Identification of the Linker-SH2 Domain of STAT as the Origin of the SH2 Domain Using Two-dimensional Structural Alignment*

Qian Gao‡§, Jian Hua†¶, Rich Kimura‖, Jeffery J. Headd**, Xin-yuan Fu‡¶, and Y. Eugene Chin**¶¶

The availability of large volumes of genomic sequences presents an unprecedented proteomic challenge to characterize the structure and function of various protein motifs. Primary structural alignment is often unable to accurately identify a given motif due to sequence divergence; however, with the aid of secondary structural prediction for analysis, it becomes feasible to explore protein motifs on a proteome-wide scale. Here we report the use of secondary structural alignment to characterize the Src homology 2 (SH2) domains of both conventional and divergent sequences and divide them into two groups, Src-type and STAT-type. In addition to the basic "αββα" structure (βB), the Src-type SH2 domain contains an extra β-strand (βE or βE-βF motif). Alternatively, the linker domain-conjugated SH2 domain in STAT contains the αB' motif. Combining BLAST data from βB core motif sequences with predicted secondary structural alignment, we have screened for SH2 domains in various eukaryotic model systems including Arabidopsis, Dicyostelium, and Saccharomyces. Two novel genes carrying the linker-SH2 domain of STAT were discovered and subsequently cloned from Arabidopsis. These genes, designated as STAT-type linker-SH2 domain factors (STATL), are found in a wide array of vascular and nonvascular plants, suggesting that the linker-SH2 domain evolved prior to the divergence of plants and animals. Using this approach, we expanded the number of putative SH2 domain-bearing genes in Dicyostelium and comparatively studied the secondary structural profiles of both typical and atypical SH2 domains. Our results indicate that the linker-SH2 domain of the transcription factor STAT is one of the most ancient and fully developed functional domains, serving as a template for the continuing evolution of the SH2 domain essential for phosphotyrosine signal transduction. Molecular & Cellular Proteomics 3:704–714, 2004.

From the ‡Whitehead Institute for Biomedical Research, Cambridge, MA 02142; †Section of Molecular and Cellular Biology, University of California, Davis, CA 95616; ¶¶Department of Pathology, Yale University School of Medicine, New Haven, CT 06510; and **Departments of Surgery Science and Molecular Biology, Cell Biology and Biochemistry, Brown University School of Medicine, Providence, RI 02903

Received, December 7, 2003, and in revised form, March 28, 2004 Published, MCP Papers in Press, April 7, 2004, DOI 10.1074/mcp.M300131-MCP200

The Src homology 2 (SH2) domain is an ~100-aa-long motif that recognizes and interacts with phosphotyrosine-containing motifs on the same or different protein molecules during signal transduction in animal cells. About 200 SH2 domain-containing genes have been identified in human cells, suggesting that this domain is one of the most rapidly expanded protein modules (1). In animal cells, SH2 domains are predominately present in signaling molecules, i.e. signaling-related enzymes including protein tyrosine kinases, protein tyrosine phosphatases, inositol phosphatase, and phospholipase and signaling adapters. However, the SH2 domain has also been found in transcription factor STAT family members (2, 3). In a signaling molecule with catalytic activity, the SH2 domain is often conjugated immediately upstream with another functional motif such as the SH3 domain, whereas in STAT the linker domain is immediate upstream of the SH2 domain. Recently, two STAT proteins have been discovered in Dicyostelium, a facultative slime mold capable of both growing as a single cell and differentiating into multicellular structures (4, 5). More recently, SHK, the SH2 domain-bearing protein kinase, has been identified in the same species (6). In both cases, the SH2 domains are the linker domain-conjugated. Because a typical SH2 domain has not been found in single-cell eukaryote yeast or microorganisms, the SH2 domain formation and its phospho-signaling were proposed to be coincident with animal evolution, perhaps critical during the transition from single-cell to multicellular animals or metazoan (7, 8).

