Bioinformatics Advance Access published September 6, 2011

Correlated evolution of transcription factors and their binding sites

Shu Yang[1], Hari Krishna Yalamanchili[1], Xinran Li[1,2], Kwok-Ming Yao[1], Pak Chung Sham[3], Michael Q. Zhang[4,5] and Junwen Wang[1,*]

[1]Department of Biochemistry, LKS Faculty of Medicine, The University of Hong Kong, 21 Sassoon Road, Hong Kong SAR, China.

[2]Current address: Department of Molecular, Cellular and Developmental Biology, University of Michigan, Ann Arbor, MI 48109-1048, USA.

[3]Department of Psychiatry and State Key Laboratory of Cognitive and Brain Sciences, LKS Faculty of Medicine, The University of Hong Kong, 21 Sassoon Road, Hong Kong SAR, China.

[4]Department of Molecular and Cell Biology, Center for Systems Biology, The University of Texas at Dallas, Dallas, TX, 75080, USA.

[5]Bioinformatics Division, TNLIST, Tsinghua University, Beijing, 100084, China.

Associate Editor: Prof. Martin Bishop

ABSTRACT

Motivation: The interaction between transcription factor (TF) and transcription factor DNA binding site (TFBS) is essential for gene regulation. Mutation in either the TF or the TFBS may weaken their interaction and thus result in abnormalities. To maintain such vital interaction, a mutation in one of the interacting partners might be compensated by a corresponding mutation in its binding partner during the course of evolution. Confirming this co-evolutionary relationship will guide us in designing protein sequences to target a specific DNA sequence or in predicting TFBS for poorly studied proteins, or even correct and rescue disease mutations in clinical applications.

Results: Based on six publicly available, experimental validated TF-TFBS binding datasets for the basic Helix-Loop-Helix (bHLH) family, Homeo family, High-Mobility Group (HMG) family and Transient Receptor Potential channels (TRP) family, we showed that the evolutions of the TFs and their TFBSs are significantly correlated across eukaryotes. We further developed a mutual information based method to identify co-evolved protein residues and DNA bases. This research sheds light on the dynamic relationship between TF and TFBS during their evolution. The same principle and strategy here can be applied to co-evolutionary studies on Protein-DNA interactions in other protein families.

Availability: All the datasets, scripts and other related files have been made freely available at: http://jjwanglab.org/co-evo.

Contact: junwen@uw.edu

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

*To whom correspondence should be addressed.

1 INTRODUCTION

The term co-evolution was first conceptualized by Charles Darwin in his study on the two species, pollinators and orchids (Darwin, 1862). At the molecular level, co-evolution has been studied in the context of protein-protein interaction (Atwell, et al., 1997; Moyle, et al., 1994; Pazos, et al., 1997). A change in one of the binding protein's interaction surface is compensated by an appropriate change in the interface of the partner protein. Studies have shown the co-evolution of interacting proteins (Pazos and Valencia, 2001) as well as of interacting protein and ligand (Goh, et al., 2000). These studies have had great impact in this area of research, such as in protein-protein interaction and drug design (Izarzugaza, et al., 2006; Tillier, et al., 2006; Tress, et al., 2005).

A protein can bind to more than one DNA sequence, and these corresponding binding sites are usually represented by a binding profile called Position Weight Matrix (PWM) (Hannenhalli, 2008). Methods have been developed to compare the similarity of different PWMs (Pietrokovski, 1996). By comparing the pairwise similarities of all PWMs from a protein family, we can study the evolution of DNA binding sites for this protein family. Thus far, the co-evolutionary relationship between a transcription factor (TF) and transcription factor DNA binding site (TFBS) has not been systematically studied.

The interaction between TFs and their TFBSs is essential for many biological processes. For instance, the interaction at the core promoter regions determines the assembly of the pre-initiation complex and the initiation of transcription (Wang and Hannenhalli, 2006; Wang, et al., 2007), whereas interactions in the distal promoter/enhancer region determine the rate of transcription in cell

type-, tissue- and developmental stage-specific manner (Juven-Gershon, et al., 2008). Therefore, study of TF-TFBS interaction is critical to our understanding of the transcriptional regulatory network of gene expression (Qin, et al., 2011).

In this study, we hypothesize that the evolution of TFs and their TFBSs are correlated. We used the basic Helix-Loop-Helix (bHLH), Homeo, High-Mobility Group (HMG) and Transisent Receptor Potential channels (TRP) protein families to illustrate this principle of co-evolutionary relationship. We first performed the analysis based on bHLH family. The bHLH family is an ancient component of transcriptional regulation with over a hundred members found in humans (Ledent, et al., 2002). This protein family has a strong selection force and plays critical roles in diverse biological processes from cell proliferation to carcinogenesis (e.g. the oncogene and iPSC reprogramming factor, cMyc). It has evolved into 125 members in humans, and 94 of them are conserved in mice (Ledent, et al., 2002). However, only a few of them have binding site profiles that are experimentally identified. Therefore, we further validated our results on three other TF families: Homeo, HMG and TRP family. We investigated the evolution of TFs by their corresponding PWMs, and quantified the correlation between them with rigorous statistical methods. We found consistently that, the evolution of TFs and their binding sites are significantly correlated in multiple datasets.

