BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 6 2014, pages 823–830
doi:10.1093/bioinformatics/btt596

Systems biology

Advance Access publication November 5, 2013

Integrating multiple resources to identify specific transcriptional cooperativity with a Bayesian approach

Pengzhan Hu, Zhongchao Shen, Haibo Tu, Li Zhang and Tieliu Shi*

Center for Bioinformatics and Computational Biology, Shanghai Key Laboratory of Regulatory Biology, the Institute of Biomedical Sciences and School of Life Sciences, East China Normal University, Shanghai 200241, China

Associate Editor: Igor Jurisica

ABSTRACT

Motivation: Limited cohort of transcription factors is capable to structure various gene-expression patterns. Transcriptional cooperativity (TC) is deemed to be the main mechanism of complexity and precision in regulatory programs. Although many data types generated from numerous experimental technologies are utilized in an attempt to understand combinational transcriptional regulation, complementary computational approach that can integrate diverse data resources and assimilate them into biological model is still under development.

Results: We developed a novel Bayesian approach for integrative analysis of proteomic, transcriptomic and genomic data to identify specific TC. The model evaluation demonstrated distinguishable power of features derived from distinct data sources and their essentiality to model performance. Our model outperformed other classifiers and alternative methods. The application that contextualized TC within hepatocarcinogenesis revealed carcinoma associated alterations. Derived TC networks were highly significant in capturing validated cooperativity as well as revealing novel ones. Our methodology is the first multiple data integration approach to predict dynamic nature of TC. It is promising in identifying tissue- or disease-specific TC and can further facilitate the interpretation of underlying mechanisms for various physiological conditions.

Contact: tieliushi01@gmail.com

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

Received on May 9, 2013; revised on September 26, 2013; accepted on October 14, 2013

1 INTRODUCTION

In the human genome, genes manifest dramatic diversity in terms of expression levels and tissue- or condition-specific expression patterns. Despite this tremendous diversity, all genes are controlled by numbered transcription factors (TFs). TFs can work as monomers, homodimers and heterodimers to facilitate (as an activator) or inhibit (as a repressor) the recruitment of RNA polymerase and display high specificity in ligand activation thereby precisely controlling the condition-dependent expression of target genes (TGs) (Bookout *et al.*, 2006; Latchman, 1997). TF activity and interaction is characterized of condition-dependence and transience. Such mutual regulation among TFs underlies major features of cellular identity and complex functions such as pluripotency (Kim *et al.*, 2008), development (Lagha *et al.*, 2012) and differentiation (Braun and Gautel, 2011; Fei and Chen, 2010) even diseases (Hoffmeyer *et al.*, 2012), which makes the study of transcriptional cooperativity (TC) of paramount importantance.

The development of numerous experimental technologies including whole-genome sequencing, large-scale expression profiling, yeast two-hybrid experiments and emerging chromatin immunoprecipitation makes it possible to understand the combinational transcriptional regulation from various perspectives. Biologically relevant synergistic interactions between different TFs are discovered in yeast by insight into gene expression data (Banerjee and Zhang, 2003). Protein–protein interaction (PPI) data is also used to infer synergistic binding of cooperative TFs (Nagamine *et al.*, 2005). Another study that can indicate the interactions between TFs is based on the genome sequence (Yu *et al.*, 2006a. b). More recently, the ENCODE Consortium provides comprehensive Chromatin immunoprecipitation coupled with high-throughput sequencing data for reliable view of TF-binding-site interaction (Gerstein *et al.*, 2012; Wang *et al.*, 2012; Whitfield *et al.*, 2012).

Previous studies on the investigation of TC are reasonable in assumptions or biology. Many TFs bind DNA preferentially at characteristic sequence motifs, thereby providing the sequence specificity required to reflect the possible spatial relationships between TFs in terms of preferences in distance and orientation (Lemon and Tjian, 2000). The interaction between two proteins indicates they contribute to the same or similar biological processes (Nagamine *et al.*, 2005). The gene profiles reveal the similarity between gene expressions, which could be used to infer synergistic relationships between TFs when their common TGs show highly correlated expression patterns (Pilpel *et al.*, 2001; Yu *et al.*, 2003). Although previous studies have made encouraging progress, the investigation of TC encounters four major challenges: (i) the time consumption and high costs of valid experiments; (ii) the vast majority of TFs not being experimentally profiled genome wide; (iii) the difficulty of interrogating the activities of multiple TFs within the same cellular environment; (iv) mainly focusing on static or global interaction but ignoring the dynamic nature of TC. All of these have underscored the need to integrate various data and construct a complementary computational approach that can assimilate them into biological models.

The TC defined in this article broadly means the functional interaction between different DNA-binding TFs in regulation of gene expression. It could be physical, genetic or regulatory interaction between TFs, competitive or synergetic co-regulation. Furthermore, TF homodimer interactions and interactions

*To whom correspondence should be addressed.

between TFs and non-DNA-binding cofactors as well as cooperativity among multiple TFs are not within our research emphasis.

In this article, we proposed a novel algorithm based on Bayesian theory for integrative analysis of diverse large-scale biological data to identify TC. Our strategy is based on hypothesis that cooperative TF pairs share shorter distance in PPI network, closer spatial relationship on promoter regions, stronger distance constraint among binding sites and higher correlation on gene expression pattern. We derived four features and tested their discrimination based on HepG2 golden standard dataset. Benchmark analyses were implemented for the assessment of model performance compared with individual feature and straight Logistic regression as well as one-feature-removed Bayesian model. Furthermore, we performed comparison with other existing methods. Bayesian approach was then applied to contextualize TC within hepatocellular carcinoma (HCC) progression.

