

Evolution of Sp Transcription Factors

Kevin J. Kolell and Douglas L. Crawford

Division of Molecular Biology and Biochemistry, School of Biological Science, University of Missouri—Kansas City

The Sp family of transcription factors binds GC-rich DNA sequences. The ubiquitously expressed Sp1 and Sp3 have been well characterized in mammals. Presented here is the characterization of the only Sp protein expressed in the liver or heart tissue of the teleost fish *Fundulus heteroclitus*. This protein, fSp3, is most similar to and homologous with mammalian Sp3 proteins. The evolution of the Sp transcription family is described, with Sp1 and Sp3 representing the most recent duplication of the Sp family. Sp4 appears to be the most ancestral member. Sp1, Sp3, and Sp4 form a monophyletic group without Sp2. Sp2 is the least similar of the Sp family and is more similar to the non-Sp transcription factors. These results suggest that Sp2 should not be considered a member of the Sp family. Only two domains (zinc fingers and B domain) share similarity outside the Sp family. The zinc fingers are homologous to other GC-binding domains, yet the B domain is homologous to protein-protein interacting domains in the CCAAT-binding/NF-Y transcription factor families. These results suggest that these different domains have different evolutionary histories.

Introduction

Transcription is a concerted effort between binding of transcription factors to specific DNA sequences, general transcription factors, and RNA polymerase II, forming a complex resulting in the initiation of transcription (Tjian and Maniatis 1994). Proximal promoters (first few hundred base pairs [bp], 5' to the start of transcription) typically have a TATA sequence at -30 bp, relative to the start of transcription. These TATA sequences bind TFIID, an essential factor for all mRNA transcription (Roeder 1991). Nonetheless, many promoters lack a TATA sequence (e.g., approximately 50% of the transcribed genes in *Drosophila* are without a TATA sequence [Arkhipova 1995]). These TATA-less promoters require Sp-binding sites (consensus sequence: GGGCGG [Dynam and Tjian 1983]) for significant activity (Azizkhan et al. 1993), and the degree of activation from Sp1 tends to be stronger in the context of the TATA-less promoters than the TATA-containing promoters (Colgan and Manley 1995). Activation of the TATA-less promoters by Sp1 most likely involves Sp1 recruitment of TFIID to TATA-less promoters (Pugh and Tjian 1991; Wiley, Kraus, and Mertz 1992). Furthermore, in many of these promoters, Sp1 binding is intimately involved in the determination of the transcription start site or sites (Joliff, Li, and Johnson 1991; Kollmar et al. 1994; Lu et al. 1994; Boam, Davidson, and Chambon 1995).

The Sp family has four members in humans: Sp1, Sp2, Sp3, and Sp4 (Suske 1999). Of these, Sp1 was identified first and is the most well characterized (Dynam and Tjian 1983). Sp1 and Sp3 are ubiquitous transcription factors that have been implicated in the control of a wide variety of genes (Hagen et al. 1994). Sp1 and Sp3 are thought to compete for similar binding sites, and the relative rate of transcription is affected by the outcome of this competition (Hagen et al. 1992; Hata et al. 1998).

Key words: modular evolution, protein polymorphism, clinal variation.

Address for correspondence and reprints: Douglas L. Crawford, Division of Molecular Biology and Biochemistry, School of Biological Science, 5007 Rockhill Road, University of Missouri—Kansas City, Kansas City, Missouri 64110. E-mail: crawforddo@umkc.edu.

Mol. Biol. Evol. 19(3):216–222, 2002

© 2002 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

Sp2 is least similar to the other Sp family members, with a 20%–27% protein sequence similarity, and very little is known about its function (Kingsley and Winoto 1992). Lastly, Sp4 has tissue-specific expression restricted to the brain and nervous tissue (Hagen et al. 1992).

The structure of the polypeptides Sp1, Sp3, and Sp4 consists of five domains (fig. 1). In Sp1, domains A, B, C, and D form both homotypic and heterotypic protein-protein interactions. Higher-order structures or multiple tetramers occur through protein-protein interactions in the A, B, or D domains (Mastrangelo et al. 1991). These higher-order structures result in super activation in which multiple-binding sites have much more transcriptional activity than the sum of single sites. Domains A and B are in the N-terminal half of the protein, which can be further divided into subdomains containing a serine and threonine-rich area followed by a glutamine-rich area. Domain C contains a highly charged region, which is responsible for the inhibition of transcription in Sp3 (Dennig, Beato, and Suske 1996). The DNA-binding domain contains three zinc fingers that recognize DNA-binding sites GGGCGG or CACCC with similar binding affinities (Hagen et al. 1992, 1994). Each zinc finger motif is composed of an alpha helix and a beta sheet structure that tetrahedrally binds a zinc atom. Finally, domain D is located at the C-terminus beyond the zinc fingers and forms protein-protein interactions with other transcription factors (Ding et al. 1999).

Presented here is a nonmammalian Sp homologue, fSp3 (teleost fish, *Fundulus heteroclitus* Sp gene that will be shown to be homologous to the Sp3 subfamily). The evolutionary relationships between this protein, fSp3, the other mammalian Sp proteins, zinc finger proteins, and nonzinc finger proteins are provided. Different domains of the Sp proteins are homologous to different families of transcription factors, indicative of modular evolution.