2 METHODS

2.1 Collection of TFBS PWMs

We first collected all available bHLH PWMs (both individual and family-wise PWMs) from JASPAR FAM database (Accession: MF0007.1) (Portales-Casamar, et al., 2010). After filtering out heterodimer members, we obtained six remaining members from human and mouse, which we used to form our mammalian dataset. We then collected all the other bHLH family PWMs from JASPAR CORE database, again with heterodimers removed. The remaining 16 members, which included the previous six, were used as our eukaryote dataset from human, mouse, yeast and Drosophilia. Both datasets had been generated through traditional experimental methods, such as Electrophoretic Mobility Shift Assay (EMSA) or Systematic Evolution of Ligands by Exponential Enrichment (SELEX). To further independently test our hypothesis, we used a recently published bHLH dataset from UniPROBE database (Grove, et al., 2009) as our CAEEL dataset. This dataset contains 20 TFs from one single species and was generated using high throughput protein array method.

Moreover, to support a general conclusion regarding the co-evolution, we further validated our results by recruiting three more TF families from JASPAR CORE database: Homeo family which contains 100 members, HMG family which contains 15 members and TRP family which contains 11 members. Members within each of the three families are coming from eukaryote and with heterodimers filtered.

The detailed information of all the PWMs involved in this study is listed in **Supplementary Table 1** (**Table S1)**.

2.2 Quantifying co-evolution and assessing the statistical significance

In order to quantify the co-evolution between TFs and TFBSs, we first calculated pairwise sequence similarities among all the members in each of our TF families. We then built a similarity matrix for all protein pairs (**Supplementary Information 1 (SI 1)**). Similarly, we calculated pairwise similarity of PWMs (Pietrokovski, 1996) and constructed a PWM similarity matrix for corresponding binding sites (**SI 2**). We then concatenated all the rows in each matrix into a vector and calculated the Pearson's Correlation Coefficient (PCC) between the two vectors (**Fig 1**).

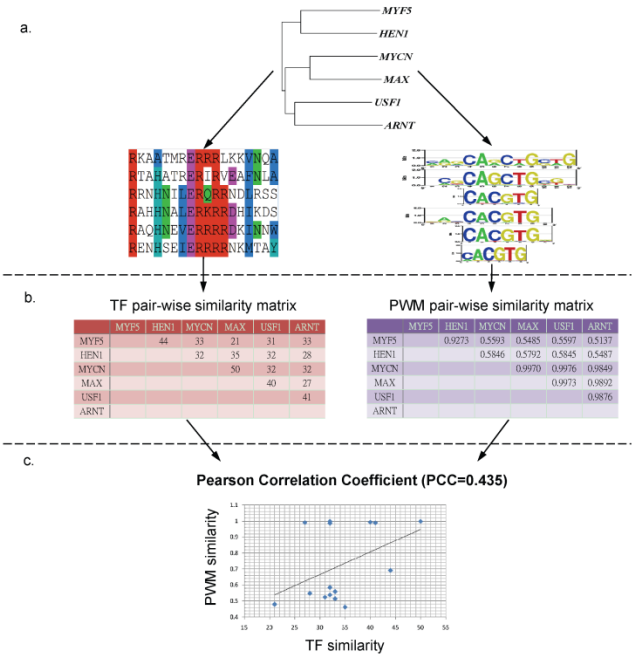

Fig. 1. Schematic pipeline to measure the co-evolution between transcription factors and their binding sites. **a)** From the phylogenetic tree of the TFs, we can assume that the TFs are co-evolved with their TFBS family-wise. The evolution of the TFs is represented by their multiple sequences alignment. The evolution of TFBS is represented by the comparison of each PWM of each TF binding sites dataset. **b)** The pairwise similarities between any TF pair and between any PWM pair were measured, respectively. Note that for both matrices, a bigger value in the cell means a higher similarity. **c)** The similarity matrix of TFs was compared with that of PWMs and PCC was calculated.