2 MATERIALS AND METHODS

2.1 Data sources

Our study started with the collection of informative data resources, which include PPI data, position weighted matrices (PWMs), TF–TG regulatory relationships and gene-expression data of specific tissue–disease. The PPI data used to calculate PPI score are publicly available from HPRD (Prasad *et al.*, 2009). The PWMs were obtained from TRANSFAC (Matys *et al.*, 2003), JASPAR (Bryne *et al.*, 2008) and researches on ENCODE data (Wang *et al.*, 2012). Promoter sequence set was built by extracting sequences 1000 bases upstream of annotated transcription starts of RefSeq genes with annotated 5' UTRs from Human genome sequence (hg19). TF–TG regulatory relationships were collected from three distinct databases: TRED (Jiang *et al.*, 2007), InnateDB (Lynn *et al.*, 2008) and HTRIdb (Bovolenta *et al.*, 2012). We used microarray data of human liver cell line (HepG2) induced by PCB153 (GSE6494) for feature and model evaluation, neuroblastomas SK–N–SH cells induced by retinoic acid (GSE9169) for method comparison and gene profiles during HCC progression (GSE25097) for application.

2.2 The PPI score

First, we constructed a PPI network based on non-redundant PPI data from HPRD that consists of 39 240 PPIs involving 9673 proteins. Then, extended Czekanowski–Dice distance was used to calculate the distances between proteins (Brun *et al.*, 2004). It can express diversity and specificity of distances between proteins adequately. For the proteins *A* and *B* with the range *l*, *D*(*A*, *B*, *l*) was defined as follows:

$$D(A,B,l)=\frac{\sum_{k=1}^{l}\frac{1}{k}\left(\left|\mathrm{Int}_k(A)\right|+\left|\mathrm{Int}_k(B)\right|\right)-2\sum_{n=1}^{l}\sum_{m=1}^{l}\frac{2}{m+n}\left|\mathrm{Int}_m(A)\cap\mathrm{Int}_n(B)\right|}{\sum_{k=1}^{l}\frac{1}{k}\left(\left|\mathrm{Int}_k(A)\right|+\left|\mathrm{Int}_k(B)\right|\right)}. \tag{1}$$

Where $\mathrm{Int}_k(A)$ is a list of proteins whose minimum number of edges needed to reach protein *A* is equal to *k* (to decrease the distance between proteins interacting with each other, $\mathrm{Int}_k(A)$ includes the protein *A* itself). *D*(*A*, *B*, *l*) is equal to Czekanowski–Dice distance.

Finally, the PPI score as a measurement of cooperativity between proteins was defined as follows:

$$\text{PPI Score} = \min_{l} D(A, B, l), \quad l \text{ in } \{1, 2\}. \tag{2}$$

2.3 Genome-wide search for PWM

Seven hundred and seventy-four unique PWMs united from TRANSFAC (Matys *et al.*, 2003), JASPAR (Bryne *et al.*, 2008) and researches on ENCODE data (Wang *et al.*, 2012) were imposed with genome-wide search for their binding occurrence and sites (the schematic is shown in Supplementary Figure S1). We performed PWM scan under the assistance of BSgenome package and only considered TF binding in promoter region that was defined as 1000 bases upstream from transcriptional start site (TSS). The minimum match score was set to 95%. Although some regulatory elements can act over large distances, up to several kilobases from TSS, we focused on sequences in the relative proximity of TSS as they are most likely to contain regulatory information for latent spatial relationships between TFs. Totally, there are ~4.8 million hits in non-repetitive promoter regions, which correspond to ~6200 hits for each PWM.

2.4 Co-occurrence score

Co-occurrence score evaluates the over-representation of TF pair occurrence in the promoters of common TGs regulated by two TFs compared to its occurrence in all promoters in human genome. It was calculated according to:

$$\mathrm{Co-occurrence\ \ score}=\sum_{k=g}^{G}\binom{k}{G}\left(\frac{n}{N}\right)^{k}\left(1-\frac{n}{N}\right)^{G-k} \tag{3}$$

where *n* is the number of the promoters of common TGs regulated by two TFs; *N* is the total number of human promoters; *g* is the occurrence of TF pair in the promoters of common TGs regulated by two TFs; and *G* is the overall occurrence of TF pair in promoters of entire human genome. The equation was used to obtain the probability of observing the TF pair *g* or more times in the *n* promoters of common TGs regulated by two TFs, given that the TF pair occurs *G* times in total *N* promoters. It is notable that some TFs have multiple binding sites in one promoter. In such cases, when calculating *g* and *G* we counted all combinations between the TF binding sites. Other researchers reported that counting all possible pairs yielded better performance than regarding these cases as a single binding-site pair occurrence (Yu *et al.*, 2006a, b). Note that one TF pair may correspond with several PWM pairs since one TF may have more than one PWM, so the co-occurrence score was taken as the minimum among them.

2.5 Distance constraint score

The distance constraint between binding sites of one TF pair in promoter regions was measured by comparing the observed distance distribution $f_{\mathrm{obv}}(d)$ with a background distribution $f(d)$ using the Kolmogorov–Smirnov test. For each TF pair, distances between their binding sites were calculated. If one or both TFs own multiple binding sites in one promoter, we included the distances between all combinations as we did for co-occurrence score. The background distribution $f(d)$ was obtained by considering binding sites of TF pair occur randomly on promoters and measuring the distances between them (in unit of bp). It can be normalized as:

$$f_L(d)=\frac{L-w_a-w_b-d+1}{\sum_{i=1}^{L-w_a-w_b+1}(L-w_a-w_b-i+1)} \tag{4}$$

where *L* is the length of one promoter sequence; w_a and w_b are the widths of the binding sites of TF_a and TF_b; *d* is the binding site distance; and $L-w_a-w_b-d+1$ means the number of all possible arrangements for the binding site pair. The background distribution $f_L(d)$ varies depending on different promoter sequence length *L*. In our approach, *L* was 1000 bp.

