DNA-Binding Specificity Changes in the Evolution of Forkhead Transcription Factors

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

Citation

Published Version
doi:10.1073/pnas.1310430110

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:12872182

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
DNA Binding Specificity Changes in the Evolution of Forkhead Transcription Factors

So Nakagawa, Stephen S. Gisselbrecht, Julia M. Rogers, Daniel L. Hart and Martha L. Bulyk

Submitted to Proceedings of the National Academy of Sciences of the United States of America

The evolution of transcriptional regulatory networks entails the expansion and diversification of transcription factor (TF) families. The forkhead family of TFs, defined by a highly conserved winged helix DNA-binding domain (DBD), has diverged into dozens of subfamilies in animals, fungi, and related protists. We have used a combination of maximum likelihood phylogenetic inference and independent, comprehensive functional assays of DNA binding capacity to explore the evolution of DNA binding specificity within the forkhead family. We present converging evidence that similar alternative sequence preferences have arisen repeatedly and independently in the course of forkhead evolution. The vast majority of DNA binding specificity changes we observed is not explained by alterations in the known DNA-contacting amino acid residues conferring specificity for canonical forkhead binding sites. Intriguingly, we have found forkhead DBDs that retain the ability to bind very specifically to two completely distinct DNA sequence motifs. We propose an alternate specificity-determining mechanism whereby conformational rearrangements of the DBD broaden the spectrum of sequence motifs that a TF can recognize. DNA binding bispecificity suggests a new source of modularity and flexibility in gene regulation and may play an important role in the evolution of transcriptional regulatory networks.

DNA binding specificity | evolution | transcription factor

Introduction

The regulation of gene expression by the interaction of sequence-specific transcription factors (TFs) with target sites (cis-regulatory elements) near their regulated genes is a central mechanism by which organisms interpret regulatory programs encoded in the genome to develop and interact with their environment. The emergence of new species has depended in part on the evolution of the network of interactions by which an organism’s TFs control gene expression. Much attention has been paid to changes in cis-regulatory sequences over evolutionary time, as these changes can result in incremental modifications of organismal phenotypes without large-scale ‘rewiring’ of transcriptional regulatory networks that would result from changes in TF DNA binding specificity (1). Nevertheless, TFs and their DNA binding specificities have changed over time (2). Gene duplication, followed by divergence of the resulting redundant TFs, has resulted in the emergence of families of paralogous TFs with diversified DNA binding specificities and functions (3). Thus, identifying mechanisms by which related DNA-binding domains (DBDs) have acquired novel specificities is important for understanding TF evolution.

The forkhead box (Fox) family of TFs spans a wide range of species, and is one of the largest classes of TFs in humans. In metazoans, Fox proteins have vital roles in development of all tissues and homeostasis. In Arabidopsis thaliana, some 116 Fox genes have been identified (4). The forkhead domain is roughly 80–100 amino acids (a.a.) in length and, like other winged helix domains, comprises a bundle of three α-helices connected via a small β-sheet to a pair of loops or “wings”. In available structures of forkhead domain-DNA complexes, helix 3 forms a canonical recognition helix positioned in the major groove of the DNA target site by the helical bundle, while the wings, which often contain a poorly alignable region rich in basic residues, lie along the adjacent DNA backbone (9-13).

Several groups have studied the evolutionary history of the Fox family using multiple sequence alignment and phylogenetic inference methods; however, the results of these studies are in many cases inconsistent. Published forkhead phylogenies lack statistical support for deep branches and the relative positions of forkhead subfamilies, especially of the fungal groups (14, 15). Thus, the relationships among Fox genes have remained unclear.