The statistical significances of co-evolution PCCs were assessed from permutations of protein and binding site sequences (SI 3). We compared the observed PCC to a null distribution of PCCs generated from permuted proteins and binding sites, and obtained an empirical *P* value (Li, et al., 2010). To permute protein sequences, we obtained all protein sequences and calculated their residue compositions. We then used the compositions to generate random sequences that matched the length of the original sequence. As an alternative to using a single uniform distribution for all positions, we also calculated the composition in a position-specific manner, and then permuted the sequence with a position-specific distribution. To permute PWM, we obtained all the original binding sites deriving the PWM and calculated the base composition. We then generated the same number of random binding sites for each TF, keeping base composition, length and label the same. The PWMs were then generated from these permuted binding sites. The PCC was recalculated using the generated PWMs and permuted protein sequences. This random process was repeated 1,000 times and the PCCs were sorted to obtain a background distribution of PCCs. Given a new PCC, we could look up its ranking in this background distribution and obtain its *P* value.

2.3 Identification of co-evolved residues using mutual information (MI).

Mutual information (MI) has been successfully used to identify contacting residues in protein-protein interactions (Weigt, et al., 2009), and coevolving residue pairs within a protein (Weil, et al., 2009). Current MI methods deal with protein sequences, where both interacting partners are in one-dimensional linear forms. However in this study, we have to deal with protein-PWM interaction, where one partner (protein) is in linear form, and the other partner (PWM) is in a 2D matrix form.

We used a straightforward way to convert the two-dimensional binding profile matrix (PWM) into a one-dimensional linear sequence based on the base composition of the PWM (**Fig. 2**). As we know, each PWM is represented as $4\times L$ matrix where 4 is the number of bases (A/T/G/C) and L is the number of binding positions. For a TF family with N PWMs, if the length of a PWM is different from the family-wise PWM's length M, we trimmed/extended the PWM to a length of M according to the method described in **SI 2**. We therefore have N PWMs, each has M positions, and thus have a total of $N\times M$ positions. We then performed chi-square tests to compare the similarity of any two of the $N\times M$ positions. For each test, we used the A/T/G/C compositions from both positions to test whether the two positions have similar base compositions, measured by P value. We then used P value to construct a pairwise similarity matrix with $N\times M$ rows and $N\times M$ columns. The pairwise similarity matrix was then used to cluster these positions into K clusters. If the positions are clustered together, they have similar base compositions. We then assigned the cluster label to the positions. By assigning a cluster label to each position, we successfully converted the two-dimensional binding profile ($4\times M$) into a linear alphabet string ($1\times M$). This string, as a consensus, is a compact way to represent a set of DNA sequences of the same length.

In this study, we had used two different values for K clusters, 7 and 15. The rationale for using this two is described in **SI 8**. Since we had observed a better performance when K = 15, we had reported all our results below with the number of clusters K = 15. The results of 7 clusters are shown in **Table S2j**. After doing this, for each position of the original PWM, a new cluster label was assigned to represent the A/T/G/C base composition of that position.

After the conversion, the MI score for protein residue and DNA alphabet could then be calculated with the following formula:

$$MI(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p_1(x)p_2(y)}\right) \qquad (1)$$

where $p(x,y)$ is the probability of x and y occurring together, and $p_1(x)$ and $p_2(y)$ are the probabilities of x occurring in X, and y occurring in Y, independently.

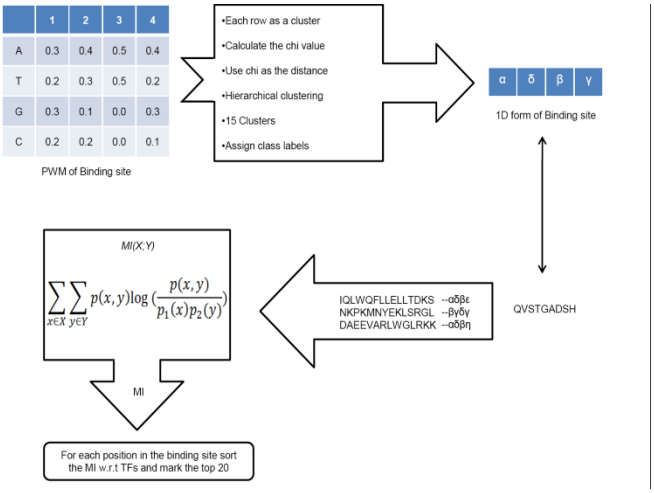

Fig. 2. Identification of interacting protein residue-DNA base pairs by mutual information. Upper row: We calculated the pairwise similarity of the profiles in any two positions, and clustered the positions based on the similarity matrix. The positions were clustered into 15 clusters and the cluster labels (in Greek alphabets) were assigned to each position. The two-dimensional PWMs were therefore converted into one-dimensional alphabets and made ready for the MI calculation. Middle row: We then used the MI formula to calculate the dependency of protein residues and DNA positions. Lower row: Finally, we ranked the residue-base pairs based on their MI scores. The top 20 residues were selected for further study.