2.6 Co-expression

For a given TF pair, assume there are *m* ($m\geq2$) common genes shared by both. Pearson correlation coefficient (PCC) based on gene-expression profiles was calculated for each of the $M = C_m^2$ pairs of genes. PCC was applied with Fisher transformation to obtain co-expression r for the following analysis:

$$r = \frac{1}{2}\ln\left(\frac{1+\mathrm{PCC}}{1-\mathrm{PCC}}\right). \tag{5}$$

2.7 Bayesian model

For each TF pair, we separated data resources into two parts: prior information and context-specific data. PPI, co-occurrence and distance constraint were considered as prior information that is independent of the cell type or physical condition and were integrated together through Logistic model as prior probability π that reflects the general propensity of cooperativity.

$$\begin{aligned}\ln\frac{\pi}{1-\pi} &= \beta_0 + \beta_1 * \mathrm{PPI\,score} + \beta_2 * \ln Co\text{-}\mathrm{occ\,score} \\ &\quad + \beta_3 * \ln\mathrm{Dis.\,Cons\,score}.\end{aligned} \tag{6}$$

Due to the possibility of numerical underflow from very small probabilities, we made use of natural logarithm for co-occurrence score and distance constraint score.

The overall model shown schematically in Figure 1 was based on Bayesian theory

$$p = \frac{P(I=1) * P(r|I=1)}{P(I=1) * P(r|I=1) + P(I=0) * P(r|I=0)} \tag{7}$$

where p represents the probability of combinational cooperativity given the prior information and specific gene profiling. Prior probability that a TF pair is cooperative was termed as $\pi = P(I=1)$. Here $1-\pi = P(I=0)$ was analogous to representing one TF pair is not cooperative. Co-expression r was assumed to be generated from one of two underlying distributions. One relates to the cooperative state of TF pair $(r|I=1)$, while the other distribution corresponds to the uncooperative state $(r|I=0)$. Co-expression r can be modelled with normal distribution as:

$$r|I\ = 1 \sim N(\mu_1, \sigma_1^2) \qquad r|I = 0 \sim N(\mu_2, \sigma_2^2)\,. \tag{8}$$

Distributions for cooperative state and uncooperative state were designated with parameters μ_1, σ_1 and μ_2, σ_2.

2.8 Parameter estimation and model computation

The parameters of the Bayesian model β_0, β_1, β_2, β_3, μ_1, σ_1, μ_2, σ_2 were estimated by maximizing the likelihood function using expectation maximization (EM) algorithm. The EM algorithm used in our model is robust and fast on convergence as long as proper initial parameters are selected. We initialized the parameters with the utilization of PPI data. Although PPI is constituent of prior information, it still contributes to an informed starting point. Detailed implementation of EM algorithm is as follows.

Step 1: Based on PPI information, extract TF pairs from the qualified TF pairs and label them with ‘co-operative’. Randomly sampled equal-sized TF pairs from the rests and label them with ‘non-co-operative’.

Step 2: Utilize the dataset constructed in Step 1 to fit $\beta's$ in Logistic model and $\mu's$, $\sigma's$ of normal distributions.

Step 3: Calculate the expected p based on parameters estimated in Step 2.

Step 4: Maximize the likelihood function to obtain all the parameters and update the initial values with the new estimations.

Step 5: Repeat regular EM steps until convergence.

The criteria for convergence are: (i) the likelihood of the model changes by 0.001 and (ii) all the parameters change by 0.001 with respect to the previous EM iteration.

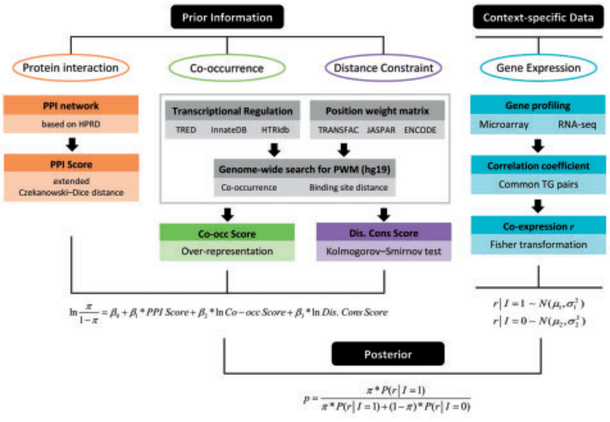

Fig. 1. Schematic of Bayesian approach for identification of TC

2.9 Golden standard dataset ofTC in HepG2

Utilizing domain-based interaction data (Miyamoto-Sato *et al.*, 2010) and TF-to-TF regulatory interactions specific in HepG2 (Neph *et al.*, 2012), we confirmed 68 TF pairs validated by experiments and hepatoblastoma-specific (HepG2) to form the golden positive dataset. Since the majority of TC remains undiscovered to date, it is difficult to derive a reasonable golden negative dataset. In this article, we used three rules to identify the negatives: (i) the TF pairs are not physical interacted; (ii) TFs are not specifically interacted in HepG2; and (iii) TF pairs cannot share identical gene ontology (GO) terms on cellular component considering TF pairs who function in different components would not be interacted. Based on these criteria, 428 TF pairs were left to form the golden negative dataset.

3 RESULTS AND DISCUSSION

3.1 Feature construction and discrimination

We derived features from various data sources and based on reasonable biological assumptions. The PPI score based on extended Czekanowski–Dice distance was used to measure the distance between proteins in the PPI network. The closeness between two TFs in PPI network suggests their contribution to the same or similar biological processes, from which, we can deduce that these two TFs possess a high probability to be cooperative. If two TFs are interacted, intuitively, one can expect their binding occurrence to be enriched in the promoters of their common TGs compared to the whole genome; furthermore, the distance between their binding sites should be significantly shorter than the random expectation. Co-occurrence score and distance constrain score were compiled from this property. Researches on various model systems have revealed that cooperative TFs have complicated patterns of significant co-expression (Wang *et al.*, 2005), besides, TGs’ expression correlation reflects the transient and dynamic natures of TC, which is potent to explain the context-specific cooperative behaviour between TFs. From this, we deduced co-expression feature (see Materials and Methods section).

