BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 1 2015, pages 17–24
doi:10.1093/bioinformatics/btu604

Genome analysis

Advance Access publication September 5, 2014

SignalSpider: probabilistic pattern discovery on multiple normalized ChIP-Seq signal profiles

Ka-Chun Wong[1,2], Yue Li[1,2], Chengbin Peng[3] and Zhaolei Zhang[1,2,4,5,*]

[1]Department of Computer Science and [2]Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada, [3]CEMSE Division, King Abdullah University of Science and Technology, Thuwal, Jeddah, K.S.A., [4]Banting and Best Department of Medical Research and [5]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Chromatin immunoprecipitation (ChIP) followed by high-throughput sequencing (ChIP-Seq) measures the genome-wide occupancy of transcription factors *in vivo*. Different combinations of DNA-binding protein occupancies may result in a gene being expressed in different tissues or at different developmental stages. To fully understand the functions of genes, it is essential to develop probabilistic models on multiple ChIP-Seq profiles to decipher the combinatorial regulatory mechanisms by multiple transcription factors.

Results: In this work, we describe a probabilistic model (SignalSpider) to decipher the combinatorial binding events of multiple transcription factors. Comparing with similar existing methods, we found SignalSpider performs better in clustering promoter and enhancer regions. Notably, SignalSpider can learn higher-order combinatorial patterns from multiple ChIP-Seq profiles. We have applied SignalSpider on the normalized ChIP-Seq profiles from the ENCODE consortium and learned model instances. We observed different higher-order enrichment and depletion patterns across sets of proteins. Those clustering patterns are supported by Gene Ontology (GO) enrichment, evolutionary conservation and chromatin interaction enrichment, offering biological insights for further focused studies. We also proposed a specific enrichment map visualization method to reveal the genome-wide transcription factor combinatorial patterns from the models built, which extend our existing fine-scale knowledge on gene regulation to a genome-wide level.

Availability and implementation: The matrix-algebra-optimized executables and source codes are available at the authors' websites: http://www.cs.toronto.edu/∼wkc/SignalSpider.

Contact: zhaolei.zhang@utoronto.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 24, 2014; revised on August 17, 2014; accepted on September 2, 2014

1 INTRODUCTION

In human and other eukaryotes, gene expression is regulated by the binding of various modulatory transcription factors (TF) onto cis-regulatory DNA elements near genes. Binding of different combinations of TFs may result in a gene being expressed in different tissues or at different developmental stages. To fully understand a gene's function in the cell, it is essential to identify the TFs that regulate the gene and their corresponding TF binding sites (TFBS) (Wong *et al*., 2013). The techniques such as protein binding microarray (Berger *et al*., 2006), microfluidic affinity analysis (Fordyce *et al*., 2010) and protein microarray assays (Ho *et al*., 2006; Hu *et al*., 2009) enable us to measure the DNA sequence binding of TFs *in vitro*. On the other hand, the technology of chromatin immunoprecipitation (ChIP) followed by microarray (ChIP-ChIP) or sequencing (ChIP-Seq) (Johnson *et al*., 2007; Ren *et al*., 2000) can efficiently measure the binding occupancy of a particular TF on a genome-wide scale *in vivo*. In a typical ChIP-Seq study, the first step is to call the peaks, i.e. determining the precise location in the genome where the TF binds. A number of peak calling tools have been developed, but recent benchmark studies have shown that the peaks from those tools are often inconsistent (Laajala *et al*., 2009; Wilbanks and Facciotti, 2010).

Because multiple transcription factors often work in cis regulatory modules to confer complex gene regulatory programs, it is necessary to develop models on multiple ChIP-Seq datasets to decipher the combinatorial DNA-binding mechanism. In the following, we briefly review some of the previous works in this area. Gerstein *et al*. used pair-wise peak overlapping patterns to construct a human regulatory network (Gerstein *et al*., 2012). Xie *et al*. proposed self-organizing map methods to visualize the co-localization of DNA-binding proteins (Xie *et al*., 2013). Giannopoulou *et al*. proposed a non-negative matrix factorization to elucidate the clustering of DNA-binding proteins (Giannopoulou and Elemento, 2013). Zeng *et al*. proposed jMOSAiCS to discover histone modification patterns across multiple ChIP-Seq datasets (Zeng *et al*., 2013). Ferguson *et al*. have described a hierarchical Bayes approach to integrate multiple ChIP-Seq libraries to improve DNA binding event predictions. Mahony *et al*. also proposed a mixture model (MultiGPS) to detect differential binding enrichment of a DNA-binding protein in different cell lines, which can improve the protein's DNA binding location predictions (i.e. Cdx2 protein in their study) (Mahony *et al*., 2014). On the other hand, Chen *et al*. proposed a statistical framework (MM-ChIP) based on MACS to perform an integrative analysis of multiple ChIP datasets to predict ChIP-enriched regions with known motifs for a given DNA-binding protein (i.e. ER and CTCF proteins in their study) (Chen *et al*.,

*To whom correspondence should be addressed.

2011). On the other hand, Ji *et al.* proposed a differential principal component analysis method on ChIP-Seq to perform unsupervised pattern discovery and statistical inference to identify differential protein–DNA interactions between two biological conditions (Ji *et al.*, 2013). Guo *et al.* described a generative probabilistic model (GEM) for high resolution DNA binding site discovery from ChIP data (Guo *et al.*, 2012). Interestingly, that model combines ChIP signals and DNA motif discovery together to achieve precise predictions of the DNA binding locations of a DNA-binding protein. The authors have further demonstrated how GEM can be applied to reveal spatially constrained transcription factor binding site pairs on a genome.