Materials and Methods

Population Cloning by Reverse Transcription PCR

The original partial Sp cDNA from *Fundulus* was isolated using degenerate primers to amplify Sp genes from liver cDNA. Degenerate primers were made on the basis of an alignment of human, mouse, and rat Sp pro-

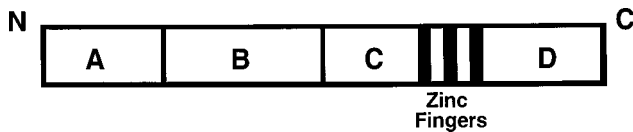


FIG. 1.—Sp protein with labeled domains. Sp proteins have five domains: A, B, C, zinc fingers, and D forming a protein that is approximately 785 amino acids long. The A, B, C, and D, domains form protein-protein interactions. The zinc fingers recognize and bind GC-rich DNA sequences.

teins (Sp1, Sp2, Sp3, and Sp4). The 5' end primers were: Sp191f = CATAYCARGTNATHCC and Sp231f = GGNGCNAAYCARCARAT. The 3' end primers were: sp744R = AANCKYTTNCCRCARTA and sp761R = TTYTCNCCNGTRTGNGT. Numbers refer to the targeted N-terminal amino acids. Gene-specific primers along with vector-specific primers were used to amplify other regions of Sp gene from a killifish *F. heteroclitus* cDNA library (Karchner, Powell, and Hahn 1999). To isolate most of the *F. heteroclitus* Sp gene cDNA, poly-A RNA was made from *F. heteroclitus* liver tissue, reverse transcribed using MMLV reverse transcriptase (Promega), and amplified with gene-specific primers. PCR products were cloned into pGEM-T easy vector (Promega).

Alignment and Phylogenetic Construction

Fundulus heteroclitus Sp nucleotide and protein sequences were used to search the GenBank database at the National Center Biological Information. Sequences matching more than one domain (fig. 1) among Sp family proteins were aligned using a Clustal alignment program (MacVector). The zinc finger regions of non-Sp genes with significant similarity ($P < 10^{-14}$) were aligned with Sp zinc finger sequences. For the B domains, non-Sp proteins with significant similarity ($P < 10^{-4}$) were aligned with Sp B domains. Aligned sequences were subjected to a heuristic search for the best possible trees using the maximum parsimony program in PAUP (Swofford 2000). Bootstrap (1,000 replicates) values were obtained from the best tree.

Proteins Used in Analyses

Sp Family Phylogeny

Proteins used for the Sp phylogenetic analysis included: human Sp3 (Q02447), mouse Sp3 (AF062567), human Sp1 (AF252284), mouse Sp1 (S79832), rat Sp1 (S25287), pig Sp1 (U57347), human Sp4 (NP_003103), mouse Sp4 (NP_0033265), and human Sp2 (Q02086).

B Domain Phylogeny

Proteins used for the phylogenetic analysis of the B domain included the Sp proteins (listed above) human OCT-1 (A47001), nuclear transcription factor Y (AAH05003), POU-domain class 2 transcription factor (NP_035267), histone H1 transcription factor (A56365), transcription factor NFY-C (AF191744), and CCAAT-binding transcription factor (I59348).

Zinc Finger Phylogeny

Zinc fingers used for the phylogenetic analysis included the Sp proteins, *Drosophila* buttonhead (Q24266), human BTEB1 transcription factor (Q13886), rat BTEB1 (Q01713), human BTEB2 (Q13887), human erythroid KLF1 (Q13351), mouse EKLF1 (P46099), human EKLF4 (O43474), mouse EKLF4 (Q60793), human ubiquitous KLF7 (O75840), human WILMS' Tumor protein (P19544), alligator WT1 (P50902), rat WT1 (P49952), mouse WT1 (P22561), human ZNF74 (Q16587), human ZF9 (Q99612), mouse ZF9 (O35819), mouse AP-2 rep transcription factor (Y14295), human TIEG1 (NP_005646), rat TIEG1 (U78875), human TIEG2 (NP_003588), *Drosophila* Sp (AAF46519), and mouse Sp5 (NP_071880).

Results

cDNA was made from *F. heteroclitus* liver mRNA. Degenerate primers for conserved amino acids among all members of the Sp family were made, used to amplify the liver cDNA, and these products were cloned. Six clones were sequenced. All six were most similar to Sp3: 48% amino acid identity to human Sp3 protein versus 32% to human Sp1. Thus, this gene was tentatively named fSp3 (*Fundulus* Sp3). The rest of the fSp3 was isolated using fSp3-specific primers with linker primers ligated onto the 5' end of cDNA or with vector-specific primers (used for heart and liver cDNA libraries). On the basis of the amino acid alignment of Sp proteins, it appears that none of the members of the Sp3 subfamily (human, mouse, and *Fundulus*) are of full length. None of the Sp3 cDNAs have the initiation codon methionine. All are approximately 50 amino acids shorter than Sp1 or Sp4. Additionally, all methods to isolate the 5' end of both the human and *F. heteroclitus* cDNAs resulted in a truncated product: the 5' end in an open reading frame, lacking an initiation codon (methionine), and lacking a 5' UTR. Genomic sequence reveals an intron >10 kbp, with no identifiable 5' exon. These results are similar to the attempts made at other laboratories to isolate the 5' end of Sp3 cDNA (Kingsley and Winoto 1992; Kennett, Udvardia, and Horowitz 1997).