To measure the extent to which these evidence resources could differentiate cooperative and non-cooperative TF pairs, we built

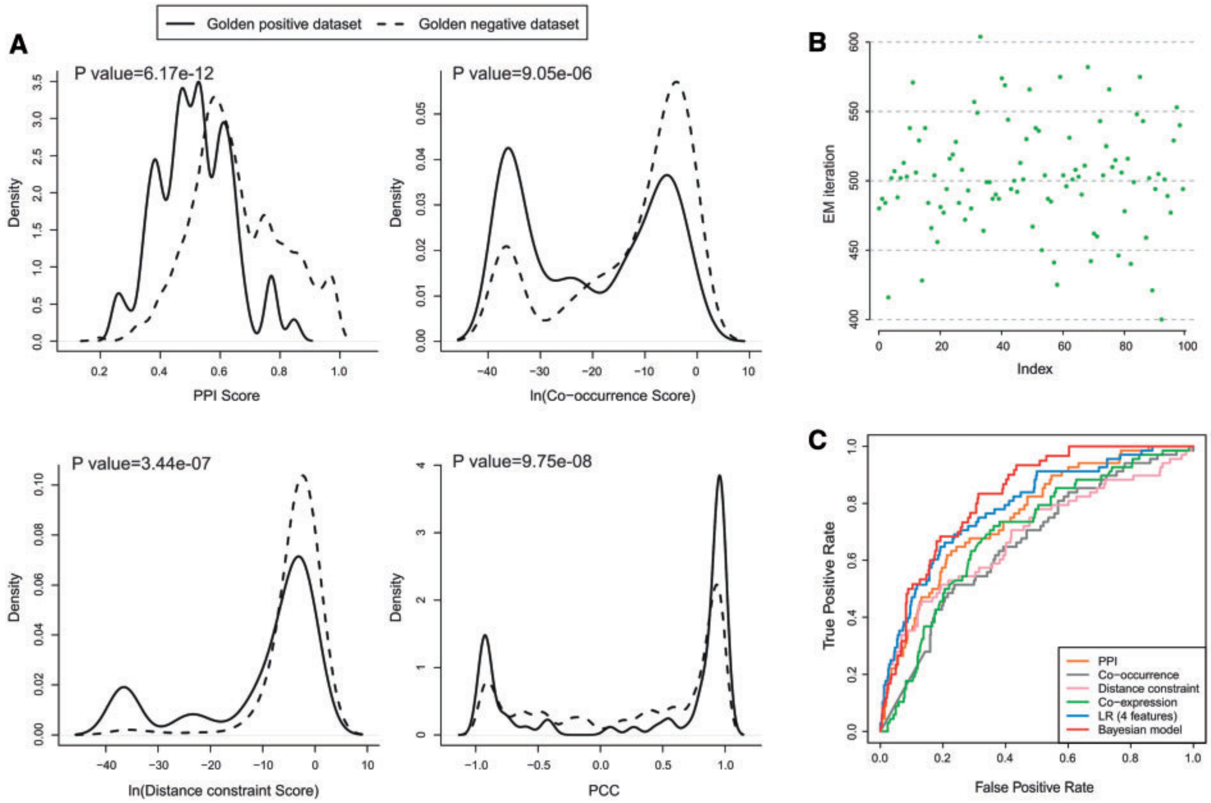

Fig. 2. The performance of Bayesian approach. (A) Density plots of four features for GPD (solid lines) and GND (dashed lines). The *P*-values inside each density plot were obtained from Wilcoxon Rank Sum test. (B) EM iteration of Bayesian model under 100 random sampling of TF pairs for plain starting point. (C) The comparison of ROC curves for Bayesian model, individual feature and straight Logistic regression model

golden standard dataset (GSD). Since combinational regulations of TFs are dynamic, reliable GSD should depend on specific conditions (considering HepG2 here). Golden positive dataset (GPD) consists of 68 cooperative TF pairs (Supplementary Table S1) and golden negative dataset (GND) includes 428 TF pairs (Supplementary Table S2). We made comparative analysis for GSD VS GND. The density plots of four features show excellent discrimination for positives and negatives (Fig. 2A). Compared with non-cooperative TF pairs, cooperative TF pairs display shorter distance in PPI network, closer spatial relationship in promoter regions, stronger distance constrains among binding sites and higher correlation on gene expression pattern, which validates our assumption. Wilcoxon Rank Sum test was employed to capture the significance of difference between two classes using the features we defined. The *P*-values (inside each density plot of Fig. 2A) indicate significantly distinguishable power of each feature, which validates the effectiveness of those features to predict TC.

3.2 Model efficiency and performance

The PPI, co-occurrence and distance constraint were considered as prior information that is independent of the cell type or physical condition and were integrated through Logistic model as prior probability to reflect the general propensity of cooperativity. Co-expression calculated from context-specific gene profiling reflects the dynamic and transient behaviour of TC. Then Bayesian model combines prior information and specific gene expression information to compute the posterior probability that a TF pair is conditionally cooperative (see Materials and Methods section). The EM algorithm was used to maximize the likelihood function for the estimation of parameters in Bayesian model. Faster convergence was achieved if we used informative initial values. Normally, GSD, if available, would be used to initialize parameters. However, in many cases, it is difficult to construct a legitimate GSD. We recommended that PPI data could be used to give an informed starting point. Using PPI-validated TF pairs labelled as ‘co-operative’ and randomly sampled TF pairs labelled as ‘non-co-operative’, we can fit the parameters in Bayesian model as starting point. Then, regular EM algorithm was conducted until convergence. With this strategy, the iteration of HepG2 dataset is 154.

As a comparison, we evaluated whether plain starting point would affect model convergence. We built an equal-sized dataset as GSD of HepG2 by random sampling without replacement for TF pairs. Then, each TF pair in the dataset was allocated with a binary variable indicating ‘co-operative’ or ‘non-co-operative’. Next, Bayesian model was trained. This process was repeated for 100 trials. The EM iteration ranges from 400 to 604 with average of 502 and SD 38 (Fig. 2B), which indicates uninformed starting point hampers the model efficiency.