In separate studies, the DNA binding specificities of various forkhead proteins have been examined. In most cases, in vitro binding has been observed to variants of the canonical forkhead target sequence RYAAAY (16-21), which we refer to as the forkhead primary (FkhP) motif (Figure 1). A similar variant, AHAACA, has been observed in in vitro selection (SELEX) (17) and protein-binding microarray (PBM) experiments (20); this specificity appears to be common to several Fox proteins, and we refer to it as the forkhead secondary (FkhS) motif (22). However, a SELEX study of the FoxN1 TF mutated in the famous nude
mouse identified an entirely different sequence, ACGC, as its preferred binding site (23). The closely related \textit{Mus musculus} FoxN4 has been shown to bind ACGC in vivo (24). A PBM survey of \textit{Saccharomyces cerevisiae} TFs identified a very similar sequence, GACGC, as the binding site of the Fox3 factor Fhl1 (19); we therefore refer to the GACGC site as the FHL motif (Figure 1).

Previous work on differences in forkhead DNA binding specificity has focused on preferential recognition of FkhP and FkhS variants by forkhead proteins (17, 18). Contrary to the common mechanism of varying specificity by changing a.a. residues that make base-specific DNA contacts (25), the positions in the forkhead recognition helix that make base-specific contacts are conserved across proteins with different binding specificities (9, 17). In sub-domain swap experiments, a 20-a.a. region immediately N-terminal to the recognition helix was shown to switch DNA-binding specificities between forkhead proteins (17). Interestingly, this region has been shown by NMR to adopt different secondary structures in forkheads with distinct DNA binding specificities (26). However, a similar analysis of sequence features conferring binding to the FHL motif has not been performed.

The observation of binding to such different sequences – RYAAAYA and GACGC –within widely diverged members of the Fox family raises the question of how the binding specificity of these proteins has evolved. We have addressed this question using a combined phylogenetic and biochemical approach. We conducted a phylogenetic analysis of Fox domains from 10 metazoans, 30 fungi, and 25 protists (Table S1). We chose these species based on their evolutionary importance and annotation level (27) (Figure S1). For example, we included \textit{Spizellomyces punctatus} and \textit{Fonticula alba}, since they are very close to the root of fungi and a closely related outgroup, respectively. We considered conserved splice junctions along with multiple sequence alignment to infer the phylogeny. We assayed DNA binding specificity \textit{in vitro} using universal PBM technology, in which a DNA-binding protein is applied to a double-stranded DNA microarray containing 32 replicates of all possible 8-bp sequences (8-mers) and is fluorescently labeled, permitting the exhaustive cataloging of the range of sequences a protein can recognize (28). We analyzed the binding specificities of 30 forkhead proteins, combining previously published data for 9 proteins with data for 21 proteins that we newly characterized for this study (Table S4). We focused on proteins from clades where we had previously observed alternate binding specificities and clades of unknown specificity. By using two orthogonal means of evaluating the same proteins, we obtain a much richer picture of the evolutionary trajectory of changes in TF DNA binding specificity than either analysis alone can provide.

Results

The published observation of roughly the same alternate binding motif (FHL) for metazoan FoxN1 and fungal Fox3 suggests the parsimonious hypothesis that they derive from a common FHL-binding ancestral protein in the last common ancestor of opisthokonts. To explore this hypothesis, we performed phylogenetic inference on a broad group of Fox domain sequences (see Materials and Methods), spanning 623 genes from 65 species (Table S1, Figure S1). We included two distantly related forkhead domains from the opisthokont sister group Amoebozoa as an outgroup. After removing partial domain sequences and those identical throughout the Fox domain, we used 529 Fox domain
sequences (340 non-redundant, Table S1). We constructed a complete maximum likelihood (ML) tree of all non-redundant Fox domain sequences (Figure S2). For each branch, the approximate Likelihood-Ratio Test (aLRT) and 100 bootstrap replicates were used to evaluate support for inferred relationships (see Materials and Methods). For presentation purposes, we constructed a ML tree of 262 (133 non-redundant) Fox domains from selected informative species (Figure 2, Table S1).