Despite the success of the methods described above, to fully understand a gene's function, it is essential to develop probabilistic models on multiple ChIP-Seq profiles to decipher the genome-wide combinatorial patterns of DNA-binding protein occupancy. Unfortunately, the majority of the previous work usually focused on large-scale clustering of called peaks, which is an intuitive and straightforward approach. However such approaches have two limitations, as (i) peak-calling ignores the contributions from weak bindings of TFs, and (ii) pair-wise analysis ignores the complex combinatorial binding pattern among the TFs. Here we propose a new approach to build fine-scale probabilistic models for directly analyzing multiple normalized ChIP-Seq signal profiles on all the promoter and enhancer regions quantitatively so that weak bindings can be taken into account (Cheng *et al.*, 2012). Especially, its computational complexity has been carefully designed to scale well with the increasing ChIP-Seq data (i.e. linear complexity). After model training, we can extract and reveal the high-order combinatorial and quantitative occurrence patterns from the trained probabilistic models for better understandings on the DNA-binding protein combinatorics.

2 METHODS

2.1 Overall approach

We propose and describe a probabilistic three-layered model called SignalSpider to model the binding mechanisms of DNA-binding proteins across different genomic regions as depicted in Figure 1. Because it is generally believed that a genome can be clustered into different types of regions (Ernst and Kellis, 2012), we have put a clustering layer to cluster the genome regions into different types at the top ($\{x^t\}$). After the clustering, we model that different DNA-binding proteins can be found on each type of region. Thus, we put a binding mode layer in the middle to represent the binding modes of each DNA-binding protein ($\{b_j^t\}$). For example, if $b_j^t = 1$ (or 0), it means the *j*-th protein binds (or does not bind) to the *t*-th region. In particular, we note that b_j^t can have more than two binding modes here (e.g. $b_j^t \in \{0, 1, 2\}$) to account for partial binding modes (Lickwar *et al.*, 2012). Based on each binding mode of each DNA-binding protein at the middle layer ($\{b_j^t\}$), we assume a Gaussian distribution for its observed ChIP-Seq profile signal ($\{s_j^t\}$) at the bottom layer. We note that such a hierarchical structure is well-justified in most gene regulation contexts. For example, its hierarchy is similar to those of iASeq (Wei *et al.*, 2012) and CorMotif (Wei and Ji, 2014) but in different contexts. SignalSpider aims at modeling and extracting genome-wide combinatorial patterns from the normalized ChIP-Seq profile signals of DNA-binding proteins; iASeq aims at inferring allele-specificity probabilities of protein–DNA binding sites from ChIP-Seq read counts; CorMotif aims at modeling the t-statistics of genes' differential expression.

2.2 Model description

Mathematically, SignalSpider is a mixture model with a three-layer hierarchy as depicted in Figure 1. The main function of SignalSpider is to extract patterns from multiple signal profiles by estimating the combinatorial interaction between the top layer cluster and the middle layer signal components. For *t*th region, the bottom layer is to represent the observed signals $\{s_1^t, s_2^t, \ldots, s_M^t\}$ across M profiles directly. We empirically observe that the normalized ENCODE ChIP-Seq profile signals (after taking natural logarithm) always follow continuous Gaussian mixture distributions. Thus, we model each profile to follow the generation of one dimensional Gaussian mixture distribution with N components, where the hidden discrete variable b_j^t represents the component that is used to generate signal s_j^t. The top layer is the cluster layer consisting of K clusters. Each cluster influences the middle layer components used for generating signals via a discrete function α. The hidden discrete variable x^t represents which cluster is used. Mathematically, a SignalSpider Model can be defined as $\theta_{SS} = (\{\pi_i\}, \{\alpha_{ijk}\}, \{(\mu_{jk}, \sigma_{jk})\})$:

$$\theta_{SS} = (\{\pi_i\}, \{\alpha_{ijk}\}, \{(\mu_{jk}, \sigma_{jk})\})$$
$$\forall i \in \{1, 2, \ldots, K\}, \forall j \in \{1, 2, \ldots, M\}, \forall k \in \{1, 2, \ldots, N\}$$

where $\pi_i = P(x^t = i)$ is the prior probability of top layer *i*th cluster. $\alpha_{ijk} = P(b_j^t = k | x^t = i)$ is the conditional probability of the *k*th middle layer component, given that the current top layer cluster is the *i*th one, for the *j*th signal profile. (μ_{jk}, σ_{jk}) are the Gaussian mean and variance of the *j*th signal profile, given that its middle layer component is the *k*th one. In essence, it is a Bayesian network in which complete data likelihood can be written as: $L = \prod_{t=1}^{T} \pi_{x^t} \prod_{j=1}^{M} \alpha_{x^t j b_j^t} \mathcal{N}(s_j^t; \mu_{j b_j^t}, \sigma_{j b_j^t})$.

2.3 Model building

By taking partial derivatives to the expected complete data likelihood $E[log\, L]$ (plus adding Lagrange multipliers to the probability sum to one constraints) with respect to parameters to zero, we can derive the expectation maximization method (details can be found in Supplementary Data).

E step:

$$p(x^t = i, b_j^t = k | D, \theta)$$
$$= \frac{\pi_i \alpha_{ijk} N(s_j^t; \mu_{jk}, \sigma_{jk})}{\sum_{i=1}^{K} \sum_{k=1}^{N} \pi_i \alpha_{ijk} N(s_j^t; \mu_{jk}, \sigma_{jk})} \qquad \forall i, j, k, t$$

$$p(x^t = i | D, \theta) = \sum_{k=1}^{N} p(x^t = i, b_j^t = k | D, \theta) \qquad \forall i, t$$

$$p(b_j^t = k | D, \theta) = \sum_{i=1}^{K} p(x^t = i, b_j^t = k | D, \theta) \qquad \forall j, k, t$$

M step:

$$\pi_i = \frac{\sum_{t=1}^{T} p(x^t = i | D, \theta)}{T} \qquad \forall i$$

$$\alpha_{ijk} = \frac{\sum_{t=1}^{T} p(x^t = i, b_j^t = k | D, \theta)}{\sum_{t=1}^{T} p(x^t = i | D, \theta)} \qquad \forall i, j, k$$

$$\mu_{jk} = \frac{\sum_{t=1}^{T} p(b_j^t = k | D, \theta) s_j^t}{\sum_{t=1}^{T} p(b_j^t = k | D, \theta)} \qquad \forall j, k$$

$$\sigma_{jk} = \frac{\sum_{t=1}^{T} p(b_j^t = k | D, \theta)(s_j^t - \mu_{jk})^2}{\sum_{t=1}^{T} p(b_j^t = k | D, \theta)} \qquad \forall j, k$$