Only a single Sp, Sp3, was found in the *Fundulus* liver and cardiac tissue. All six PCR products using primers to conserved regions among Sp1, Sp2, Sp3, and Sp4 were fSp3. The sequence variations among these six PCR products are within the range of sequence variations found among 14 different *Fundulus* Sp alleles (unpublished data). Amplification with Sp1-specific primers did not result in any product. No other Sp cDNAs were isolated among greater than 6,000 cardiac cDNAs from *F. heteroclitus* (Oleksiak, Kolell, and Crawford 2001). Northern analysis on the liver and cardiac mRNA revealed a single product with an Sp3 probe. Southern analysis with human Sp1 probe did not hybridize to *Fundulus* genomic DNA. In Western analyses and gel shift analyses, Sp3-specific, but not Sp1-specific, antibodies cross-reacted with proteins in *Fun-*

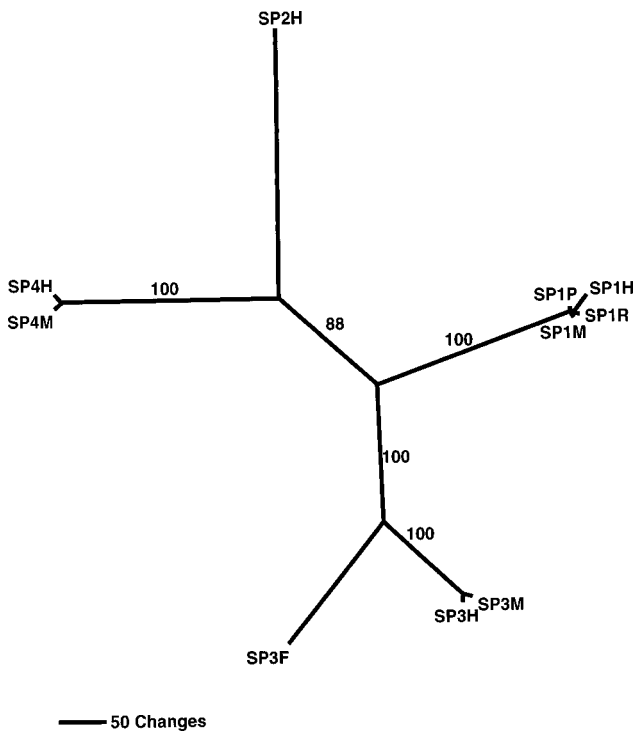


FIG. 2.—Unrooted phylogeny among Sp family. The entire protein sequences of Sp proteins were aligned and subjected to maximum parsimony analysis.

dulus nuclear extract (data not shown). Thus, unlike mammals, there is only one ubiquitously expressed Sp.

The fSp3 protein sequence was aligned with Sp1, Sp2, Sp3, and Sp4 protein sequences from human, mouse, and rat. These are the only genes in the GenBank that are similar to more than one Sp domain. The Sp proteins were subjected to maximum parsimony analysis (fig. 2). Each of the three of the Sp subtypes (Sp1, Sp3, Sp4) is supported by 100% bootstrap values (1,000 replicates). The *F. heteroclitus* Sp protein groups with the mammalian Sp3 proteins, supporting the hypothesis that this transcription factor is an Sp3 gene.

The evolutionary relationships of the protein domains (fig. 1) within the Sp family were examined. These domains were identified in human Sp3 and Sp1 (Pascal and Tjian 1991; Gill et al. 1994) and used to define regions in fSp3. The fSp3 domain A consists of amino acids 1–202, domain B of amino acids 203–387, domain C of amino acids 388–479, the zinc finger of amino acids 480–574, and domain D of amino acids 575–624. All five fSp3 domains were used to search the GenBank database. Only the zinc finger and domain B had significant similarities to non-Sp proteins in the database.

The three contiguous fSp3 zinc fingers were used in a BLAST similarity search of the GenBank database. Proteins with significant similarity ($P < 10^{-14}$) to the zinc finger were aligned using CLUSTAL. Maximum parsimony analyses (PAUP, fig. 3) was applied to these zinc finger proteins using the invertebrate *Drosophila* Sp buttonhead protein as the outgroup. There were 208 minimum trees. The consensus tree (fig. 3) of these 208