Various portions of the phylogeny could be determined with high confidence. Our analysis recovered the previously identified subfamily relationships between Fox proteins, as well as identifying a new fungal group (Fox3) not previously observed because it is not represented in S. cerevisiae. However, the structure of the deep portions of the Fox tree could not be resolved for two major reasons. First, the number of alignable positions within the Fox domain is too small to resolve the phylogenetic history of such a broadly and deeply diverged family, and regions outside the domain are not alignable among distantly related members. Second, some Fox genes appear to have evolved through gene conversion and/or crossover events (15), as evinced by the appearance of species-specific Fox domain signatures.

The ML tree inferred here strongly supports the hypothesis of Larroux et al. that a monophyletic group of forkhead domains (which they refer to as clade I) emerged in the common ancestor of metazoans (14) (aLRT value = 0.9999, bootstrap value = 4%) (Figure 2). Additionally, there is a splice site between a.a. positions 46 and 47 in the Pfam Forkhead domain hidden Markov model (29) conserved in various clade II forkhead proteins across kingdoms; no clade I genes share this splice site, further supporting the monophyly of clade I in metazoans.

Surprisingly, there is no support for a tree topology in which metazoan FoxN and fungal Fox3 subfamilies form a monophyletic, FHL-binding clade. A tree containing a FoxN+3 clade (Figure S3A) is significantly less likely than the observed tree (p < 10^-8, likelihood ratio test), and likelihood maximization using this as a starting tree separates the FoxN and Fox3 clades (Figure S3B,C). Moreover, we see separate, well-supported clades (aLRT values ≥ 0.99) combining each of these groups with others that bind only the FkhP,S motifs, cluster 2 proteins are specific only for FHL variants, and cluster 3 proteins have more complex specificity; see text for details. Sequence motifs shown were generated by alignment of the indicated clusters of 8-mers and are for visualization purposes only.

A phylogenetic tree constructed from only fungal Fox3 domains (Figure 3A) is much more stable than the larger, more complex tree, with acceptable bootstrap support at major branch points; moreover, it follows the species tree closely (see Figure S1), suggesting radiation of a family of orthologs. The most basally diverged member of this group, Allomyces macrogynus Fox3, binds only the canonical FkhP and FkhS motifs (Figure S3A and Figure 4), providing experimental support for the hypothesis that FHL binding arose within the Fox3 clade after its divergence from other forkhead domains. The remaining Fox3 proteins considered here fall into two distinct groups. Those most closely related to Fhl1 (S. cerevisiae Fox3) show the same FHL-binding specificity, binding the FkhP5S motifs better than non-forkhead proteins (percent signs in Figure 3A). Members of the other group, including Aspergillus nidulans Fox3, bind another motif entirely, which we term the Forkhead Variant Helix (FVH) motif (dollar signs in Figure 3A; see Figure 1), with no specific binding to either the FkhP5S or FHL motifs.

Similarly, the phylogeny of the holozoan FoxN subfamily is relatively stable (Figure 3B). Our analysis supports the existence...
of a fundamental split into FoxN1/4 and N2/3 clades, with FoxR (initially called N5 (30)) placed within the N1/4 group (14). As expected, FoxN1 and other N1/4 proteins are highly specific for the FHL motif. Surprisingly, all FoxN2/3 proteins assayed by PBMs exhibited high sequence specificity for both the FkhPS and FHL motifs (see Figure 4). For example, the top two 8-mers (ranked by PBM enrichment (E) score), which indicates the preference of a protein for every possible 8-mer (28) bound by each of these two groups. Specifically, all tested FHL-binding Fox3 proteins strongly prefer A immediately 3’ to the core GACGC, which we refer to as the FHL-3 motif, while FHL motifs from FoxN/R proteins all strongly disfavor A in that position, a variant we refer to as the FHL-N motif (Figure 1).

The unexpected variety in Fox domain binding specificity led us to perform additional PBM experiments on a range of Fox domains, focusing on representative proteins from other clade II groups such as Fox4 and FoxM, and assemble them with previously published PBM data (Figure S4, Table S2, Table S3).

