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\textsuperscript{a,b,1,2}, Stephen S. Gisselbrecht\textsuperscript{a,1}, Julia M. Rogers\textsuperscript{c,d,3}, Daniel L. Hart\textsuperscript{a,3} and Martha L. Bulyk\textsuperscript{c,d,e,3}

\textsuperscript{a}Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, MA 02138, USA
\textsuperscript{b}Center for Information Biology, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan
\textsuperscript{c}Division of Genetics, Department of Medicine, \textsuperscript{d}Department of Pathology, Brigham and Women’s Hospital and Harvard Medical School, Boston, MA 02115, USA.
\textsuperscript{e}Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, MA 02132, USA.
\textsuperscript{1}S.N., S.S.G., and J.M.R. contributed equally to this work.
\textsuperscript{2}Present address: Department of Molecular Life Science, Tokai University School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1193, Japan
\textsuperscript{3}To whom correspondence should be addressed.

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 alternations 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 a variety of organ systems, metabolic homeostasis, and regulation of cell cycle progression, while fungal Fox proteins are involved in cell cycle progression and the expression of ribosomal proteins. The Fox family of TFs shares a conserved DBD that is structurally identifiable as a subgroup of the much larger winged helix superfamily, which includes both sequence-specific DNA-binding proteins and linker histones, which appear to bind DNA nonspecifically (4, 5). Proteins with unambiguous sequence homology to the forkhead domain are present throughout opisthokonts—the phylogenetic grouping which includes all descendants of the last common ancestor of animals and fungi—but have diverged so extensively over approximately one billion years of evolution that distantly related Fox proteins are not generallyalignable outside of the forkhead domain (6, 7). Moreover, distantly related Fox-like domains have been found in Amoebozoa, a sister group to opisthokonts (8). Three distinct subfamilies (Fox1 through Fox3) of fungal Fox proteins have been identified. Metazoan Fox proteins are classified into 19 subfamilies (FoxA through FoxS), some of which have been further subdivided on phylogenetic grounds. The Fox domain itself 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 RYAAAYA (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

Reserved for Publication Footnotes
mouse identified an entirely different sequence, ACGC, as its preferred binding site (23). The closely related *Mus musculus* FoxN4 has been shown to bind ACGC in vivo (24). A PBM survey of *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 conformations in various context (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 – RYAAAAYA 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 metazoa, 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 *Spizellomyces punctatus* and *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 in vitro using universal PBM technology, in which a DNA-binding protein is applied to a double-stranded DNA microarray containing 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 obtained 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 FoxN1/4 and fungal Fox3 suggests the parsimonious hypothesis that they derive from a common FHL-binding ancestral protein in the last common ancestor of metazoans and fungi (23; 19). Since the FHL motif (GACGC) has been shown to be preferentially recognized by Fox3 in a variety of species (24; 19), we therefore refer to the GACGC site as the FHL motif.

Fig. 1. DNA binding site motifs bound by forkhead domain proteins. A representative member of each class of binding site discussed in the text is shown. Bold symbols are used to represent binding specificities in subsequent figures.

Fig. 2. Maximum likelihood phylogenetic tree of forkhead domains. This compact tree was constructed for presentation purposes from a representative subset of phylogenetically informative species: metazoa mouse, fly and sponge; choanoflagellate *Salpingoeca rosetta* and *Monosiga brevicollis*; *Capsaspora owczarzaki* and *Sphaeroforma arctica* from Ichthyosporea, *Saccharomyces cerevisiae*; *Allomyces macrogynus* from Blastocladiomycota; *Spizellomyces punctatus* from Chytridiomycota; *Maritella verticillata* from Mortierellomycota; *Fonticula alba* from Magnaporthe; and *Capsaspora owczarzaki* from Ichthyosporea (23). Nodes supported with strong likelihood ratios are indicated with red circles (aLRT ≥ 95%) or blue circles (aLRT ≥ 99%). Bootstrap support values are shown for nodes with <80% support. Clades containing alternate binding specificities are highlighted in color (see text). Importantly, the groupings of subfamilies in this tree and the complete tree with all Fox domains are almost identical to each other (see Figure S2).

Results

The published observation of roughly the same alternate binding motif (FHL) for metazoan FoxN1/4 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 (Fox4) 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 widely and deeply diverged family, and regions outside the do- main 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 different motifs as defined in Figure 1.

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+ clade (Figure S3A) is significantly less likely than the observed tree (p < 10^-5, 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.
of a fundamental split into FoxN1/4 and N2/3 clades, with FoxR
(subfamily). This cluster is further divided into holozoan
family. 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
preferences (Figure 1). As in the case of FoxN and Fox3, ML
preference (Figure S3E,F).