3 RESULTS

3.1 Correlated evolution between protein sequences and DNA binding sites.

For the bHLH family, as illustrated in **Fig. 1**, we first built the similarity matrix for all protein pairs and the similarity matrix for corresponding binding sites PWM pairs. We then calculated PCC between these two matrices and assessed its statistical significance as described in the method section. The protein sequence similarity can be calculated from either the whole protein sequences of the TFs or only the binding domain sequences. The DNA binding domain refers to the basic region of bHLH that is directly involved in DNA binding, which might be most relevant to its DNA binding motifs. However, the rest of the protein sequence might have long-range effects on DNA recognition and thus the usage of whole protein sequence might provide additional evolutionary information. We, therefore, used both the whole protein sequence and the binding domain to measure the evolution of the TFs.

As shown in **Table 1**, protein domain sequences are significantly co-evolved with their binding sites in all three datasets. The PCC values for mammalian, eukaryotes, and CAEEL datasets are 0.435, 0.353 and 0.329, respectively, with P values of 0.029, 0.002 and 2×10^{-6}. Even though the whole protein sequences did not show significant co-evolution with their binding sites in the mammalian and eukaryotes datasets, they are significant in the CAEEL dataset, PCC = 0.309 (P value = 7×10^{-6}). In general, the protein domain sequences show stronger co-evolution than the whole protein sequences. The reasons might be as follows: (1) Although the TF family is found in almost all eukaryotes, the one defining feature of a bHLH protein is that it contains the basic region of around fifteen conserved basic amino acids that bind to the specific DNA sequence. The rest of the sequence is more diverse, and although it is necessary for, it is independent of the DNA binding activity. (2) The protein-DNA complex structure (**Fig. S1**) for bHLH protein (PDB id: 1mdy) shows that the interacting surfaces are mainly on the bHLH DNA binding domain (Ma, et al., 1994). To take into account the position effect on amino acid composition of the TF family, we permuted the protein sequences in a position-specific way (**SI 3**) to calculate the P value. We have also performed a parametric test that is based on t-statistics (**SI 4**) and a nonparametric test based on Spearman's rank correlation (**SI 5**) to assess the significances of the correlations. The results from all these methods produce significant P values, as shown in **Table S2**.

Next, in order to draw a more general conclusion regarding the co-evolution, we further validated our hypothesis in three more protein families: Homeo (100 members), HMG (15 members) and TRP (11 members). We performed correlation analysis on these

three families. The results are shown in **Table 2**. In all three families, both whole and domain sequences are significantly co-evolved with their binding sites. Especially for Homeo family, 100 members sample size grants more confidence on the results. Its domain sequences show stronger co-evolution (PCC = 0.416) than the whole protein sequences (PCC = 0.245), which is consistent with what we have observed in bHLH family. For the other two families, whole protein sequences show higher (yet not much higher) correlation than domain sequences. In summary, the validation was based on much larger datasets and more diverse protein families. The results indicate that our discovery on the co-evolution between TFs and their binding sites is general and reliable.

Table 1. The co-evolution test of bHLH TFs and their TFBS at whole protein sequence and domain sequence levels

Data source	Dataset* (name, size)	Protein sequence			
		Whole		Domain	
		PCC†	*P* value‡	PCC	*P* value
JASPAR	I (mammals, 6)	0.134	0.443	0.435	**0.029**
	II (eukaryotes, 16)	0.036	0.143	0.353	**0.002**
UniPROBE	III(CAEEL, 20)	0.309	**7×10^{-6}**	0.329	**2×10^{-6}**

*TFs in dataset I are from JASPAR FAM definition, dataset II include new additions from JASPAR CORE, and dataset III are collected from UniPROBE database. All three datasets exclude heterodimers.
†PCC: Pearson's correlation coefficient.
‡*P* value was calculated from permutated protein sequences and PWMs (1,000 times); *P* values < 10-3 were obtained from model based t-test (see supplementary notes for details). *P* values <0.05 are considered to be significant and are highlighted in bold.

Table 2. The co-evolution test of Homeo, HMG and TRP familyTFs and their TFBS at whole protein sequence and domain sequence levels

Data source	Dataset	Protein sequence			
		Whole		Domain	
		PCC†	*P* value‡	PCC	*P* value
JASPAR	IV (Homeo, 100)	0.245	**$<10^{-7}$**	0.416	**$<10^{-7}$**
	V (HMG, 15)	0.194	**0.024**	0.191	**0.025**
	VI (TRP, 11)	0.441	**3.8×10^{-4}**	0.345	**0.005**

†,‡Annotations are the same as Table 1.

3.2 Control for the effects of speciation

As our datasets are from multiple species, we next asked whether the observed correlations were due to the effects of speciation. For this purpose we calculated the PCC between the bHLH's binding profiles and the domain sequences from Homeo family and HMG family (**Table S3**). The PCCs from both families were not significant (**Table S4**). We further computed a similarity matrix from synonymous changes among the bHLH family members and correlated this matrix with that of the bHLH binding profiles. The PCC value between bHLH binding profiles and inverse synonymous change distance matrix was 0.078, which is not significant (*P* value = 0.462, **SI 6**). Both tests suggested that the higher correlation between bHLH domains and their PWMs was not due to speciation.