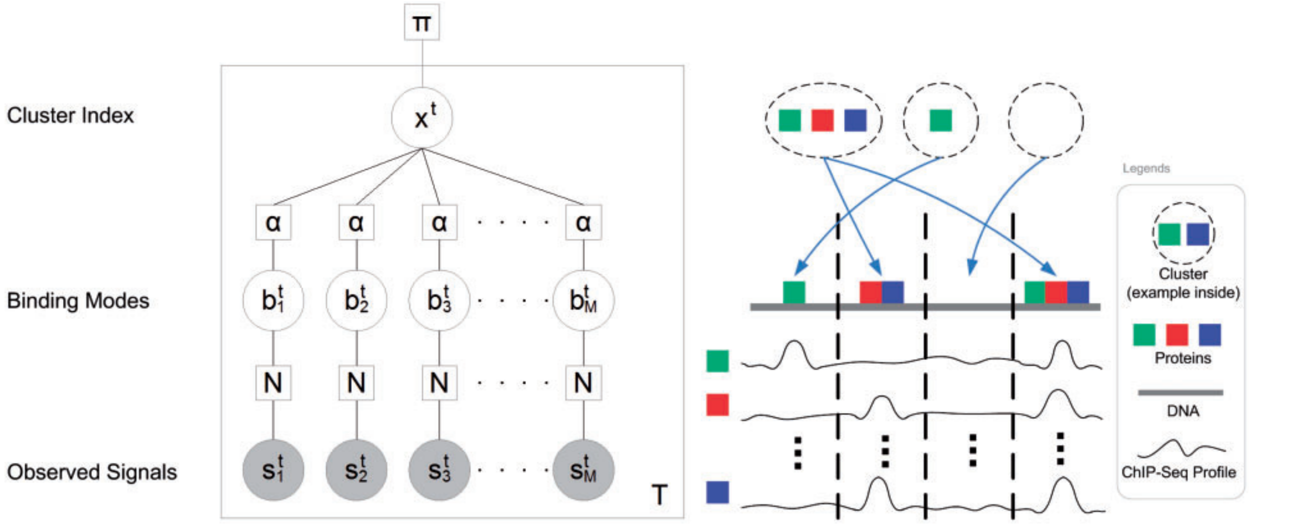


Fig. 1. Factor graph representations of SignalSpider in plate notation. Briefly, circles are variables and squares are functions. White and grey circles denote hidden and observed variables respectively

Briefly, all the model parameters ($\{\pi_i\}$, $\{\alpha_{ijk}\}$, $\{(\mu_{jk}, \sigma_{jk})\}$) are randomly initialized at the beginning. E-step and M-step are then alternated and repeated until the percentage of changes in the model parameters is numerically negligible (i.e. $<0.001\%$). Multiple runs are used to avoid local optima.

2.4 Complexity analysis

The prior probability of top layer ith cluster $\{\pi_i\}$ is $O(K)$. The conditional probability of the jth signal profile being in its kth middle layer component, given that the current top layer cluster is the ith one $\{\alpha_{ijk}\}$, is $O(KMN)$. The Gaussian mean and variance of the kth middle layer component for the jth signal profile $\{(\mu_{jk}, \sigma_{jk})\}$ is $O(MN)$. Limited by the parameters between top layer and middle layer (i.e. $\{\alpha_{ijk}\}$), the overall model complexity is $O(KMN)$.

On the other hand, the overall time complexity of each iteration of the expectation maximization method is $O(KMNT)$ because it is limited by computing the posterior probabilities of hidden variables $\{p(x^t = i, b_j^t = k|D, \theta)\}$ in the E-step, although we can pre-compute all the numerators before summing them up to obtain the denominators for speedup. In other words, its model complexity and building time complexity are both linear to the input data size. It can scale well with the increasing ChIP-Seq datasets.

3 RESULTS

3.1 Data sources

The primary data sources are the normalized ChIP-Seq signal profiles processed from the ENCODE consortium (ENCODE, 2012). They have been normalized by replicate aggregation and kernel smoothing by Wiggler (Hoffman *et al.*, 2013). In particular, we have downloaded all the available normalized ChIP-Seq signal profiles in the K562 cell line. For illustrative purpose, some of the ChIP-Seq normalized signal profiles for E2F4, CFOS, GATA2, JUNB and TBP are plotted in Supplementary Figure S1. From the figure, we can observe that the signal magnitude and variation are quite different among the profiles. It is challenging to extract any combinatorial pattern of the signal profiles by naive methods.

In particular, we limit our study to the human genome (hg19) regions designated as 'Promoter' or 'Enhancer' regions by both ChromHMM and Segway at 1000 base pairs resolution (Hoffman *et al.*, 2012). Maximal normalized ChIP-Seq signals are taken at those regions; natural logarithm has also been taken for efficient Gaussian modeling.

3.2 Tests on simulated ChIP-Seq profiles

3.2.1 Performance comparison To verify the robustness of SignalSpider, we seek to compare it with ChromHMM (Ernst and Kellis, 2012) and jMOSAiCS (Zeng *et al.*, 2013) on their genome clustering abilities in different parameter settings. We would like to note that, in addition to its ability to cluster genome regions, SignalSipider can also extract higher-order binding patterns among the DNA-binding proteins of interests from their ChIP-Seq data. Such ability is useful for interpreting the combinatorial transcription factor binding patterns encoded in multiple ENCODE ChIP-Seq data. It will be shown in the following section.