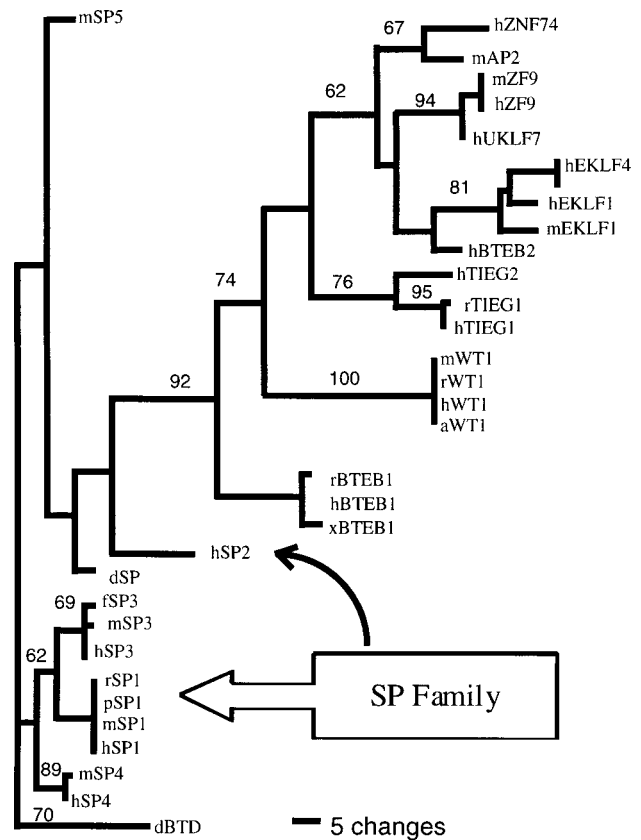


FIG. 3.—Zinc finger phylogeny. Evolutionary relationship among zinc finger proteins (consensus of 208 shortest trees). Overall tree is significant ($P < 0.01$, PTP). All proteins in GenBank with significant similarity to fSp3 zinc fingers were used in maximum parsimony analyses. Numbers above branches are bootstrap values $>50\%$ for 1,000 replicates. A smaller case letter before the gene name denotes the organism: a, alligator; d, *Drosophila*; f, *Fundulus heteroclitus*; h, human; m, mouse; r, rat; x, *Xenopus laevis*.

trees has a significant phylogenetic signal ($P < 0.001$, permutation tail probability [PTP] test [Faith and Cranston 1991]). That is, maximum parsimony analysis was applied to each if the 1,000 random permutations of the zinc finger proteins, and these analyses of the randomized data sets never produced as short a tree (280 characters). The shortest random permuted tree had 615 characters. The zinc fingers from *F. heteroclitus* groups with the mammalian Sp3 proteins, supporting the hypothesis that this transcription factor is an Sp3 gene. Similarly, Sp1 and Sp4 each form separate clades within the Sp zinc finger family. Among zinc finger regions, Sp1, Sp3, and Sp4 form a monophyletic group (figs. 3 and 4). This group never includes Sp2 among the 208 minimum trees. Sp2 shares only 8 of the 16 unique, derived amino acids (fig. 4). Although the bootstrap value for monophyletic Sp1, Sp3, and Sp4 is not $>50\%$, fewer than 10% of bootstrap trees included Sp2 with the other Sp proteins. The T-PTP test (topology-dependent permutation tail probability [Faith 1991]), with the constraints that Sp1, Sp3, and Sp4 are monophyletic and Sp2 is monophyletic with the other zinc fingers, is significant ($P < 0.01$, T-PTP). Finally, the trees that are constrained so that Sp2 groups with the other Sp pro-

	7	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85
ALL:	xxCxxxgCxxkxyxKxShLxxHxRxHs	GE	r _t f _r	x _w Cx _e	xxCxx _r	FxxRSD _e	LxxRHxRxHTG _x	xxfxCxxC _r	x _m F _s	DHLxxHx	x _r t _r H					
SpS	I	G T	R	W	R	I	MF	T	Q			E SK				
Sp4	E				T	I	MF			R	R E P S	M			S	V
Sp3	P				S	x	Mx			R	K V x S	M			A	I
Sp1	Q				T	M	SY			K	K A P P	x			S	I
Sp2	P	R			<u>L</u>	<u>T</u>	<u>V</u>	<u>FF</u>		<u>A</u>	K E <u>AQ</u> <u>Q</u>	M			<u>T</u>	Y

FIG. 4.—Conserved amino acids in the three zinc fingers. For all 33 proteins, identical or nearly identical amino acids (lower case letters, shared by 31 of the 33 zinc finger proteins) are displayed next to ALL. Sites that have only two conserved amino acids are shown as a ratio, and bold letters are amino acids that are proposed to interact with DNA in Sp1 (Narayan, Kriwacki, and Caradonna 1997). SpS lists amino acids that are unique, derived amino acids for Sp1, Sp3, and Sp4 versus the other major clades (i.e., ignoring dSp, Buttonhead, and Sp2). Unique, derived amino acids within each subfamily of Sp proteins are listed next to each group. The amino acids for Sp2 that are in a variable region are listed next to Sp2. Derived Sp amino acids that are not shared with Sp2 are underlined.

teins are four steps longer than the tree with Sp2 excluded. This analysis supports the observation (fig. 3) that Sp2 groups are outside the Sp family of transcription factors. Sp2 and the newly identified Sp5 (Harrison et al. 2000) have considerably less sequence similarity with other Sp family members (i.e., 27% and 20%, respectively). The zinc finger data suggest that Sp2 and Sp5 proteins represent the most divergent proteins of the Sp family and may not be members of the Sp family.

The zinc finger domain of *Fundulus* was 99% identical to human Sp3 with only one amino acid change, a fish serine to mammalian proline. The amino acids for zinc fingers in the Sp family, and all significantly similar proteins, are depicted in figure 4. Notice, these amino acids are for the three contiguous zinc fingers found among all the 33 proteins (fig. 3) and are similar to the conserved amino acids found within and between the

zinc finger proteins: (FYH)xC_{x2-4}C_{x3}F_{x5}L_{x2}H_{x3}H_{x5} (where bold letters are metal-binding residues [Narayan, Kriwacki, and Caradonna 1997]). Among the triplicate zinc finger proteins used in this analysis (ALL, fig. 4), 32 of 87 amino acids are invariant, and an additional 18 positions possess only one of the two amino acids. Thus, 50 of 87 amino acids are invariant or virtually invariant. These proteins not only have conserved amino acids but also share function by binding to GC-rich sequences.