To evaluate the effectiveness of our approach, we first performed comparison with individual feature and straight

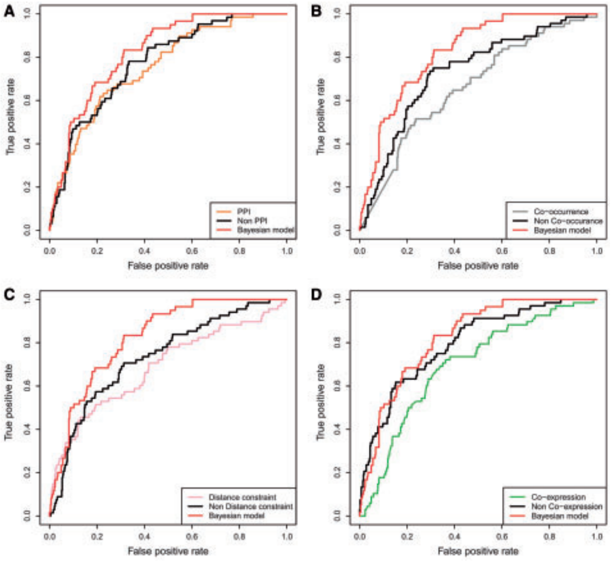

Fig. 3. The comparison of ROC curves for one-feature-single prediction, one-feature-removed model and full-feature model. (A) PPI, (B) co-occurrence, (C) distance constraint and (D) co-expression

Table 1. Comparison with other methods based on benchmark datasets (in terms of AUC value)

Benchmark dataset	Our model	Banerjee and Zhang, 2003	Nagamine *et al.*, 2005	Yu *et al.*, 2006
HepG2	0.83	0.54	0.61	0.70
SK–N–SH	0.80	0.52	0.63	0.71

Logistic regression that includes four features, using the HepG2 dataset we described before. For individual feature, we simply used the feature score to predict TF cooperativity. For Bayesian model and straight Logistic regression, we trained them with 5-fold cross-validation. Receiver operating characteristic (ROC) curves of these classifiers (Fig. 2C) show that our proposed Bayesian model consistently outperformed the other tested classifiers. Not surprisingly, multi-feature models performed better than single feature. At sensitivity of 80%, the false positive rate of Bayesian model is ~30%, 10% < Logistic regression. Moreover, Bayesian model has the biggest AUC (area under ROC curve) of 0.83 while Logistic's AUC is 0.78. We further investigated the importance of each feature for Bayesian model performance. Four features were removed from the model successively. Each time, only one feature was removed and Bayesian model was trained on HepG2 dataset. Note that when co-expression is excluded, the Bayesian model turns to typical Logistic regression with three features. Figure 3 displays their performance in terms of ROC. As expected, one-feature-removed model substantially transcends one-feature-only model but is inferior to full-feature model. It demonstrates the essentiality of each feature. Although no data type is fully informative when taken alone, by combination, Bayesian approach is capable to confidently infer quantitative TC.

3.3 Comparison with other methods

We performed extensional comparison with other existing methods mentioned in the section 1 to see whether integrated Bayesian model can predict TC more accurately. Each method was replayed and its prediction performance was assessed based on two independent, high-quality benchmark datasets. The first one is HepG2 dataset described before and second one is SK-N-SH (neuroblastoma specific) dataset which consists of 68 positive TF pairs (Supplementary Table S3) and 421 negative TF pairs (Supplementary Table S4) using the same criteria as HepG2 dataset. Table 1 lists the AUCs (ROC curves in Supplementary Figure S2), which show that our Bayesian model is better in predictions than existing algorithms. This favourable results are due to at least one of the following reasons: (i) the predictive capability of single data source is inadequate and unilateral while Bayesian model provide comprehensive prediction by information fusion; (ii) although genome-wide TC were predicted by some methods (Nagamine *et al.*, 2005; Yu *et al.*, 2006a, b), they neglected the transient and dynamic behaviour of TC; (iii) we integrated generic features (such as PPI score) and specific feature (co-expression) to capture the true cooperative nature of TFs.

A previous study (Wang *et al.*, 2009) exploited Bayesian network integration of genome-wide data to evaluate TC, which is conceptually similar to our model. However, there are several technical differences in methodology. First, among 15 features defined by Wang *et al.*(2009), some features represent the same characteristic only based on different datasets (such as ChIP–chip-based TG overlap, literature-based TG overlap and motif occurrence-based TG overlap), which inevitable brings redundant information and dependences among them. We compiled features from distinct sources, ensuring their approximate uncorrelation (Supplementary Figure S3) as well as complementation to each other. Second, the Bayesian network structure is predetermined as approximate conditional independences and inter-dependencies and the parameters are learned from GSD. As a comparison, our model only needs informative initial values and accurate parameter estimation can be achieved by EM algorithm. This difference will be notable if GSD is scarce for some organisms. Third, Wang *et al.* (2009) labelled all features with conditional probability and predict the general TF cooperativity in yeast while we separated features into prior and context-specific, identifying the dynamic nature of TC in human. Actually, it is reported that TF activity and TF synergy are condition-dependent (Luscombe *et al.*, 2004), which means our model is more biologically meaningful. Furthermore, our reduced yet complementary features will make it more feasible to extend the model to other organisms.