In addition to finding more examples of proteins that exhibit the sequence preferences described above, we also discovered a third instance of binding to an FHL-like motif. Two metazoan FoxM proteins exhibit high specificity for the FkhP and FkhS motifs, and for a third FHL variant, GATGC, which we refer to as FHM-4. The most preferentially bound 8-mer matching this motif is an overlapping inverted repeat, GATGCATC; human FoxM1 has previously been shown to bind overlapping multimers of the FkhP motif in vitro, which suggests that these two FoxM proteins might bind as dimers to GATGCATC. Phylogenetic analysis strongly supports an independent origin of the FoxM subfamily from FoxN (p < 10^{-4}, likelihood ratio test, Figure S3D), in that each subfamily is more closely related to proteins that bind only FkhP and FkhS than to each other, suggesting that this represents yet a third independent emergence of a form of FHL binding (FHL-M), with each one characterized by slight differences in DNA sequence preference (Figure 1). As in the case of FoxN and Fox3, ML inference with a starting tree containing a FoxM+N clade leads us to perform additional PBM experiments on a range of Fox domains, focusing on representative proteins from other clade II groups such as Fox4 and FoxM, and assemble them with previously published PBM data (Figure S4, Table S2, Table S3).

The unexpected variety in Fox domain binding specificity led us to perform additional PBM experiments on a range of Fox domains, focusing on representative proteins from other clade II groups such as Fox4 and FoxM, and assemble them with previously published PBM data (Figure S4, Table S2, Table S3).

In addition to finding more examples of proteins that exhibit the sequence preferences described above, we also discovered a third instance of binding to an FHL-like motif. Two metazoan FoxM proteins exhibit high specificity for the FkhP and FkhS motifs, and for a third FHL variant, GATGC, which we refer to as FHM-4. The most preferentially bound 8-mer matching this motif is an overlapping inverted repeat, GATGCATC; human FoxM1 has previously been shown to bind overlapping multimers of the FkhP motif in vitro, which suggests that these two FoxM proteins might bind as dimers to GATGCATC. Phylogenetic analysis strongly supports an independent origin of the FoxM subfamily from FoxN (p < 10^{-4}, likelihood ratio test, Figure S3D), in that each subfamily is more closely related to proteins that bind only FkhP and FkhS than to each other, suggesting that this represents yet a third independent emergence of a form of FHL binding (FHL-M), with each one characterized by slight differences in DNA sequence preference (Figure 1). As in the case of FoxN and Fox3, ML inference with a starting tree containing a FoxM+N clade leads to separation of the subfamilies (Figure S3E).
One of the forkhead-like domains from the non-opisthokont Acanthamoeba castellanii did not fall into any of these three clusters, as it binds another distinct motif (see Figure 1, Figure S4). These binding differences are associated with widespread differences in the recognition helix (Figure 5A). Indeed, altered recognition positions (Figure 5B) can clearly explain the non-FkhPS/N PS specificity of the forkhead-related protein from A. castellanii and A. nidulans Fox3; furthermore, there are sufficient differences in the recognition helix of H. sapiens FoxR1 that it is perhaps surprising that its specificity is so similar to that of other Fox proteins. Surprisingly, however, the majority of specificity changes in the Fox family, including FHL binding and bispecificity, do not correlate with changes in canonical specificity-determining positions. Indeed, although H. sapiens FoxN4 is highly specific for only the FHL motif, and H. sapiens FoxN2 is bispecific and robustly binds FkhP and FkhS sites as well as the FHL motif, these two FoxNs are identical throughout the entire recognition helix; thus, the inability of FoxN4 to recognize FkhP sites is not strictly a function of the canonical DNA-contacting residues in the recognition helix.