3.3 Correlated evolution between protein residues and DNA binding sites

All the above datasets exhibited significant TF-TFBS co-evolutions at the binding domain level. We then wanted to computationally pinpoint the residues/bases pairs that are co-evolved. We used a MI based method for this purpose. Here we used bHLH family as our case study. Previous study has shown that the accuracy of MI critically depends on the number of sequences in the alignment (Fernandes and Gloor, 2010). If there are very few sequences, top MI values could be occurring by chance. Therefore, the mammalian dataset was not used for MI analysis.

Here the challenge is how to measure MI between sequences and matrices (PWMs), which differs in protein-protein interaction where MI between sequences is measured. To calculate MI between protein and PWM, we first developed an algorithm, as described in methods section, to convert the two-dimensional binding profile for each TF into one-dimensional alphabets, based on the similarities of the DNA base composition at each position (**Fig. 2**). We then downloaded the protein domain alignment from the Pfam database (Accession: PF00010). Out of the 16 members in our eukaryote dataset 13 were annotated in this alignment, which formed the basis of the MI analysis. The remaining three members were manually aligned onto the alignment with the help of ClustalW. And for our *C. elegans* dataset, all 20 members were directly extracted from Pfam alignment. Then we calculated the MI for each residue position in the protein domains and for each binding position in the profiles for both eukaryote and *C. elegans* alignments. We took the top 20 residues with the highest MI scores for each DNA profile position. After removing the redundancy, we identified 36 highly coevolving residue positions from 111 binding positions in our eukaryote alignment, and 28 residue positions from 68 binding positions in our *C. elegans* alignment.

Next, we used protein-DNA complexes with solved 3D structure to test whether structurally determined interacting residue-base pairs are co-evolved. We collected structurally annotated interactions from the BIPA database (Atchley and Fitch, 1997) (see **SI 7** for details of BIPA annotation). Three members in dataset II have solved protein-DNA structures in the PDB database: USF (1AN4) (Ferre-D'Amare, et al., 1994), PHO4 (1A0A) (Shimizu, et al., 1997) and MAX (1HLO) (Brownlie, et al., 1997). As the annotated interacting residues/bases are similar among them, we chose the most ancient bHLH TF USF (Atchley and Fitch, 1997) as a template. Positions in the binding domain alignment corresponding to hydrogen bonds, water-mediated hydrogen bonds and van der Waals contacts were selected (**SI 7**). Especially, for the eukaryote alignment, out of a total of 19 interacting position pairs, 14 were also identified as highly coevolving position pairs by the MI method, which showed a significant overlap with *P* value <0.001 (**Fig. S2a**). For the *C. elegans* alignment, the overlapped position amount is 10 (*P* value = 0.107, **Fig. S2b**).

Table 3. The co-evolution test of TFs and their TFBS based on mutual information identified residues, and BIPA annotated interacting residues

Data source	Dataset*	Protein residues			
		MI		Interacting (BIPA)	
		PCC	*P* value	PCC	*P* value
JASPAR	II (eukaryotes, 16)	0.346	**0.003**	0.369	**0.002**
UniPROBE	III(CAEEL, 20)	0.177	**0.007**	0.323	**2.5×10^{-6}**

*All the annotations are the same as Table 1.

Both MI selected residues and BIPA annotated residues showed significant correlations with TFBSs: in eukaryote dataset, *P* value = 0.003 for MI, and 0.002 for BIPA; in *C. elegans* dataset, *P* value = 0.007 for MI, and 2.5×10^{-6} for BIPA (**Table 3**), which indicates the critical roles the important residues play in TF-TFBS interaction. Moreover, the BIPA annotated residues group showed higher correlations (PCC = 0.369 for eukaryote dataset, and 0.323 for *C. elegans* dataset) than MI residues (PCC = 0.346 and 0.177). Especially, for eukaryote dataset, the correlation at interacting residues level is even higher than binding domain level (PCC = 0.353). Also, one interesting position at eukaryote alignment that was identified by both BIPA and MI in this study is position 13. In general, most bHLH TFs prefer to recognize and specifically bind to a 6mer DNA sequence, 5'-CANNTG-3', which is termed as an E-box motif (Chaudhary and Skinner, 1999). As shown in **Fig. 3**, residues at position 13 dramatically correlated with different binding preferences in the E-box region. When arginine is at this position, the corresponding TF will bind to canonical CACGTG E-box; for any other amino acid, the corresponding TF will bind to different E-box motifs other than canonical CACGTG. Our observation is consistent with previous studies on the important role of this position in E-box binding affinity (Ma et al. 1994; Shimizu et al. 1997).