The phylogenetic analysis of the B domain included four nonzinc finger proteins (fig. 5). POU domain transcription factor was used as an outgroup because it had the lowest similarity score to fSp3. This analysis produced three minimum trees. The strict consensus had two major clades: one for the Sp family and the other for Sp2 and CCAAT-binding/NF-Y transcription factor families. This tree is well supported (*P* < 0.01, PTP), and the bootstrap values for these two clades are >75%. Sp1, Sp3, and Sp4 form a monophyletic group that does not include Sp2. Forced topology in which Sp2 is grouped with the other SpS is six steps longer than the topology of SpS without Sp2.

Within the Sp family, the phylogeny of domains A and C are topologically similar to the zinc finger domain and whole-protein phylogeny (fig. 6). These data indicate that Sp1 and Sp3 are more closely related and suggest that they arose by duplication from Sp4 (figs. 3 and 6).

Unlike whole protein or the other three domains, for both the B and D domains (figs. 5 and 6), there is

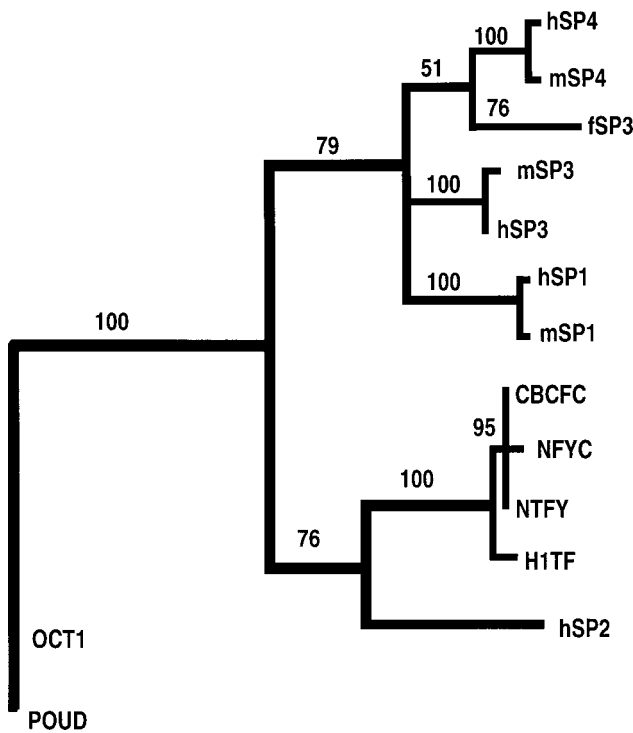


FIG. 5.—Evolutionary relationship for B domain. The B domain (aa203–387) from Sp proteins and proteins with significant similarity were aligned with Clustal and analyzed using maximum parsimony. Bootstrap values >50% from 1,000 replicates are listed above branches. Protein sequences represent three major families: (1) CCAAT-binding/nuclear transcription factor Y family, (2) POU domain/octamer 1 transcription factor family, and (3) Sp transcription factor family.

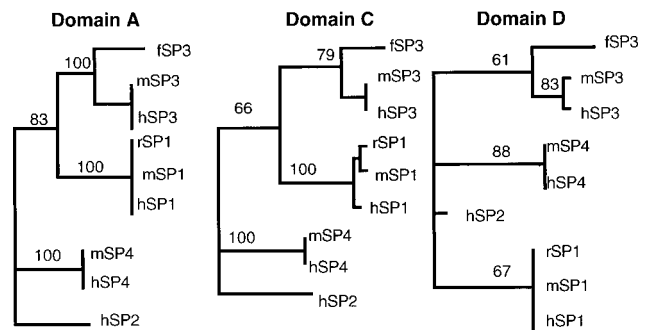


FIG. 6.—Evolutionary relationships for different domains within the Sp family. The three domains (A, C, and D) are listed above the trees. Domain A is composed of amino acids 1–202, domain C is composed of amino acids 388–479, and domain D is composed of amino acids 575–624. Maximum parsimony analyses were done using PAUP (Version 4.0b8). Numbers above branches are bootstrap values >50% for 1,000 replicates.

little resolution among members of the Sp subfamilies. This is not because of a lack of derived characters: the D domain tree has 20 informative characters and the B domain has 75. However, the difference in topology between the B and D domains versus the other domains is not supported by partition homogeneity test ($P > 25$; 1,000 replicate, PAUP). Partition homogeneity is based on the observed incongruent tree length difference resulting from combining the data matrices. Thus, because the B and D domains lack resolution among Sp subfamilies, it is not surprising that these two domains do not provide significant heterogeneity. Finally, for the B domain, unlike all other domains, fSp3 shares more derived characters with the mammalian Sp4.

Discussion

Fundulus Sp3

Four members of the Sp transcription family have been characterized in mammals (Sp1, Sp2, Sp3, and Sp4). The *Fundulus* Sp protein (fSp3) is most similar to, and shares the most derived characters with, other Sp3 proteins (figs. 2–4 and 6). The phylogenetic relationship between Sp1 and Sp3 suggests that they arose by duplication from Sp4 (figs. 3 and 6). In mammals, Sp1 and Sp3 are ubiquitously expressed and can have antagonistic effects (Hata et al. 1998; Suske 1999). Thus, one might expect to find several Sp subtypes in the teleost. However, *Fundulus* Sp3 was the only Sp gene expressed in the liver or cardiac tissue (see *Results*). This could be because of a more restrictive pattern of expression for the other Sp genes (Sp1 and Sp4). Sp4 has been shown to be restricted to the brain and nervous tissue in mammals. However, Sp1 is thought to be ubiquitous. For the teleost fish, Sp1 may have a strictly developmental role or may be tissue restricted. Thus, Sp3 may serve as the only ubiquitous Sp factor in teleosts.