Discussion

The previously unappreciated diversity in DNA binding specificity of Fox domain TFs that we have discovered raises the question of how specificity has evolved in this family. We have presented evidence that major changes in specificity have occurred separately in three different Fox subfamily lineages. In fungal Fox3 proteins, two different alternate specificities (FHL-3 and FVH) have arisen, with alteration of the canonical recognition positions in the FVH-binding but not the FHL-3-binding proteins. In metazoan FoxM proteins, binding to the canonical FkhP and FkhS sites has been supplemented with binding to a very different site, the FHL-M motif, with the same proteins binding well to both motifs. In addition, in the holozoan FoxN subfamily some proteins (FoxN2/3) exhibit this kind of bispecificity for two very different motifs (FkhPS and FHL-N), while others (FoxN1/4) have completely lost the ability to bind the classic forkhead site (FkhPS) in favor of the FHL-N motif. Finally, a derived subfamily unique to vertebrates (FoxR) appears to have regained specificity for a variant of the canonical FkhP motif from a more recent, exclusively FHL-specific ancestor. Formally, it is possible that lineages containing only proteins that bind only the FkhPS sites are derived from a more promiscuously binding ancestor with loss of FHL binding; however, this model would require a much larger number of specificity changes than the model we put forth here. Moreover, each instance of specificity change inferred from phylogenetic analyses is corroborated by minor but consistent differences in the motifs that have arisen; for example, all FoxN proteins bind to a version of the FHL motif that is distinguishable from the very similar FHL motif of fungal Fox3 proteins by preferences at a flanking position.

Our strategy of combining phylogenetic inference with comprehensive assays of DNA binding specificity permits us to study the evolution of DNA binding specificity in more detail using information from these complementary approaches. The monophyly of clade I, for example, is supported both by a high-confidence node in the inferred phylogeny and by the observed uniformity of binding specificity within this group. In the absence of phylogenetic analyses, the observation of an alternate specificity (GAYGC) appearing three times in different Fox domain subfamilies would lead to a parsimonious hypothesis that one ancestral FHL-binding forkhead domain arose before the last common ancestor of metazoa and fungi and gave rise to fungal Fox3 and metazoan FoxM and N groups. However, this hypothesis is strongly refuted by ML phylogenetic inference, which instead suggests independent origins of all three groups of alternate-specificity proteins. Further support for this surprising model comes from the observation that fine differences in FHL specificity distinguish these three groups, as discussed above.

This model raises the question of how such similar alternate specificities could have arisen independently in three different forkhead superfamilies. In the group of Fox3 proteins from fungi related to A. nidulans, the alternate specificity in the FVH motif with concomitant loss of binding to FkhPS sequences might be due to the extensive changes observed in the recognition helix. However, the appearances of the FHL motif variants during forkhead evolution, whether along with FkhP binding in bispecific proteins or as a replacement, do not correlate with any changes at a.a. positions known to specify FkhP binding, and suggest an alternate mechanism for changes in DNA binding specificity.

We propose that the existence of bispecific proteins that bind both FkhPS and FHL sequences with high specificity points to a possible explanation—that some Fox domain proteins which bind strongly to the FkhP site can achieve an alternate conformation which supports recognition of the FHL motif. It is intriguing, in the context of this observation, that both M. musculus Fox3 and A. macrognous Fox3 show weak binding to a subset of FHL-containing 8-mers, and exhibit binding similarity to alternate protein interaction surfaces of a TF, thus creating a new regulatory role for the alternate binding motifs as allosteric effectors of interactions with cofactors (31, 32). Exploring the mechanisms of such regulatory consequences will require an approach combining structural studies of distinct TF-DNA complexes, such as those identified here, with in vitro analyses of binding site utilization and function. This newly discovered phenomenon of DNA binding bispecificity suggests a novel source of modularity and flexibility in the structure of TFs and transcriptional regulatory networks. Improved understanding of the evolution of TF binding specificity will provide insights into the evolution of transcriptional regulatory networks, which ultimately will shed light on the processes underlying the evolution of new body plans and environmental responses.