Biclustering of the 30 total Fox proteins and bound 8-mers
according to PBM enrichment (E) scores reveals three major
functional protein classes (Figure 4). The first prominent cluster
proteins characterized by specificity only for the FkhP
and FkhS motifs. Binding to these motifs tracks together across
proteins; the motif constructed from these 8-mers is an average
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 FHL-
M. 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
(Figure 4 and Figure S4), in that each subfamily
bind as dimers to GATGCATC. Phylogenetic analysis strongly
supports an independent origin of the FoxM subfamily from FoxN
(Figure 4 and Figure S4), in that each subfamily
bind as dimers to GATGCATC. Phylogenetic analysis strongly
supports an independent origin of the FoxM subfamily from FoxN
(Figure 4 and Figure S4), in that each subfamily
bind as dimers to GATGCATC. Phylogenetic analysis strongly
supports an independent origin of the FoxM subfamily from FoxN
(Figure 4 and Figure S4), in that each subfamily
bind as dimers to GATGCATC. Phylogenetic analysis strongly
supports an independent origin of the FoxM subfamily from FoxN
(Figure 4 and Figure S4), in that each subfamily
bind as dimers to GATGCATC. Phylogenetic analysis strongly
supports an independent origin of the FoxM subfamily from FoxN
(Figure 4 and Figure S4), in that each subfamily
bind as dimers to GATGCATC. Phylogenetic analysis strongly
supports an independent origin of the FoxM subfamily from FoxN
(Figure 4 and Figure S4), in that each subfamily
bind as dimers to GATGCATC. Phylogenetic analysis strongly
supports an independent origin of the FoxM subfamily from FoxN
(Figure 4 and Figure S4), in that each subfamily
bind as dimers to GATGCATC. Phylogenetic analysis strongly
supports an independent origin of the FoxM subfamily from FoxN
(Figure 4 and Figure S4), in that each subfamily
bind as dimers to GATGCATC. Phylogenetic analysis strongly
supports an independent origin of the FoxM subfamily from FoxN
(Figure 4 and Figure S4), in that each subfamily
One of the forkhead-like domains from the non-opisthokont
Acanthamoeba castellanii did not fall into any of these three
classes, as it binds another distinct motif (see Figure 1, Figure S4). The
dag differences are associated with widespread
differences in the recognition helix (Figure S5A). Indeed, altered
recognition positions (Figure S5B) can clearly explain the non-
FkhPS specificities of the forkhead-related protein from A. castel-
nani 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 bispeci-
ficity, 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
bspecific 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 speci-
ficity of Fox domain TFs that we have discovered raises the
question of how specificity has evolved in this family. We have pre-
sented 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 pro-
teins. In metazoan FoxM proteins, binding to the canonical FkhF
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 sequences are derived from a more promiscuously bind-
ing 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 com-
prehensive 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 mon-
ophyly 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 ab-

ence of phylogenetic analyses, the observation of an alternate
specificity (GAYGC) appearing three times in different Fox do-
main 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 lineages. In the group of Fox3 proteins from fungi
related to A. nidulans, the alteration in specificity to 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
direct 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 conforma-
tion which supports recognition of the FHL motif. It is intriguing,
in the context of this observation, that both M. musculus Fox31
and A. macrognus Fox3 show weak binding to a subset of FHL-
containing 8-mers, and exhibit binding similarity to alternate
proteins interaction surfaces of a TF, thus creating a new
regulatory role for the alternate binding motifs as allosteric effec-
tors of interactions with cofactors (31, 32). Exploring the mecha-
nisms of such regulatory consequences will require an approach
combining structural studies of distinct TF-DNA complexes, such
as those identified here, with in vivo 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 reg-
ulatory 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 summa-
rized in Table S1. For each nonredundant protein sequence, we performed a
hidden Markov model (HMM) search using HMMER3 (33) with the Fork
head 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
multiple alignment of forkhead sequences, 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, 16).

For phylogenetic analyses, each a.a. sequence of Fox domains was
aligned using five multiple sequence alignment programs: a) L-INS-i program
of T-Coffee (36), b) Clustal Omega (38), and c) 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 a.a. replacement models of LG (41) with gamma-distributed rate
variation (o = 0.881) were selected for whole forkhead domains, using the
Akaike information criterion implemented in PROTEST 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 (a.LRT) (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
S1B, respectively). RAxML was applied to optimize the lengths of branches

Footline Author
Cloning and protein expression

The DBDs of the forhead proteins, flanked by at&b recombination sites, were constructed by gene synthesis and cloned into the puC57 vector (GenScript USA, Inc.). Constructs were transfected to the pDEST15 vector, which provides an N-terminal glutathione S-transferase (GST) tag, using the Gateway recombining cloning system (Invitrogen). All cloned forhead 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-designed “all 10-mer” arrays in the 4x4K (Agilent Technologies, Inc.; AMADID #015681 or 8x60K (Agilent Technologies, Inc.; AMADID #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 heatmap.2 function in the gplots R package with the Manhattan distance metric.


The authors declare no conflicts of interest.

Acknowledgements.