We have generated simulated data based on ENCODE ChIP-Seq data properties. In particular, the number of data (T) was set to the same size as in the following section (i.e. $T = 92, 485$); the number of clusters (K) was selected from $\{2, 3, \ldots, 10\}$; the number of profiles (M) was selected from $\{3, 4, 5, 10\}$ to reflect the usual clique sizes in the following section; the number of ChIP-Seq signal components of each DNA-binding protein (N) was set to 2 to accommodate the binary clustering nature of ChromHMM and jMOSAiCS (although SignalSpider can handle 3 or more in this aspect). For the means and variance of each simulated profile, we randomly paired it with an actual profile in the ENCODE ChIP-Seq data. For the actual profile paired, we fitted a one dimensional Gaussian mixture model on it with N components in 10 replicates because we observed Gaussian mixtures fit well to the actual ChIP-Seq normalized profile data after taking natural logarithm. The fitted model with the highest likelihood was then used as the basis for the simulated profile paired. By doing all the above, we obtained simulated datasets with known cluster labels in different parameter settings.

For each parameter setting, we generated 10 datasets. For each dataset, we ran SignalSpider, ChromHMM and jMOSAiCS on it with the corresponding parameter setting. Ten random runs with random initialization were used for SignalSpider and ChromHMM for each dataset. The run with the highest likelihood was taken as the final model for both SignalSpider and ChromHMM on each dataset. The convergence graphs are shown in the Supplementary Data. In summary, it can be observed that SignalSpider converges within 20 iterations for most cases. For example, the convergence graph for the 10 runs with the setting $K = 10$ and $M = 10$ is depicted in Supplementary Figure S2.

For each dataset, we evaluate the clustering ability of each method based on two metrics: Rand Statistics (RS) (Halkidi *et al.*, 2001) and Purity (Zhao and Karypis, 2002). Rand Statistics is based on the intra-cluster similarity and inter-cluster dissimilarity. For the intra-cluster similarity, if a pair of data vectors is in the same cluster in both the target result and the clustering result, then the score will be increased by one. For the inter-cluster dissimilarity, if a pair of vectors is in different clusters in both the target result and the clustering result, then the score will also be increased by one. On the contrary, if a pair of data vectors is in the same cluster in the target result, but not in the clustering result, the score will not be increased. After we have checked all the possible pairs, the score is normalized by the total number of possible pairs. On the other hand, purity solely measures the intra-cluster similarity. Nevertheless, it is useful in the sense that we only care about the quality of individual clusters. Their mathematical definitions can be found in Supplementary Data. The results are depicted in Figure 2. It can be observed that SignalSpider has a better performance than ChromHMM and jMOSAiCS on the given datasets. We reason that it is because ChromHMM and jMOSAiCS involve peak-calling, ignoring the weak DNA-binding events.

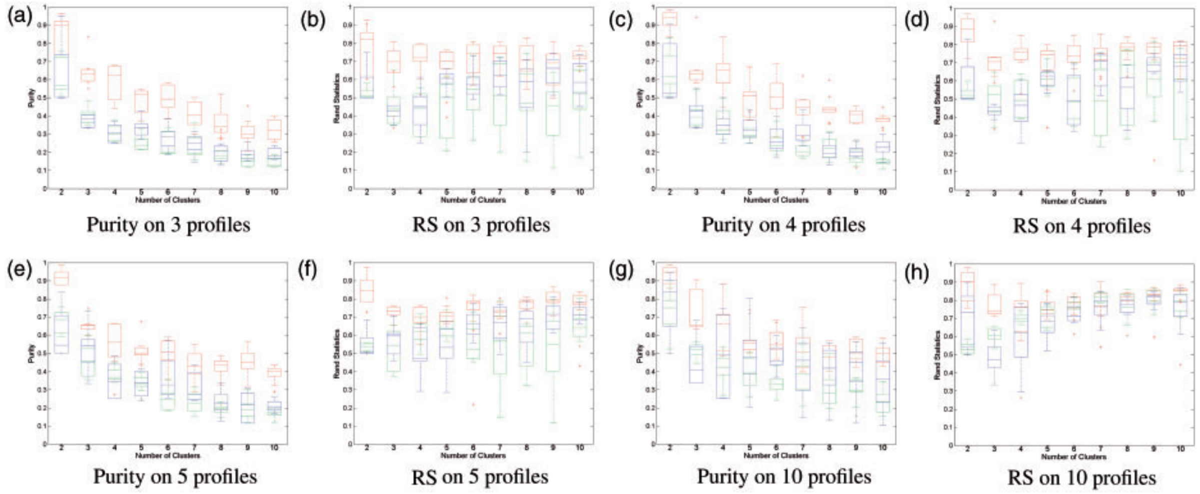
3.2.2 Sensitivity analysis on cluster prior probabilities It has been reported that weak clusters may be absorbed into strong clusters during a clustering process (Wei and Ji, 2014). To examine its effect, we have conducted a sensitivity analysis on cluster prior probabilities under two simulations (Scenario 1 and 2; details can be found in Supplementary Data). Briefly, for each simulation, we examine how often a weak cluster (with low prior cluster probability) can be discovered in the presence of another weak cluster and a strong cluster (with high prior cluster probability) in 100 random runs. From the simulations, it can be observed that weak clusters are more difficult to be discovered than strong clusters because weak clusters tend to be merged into a discovered strong cluster. Nonetheless, from the results, we can observe that such phenomenon is not severe (e.g. $>70\%$ discovery rate for clusters with 0.2 prior probabilities or higher). Multiple runs of SignalSpider with different initializations can be adopted to circumvent such problems.

3.3 Complex pattern discovery

To demonstrate the higher-order (beyond pair-wise) pattern discovery of SignalSpider, the available protein–protein pair-wise interaction data is used as the basis so that we don't have to enumerate all the possible sets which are computationally infeasible.

3.3.1 Data acquisition For the DNA-binding proteins we have identified from the ENCODE ChIP-Seq data, we mapped them onto the BioGRID protein–protein interaction database (Chatr-Aryamontri *et al.*, 2013). Based on the network mapped, we extract the maximal cliques with $k = 3$ (Palla *et al.*, 2005). Note that maximal cliques are defined as the cliques with at least three nodes which subgraphs are also cliques. By choosing $k = 3$, we ensure that the output cliques have at least three nodes for discovering higher-order combinatorial patterns on at least three profiles (beyond pair-wise patterns on two profiles which have already been given from BioGRID). The clique list is shown Supplementary Table S1.