3.4 Prediction of TC during hepatocarcinogenesis

We applied Bayesian model to study the transcriptional regulatory mechanism underlying HCC progression. Gene-profiling data of HCC progression were obtained from GEO database (GSE25097). Since we are highly interested in the cancerization from cirrhosis to HCC, six samples for healthy liver, 40 for cirrhosis, 129 for both HCC with cirrhosis background (short as cirr-HCC) and adjacent non-tumours were extracted for the

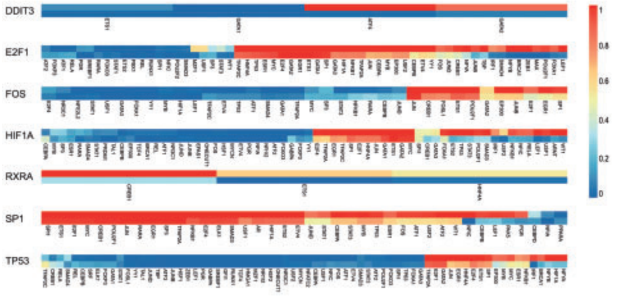

Fig. 4. Transcriptional cooperativity variability map for partners of seven HCC-related TFs over two contexts. For each TF, columns denote partners, up row and down row denote the likelihood of TC under cirrhosis and cirr-HCC, respectively

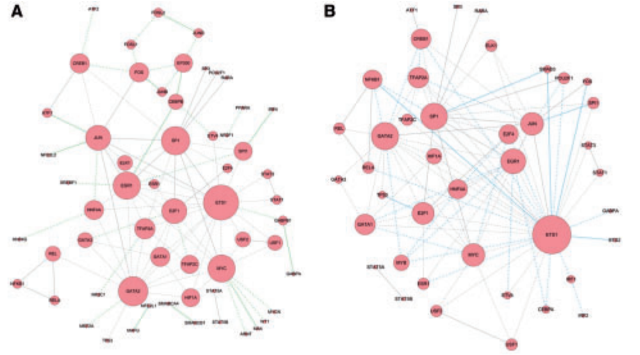

Fig. 5. Transcriptional cooperativity networks using TF pairs with top 100 highest posterior probabilities. (A) Network for cirrhosis. (B) Network for cirr-HCC. Node denotes TF whose size is proportion to the degree in the network. Edge represents cooperative relationship. All validated interactions are in solid lines and novel ones are in dashed lines. Grey, green and blue lines designate interactions shared by two contexts, specific in cirrhosis and specific in cirr-HCC, respectively

prediction of TC under two contexts: cirrhosis and cirr-HCC. Among 78 210 TF pairs from 396 TFs with available PWMs, only 3474 pairs have ≥ 2 common TGs. Additionally, after the genome-wide searching for TF binding sites at a strict match score 95%, 1703 pairs were kept under the need of co-occurrence on the same promoter. We defined these 1703 TF pairs as 'qualified' before the application of Bayesian model. Due to the removal of some TF pairs whose TGs were not expressed in our gene profiling, posterior probability of 1665 and 1662 TF pairs are calculated for the two contexts.

To discover the carcinoma associated TC alterations over two contexts, seven HCC-related TFs were chosen from Cancer Resource dataset (Ahmed *et al.*, 2011). The variability map (Fig. 4) presents the degree of variability among the partners of a given TF from cirrhosis to cirr-HCC. Cooperative partners almost retained the same for FOS, SP1 and TP53 over two contexts (for instance, JUN and ETS1 for FOS, E2F1 and POU2F1 for SP1, TFAP2A and GATA2 for TP53). However, dominate tendency was observed among DDIT3, E2F1, HIF1A and RXRA that some of their partners were only cooperative under cirrhosis.

3.5 Recasting TC network of hepatocarcinogenesis

To obtain a systematic-level perspective, we recast TC as a network. For comparison between cirrhosis and cirr-HCC, we framed same scale networks of TF pairs with top 100 highest posterior probability (all >0.9, except ETS1–FOS in cirr-HCC with 0.89), wherein the nodes are TFs and the edges designate interactions (Fig. 5 and Supplementary Tables S5 and S6). We then mapped the interactions in the derived networks to non-redundant protein complex dataset based on CORUM (Ruepp *et al.*, 2010). Seven and six TF pairs were found co-occur in at least one protein complex for cirrhosis and cirr-HCC, respectively. Besides, we also used TF–TF interaction data extracted from TcoF database (Schaefer *et al.*, 2011) to validate our predictions, which resulted in 30 and 26 TF pairs with interaction evidences. In total, 31 and 27 interactions were validated in cirrhosis and cirr-HCC networks separately. To evaluate the significance of predicted networks, we randomly sampled 100 TF pairs for 2000 trials and averagely 11 pairs were validated by protein complex and TF–TF interaction information, demonstrating that our derived TC networks are highly significant (P-value $< 5 \times 10^{-4}$ for cirrhosis and $< 1 \times 10^{-3}$ for cirr-HCC). Both our results recapitulated previously reported observations and revealed novel potential interactions, which could provide novel insight into the molecular mechanism controlling formative carcinogenesis.

Each TC network consists of 100 interactions which involve 59 TFs and 41 TFs for cirrhosis (Fig. 5A) and cirr-HCC (Fig. 5B), respectively. Comparisons between TC networks revealed that when the disease progression switches from cirrhosis to HCC, TC changes correspondingly. Nevertheless 68 relationships (grey lines) were observed consistently under different conditions, which can indicate the development and progression of carcinogenesis share highly common regulations. Moreover, ETS1, GATA2, SP1, EGR1 and MYC are consistently hub TFs over two contexts.

The interactions among these five hub TFs are conserved during the progression. However, hub TFs have more specific interactions with other TFs under cirr-HCC. For instance, there are 13 specific interactions with ETS1 under cirr-HCC while only one under the context of cirrhosis. These five hub TFs could be regarded as pivotal factors to understand the molecular mechanism that underlies the development and progression of carcinogenesis.

3.6 Functional analysis of specific cooperativity

We took specific cooperativity of E2F1–TFAP2C versus E2F1–TP53 and specific transcriptional complex FOS–EP300–JUNB versus FOS–ETS1–NFKB1, all extracted from the TC networks, as the instances to investigate the potential molecular mechanism underlying the cancerization from cirrhosis to HCC.