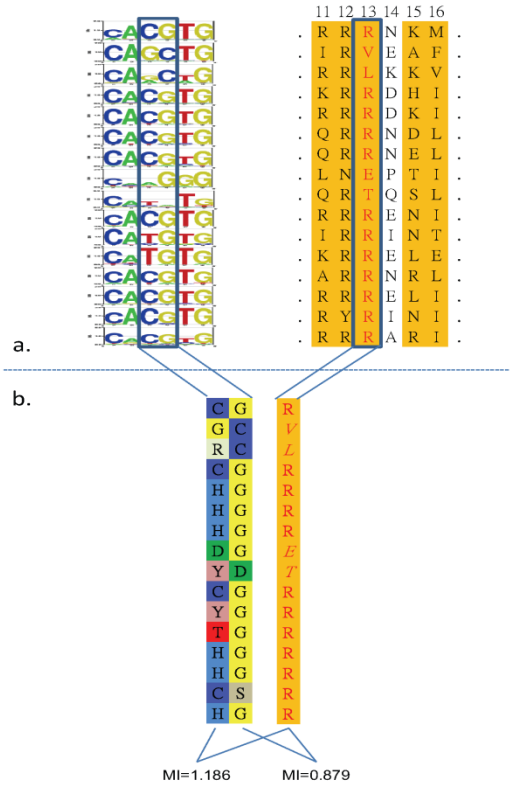

Fig. 3. The amino acid at position 13 in the bHLH protein's basic region is significantly co-evolved with the central two positions of E-box binding sites. **a)** On the left is one segment containing the basic region of bHLH proteins from the Pfam multiple sequence alignment of the eukaryote bHLH dataset. Interacting residues annotated by the BIPA database are highlighted and the critical position 13 is marked. On the left is the schematic alignment of DNA binding profiles corresponding to the bHLH TFs on the right. The central two positions of the E-box binding motif (CANNTG), which vary from different bHLH subgroups, are also marked. **b)** The position 13 of the bHLH protein's basic region and central two positions of E-box binding sites can be closely compared with each other. Here the positions on DNA binding profiles are presented as our novel clustered sequences derived from the base composition in PWMs. Significant co-varying patterns are shown with the corresponding calculated high Mutual Information (MI) scores.

The results indicate that co-evolved residues and structurally important residues are not identical but related; they both play important roles in a protein's function. This observation may also indicate the potential of applying such criteria to filter out less important residues for a long binding domain. The results also show that PCCs from MI identified residues are not necessarily higher than the PCCs from the domain sequences (Tables 1 and 3). This is because different perspectives MI and PCC focusing on: MI considers individual residue while PCC considers entire sequence. MI is computed in a position-independent manner from alignment of protein binding domain sequences and alignment of our clustered strings from PWMs. It assumes each position in the alignment is independent. As a result, a co-evolved position (residue) pinpointed by MI indeed highly correlates with one or several posi-

tions in the PWM-converted string alignment. While PCC measures the correlation at sequence level, which means the joint of best hits (i.e. residues) pinpointed by MI may not guarantee a best PCC (though they are fairly related).

3.4 DNA binding sites symmetry

The bHLH proteins bind to DNA as homodimers as shown in **Fig. S1**, and their DNA binding sites and E-boxes are indeed symmetrical motifs. Such a symmetric feature has been widely utilized in current studies for TF binding motif discovery (Tan, et al., 2005). Here, we have also performed the same correlation analysis based on such symmetrical motifs. A strong correlated evolution between these symmetrical PWMs and their TFs at the domain and residue levels is observed (**Table S2g** and **S2h**). In general, the correlations here follow similar distribution and are even ~25% greater than using original PWMs. This further supports our observations of co-evolution.

4 DISCUSSION

We used the bHLH, Homeo, HMG and TRP family TFs and their TFBSs to test the hypothesis that the evolution of TFs and their binding sites are correlated. For bHLH family, we used three datasets: the mammalian dataset and eukaryote dataset from the JASPAR database, and the CAEEL dataset from the UniProbe database. In all families and datasets we tested, the co-evolutionary relationship between TF and TFBS can be clearly seen, showing that the observations were not dependent on the platform used to obtain the TFBSs. Moreover, irrespective of whether the TFs are among different species (bHLH family mammalian and eukaryotes) or within one species (bHLH family *C. elegans*), we have observed similar co-evolution relationship. This observation strengthens our hypothesis that such an evolutionary relationship exists between TFs and their TFBSs.

