ZFPM1). However, we can not find the pair of H3K9Ac and ZFPM1 in the list of pairs generated from the information extraction system. Based on the search keyword "ZFPM1" and "Histone", no results are even retrieved from the PubMed. Moreover, no information from microarray data or gene ontology are found to support the high score region. Based on our proposed framework, the region's score is decreased and more attention will be paid to the region and the related gene.

For the regions with type IV, a detailed example is shown in Figure 4. Firstly, thousands of enriched regions are discovered from ChIP-seq data based on a peak finding algorithm. Among the output regions, one region is initially not considered as an enriched region because of its low score generated from the peak finding algorithm. However if we check the related gene against other biologists' findings based on the microarray data, experimental results described in biomedical literature, and the Gene Ontology, the region would be enriched by H3K27em3. Especially, based on the sentence The results from these studies showed that H3K27me3 is associated primarily with the INK4A, and not the ARF, locus in the explanted fibroblasts, the pair of H3K27me3 and INK4A is extracted based on the information extraction system mentioned above. Generating such an error may be ascribed to the peak finding algorithm's inability of processing diffuse data. By employing the reranking model, the region is assigned a new score which will be considered as an enriched region. From this example, we speculate that employing the reranking model based on the multiple knowledge sources can improve the recall and reliability of the enriched region detection results.

V. CONCLUSION

In this paper, we have presented a novel framework of incorporating multiple knowledge sources in order to precisely analyze ChIP-seq data for histone modification. Information extracted from text, Gene Ontology, and knowledge mined from microarray data are combined in

TABLE 2
ENRICHED REGIONS BEFORE AND AFTER RE-RANKING

Region's score generated by Peak finding algorithm (Before re-ranking)	*information from multiple knowledge sources*	*Region's score (After re-ranking)*
High score (I)	Information supports the high score	High score
High score (II)	Information weakens the high score or no information supports the high score	Medium score
Low score (III)	Information supports the low score	Low score
Low score (IV)	Information weakens the low score or no information supports the low score	Medium score

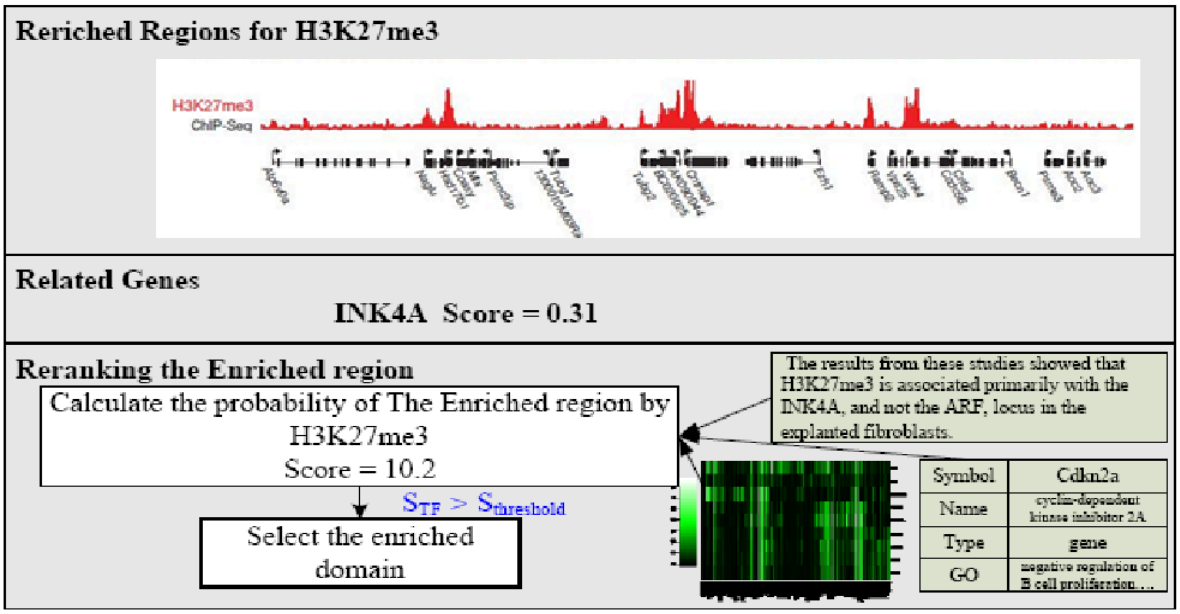

Fig. 4 An example of reranking the discovered enriched regions based on multiple knowledge sources.

a unified probabilistic model to rerank the enriched regions detected from peak finding algorithms. By filtering the reranked enriched regions, more reliable and precise results are generated. A case study has been presented to illustrate its feasibility. In future work we will continue on the development of the gene expression data clustering component and the gene ontology inference component and conduct a large scale of experiments to evaluate the system performance.

