Correlated evolution of transcription factors and their binding sites

Shu Yang1, Hari Krishna Yalamanchili1, Xinran Li1,†, Kwok-Ming Yao1, Pak Chung Sham2, Michael Q. Zhang3,4 and Junwen Wang1,∗

1Department of Biochemistry, LKS Faculty of Medicine, The University of Hong Kong, 21 Sassoon Road, Hong Kong SAR, China, 2Department 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, 3Department of Molecular and Cell Biology, Center for Systems Biology, The University of Texas at Dallas, Dallas, TX, 75080, USA and 4Bioinformatics Division, TNLIST, Tsinghua University, Beijing, 100084, China

ABSTRACT
Motivation: The interaction between transcription factor (TF) and transcription factor 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 correcting and rescuing disease mutations in clinical applications.

Results: Based on six, publicly available, experimentally 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 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.

Received on June 11, 2011; revised on August 1, 2011; accepted on August 16, 2011

1 INTRODUCTION
The term co-evolution was first conceptualized by Charles Darwin (1862) in his study on the two species, pollinators and orchids. At the molecular level, co-evolution has been studied in the context of protein–protein interaction (Atwell et al., 1997; Motyle et al., 1994; Pazaros 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 (Pazaros and Valencia, 2001), as well as of interacting protein and ligand (Geb et al., 2000). These studies have had great impact in this area of research, such as in protein–protein interaction and drug design (Izaurzegaza 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 evolutions of TFs and their TFBSs are correlated. We used the basic helix-loop-helix (bHLH), Homeo, high-mobility group (HMG) and transient receptor potential channels (TRP) protein families to illustrate this principle of co-evolutionary relationship. We first performed the analysis based on the bHLH family. The bHLH family is an ancient component of transcriptional regulation with more than 100 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...
We then used the compositions to generate random sequences that matched the PWMs and permuted protein sequences. This random process was repeated 1000 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

Mutual information (MI) has been successfully used to identify contacting residues in protein–protein interactions (Weigt et al., 2009), and co-evolving residue pairs within a protein (Weil et al., 2009). Current MI methods deal with protein sequences, where both interacting partners are in 1D 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 2D binding profile matrix (PWM) into a 1D linear sequence based on the base composition of the PWM (Fig. 2). As we know, each PWM is represented as a 4 × 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 Supplementary Material 2. We therefore have N PWMs, each has M positions and thus, have a total of N × M positions. We then performed chi-square tests to compare the similarity of any two of the N × 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 × M rows and N × M columns.
The rationale for using these two is described in Supplementary Material 8. The pairwise similarity matrix was then used to cluster these positions. Since we had observed a better performance when the A/T/G/C base composition of that position could then be calculated with the following formula:

\[ p_{xy} = \frac{\sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right) }{\text{Total number of pairs}} \]  

where \( p(x,y) \) is the probability of \( x \) and \( y \) occurring together, and \( p_x \) and \( p_y \) are the probabilities of \( x \) occurring in \( X \), and \( y \) occurring in \( Y \), independently.

### 3 RESULTS

#### 3.1 Correlated evolution between protein sequences and DNA binding sites

For the bHLH family, as illustrated in Figure 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 Section 2, Methods. 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 \( 2 \times 10^{-6} \). 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, with \( P \)-value of 0.309 (\( P = 7 \times 10^{-6} \)). In general, the protein domain sequences show stronger co-evolution than the whole protein sequences. The reasons might be as follows: (i) 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 15 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; (ii) The protein-DNA complex structure (Supplementary Fig. S1) for bHLH protein (PDB id: 1mysql) 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 (Supplementary Material 3) to calculate the \( P \)-value.

<table>
<thead>
<tr>
<th>Dataset Description</th>
<th>Protein sequence</th>
<th>Domain sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>JASPAR I (mammals, 6)</td>
<td>0.134 0.443 0.029</td>
<td>0.435 0.353 0.329</td>
</tr>
<tr>
<td>II (eukaryotes, 16)</td>
<td>0.036 0.143 0.002</td>
<td>( 2 \times 10^{-6} )</td>
</tr>
<tr>
<td>UniPROBE III (CAEEL, 20)</td>
<td>0.309 7 \times 10^{-6} 0.329</td>
<td></td>
</tr>
</tbody>
</table>

*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 permuted protein sequences and PWMs (1000 times).

**\( P \)-value of \( 2 \times 10^{-6} \):** obtained from a model-based \( t \)-test (see Supplementary Material for details). **\( P < 0.05 \) are considered to be significant and are highlighted in bold.**

![Identification of interacting protein residue–DNA base pairs by co-evolution](image)

**Fig. 2.** Identification of interacting protein residue–DNA base pairs by co-evolution. The MI score for protein residue and DNA alphabet can be calculated with the following formula:

\[ MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right) \]
All the above datasets exhibited significant TF–TFBS co-evolutions. As our datasets are from multiple species, we next asked whether the 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 the 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.

### 3.2 Control for the effects of speciation

As our datasets are from multiple species, we next ask 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 (Supplementary Table S3). The PCCs from both families were not significant (Supplementary 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 = 0.462, Supplementary Material 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/base 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 Section 2, Methods, to convert the 2D binding profile for each TF into 1D alphas, 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. For our Caenorhabditis elegans dataset, all 20 members were directly extracted from Pfam alignment. 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 alignment. After removing the redundancy, we identified 36 highly co-evolving 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 Supplementary Material 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), PBR4 (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 (Supplementary Material 7). Specifically, for the eukaryote alignment, out of a total of 19 interacting position pairs, 14 were also identified as highly co-evolving position pairs by the MI method, which showed a significant overlap with P < 0.005 (Supplementary Fig. S2a). For the C.elegans alignment, the overlapped position amount is 10 (P = 0.107, Supplementary Fig. S2b).

Both MI selected residues and BIPA annotated residues showed significant correlations with TFBSs: in eukaryote dataset, the correlation at interacting residues level is 0.345, which indicates the critical roles the important residues that 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). Specifically, 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 Figure 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).

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 of MI and PCC focusing on: MI considers individual residue, whereas, 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 may indeed highly correlates with one or several positions 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 Supplementary Figure S1, and their DNA binding sites and E-boxes are indeed symmetrical motifs. A strong correlated evolution between these symmetrical PWMs and their TFs at the domain and residue levels is observed [Supplementary Table S2 (g and h)]. 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 evolutions 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 that we tested, the co-evolutionary relationship between TF and TFBS can be seen clearly, 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-evolutionary relationship. This observation strengthens our hypothesis that such an co-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 non-parametric test that is based on background distribution of PCCs obtained by permuting both protein sequences and DNA binding sites (see Section 2. Methods). To allow for different amino acid compositions at different positions, we also permuted the protein sequence in a position-specific way (Supplementary Material 3). We further used two alternative tests, one is a parametric test assuming that null PCCs form a r-distribution (Supplementary Material 4) and another is a non-parametric test by Spearman’s rank correlation coefficient (Supplementary Material 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 C AeL 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 (0 TFBS out of 161) 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 TFBSs 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 co-evolving 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 interactions and transcriptional regulation.

ACKNOWLEDGEMENTS

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

Funding: The Research Grants Council of Hong Kong (781511M, 778609M, N_HKU752/10, AoE M-04/04); Food and Health Bureau of Hong Kong (10091262); the National Science Foundation of China (31061160497).

Conflict of Interest: none declared.

REFERENCES