We applied multiple statistical tests to quantify the statistical significance of any co-evolutionary relationship and all tests were positive. By default, we used a nonparametric test that is based on background distribution of PCCs obtained by permuting both protein sequences and DNA binding sites (see methods section). To allow for different amino acid compositions at different positions, we also permuted the protein sequence in a position-specific way (**SI 3**). We further used two alternative tests, one is a parametric test assuming that null PCCs form a t-distribution (**SI 4**) and another is a nonparametric test by Spearman's rank correlation coefficient (**SI 5**). All these tests showed significant co-evolution between TFs and TFBSs.

The observed co-evolutionary relationship was not due to speciation. When we calculated the PCC between the PWMs from bHLH and protein sequence of other families, such as Homeo and HMG from the same species, the co-evolutionary relationship was no longer significant. Furthermore, we observed very significant co-evolution from the bHLH CAEEL dataset, in which the TFs are from only the single species *C. elegans*.

As the advances of new technologies, such as high-throughput protein arrays, more binding sites for a TF will be identified and as a result, the PWM will be continuously refined. To evaluate the effect of PWM's stability (i.e. the number of TFBS used to construct a PWM) on our observation, we performed a bootstrap experiment. We downloaded all the available TFBS original sites of bHLH eukaryote dataset (9 TFs out of 16) from JASPAR database and performed the evaluation based on them. We sampled 80% of the original binding sites for each TF to rebuild its PWM, and computed the PCC with all new PWMs; we repeated the sampling procedure for 1000 times to get a distribution of PCC so that we evaluated the effect of PWMs' stabilities by assessing the standard deviation of PCC. As a result, the sample standard deviation is as small as 0.011, which suggests the robustness of PWMs used in this study.

We carried out co-evolutionary tests at both whole protein sequence level and binding domain sequence level. Our results showed that domain level of bHLH TFs have significant co-evolution with TFBSs. These results are also consistent with previous studies (Atchley and Fitch, 1997; Jones, 2004; Ledent, et al., 2002) showing that sequences outside the conserved bHLH domain are extensively rearranged family-wide, and indicates that their evolution may be relatively more independent of TFBSs. For the other three families, both domain and whole sequence levels showed significant co-evolution with TFBSs which further validated our results. Our analyses also showed that at residue level, highly co-evolved amino acid residues are not entirely the same as DNA-interacting residues but they are strongly related.

Previous studies have observed correlated TF residues and TFBS bases in prokaryotes (Huang, et al., 2009) and metazoans (Noyes, et al., 2008). However, TF-TFBS interactions are not only determined by single residue-base pairs but also by long-range context of the surrounding residues (Pazos and Valencia, 2008). Previous analyses have not taken into account this context information for coevolving residues, and the relationship between TFs and TFBSs, as a whole, has not been systematically studied.

This research establishes the dynamic, co-evolutionary relationship between TFs and TFBSs. The confirmation of this relationship has long-term implications that will impact the field in many ways, similar to the discovery of co-evolutionary relationship of protein-protein interaction (Pazos and Valencia, 2001) and protein-ligand interaction (Goh, et al., 2000). For example, using this co-evolutionary relationship, we can rationally design proteins to specifically target a DNA sequence, similar to the design of zinc finger nucleases (Urnov, et al., 2005). On the other hand, we can use this relationship to predict TFBS on the DNA for a novel protein (Alleyne, et al., 2009). Clinically it is also conceivable to correct disease mutations by rational design of compensatory "repairs".

Although the current study is based on four families, the principle and the framework established here could be readily applied to other TF families. As more data sources become available in the future, subsequent studies will lead to a more thorough understanding of protein-DNA interaction and transcriptional regulation.

ACKNOWLEDGEMENTS

We thank Profs. Terry Hwa, Gary Stormo, Sridhar Hannenhalli, and Li-San Wang, the anonymous reviewers and members in the JJWang lab for critical comments on the manuscript.

Funding: This work was supported by grants from the Research Grants Council (781511M, 778609M, N_HKU752/10, AoE M-04/04) and Food and Health Bureau (10091262) of Hong Kong, and grants from the National Science Foundation of China.

Conflict of Interest: none declared.

REFERENCES

Alleyne, T.M., *et al.* (2009) Predicting the binding preference of transcription factors to individual DNA k-mers, *Bioinformatics*, **25**, 1012-1018.

Atchley, W.R. and Fitch, W.M. (1997) A natural classification of the basic helix-loop-helix class of transcription factors, *Proc Natl Acad Sci U S A*, **94**, 5172-5176.

Atwell, S., *et al.* (1997) Structural plasticity in a remodeled protein-protein interface, *Science*, **278**, 1125-1128.

Brownlie, P., *et al.* (1997) The crystal structure of an intact human Max-DNA complex: New insights into mechanisms of transcriptional control, *Structure*, **5**, 509-520.

Chaudhary, J. and Skinner, M.K. (1999) Basic helix-loop-helix proteins can act at the E-box within the serum response element of the c-fos promoter to influence hormone-induced promoter activation in Sertoli cells, *Mol Endocrinol*, **13**, 774-786.