3.3.2 Model building For each clique, we trained SignalSpider on the ChIP-Seq signal profiles of the corresponding DNA binding proteins belonging to the clique. In the parameter settings,



Purity on 3 profiles

RS on 3 profiles

Purity on 4 profiles

RS on 4 profiles

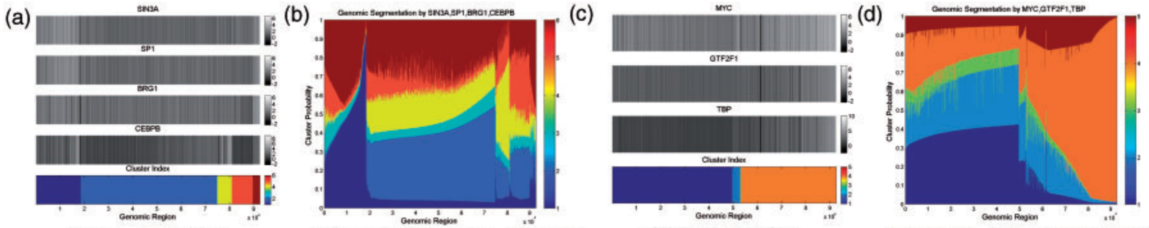
Purity on 5 profiles

RS on 5 profiles

Purity on 10 profiles

RS on 10 profiles

Fig. 2. Performance comparison for SignalSpider, ChromHMM and jMOASiCs. The red colour corresponds to SignalSpider; the blue colour corresponds to ChromHMM; the green colour corresponds to jMOASiCs. RS stands for Random Statistics (Halkidi *et al.*, 2001)



Clique 11 Profiles

Clique 11 Clustering Probabilities

Clique 31 Profiles

Clique 31 Clustering Probabilities

Fig. 3. Examples of clustering performed by SignalSpider. In all the panels, the horizontal axis represents the bins $\{1, 2, \ldots, T\}$. For panels (**a**) and (**c**), the grey scheme represents ChIP-Seq signal values while the colour scheme indicates the cluster index x^t. For panels (**b**) and (**d**), the vertical height represents each cluster probability while the colour scheme indicates the cluster index x^t. The clustering probabilities $P(x^t)$ are calculated by computing belief propagation (sum product algorithm) from the leaves to the root of the graph topology which is outlined in Figure 1. The horizontal region order (from left to right) of the panels is sorted by the ascending cluster index i which is determined by $\arg\max_i\ P(x^t = i)$ for all bins (i.e. $\forall\ t \in \{1, 2, \ldots, T\}$)

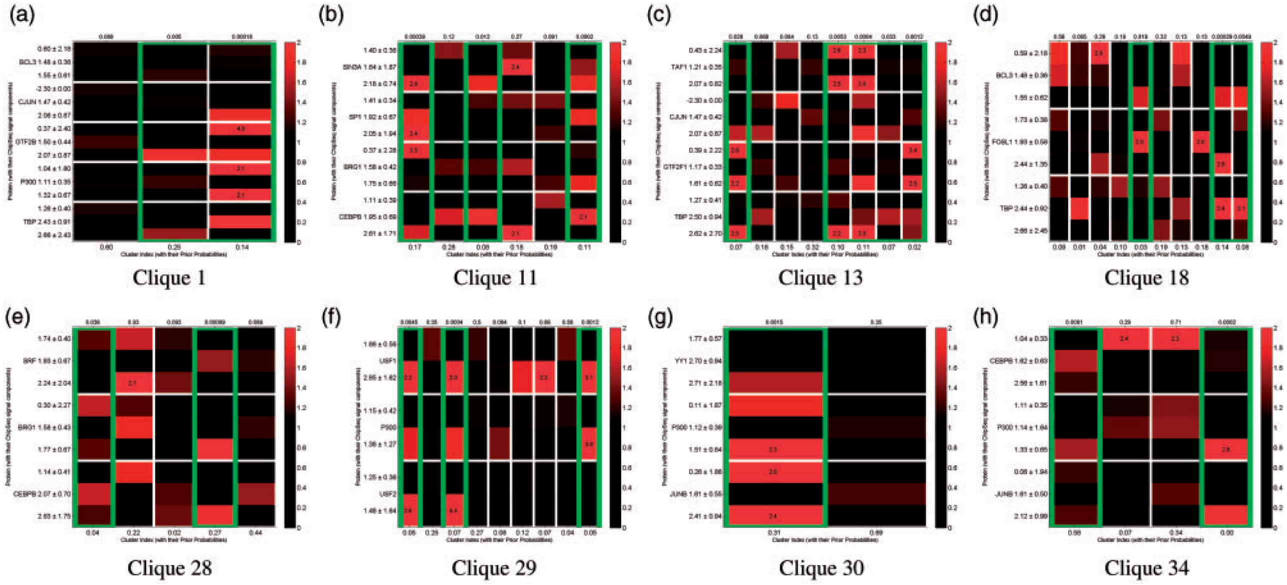
the number of (top layer) clusters (K) was varied from 2 to 10; number of (middle layer) ChIP-Seq signal components of each DNA-binding protein (N) was varied from 2 to 3; 10 replicate runs were used for each parameter setting. The SignalSpider model with the highest marginal data likelihood was then taken as the representative model for each clique. Note that the SignalSpider model has equality constraints for regularization (more details can be found in the Supplementary Data). To verify its modeling, we can marginalize and compute individual (middle layer) ChIP-Seq signal component occurring probability using $P(b_j = k) = \sum_{i=1}^{M} P(b_j = k|x = i)P(x = i) = \sum_{i=1}^{M} \alpha_{ijk}\pi_i$ and plot its marginalized Gaussian mixture distribution, some of which are depicted in Supplementary Figure S4. From the figure, we can observe that SignalSpider can model the signal distributions precisely. Another practical aspect of SignalSpider is its soft clustering ability. We can verify its modeling by looking at the top layer cluster probabilities $P(x^t)$ as shown in Figure 3. From the figure, we can observe that the profiles are well clustered by SignalSpider into different combinations according to their signal values.