In the B domain, unlike all other domains, fSp3 does not group with the mammalian Sp3s (fig. 5). Instead, fSp3 clades with the more ancestral Sp4. The grouping of fSp3 with Sp4 may reflect a difference between mammals and teleosts in the pattern of Sp expression. The B domain interacts with the essential transcription factor TFIID (TATA-binding factor) via hTAF130. Among TATA-less promoters, Sp-binding sites (consensus sequence: GGGCGG; [Dynam and Tjian 1983]) are required for significant activity (Azizkhan et al. 1993), and the degree of activation from Sp tends to be stronger in the context of TATA-less promoters than TATA-containing promoters (Colgan and Manley 1995). In mammals, both Sp1 and Sp3 are most often expressed together and are often antagonistic (Hagen et al. 1992; Hata et al. 1998). Thus, among mammals, because both Sp1 and Sp3 are affecting transcription via the interaction of the B domain with TFIID, selection may favor divergence among B domains. However, in *Fundulus* there is only one Sp, fSp3, which is expressed ubiquitously. This expression pattern could cause selection to maintain a more general or ancestral function, and thus the fSp3 B domain would lack the derived characters

found in the rest of the protein. This idea is supported by the lack of polymorphism in the B domains among *Fundulus* populations (unpublished data). If the difference in the phylogeny of the B domains versus the other domains is in response to the pattern of expression for Sp1, it suggests that this domain plays an important role in defining the different roles of Sp3 and Sp1.

Sp Family

Among Sp proteins, only Sp1, Sp3, and Sp4 have all five SP domains (fig. 1). Sp2 has four of the five domains. Other proteins with Sp nomenclature (e.g., dSp, mSp5) have only one of these domains—the three consecutive zinc finger domains. The functions of Sp proteins are to bind GC-rich DNA and interact with TAFs as well as other proteins. The four protein–protein interactive domains as well as the zinc fingers define these functions. If one defines Sp proteins by their functions, then the other proteins which lack all the five domains will not be functionally similar. These functional relationships are a reflection of their evolutionary history.

In the evolutionary analyses in which other transcription factors can be aligned with Sp proteins (figs. 3 and 5), Sp2 does not group with other Sp transcription factors. Using the DNA-binding zinc fingers (Nardelli et al. 1991; Pavletich and Pabo 1991), Sp2 does not form a monophyletic group with other Sp proteins (fig. 3). In the B domain, the Sp2 groups with CCAAT/NF-Y transcription factors, not with the other Sp proteins (fig. 5). For Sp1, Sp3, and Sp4 proteins, the B domains form a monophyletic group. These data suggest that Sp2 is not a member of the Sp family.

The zinc finger proteins used in this analysis (fig. 3) have three consecutive zinc fingers ($Cx_{2-4}Cx_3Fx_5Lx_2Hx_3Hx_5$, where bold letters are metal-binding residues [Narayan, Kriwacki, and Caradonna 1997]). Most of the amino acids (50 of 87) are invariant or have only one alternative amino acid (ALL, fig. 4). Of the remaining 37 amino acids, 16 derived amino acids distinguish all Sp1, Sp3, and Sp4 proteins from the other major clades of similar zinc finger proteins (Sps, fig. 4). Sp2 shares only 8 of the 16 derived amino acids. Thirteen amino acids vary among Sp1, Sp3, and Sp4 (fig. 4). The most interesting variable site is amino acid 81, one of the amino acids that interact with DNA (Narayan, Kriwacki, and Caradonna 1997). For this interactive site, all the Sps have a serine or alanine: Sp3s have an alanine, Sp4 and Sp1 have a serine. The threonine found in Sp2 is found in several other zinc fingers. It seems likely that these amino acids could affect the Sp binding to DNA.

These data (the presence of all the five domains, the evolutionary relationship of the whole protein or the Sp domains, and the derived amino acids in zinc fingers) all indicate that Sp1, Sp3, and Sp4 form a monophyletic group and thus describe members of the Sp family. The observation that other proteins have similar zinc fingers (e.g., dSp or mSp5) or B domains (e.g., CCAAT/NF-Y-

binding proteins) should not necessarily allow them to be classified as Sp proteins.

Modular Evolution

The Sp family is often grouped with other proteins that contain zinc fingers. The zinc finger tree shows (fig. 3) that Sp transcription factors can be grouped with a myriad of transcription factors that (1) contain three zinc fingers and (2) bind GC-rich templates. The zinc fingers share homology with several proteins, including Kruppel factors. Although these proteins share an ability to bind to GC-rich sequences, this does not adequately describe the evolution or homology of the Sp transcription factors. Instead, it may represent the evolution of that single domain. In contrast to the zinc fingers, the B domain suggests a different phylogeny: Sp, CCAAT/NF-Y-binding proteins, and POU/OCT1 transcription factors that bind to different promoter sequences. The significant similarity ($P < 10^{-4}$) of the B domains and the apparent homology among the B domains (fig. 5) for these proteins suggest that these proteins may be grouped based on protein-protein interactions much like grouping proteins based on the zinc finger DNA-binding region.