Materials and Methods

Forkhead sequences

The genome sequences and annotations used in this study are summarized in Table 5. For each described protein sequence, we performed a hidden Markov model (HMM) search using HHM (33) with the Forkhead domain (PF00250) in the Pfam database (E-value < 10−10) (29). Using the hit sequences as queries, we conducted iterative homology search using PSI-BLAST (E-value < 10−10) (34). We then constructed a HMM from each sequence alignment, and searched against all protein sequences again. All obtained genes are described with their identification method in Table S1. All sequences used for the phylogenetic analysis contain five alpha-helices and three beta-sheets as in human Fox2 (11).

For phylogenetic analyses, each a.a. sequence of Fox domains was aligned using five multiple sequence alignment programs: a) LINS-i program (78), b) T-Coffee (79), c) Clustal Omega (80), d) muscle (81), and e) Clustal W (39). The accuracies of multiple sequence alignments were evaluated by FastSP (40), and the MAFFT alignment was selected by the number of homologous a.a. sites.

Phylogenetic inference

The non-opisthokont models of LG (41) with gamma-distributed rate variation (α = 0.881) were selected for whole forkhead domains, using the Akaike information criterion implemented in PROTEKT 3 (42). Phylogenetic trees were constructed using the maximum-likelihood method in PhyML 3.0 (43) with robustness evaluated by bootstrapping (100 times) (44) and by approximate likelihood-ratio test (aLRT) (45, 46). The starting tree for branch swapping was obtained using a ML tree constructed by RAxML (47). For likelihood ratio tests, two ML trees were constructed from the ML tree in Figure 2, changing the branching pattern of Fox3 and FoxM (Figure S3A and S2B, respectively). RAxML was applied to optimize the lengths of branches.
Comparing the ML score obtained from the tree in Figure 2 (-13406.2), p-values were calculated based on the chi-square distribution with one degree of freedom. 

Cloning and protein expression 

The DBDs of the forkhead proteins, flanked by attB recombination sites, were constructed by gene synthesis and cloned into the pUC57 vector (GenScript USA, Inc.). Constructs were transferred to the pDEST15 vector, which provides an N-terminal glutathione S-transferase (GST) tag, using the Gateway recombinational cloning system (Invitrogen). All cloned forkhead domain sequences are provided in Table S4. Proteins were expressed by in vitro transcription and translation (IVT) using the PUREXpress in vitro Protein Synthesis kit (New England Biolabs, Inc.). Concentrations of the expressed GST-fusion proteins were determined by Western blots in comparison to a dilution series of recombinant GST (sigma).

PBM experiments and analysis 

Double-stranding of oligonucleotide arrays and PBM experiments were performed essentially as described previously, except where noted in Table S4, using custom-designs "all 10-mer" arrays in the 4×4K (Agilent Technologies, Inc.; AMD ID #030236) array format (28, 48). Microarray data quantification, normalization, and motif derivation were performed as described previously (28, 48); some published PBM data (21) were reanalyzed for this study. DNA binding site motif sequence logos were generated using enoLOGOS (49). 8-mer E-score data were collected for any contiguous 8-mer bound (E-score ≥ 0.35) by at least one assayed Fox protein and clustered using the heatmap2 function in the ggplot2 package with the Manhattan distance metric. 


The authors declare no conflicts of interest.

Acknowledgements.

We thank S. Brown, S. Brown, and J. Ruiz-Trillo for sharing pre-publication forkhead sequences from F. alba and A. castellanii, Anastasia Vedenko and Leila Sokhi for technical assistance, and Anton Aboukhalil, Shamir S.R. (2011) for readdressed PBM data for helpful discussion. This study was supported by a Research Fellowship for Young Scientists from the Japan Society for the Promotion of Science to S.N., and by National Institutes of Health grant # R01 HG003955 to M.L.B. J.M.R. was supported in part by the Molecular Biophysics Training Grant # T32 GM080313 from the National Institutes of Health. This article contains Supporting Information online.