VI. ACKNOWLEDGEMENT

We would like to thank Augusto Rendon and Peter Smethurst for constructive suggestions on the proposed framework and Sylvia Nünberg for providing the ChIP-seq data.This article is supported by social science fund project in Jiangsu ,whose ID is 12DDB011.

REFERENCES

[1] Alejandro Vaquero, Alejandra Loyola, and Danny Reinberg. The constantly changing face of chromatin. Sci.Aging Knowl. Environ, 2003, 2003.
[2] Elaine R Mardis. Chip-seq: welcome to the new frontier. Nature Methods, (4):613 – 614, 2007.
[3] Hongkai Ji, Hui Jiang, Wenxiu Ma, David S Johnson, Richard M Myers, and Wing H Wong. An integrated software system for analyzing chip-chip and chip-seq data. Nature Biotechnology, 26:1293–1300, 2008.
[4] Yong Zhang, Tao Liu, Clifford Meyer, Jerome Eeckhoute, David Johnson, Bradley Bernstein, Chad Nussbaum, Richard Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based analysis of chip-seq (macs).Genome Biology, 9(9):R137, 2008.
[5] Joel Rozowsky, Ghia Euskirchen, Raymond K Auerbach, Zhengdong D Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B Gerstein. Peakseq enables systematic scoring of chip-seq experiments relative to controls. Nature Biotechnology, (27):66 – 75, 2009.
[6] Anton Valouev, David S Johnson, and Andreas Sundquist. Genome-wide analysis of transcription factor binding sites based on chip-seq data. Nature Methods, 5:829–834, 2008.
[7] Peter V Kharchenko, Michael Y Tolstorukov, and Peter J Park. Design and anlysis of chip-seq experiments for dna-binding proteins. Nature Biotechnology, 26:1351-1359, 2008.
[8] Samir J Courdy David A Nix and Kenneth M Boucher. Empirical methods for controlling false positives and estimating confidence in chip-seq peaks. BMC Bioinformatics, 9(523), 2008.
[9] Artem Barski Kairong Cui Raja Jothi, Suresh Cuddapah and Keji Zhao. Genome-wide identification of in