Darwin, C. (1862) On the Various Contrivances by which British and Foreign Orchids are Fertilised by Insects, and on the Good Effects of Intercrossing, *London: John Murray*.

Fernandes, A.D. and Gloor, G.B. (2010) Mutual information is critically dependent on prior assumptions: would the correct estimate of mutual information please identify itself?, *Bioinformatics*, **26**, 1135-1139.

Ferre-D'Amare, A.R., *et al.* (1994) Structure and function of the b/HLH/Z domain of USF, *EMBO J*, **13**, 180-189.

Goh, C.S., *et al.* (2000) Co-evolution of proteins with their interaction partners, *J Mol Biol*, **299**, 283-293.

Grove, C.A., *et al.* (2009) A Multiparameter Network Reveals Extensive Divergence between C. elegans bHLH Transcription Factors, *Cell*, **138**, 314-327.

Hannenhalli, S. (2008) Eukaryotic transcription factor binding sites--modeling and integrative search methods, *Bioinformatics*, **24**, 1325-1331.

Huang, N., *et al.* (2009) Structure and function of an ADP-ribose-dependent transcriptional regulator of NAD metabolism, *Structure*, **17**, 939-951.

Izarzugaza, J.M.G., *et al.* (2006) TSEMA: interactive prediction of protein pairings between interacting families, *Nucleic Acids Res*, **34**, W315-W319.

Jones, S. (2004) An overview of the basic helix-loop-helix proteins, *Genome Biology*, **5**, 226.

Juven-Gershon, T., *et al.* (2008) The RNA polymerase II core promoter - the gateway to transcription, *Curr Opin Cell Biol*, **20**, 253-259.

Ledent, V., Paquet, O. and Vervoort, M. (2002) Phylogenetic analysis of the human basic helix-loop-helix proteins, *Genome Biology*, **3**, RESEARCH0030.

Li, M.J., Sham, P.C. and Wang, J.W. (2010) FastPval: a fast and memory efficient program to calculate very low P-values from empirical distribution, *Bioinformatics*, **26**, 2897-2899.

Ma, P.C., *et al.* (1994) Crystal structure of MyoD bHLH domain-DNA complex: perspectives on DNA recognition and implications for transcriptional activation, *Cell*, **77**, 451-459.

Moyle, W.R., *et al.* (1994) Co-evolution of ligand-receptor pairs, *Nature*, **368**, 251-255.

Noyes, M.B., *et al.* (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites, *Cell*, **133**, 1277-1289.

Pazos, F., *et al.* (1997) Correlated mutations contain information about protein-protein interaction, *J Mol Biol*, **271**, 511-523.

Pazos, F. and Valencia, A. (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction, *Protein Eng*, **14**, 609-614.

Pazos, F. and Valencia, A. (2008) Protein co-evolution, co-adaptation and interactions, *EMBO J*, **27**, 2648-2655.

Pietrokovski, S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments, *Nucleic Acids Res*, **24**, 3836-3845.

Portales-Casamar, E., *et al.* (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles, *Nucleic Acids Res*, **38**, D105-D110.

Qin, J., *et al.* (2011) ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor, *Nucleic Acids Research*, **39**, W430-W436.

Shimizu, T., *et al.* (1997) Crystal structure of PHO4 bHLH domain-DNA complex: flanking base recognition, *Embo J*, **16**, 4689-4697.

Tan, K., McCue, L.A. and Stormo, G.D. (2005) Making connections between novel transcription factors and their DNA motifs, *Genome Research*, **15**, 312-320.

Tillier, E.R.M., *et al.* (2006) Codep: Maximizing co-evolutionary interdependencies to discover interacting proteins, *Proteins*, **63**, 822-831.

Tress, M., *et al.* (2005) Scoring docking models with evolutionary information, *Proteins*, **60**, 275-280.

Urnov, F.D., *et al.* (2005) Highly efficient endogenous human gene correction using designed zinc-finger nucleases, *Nature*, **435**, 646-651.

Wang, J.W. and Hannenhalli, S. (2006) A mammalian promoter model links cis elements to genetic networks, *Biochemical and Biophysical Research Communications*, **347**, 166-177.

Wang, J.W., *et al.* (2007) MetaProm: a neural network based meta-predictor for alternative human promoter prediction, *BMC Genomics*, **8**, -.

Weigt, M., *et al.* (2009) Identification of direct residue contacts in protein-protein interaction by message passing, *Proc Natl Acad Sci U S A*, **106**, 67-72.

Weil, P., Hoffgaard, F. and Hamacher, K. (2009) Estimating sufficient statistics in co-evolutionary analysis by mutual information, *Comput Biol Chem*, **33**, 440-444.