The diverse B domains (fig. 5) have a conserved function; they bind to the same TAFs (TFIID associated factors [Coustry et al. 1998; Wolstein et al. 2000]). Additionally this function is phylogenetically conserved. The B domain of hSp1 binds both the human TAF130 (Tanese et al. 1996; Rojo-Niersbach, Furukawa, and Tanese 1999) and its *Drosophila* homologue dTAF110 (Gill et al. 1994). This occurs even though *Drosophila* lacks an Sp protein (Pugh and Tjian 1990). Similarly, OCT1 and CCAAT-binding factors have been shown to bind both hTAF130 and dTAF110 (Coustry et al. 1998; Wolstein et al. 2000). The observation that the homologous B domain proteins interact with TFIID subunits in phylogenetically distant species appears to reflect a more ancestral state than the homology among Sp family members. This functional similarity among different families of transcription factors either causes the similarity in the B domain's amino acid sequence (convergent evolution) or these transcription factors have evolved from different domains or modules, so that the B domain enjoys a separate evolutionary history from the zinc fingers.

Modular evolution occurs when protein domains are created by gene duplication followed by domain shuffling to evolve novel transcription factors. Modular evolution is thought to contribute to the evolution of the basic helix-loop-helix transcription factor family (Morgenstern and Atchley 1999). For modular evolution to occur, the expectation would be that these individual domains would have different ancestry. This appears to be true for the Sp family: zinc fingers and the B domains are related to different factors (compare figs. 3 and 5). Consistent with a modular theory of evolution, the B domain is a different exon from the DNA-binding domain (Suske 1999). These data suggest that these domains had different evolutionary histories prior to the

creation of the Sp family and that the Sp family of transcription factor is a mosaic of different protein modules.

In summary, in *Fundulus*, the only member of the Sp family to be widely expressed is homologous to the Sp3 subtype. Among the Sp transcription factors, only Sp1, Sp3, and Sp4 form a monophyletic group. Yet the domains of these Sp proteins have different evolutionary histories: the zinc fingers to GC-binding proteins and B domain to TAF-binding proteins. These data suggest that for these three members, the ancestral form was formed by mosaic evolution: the assortment of different domains to form a new functional protein.

Acknowledgments

This research was benefited by Dr. J. Segal's initial investigation in the transcriptional regulation of genes in *Fundulus*. Many insights and the Sp-antibody work were provided by Dr. Marjorie F. Oleksiak. This research was supported by NSF IBN grant number 9986602.

LITERATURE CITED

- ARKHIPOVA, I. R. 1995. Promoter elements in *Drosophila melanogaster* revealed by sequence analysis. *Genetics* **139**:1359–1369.
- AZIZKHAN, J. C., D. E. JENSEN, A. J. PIERCE, and M. WADE. 1993. Transcription from TATA-less promoters: dihydrofolate reductase as a model. *Crit. Rev. Eukaryot. Gene Expr.* **3**:229–254.
- BOAM, D. S., I. DAVIDSON, and P. CHAMBON. 1995. A TATA-less promoter containing binding sites for ubiquitous transcription factors mediates cell type-specific regulation of the gene for transcription enhancer factor-1 (TEF-1). *J. Biol. Chem.* **270**:19487–19494.
- COLGAN, J., and J. L. MANLEY. 1995. Cooperation between core promoter elements influences transcriptional activity in vivo. *Proc. Natl. Acad. Sci. USA* **92**:1955–1959.
- COUSTRY, F., S. SINHA, S. N. MAITY, and B. CROMBRUGGHE. 1998. The two activation domains of the CCAAT-binding factor CBF interact with the dTAFII110 component of the *Drosophila* TFIID complex. *Biochem. J.* **331**:291–297.
- DENNIG, J., M. BEATO, and G. SUSKE. 1996. An inhibitor domain in Sp3 regulates its glutamine-rich activation domains. *EMBO J.* **15**:5659–5667.
- DING, H., A. M. BENOTMANE, G. SUSKE, D. COLLEN, and A. BELAYEW. 1999. Functional interactions between Sp1 or Sp3 and the helicase-like transcription factor mediate basal expression from the human plasminogen activator inhibitor-1 gene. *J. Biol. Chem.* **274**:19573–19580.
- DYNAN, W. S., and R. TJIAN. 1983. The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. *Cell* **35**:79–87.
- FAITH, D. P. 1991. Cladistic permutation test for monophyly and nonmonophyly. *Syst. Zool.* **40**:366–375.
- FAITH, D. P., and P. S. CRANSTON. 1991. Could a cladogram this short have arisen by chance alone? On the permutation tests for cladistic structure. *Cladistics* **7**:1–28.
- GILL, G., E. PASCAL, Z. H. TSENG, and R. TJIAN. 1994. A glutamine-rich hydrophobic patch in transcription factor Sp1 contacts the dTAFII110 component of the *Drosophila* TFIID complex and mediates transcriptional activation. *Proc. Natl. Acad. Sci. USA* **91**:192–196.
- HAGEN, G., S. MULLER, M. BEATO, and G. SUSKE. 1992. Cloning by recognition site screening of two novel GT box bind-