E2F1 possesses a specific cooperativity with pro-inflammatory factor TFAP2C under cirrhosis and tumour suppressor TP53 under cirr-HCC. We found that many differentially expressed TGs regulated by this TF pair are enriched in regulation of biosynthetic and metabolic process for cirrhosis while cell cycle, apoptotic and cell death for cirr-HCC. For instance, the cooperativity of E2F1 and TFAP2C regulates the folic acid-containing compound biosynthetic and metabolic process (Bonferroni-adjusted P-value $< 1.4 \times 10^{-5}$), which has been linked with the effect of alcohol consumption on liver damage and progression towards HCC, by inhibiting formyltetrahydrofolate synthetase (*MTHFD1*) (Persson *et al.*, 2013). As a comparison, E2F1 and TP53 cooperate to promote cell-cycle progression (Bonferroni-adjusted P-value $< 2.4 \times 10^{-4}$) by activating G1/S-related gene *CCNE1* and *PCNA*, G2 phase-related gene *TOP2A*, G2/M-related gene *TUBA4A* and G2/M-, M/G1-related gene *CCND1* (Whitfield *et al.*, 2002). The uncontrolled cell cycle is one of the major aspects for cancers. Additionally, there exists evidence that the synergy of E2F1 and TP53 can induce apoptosis (Wu and Levine, 1994), which may be achieved through co-regulating apoptosis factor *TP53BP2* according to our analysis.

Specific cooperativity FOS–EP300–JUNB is observed under cirrhosis. It was shown that EP300 can stimulate activity of the FOS–JUN complex *in vivo* and part of this response was conveyed through direct interaction of EP300 with FOS (Bannister and Kouzarides, 1995). There were reports that dentin matrix protein 1 (*DMP1*) was transcriptionally regulated by TF FOS (Narayanan *et al.*, 2002) and the cooperative interaction of JUNB with EP300 remarkably mediates the *DMP1*-promoter activity during mineralization, which was confirmed *in vivo* by immunoprecipitation and chromatin immunoprecipitation analysis. Further, phosphorylation of JUNB was found to be essential for its interaction with EP300 (Narayanan *et al.*, 2004). Although DMP1 is an Arg–Gly–Asp-containing acidic phosphoprotein that was originally thought to be a mineralized extracellular matrices components (George *et al.*, 1993), DMP1 mRNA was later detected in non-mineralized mouse tissues including liver, muscle, brain, pancreas and kidney by RT–PCR (Terasawa *et al.*, 2004). More recently, it was demonstrated that DMP1 can block proliferation, migration, tubulogenesis responses and tumour-associated angiogenesis by modulating vascular endothelial growth factor (*VEGF*), which expands the role of DMP1 in angiogenic field for the first time (Pirotte *et al.*, 2011). Our predicted cooperativity may provide upstream transcriptional mechanism for the angiogensic function of DMP1, which is implicated in the regeneration of hepatocytes and contributes to the development of fibrous bands in hepatic cirrhosis (Corpechot *et al.*, 2002; Fernandez *et al.*, 2009). Specific cooperativity ETS1–FOS and ETS1–NFKB1 forming complex FOS–ETS1–NFKB1 are observed under cirr-HCC. *Cis*-acting elements with AP1, ETS-like and NFKappaB binding motifs have been identified in the promoter region of the granulocyte-macrophage colony stimulating factor (*CSF2*) gene (Thomas *et al.*, 1997), which encodes GM–CSF cytokine involved in differentiation, proliferation and activation of the hematopoietic system. The results suggest constitutive ETS1 as well as inducible NFKB1 and FOS cooperate as a transcriptional complex in activated T cells. Additionally, pathway enrichment analysis for common TGs shared by FOS–ETS1–NFKB1 showed that the most significant one is ‘pathways in cancer’ (Bonferroni-adjusted P-value $< 1.8 \times 10^{-4}$). Possibly, specific complex FOS–ETS1–NFKB1 can function in activated T cells by regulating *CSF2* which cooperates with other genes involved in cancer pathways and result in promoting hepatocyte necrosis, inflammation and regeneration, ultimately carcinogenesis (Block *et al.*, 2003; Zhao *et al.*, 2012).

Bayesian approach can penetrate the limitations of qualitative inference from experiments and inadequate capability of individual data source. It provided rigorously quantitative measurement of TC. Furthermore, the prior information reflects the general propensity of TC, which sets the stage for tissue- or disease-specific analysis of regulatory programs. This approach is compatible for discretionary gene profiling, ensuring its adaptability. However, the Bayesian approach has its own limitations. Owing to the size of the TF repertoire and a small portion of reported PWMs so far, the overall landscape of TC is difficult to be identified. On the other hand, to assure the specificity of this model, 1000 bases upstream from TSS was considered in our study and a strict threshold was set for the pattern search of PWMs, which would miss a considerable portion of qualified TF pairs. Nevertheless, this model is able to capture the dynamic nature of TC by comprehensive analysis of data from multiple cascades. We believe that our proposed Bayesian approach is a promising application in identifying tissue- or disease-specific TC and will further facilitate the interpretation of underlying mechanisms for various physiological conditions.

ACKNOWLEDGEMENTS

We thank Prof. Wenxin Qin from Shanghai Cancer Institute for patient and helpful discussions. We thank Chen Zhao, Peng Li, Geng Chen and Huadong Zhao for conversations. We also thank the anonymous reviewers for their criticism and valuable suggestions.

Funding: This work was supported by the National 973 Key Basic Research Program (Grant No. 2010CB945401 and 2013CB127000), the National Natural Science Foundation of China (Grant No. 31240038, 31071162 and 31000590) and the Science and Technology Commission of Shanghai Municipality (11DZ2260300).

Conflict of Interest: none declared.

REFERENCES

Ahmed,J. *et al.* (2011) CancerResource: a comprehensive database of cancer-relevant proteins and compound interactions supported by experimental knowledge. *Nucleic Acids Res.*, **39**, D960–967.

Banerjee,N. and Zhang,M.Q. (2003) Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res.*, **31**, 7024–7031.