3.3.3 Higher-order combinatorial pattern discovery In particular, we are interested in the patterns enriched in each clique model. Thus, we calculated the fold enrichment for each ChIP-Seq signal component in each cluster $P(b_j = k|x = i)$, comparing to its own marginal occurring probability $P(b_j = k)$. Mathematically, it is calculated as:

$$FOLD_{ijk} = \frac{P(b_j = k, x = i)}{P(b_j = k)P(x = i)} = \frac{\alpha_{ijk}}{\sum_{i=1}^{M} \alpha_{ijk}\pi_i}$$

where $FOLD_{ijk}$ is the fold enrichment value for the kth ChIP-Seq signal component of the jth DNA-binding protein in the ith cluster. To visualize all the clique models, we have plotted the enrichment maps with full discussions in Supplementary Data. Representative examples are discussed and depicted in Figure 4.

For Clique 1, we observe the prominent co-occurrence of CJUN, GTF2FB, P300 and TBP in 14% of genomic ChIP-Seq signal regions. These proteins are known to be part of the preinitiation complex (PIC) for initiating gene transcription as shown in Figure 8 of Martens *et al*. (2003). On the other hand, GTF2B and TBP are further found to be co-associative in another 25% genomic regions, which provides genome-wide evidence on their direct binding as revealed by a recent PIC crystal structure study (Murakami *et al*., 2013). For Clique 11, we can observe there is a strong enriched co-occurrence pattern between SIN3A, SP1 and CEBPB. It has been shown in a previous study that SIN3A forms complexes with p53 and HDAC1 to repress the ERalpha promoter on which SP1 and the CCAAT binding site bound by CEBPB are found. Our model provides additional genome-wide evidence for such interactions (De Amicis *et al*., 2011). For Clique 13, we observe the same PIC pattern as in clique 1. Notably, Clique 13 reveals more details in the formation of PIC than clique 1. For instance, TAF1 and TBP show a strong co-association pattern because both of them are part of Transcription factor II D (TFIID). On the other hand, GTF2F1 is part of Transcription factor II F (TFIIF). TFIID and TFIIF are known to form the RNA polymerase II preinitiation complex with other transcription factors (including CJUN) to initialize gene transcription, inducing the co-occurrence pattern observed by our model (Ruppert and Tjian, 1995). For Clique 18, we observe that TBP shows enriched occurrence in 22% of ChIP-Seq signal regions. It is consistent with the estimate that around 24% of human genes contain a TATA box within the core promoter (Yang *et al*., 2007). Among those, as suggested by the co-occurrence pattern, half of them are associated with the activating protein-1 (AP-1) which is composed of FOSL1 and activated by BCL3 (Na *et al*., 1999). For Clique 28, we observe different pair-wise co-occurrence patterns of SRF, BRG1 and CEBPB. It is expected since each pair of them is shown to work in different tissues. For instance, SRF and BRG1 are shown in complex to regulate muscle-specific gene expression (Zhang *et al*., 2007); BRG1 and CEBPB are shown to interact for regulating the beta- and gamma-casein promoters (Xu *et al*., 2007); SRF and CEBPB are demonstrated to regulate the serum response element (Hanlon and Sealy, 1999). For Clique 29, we observe that P300 co-occurs with USF1 and USF2. In the former study by Huang *et al*., it was shown that the USF family can recruit histone modification complexes including P300 for maintaining chromatin barriers (Huang *et al*., 2007). Taken it together with our results, it suggests that such process may be prevalent across the genome. For Clique 30, we observe that YY1, P300 and JUNB co-occur in 31% of ChIP-Seq regions. Independent of this study, Wang *et al*. have proposed a model explaining their co-occurrence (Wang *et al*., 2007). In that model, YY1 binds to the promoter region of the gene HLJ1 while AP-1 (which is



Clique 1 Clique 11 Clique 13 Clique 18

Clique 28 Clique 29 Clique 30 Clique 34

Fig. 4. Fold enrichment maps for the SignalSpider models we have learnt for different cliques. Each clique is composed of multiple DNA-binding proteins, each of which has its own normalized ChIP-Seq signal profile. For each figure, the bottom horizontal axis denotes different prior clusters, while the vertical axis denotes different ChIP-Seq signal components for each DNA-binding protein. The top axis denotes the *P*-values of the cluster patterns. The cluster patterns with $P \leq 0.05$ are highlighted in green rectangles (see Supplementary Data for the *P*-value calculation method). The red and black colour scheme indicates different levels of fold enrichment for each ChIP-Seq signal component across different prior clusters

composed of JUNB) binds to the enhancer region of the same gene. P300 is used to mediate the interaction between YY1 and AP-1 through DNA bending or looping mechanisms. Such a local mechanism supports the biological validity of our results. On the other hand, our results provide genome-wide evidences on them. For Clique 34, we observe that JUNB is more enriched with P300 than CEBPB. It is interesting because P300 and CEBPB have similar structures and functions as gene activators via acetylating histones and recruiting other transcription factors (Liu *et al.*, 2008). Taken together with our results, it suggests that P300 is preferred more by JUNB to interact than CEBPB although P300 and CEBPB are functionally similar on a genome-wide basis.

3.3.4 Comparison with jMOSAiCs We have run jMOSAiCS (Zeng *et al.*, 2013) (with its default setting) on the same clique datasets. The enrichment patterns discovered by jMOSAiCS are depicted in Supplementary Figures S7 and S8. In general, we observe that the patterns discovered by SignalSpider are largely consistent with those discovered by jMOSAiCS. However, the patterns discovered by SignalSpider are more informative than those by jMOSAiCS according to the enrichment analysis because SignalSpider can consider the ChIP-Seq signal contributions from weak bindings of TFs. For instance, the weak bindings of the P300 protein can be quantitatively revealed by SignalSpider as depicted in Supplementary Figures S7 and S8. In addition, SignalSpider can learn the depletion patterns which most existing methods are difficult to capture.