- ing proteins: a family of Sp1 related genes. *Nucleic Acids Res.* **20**:5519–5525.
- . 1994. Sp1-mediated transcriptional activation is repressed by Sp3. *EMBO J.* **13**:3843–3851.
- HARRISON, S., D. HOUZELSTEIN, S. DUNWOODIE, and R. BEDDINGTON. 2000. Sp5, a new member of the Sp1 family, is dynamically expressed during development and genetically interacts with Brachyury. *Dev. Biol.* **227**:358–372.
- HATA, Y., E. DUH, K. ZHANG, G. S. ROBINSON, and L. P. AIELLO. 1998. Transcription factors Sp1 and Sp3 alter vascular endothelial growth factor receptor expression through a novel recognition sequence. *J. Biol. Chem.* **273**:19294–19303.
- JOLLIFF, K., Y. LI, and L. F. JOHNSON. 1991. Multiple protein-DNA interactions in the TATAA-less mouse thymidylate synthase promoter. *Nucleic Acids Res.* **19**:2267–2274.
- KARCHNER, S. I., W. H. POWELL, and M. E. HAHN. 1999. Identification and functional characterization of two highly divergent aryl hydrocarbon receptors (AHR1 and AHR2) in the teleost *Fundulus heteroclitus*. Evidence for a novel subfamily of ligand-binding basic helix loop helix-Per-ARNT-Sim (bHLH-PAS) factors. *J. Biol. Chem.* **274**:33814–33824.
- KENNETT, S. B., A. J. UDVADIA, and J. M. HOROWITZ. 1997. Sp3 encodes multiple proteins that differ in their capacity to stimulate or repress transcription. *Nucleic Acids Res.* **25**:3110–3117.
- KINGSLEY, C., and A. WINOTO. 1992. Cloning of GT box-binding proteins: a novel Sp1 multigene family regulating T-cell receptor gene expression. *Mol. Cell. Biol.* **12**:4251–4261.
- KOLLMAR, R., K. A. SUKOW, S. K. SPONAGLE, and P. J. FARNHAM. 1994. Start site selection at the TATA-less carbamoylphosphate synthase (glutamine-hydrolyzing)/aspartate carbamoyltransferase/dihydroorotase promoter. *J. Biol. Chem.* **269**:2252–2257.
- LU, J., W. LEE, C. JIANG, and E. B. KELLER. 1994. Start site selection by Sp1 in the TATA-less human Ha-ras promoter. *J. Biol. Chem.* **269**:5391–5402.
- MASTRANGELO, I. A., A. J. COUREY, J. S. WALL, S. P. JACKSON, and P. V. HOUGH. 1991. DNA looping and Sp1 multimer links: a mechanism for transcriptional synergism and enhancement. *Proc. Natl. Acad. Sci. USA* **88**:5670–5674.
- MORGENSTERN, B., and W. R. ATCHLEY. 1999. Evolution of bHLH transcription factors: modular evolution by domain shuffling? *Mol. Biol. Evol.* **16**:1654–1663.
- NARAYAN, V., R. KRIWACKI, and J. CARADONNA. 1997. Structures of zinc finger domains from transcription factor Sp1. Insights into sequence-specific protein-DNA recognition. *J. Biol. Chem.* **272**:7801–7809.
- NARDELLI, J., T. J. GIBSON, C. VESQUE, and P. CHARNAY. 1991. Base sequence discrimination by zinc-finger DNA-binding domains. *Nature* **349**:175–178.
- OLEKSIK, M. F., K. KOLELL, and D. L. CRAWFORD. 2001. The utility of natural populations for microarray analyses: isolation of genes necessary for functional genomic studies. *Mar. Biotechnol.* **3**:S203–S211.
- PASCAL, E., and R. TJIAN. 1991. Different activation domains of Sp1 govern formation of multimers and mediate transcriptional synergism. *Genes Dev.* **5**:1646–1656.
- PAVLETICH, N. P., and C. O. PABO. 1991. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* **252**:809–817.
- PUGH, B. F., and R. TJIAN. 1990. Mechanism of transcriptional activation by Sp1: evidence for coactivators. *Cell* **61**:1187–1197.
- . 1991. Transcription from a TATA-less promoter requires a multisubunit TFIID complex. *Genes Dev.* **5**:1935–1945.
- ROEDER, R. G. 1991. The complexities of eukaryotic transcription initiation: regulation of preinitiation complex assembly. *Trends Biochem. Sci.* **16**:402–408.
- ROJO-NIERSBACH, E., T. FURUKAWA, and N. TANESE. 1999. Genetic dissection of hTAF(II)130 defines a hydrophobic surface required for interaction with glutamine-rich activators. *J. Biol. Chem.* **274**:33778–33784.
- SUSKE, G. 1999. The Sp-family of transcription factors. *Gene* **238**:291–300.
- SWOFFORD, D. L. 2000. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4.0b8. Sinauer Associates, Sunderland, Mass.
- TANESE, N., D. SALUJA, M. F. VASSALLO, J. L. CHEN, and A. ADMON. 1996. Molecular cloning and analysis of two subunits of the human TFIID complex: hTAFII130 and hTAFII100. *Proc. Natl. Acad. Sci. USA* **93**:13611–13616.
- TJIAN, R., and T. MANIATIS. 1994. Transcriptional activation: a complex puzzle with few easy pieces. *Cell* **77**:5–8.
- WILEY, S. R., R. J. KRAUS, and J. E. MERTZ. 1992. Functional binding of the “TATA” box binding component of transcription factor TFIID to the –30 region of TATA-less promoters. *Proc. Natl. Acad. Sci. USA* **89**:5814–5818.
- WOLSTEIN, O., A. SILKOV, M. REVACH, and R. DIKSTEIN. 2000. Specific interaction of TAFII105 with OCA-B is involved in activation of octamer-dependent transcription. *J. Biol. Chem.* **275**:16459–16465.

ANTONY DEAN, reviewing editor

Accepted September 26, 2001