Bannister,A.J. and Kouzarides,T. (1995) Cbp-induced stimulation of C-Fos activity is abrogated by E1a. *Embo. J.*, **14**, 4758–4762.

Block,T.M. *et al.* (2003) Molecular viral oncology of hepatocellular carcinoma. *Oncogene*, **22**, 5093–5107.

Bookout,A.L. *et al.* (2006) Anatomical profiling of nuclear receptor expression reveals a hierarchical transcriptional network. *Cell*, **126**, 789–799.

Bovolenta,L.A. *et al.* (2012) HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics*, **13**, 405.

Braun,T. and Gautel,M. (2011) Transcriptional mechanisms regulating skeletal muscle differentiation, growth and homeostasis. *Nat. Rev. Mol. Cell Biol.*, **12**, 349–361.

Brun,C. *et al.* (2004) Clustering proteins from interaction networks for the prediction of cellular functions. *BMC Bioinform.*, **5**, 95.

Bryne,J.C. *et al.* (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–106.

Corpechot,C. *et al.* (2002) Hypoxia-induced VEGF and collagen I expressions are associated with angiogenesis and fibrogenesis in experimental cirrhosis. *Hepatology*, **35**, 1010–1021.

Fei,T. and Chen,Y.G. (2010) Regulation of embryonic stem cell self-renewal and differentiation by TGF-beta family signaling. *Sci. China Life Sci.*, **53**, 497–503.

Fernandez,M. *et al.* (2009) Angiogenesis in liver disease. *J. Hepatol.*, **50**, 604–620.

George,A. *et al.* (1993) Characterization of a novel dentin matrix acidic phosphoprotein. Implications for induction of biomineralization. *J. Biol. Chem.*, **268**, 12624–12630.

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

Hoffmeyer,K. *et al.* (2012) Wnt/beta-catenin signaling regulates telomerase in stem cells and cancer cells. *Science*, **336**, 1549–1554.

Jiang,C. *et al.* (2007) TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res.*, **35**, D137–140.

Kim,J. *et al.* (2008) An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, **132**, 1049–1061.

Lagha,M. *et al.* (2012) Mechanisms of transcriptional precision in animal development. *Trends Genet.*, **28**, 409–416.

Latchman,D.S. (1997) Transcription factors: an overview. *Int. J. Biochem. Cell B*, **29**, 1305–1312.

Lemon,B. and Tjian,R. (2000) Orchestrated response: a symphony of transcription factors for gene control. *Gene Dev.*, **14**, 2551–2569.

Luscombe,N.M. *et al.* (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–312.

Lynn,D.J. *et al.* (2008) InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Mol. Syst. Biol.*, **4**, 218.

Matys,V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.

Miyamoto-Sato,E. *et al.* (2010) A comprehensive resource of interacting protein regions for refining human transcription factor networks. *PLoS One*, **5**, e9289.

Nagamine,N. *et al.* (2005) Identifying cooperative transcriptional regulations using protein-protein interactions. *Nucleic Acids Res.*, **33**, 4828–4837.

Narayanan,K. *et al.* (2002) Transcriptional regulation of dentin matrix protein 1 (DMP1) by AP-1 (c-fos/c-jun) factors. *Connect Tissue Res.*, **43**, 365–371.

Narayanan,K. *et al.* (2004) Transcriptional regulation of dentin matrix protein 1 by JunB and p300 during osteoblast differentiation. *J. Biol. Chem.*, **279**, 44294–44302.

Neph,S. *et al.* (2012) Circuitry and dynamics of human transcription factor regulatory networks. *Cell*, **150**, 1274–1286.

Persson,E.C. *et al.* (2013) Alcohol consumption, folate intake, hepatocellular carcinoma incidence and liver disease mortality. *Cancer Epidemiol. Biomarkers. Prev.*, **22**, 415–421.

Pilpel,Y. *et al.* (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.

Pirotte,S. *et al.* (2011) Dentin matrix protein 1 induces membrane expression of VE-cadherin on endothelial cells and inhibits VEGF-induced angiogenesis by blocking VEGFR-2 phosphorylation. *Blood*, **117**, 2515–2526.

Prasad,T.S.K. *et al.* (2009) Human protein reference database-2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

Ruepp,A. *et al.* (2010) CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids. Res.*, **38**, D497–D501.

Schaefer,U. *et al.* (2011) TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins. *Nucleic Acids Res.*, **39**, D106–D110.

Terasawa,M. *et al.* (2004) Expression of dentin matrix protein 1 (DMP1) in nonmineralized tissues. *J. Bone Miner. Metab.*, **22**, 430–438.

Thomas,R.S. *et al.* (1997) ETS1, NFkappaB and AP1 synergistically transactivate the human GM-CSF promoter. *Oncogene*, **14**, 2845–2855.

Wang,J. *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798–1812.

Wang,W. *et al.* (2005) Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. *Proc. Natl Acad. Sci. USA*, **102**, 1998–2003.

Wang,Y. *et al.* (2009) Predicting eukaryotic transcriptional cooperativity by Bayesian network integration of genome-wide data. *Nucleic Acids Res.*, **37**, 5943–5958.

Whitfield,M.L. *et al.* (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.

Whitfield,T.W. *et al.* (2012) Functional analysis of transcription factor binding sites in human promoters. *Genome Biol.*, **13**, R50.

Wu,X. and Levine,A.J. (1994) p53 and E2F-1 cooperate to mediate apoptosis. *Proc. Natl Acad. Sci. USA*, **91**, 3602–3606.

Yu,H.Y. *et al.* (2003) Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet.*, **19**, 422–427.

Yu,X. *et al.* (2006a) Genome-wide prediction and characterization of interactions between transcription factors in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **34**, 917–927.

Yu,X. *et al.* (2006b) Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res.*, **34**, 4925–4936.

Zhao,X. *et al.* (2012) The role and clinical implications of microRNAs in hepatocellular carcinoma. *Sci. China Life Sci.*, **55**, 906–919.