vivoprotein-dna binding sites from chip-seq data. Nucleic Acids Research, 36:5221–5231, 2008.
[10] Han Xu, Chia-Lin Wei, Feng Lin, and Wing-Kin Sung. An hmm approach to genome-wide identification of differential histone modification sites from chip-seq data. Bioinformatics, 24(20):2344–2349, October 2008.
[11] Chongzhi Zang, Dustin E. Schones, Chen Zeng, Kairong Cui, Keji Zhao, andWeiqun Peng. A clustering approach for identification of enriched domains from histone modification chip-seq data. Bioinformatics, 25(15):1952–1958, August 2009.
[12] Corinna Kolarik, Roman Klinger, and Martin Hofmann-Apitius. Identification of histone modifications in biomedical text for supporting epigenomic research. BMC Bioinformatics, 10:S28, 2009.
[13] L. Wong. PIES, a protein interaction extraction system. In Proceedings of the Pacific Symposium on Biocomputing., pages 520–531, Hawaii, U.S.A, 2001.
[14] Christian Blaschke and Alfonso Valencia. The Frame-Based Module of the SUISEKI Information Extraction system. IEEE Intelligent Systems, 17(2):14–20, 2002.
[15] I. Donaldson, J. Martin, B. de Bruijn, and C. Wolting. PreBIND and Textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics, 4(11), 2003.
[16] Jung-Hsien Chiang, Hsu-Chun Yu, and Huai-Jen Hsu.GIS: a biomedical text-mining system for gene information discovery. Bioinformatics, 20(1):120–121, 2004.
[17] Syed Toufeeq Ahmed, Deepthi Chidambaram, Hasan Davulcu, and Chitta Baral. IntEx: A Syntactic Role Driven Protein-Protein Interaction Extractor for BioMedical Text. In Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Database 2005, pages 54–61, 2005.
[18] TC Rindflesch, L Tanabe, JN.Weinstein, and L. Hunter. EDGAR: extraction of drugs, genes and relations from the biomedical literature. In Proceedings of Pacific Symposium Biocomputing, pages 517–28, 2000.
[19] David P. A. Corney, Bernard F. Buxton, William B. Langdon, and David T. Jones. BioRAT: extracting biological information from full-length papers. Bioinformatics, 20(17):3206–3213, 2004.
[20] Rzhetsky A, Iossifov I, Koike T, Krauthammer M, KraP, Morris M, Yu H, DuboulĘ PA, Weng W, Wilbur WJ,Hatzivassiloglou V, and Friedman C. GeneWays: a system for extracting, analyzing, visualizing, and integrat ing molecular pathway data. Journal of Biomedical Informatic, 37(1):43–53, February 2004.
[21] Hongbo Deng, Michael R. Lyu, and Irwin King. Effective latent space graph-based re-ranking model with global consistency. In Proceedings of the Second ACM International Conference on Web Search and Data Mining, pages 212–221, Barcelona, Spain, 2009.
[22] M. Collins. Ranking algorithms for named-entity extraction: Boosting and the voted perceptron. In Proceedings of the Annual meeting of the Association for Computational Linguistics (ACL) 2002, pages 489–496, 2002.
[23] Ruifang Ge and Raymond J. Mooney. Discriminative reranking for semantic parsing. In Proceedings of the conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL) 2006, pages 263–270, 2006.
[24] Kristina Toutanova, Aria Haghighi, and Christopher D. Manning. Joint learning improves semantic role labeling. In Proceedings of the Annual meeting of the Association for Computational Linguistics (ACL) 2005, pages 589 – 596, 2005.
[25] Y. He and S. Young. Semantic processing using the hidden vector state model. Computer Speech and Language,19(1):85–106, 2005.
[26] Deyu Zhou, Yulan He, and Chee Keong Kwoh. Extracting Protein-Protein Interactions from the Literature using the Hidden Vector State Model. International Journal of Bioinformatics Research and Applications, 4:64–80, 2008.
[27] Xiao peng hua and Shifei Ding.Incremental Learning Algorithm for Support Vector Data Description. Journal of Software, Vol 6, No 7 (2011), 1166-1173, Jul 2011
[28] Deyu Zhou and Yulan He. Discriminative Training of theHidden Vector State Model for Semantic Parsing. IEEE Transaction on Knowledge and Data Engineering, page In Press, 2008.
[29] Xixiang Zhang,Guangxue Yue, Xiajie Zheng and Fei Yu.Assigning Method for Decision Power Based on Linguisitc 2-tuple Judgment Matrices.Journal of Software, Vol 6, No 3 (2011), 508-515, Mar 2011.
[30] S. M. Masud Karim.Data Exchange: Algorithm for Computing Maybe Answers for Relational Algebra Queries.Journal of Software, Vol 6, No 1 (2011), 3-9, Jan 2011

Dafeng Chen , Male, was born in 1977, received the master degree of Engineering from Southeast University, China. He is a lecturer in Institute of Information Science and Technology, Nanjing Audit University since December 2000. As a primary principal or researcher, he has finished 3 national or ministry projects successively. He has wide research interests, mainly including computer audit, measuring and testing techniques, and Intelligent Control.

Deyu Zhou, Male,received the BS degree in mathematics and ME degree in computer science from Nanjing University, China, in 2000 and 2003, respectively. In 2009, he got the PhD degree in School of System Engineering, University of Reading, United Kingdom. Currently, he worked at School of Computer Science and Engineering, Southeast University. His interests are statistical methods for mining knowledge from biomedical data. includes the biography here.

Yuliang Zhuang, Male, Professor, Ph.D., Supervisor of Ph.D. Candidates. As a primary principal or researcher ,a lot of national or ministry projects have been finished successively. He has wide research interests and engages in the study of management information systems, electronic commerce, logistics and supply chain management etc.