3.3.5 Gene ontology enrichment analysis For each cluster in each clique, we ran Gene Ontology (GO) enrichment analysis using Fisher's test. For each promoter or enhancer region, we take the gene within 1000 bp downstream as the target gene. Different statistical significant GO terms (Biological Process) are found for each cluster (Supplementary Data). For example, for Clique 1, we observed that the GO term 'cellular response to external stimulus' (GO:0071496) is statistically enriched ($P = 0.00028$) for the cluster 3 regions (with enrichment of CJUN) but not in cluster 2 regions (without enrichment of CJUN). It is consistent with the known functional role of CJUN which binds with CFOS to form the AP-1 early response transcription factor (Hess *et al.*, 2004). More examples can be found in Supplementary Data.

3.3.6 Evolutionary conservation enrichment analysis For each cluster in each clique, we checked their conservation using PhastCons with other 99 vertebrate genomes (phastCons100way) (Siepel *et al.*, 2005). For each cluster, we computed the mean of the PhastCons scores across its regions. Interestingly, we can observe that, if a cluster involves enriched co-occurrence of at least three proteins (active cluster regions), it is usually more conserved than the background (i.e. all the regions we have considered in this study). Their statistical significances (*P*-values) are justified by both Wilcoxon rank sum test and *t*-test ($P < 0.0001$). Examples are depicted in Figure 5. Complete results can be found in Supplementary Data.

3.3.7 Chromatin interaction enrichment analysis We next investigated whether the regions in each clique are more likely to form three-dimensional chromatin interactions inside the K562 cell nuclei. We downloaded and extracted the K562 high-confidence Hi-C interaction data from Fit-Hi-C (Ay *et al.*, 2014). To accommodate the Hi-C data resolution, each interaction region midpoint is extended to 50 000 bp in both directions. We set a q-value threshold to 0.05, resulting in 37 893 very high-confidence chromatin interactions (shown in Supplementary Data) among

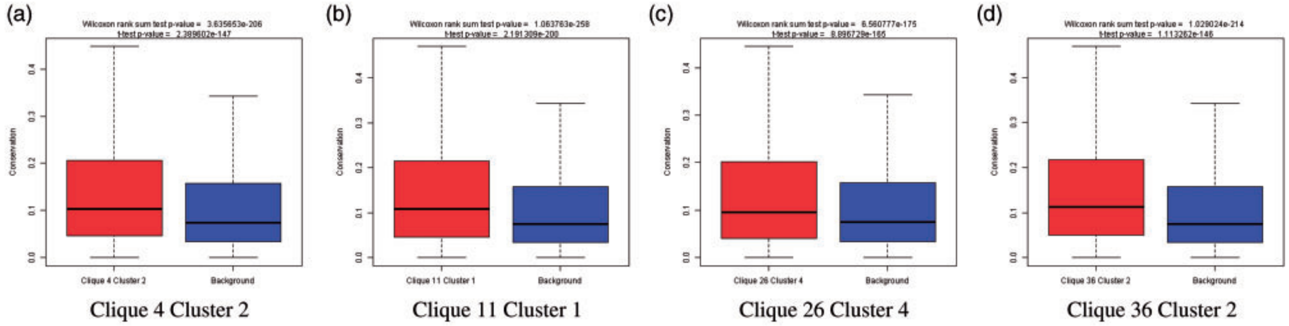


Fig. 5. Examples of conserved regions. The vertical axis is the conservation (PhastCons scores), while the horizontal axis denotes two regions. The left boxes correspond to the cluster regions of interest, while the right boxes denote the background region. Outliers are omitted. *P*-values for Wilcoxon rank sum test and *t*-test are shown at the title

the 92 485 regions we have considered (4 276 691 370 possible interactions). After that, for each possible interaction pair in each cluster, we check whether it belongs to those 37 893 very high-confidence chromatin interactions. Interestingly, we found that those active clusters in the last paragraph are also enriched in Hi-C interaction pairs. For instance, all the clusters shown in Figure 5 are enriched for Hi-C chromatin interactions using Fisher exact test ($P < 0.0001$). Details can be found in Supplementary Data.

3.4 Revealing DNA-binding protein machineries

Because SignalSpider is linear in complexity, we have built a SignalSpider model on all the ChIP-Seq profiles studied here using 100 cluster indices with 2 binding modes. The cluster centroids are depicted in Supplementary Figure S9. We observe that SignalSpider can reveal different DNA-binding protein machineries across different regions. For example, we observe that the chromatin structure protein profiles (e.g. CTCF) are well-separated from the other clusters, and thus consistent with their insulator roles.

4 DISCUSSION

With the prevalence of ChIP-Seq technology, massive ChIP-Seq data have been accumulated. The current ChIP-Seq data analysis studies usually involve single profile peak calling. Two areas of improvement are currently being explored in the field: how to integrate binding profiles of multiple transcription factors and how to include the signals from the areas that are not called as peaks. In light of that, we have developed a probabilistic model to analyse multiple ChIP-Seq signal profiles directly. The model (SignalSpider) not only can cluster genome into several region types, but also can extract subtle binding combinations from multiple DNA-binding proteins. The results show that the proposed model demonstrated performance better than the other existing methods on the simulated data derived from the available ChIP-Seq dataset. With the support of GO enrichment analysis, evolutionary conservation, chromatin interaction enrichment analysis and available wet-lab studies, we found that the discovered patterns show significant genome-wide insights. In particular, the patterns indicate that the genome-wide DNA-binding protein occurrence patterns are far from simple co-associative pair-wise interactions. Instead, we observed different higher-order enrichment and depletion patterns across different proteins. With increasing available ChIP-Seq data, we envision that SignalSpider will be useful in understanding the genome-wide DNA-binding dynamics.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their positive and constructive comments.

Funding: Discovery Grant from Natural Sciences and Engineering Research Council, Canada (NSERC), grant number [327612-2009 RGPIN to Z.Z.]; Acres Inc. - Joseph Yonan Memorial Fellowship, Kwok Sau Po Scholarship, and International Research and Teaching Assistantship from University of Toronto (to K.W.).

Conflict of interest: none declared.

REFERENCES

Ay,F. *et al.* (2014) Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.*, **24**, 999–1011.

Berger,M.F. *et al.* (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.

Chatr-Aryamontri,A. *et al.* (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D816–D823.

Chen,Y. *et al.* (2011) MM-ChIP enables integrative analysis of cross-platform and between-laboratory ChIP-ChIP or ChIP-seq data. *Genome Biol*, **12**, R11.

Cheng,C. *et al.* (2012) Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.*, **22**, 1658–1667.

De Amicis,F. *et al.* (2011) Resveratrol, through NF-Y/p53/Sin3/HDAC1 complex phosphorylation, inhibits estrogen receptor alpha gene expression via p38MAPK/CK2 signaling in human breast cancer cells. *FASEB J.*, **25**, 3695–3707.

ENCODE. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.

Fordyce,P.M. *et al.* (2010) *De novo* identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nat. Biotechnol.*, **28**, 970–975.

Gerstein,M.B. *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**, 91–100.

Giannopoulou,E.G. and Elemento,O. (2013) Inferring chromatin-bound protein complexes from genome-wide binding assays. *Genome Res.*, **23**, 1295–1306.

Guo,Y. *et al.* (2012) High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput. Biol.*, **8**, e1002638.

Halkidi,M. *et al.* (2001) On clustering validation techniques. *J. Intell. Inf. Syst.*, **17**, 107–145.

Hanlon,M. and Sealy,L. (1999) Ras regulates the association of serum response factor and CCAAT/enhancer-binding protein beta. *J. Biol. Chem.*, **274**, 14224–14228.

Hess,J. *et al.* (2004) AP-1 subunits: quarrel and harmony among siblings. *J. Cell. Sci.*, **117** (Pt 25), 5965–5973.

Ho,S.W. *et al.* (2006) Linking DNA-binding proteins to their recognition sequences by using protein microarrays. *Proc. Natl Acad. Sci. USA*, **103**, 9940–9945.

Hoffman,M.M. *et al.* (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**, 473–476.

Hoffman,M.M. *et al.* (2013) Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.*, **41**, 827–841.

Hu,S. *et al.* (2009) Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling. *Cell*, **139**, 610–622.

Huang,S. *et al.* (2007) USF1 recruits histone modification complexes and is critical for maintenance of a chromatin barrier. *Mol. Cell. Biol.*, **27**, 7991–8002.

Ji,H. *et al.* (2013) Differential principal component analysis of ChIP-seq. *Proc. Natl Acad. Sci. USA*, **110**, 6789–6794.

Johnson,D.S. *et al.* (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*, **316**, 1497–1502.

Laajala,T.D. *et al.* (2009) A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics*, **10**, 618.

Lickwar,C.R. *et al.* (2012) Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature*, **484**, 251–255.

Liu,X. *et al.* (2008) The structural basis of protein acetylation by the p300/CBP transcriptional coactivator. *Nature*, **451**, 846–850.

Mahony,S. *et al.* (2014) An integrated model of multiple-condition ChIP-Seq data reveals predeterminants of Cdx2 binding. *PLoS Comput. Biol.*, **10**, e1003501.

Martens,J.H. *et al.* (2003) Cascade of distinct histone modifications during collagenase gene activation. *Mol. Cell. Biol.*, **23**, 1808–1816.

Murakami,K. *et al.* (2013) Architecture of an RNA polymerase II transcription preinitiation complex. *Science*, **342**, 1238724.

Na,S.Y. *et al.* (1999) Bcl3, an IkappaB protein, stimulates activating protein-1 transactivation and cellular proliferation. *J. Biol. Chem.*, **274**, 28491–28496.

Palla,G. *et al.* (2005) Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435**, 814–818.

Ren,B. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Ruppert,S. and Tjian,R. (1995) Human TAFII250 interacts with RAP74: implications for RNA polymerase II initiation. *Genes Dev.*, **9**, 2747–2755.

Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

Wang,C.C. *et al.* (2007) Synergistic activation of the tumor suppressor, HLJ1, by the transcription factors YY1 and activator protein 1. *Cancer Res.*, **67**, 4816–4826.

Wei,Y. and Ji,H. (2014) Joint analysis of differential gene expression in multiple studies using correlation motifs. *Biostatistics*, doi:10.1093/biostatistics/kxu038.

Wei,Y. *et al.* (2012) iaseq: integrative analysis of allele-specificity of protein-DNA interactions in multiple ChIP-seq datasets. *BMC Genomics*, **13**, 681.

Wilbanks,E.G. and Facciotti,M.T. (2010) Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One*, **5**, e11471.

Wong,K.C. *et al.* (2013) DNA Motif Elucidation using Belief Propagation. *Nucleic Acids Research*, **41**, e153.

Xie,D. *et al.* (2013) Dynamic trans-acting factor colocalization in human cells. *Cell*, **155**, 713–724.

Xu,R. *et al.* (2007) Extracellular matrix-regulated gene expression requires cooperation of SWI/SNF and transcription factors. *J. Biol. Chem.*, **282**, 14992–14999.

Yang,C. *et al.* (2007) Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene*, **389**, 52–65.

Zeng,X. *et al.* (2013) jMOSAiCS: joint analysis of multiple ChIP-seq datasets. *Genome Biol.*, **14**, R38.

Zhang,M. *et al.* (2007) A novel role of Brg1 in the regulation of SRF/MRTFA-dependent smooth muscle-specific gene expression. *J. Biol. Chem.*, **282**, 25708–25716.

Zhao,Y. and Karypis,G. (2002) Criterion functions for document clustering: experiments and analysis. In: *Technical report*. Department of Computer Science, University of Minnesota.