The characteristic structure of the SH2 domain is three β-strands flanked by two α-helices (αββα). The first β-strand (βB), conserved in its sequence GXF/YBBR (9), is the core motif critical for binding phosphotyrosine (pTyr) (10). This sequence is required for the normal function of the SH2 domain and conveniently serves as the fingerprint structure in SH2 domain recognition. While βB core motif and motif-like sequences exist widely in different genes, some perfect βB sequences are not necessarily indicative of the SH2 domain.
(11). In these cases, secondary structural alignment clarifies the confusion caused by sequence alignment alone. Additionally, secondary structural analysis can provide reliable structural evidence for SH2 domains with ambiguous βB and βE-flanking sequences. A careful analysis of the amino acid sequence, motif orientation, and secondary structural features indicate that STAT SH2 domains differ from those involved in signal transduction. In a STAT protein, the SH2 domain is the immediate extension from five continuing α-helices representing the linker domain. Moreover, all STAT-like SH2 domains carry a αB′ motif between βD and αB sequences. We combined βB core motif sequence BLAST with secondary structural screening to identify SH2 domains in genome databases of various eukaryotic model systems. We identified and analyzed a novel gene family, which carries the linker-SH2 domain of STAT from the genome database of Arabidopsis as well as other plants. Two of these linker-SH2 domain-carrying genes were cloned from the cDNA library of Arabidopsis and sequenced. Using secondary structural alignment, we comprehensively analyzed the typical and atypical SH2 domains found in plants, yeast, and Dictyostelium. According to our secondary structural analysis, the linker-SH2 domain of SHK in Dictyostelium, representing the modern SH2 domain in signal transduction, may share the common ancestor with or even directly evolved from the linker-SH2 domain of STAT. The discovery of the linker-SH2 domain of STAT in plants supports the notion that this domain had been developed prior to the divergence of plant and animal kingdoms.

MATERIALS AND METHODS

Pat-match Search and Secondary Structural Analysis—The consensus sequence used for Pat-match (www.arabidopsis.org/cgi-bin/patmatch/nph-patmatch.pl) for TAIR Arabidopsis genome and seq. yeastgenome.org/cgi-bin/SGD/PATMATCH/nph-patmatch for yeast genome) screening is “GX/YBBR” (X = any amino acid; B = hydrophobic amino acid). This consensus represents 98% of the βB core motif sequences based upon analysis of 775 SH2 domain sequences in the SMART database using the SMART Simple Modular Architecture Research Tool (smart.embl-heidelberg.de). The sequences flanking the βB motif in these putative genes were analyzed with the secondary structure prediction program 3D-PSSM web tool version 2.5.6 (www.sbg.bio.ic.ac.uk/~3dpssm/), which predicts α-helices and β-strands. When overall α-helix/β-strand arrangement, but not individual amino acid residue around the motif sequences was evaluated, all predicted α-helix and β-sheet were correct comparing with the structural information obtained from crystallization analysis.

Cloning and Northern blot of at-STATLa and at-STATLb—Both at-statla and at-statlb genes were cloned from the Arabidopsis cDNA library. Primers from known, flanking regions were used in order to clone the full-length sequences of at-statla and at-statlb. Various primer combinations were designed in order to cover with overlapping the full-length sequences of at-statla and at-statlb. PCR products were subsequently sequenced and the gene sequences of both at-statla and at-statlb have been deposited in GenBank. Total RNAs prepared from different parts of Arabidopsis were used for Northern blot analysis.

Western Blotting and Immuno-precipitation—Arabidopsis thaliana, ecotype Columbia, was grown on Murashige and Skoog agar medium at 22 °C under constant light for 3 weeks. Sodium orthovanadate (100 μM) and hydrogen peroxide (1 mM) were added for the indicated times. The whole plant of Arabidopsis was homogenized in ice-cold extraction buffer (30 mM Tris-HCl, pH 8.5, 150 mM NaCl, 1 mM EDTA, 20% glycerol, 1 mM dithiothreitol, and protease inhibitors). Cell debris was separated from soluble material by centrifugation at 18,000 × g for 5 min. GST-at-STATLa in full length and GST-at-STATLb-SH2 domain (596–692) were constructed, expressed, and purified as described previously (12). Purified glutathione S-transferase (GST) recombinant proteins were incubated with the above-prepared extracts. Extensively washed GST protein precipitates were then subjected to standard Western blotting procedures using horse-radish peroxidase-conjugated anti-pTyr (pY20) and enhanced chemiluminescence (Amersham Biosciences, Piscartaway, NJ) for detection.

RESULTS AND DISCUSSION

STAT-type SH2 Domain Differs from Src-type SH2 Domain at Secondary Structural Level—Using the secondary structural prediction program, we analyzed the SH2 domains available in SMART protein database. As expected, all the SH2 domains examined contain the basic “αβββα” structure. The SH2 domains of signaling factors including enzymes and adapters exclusively contain the predicted βE or βE-βF motif, consistent with crystallographic findings (Fig. 1A) (13–16). The small β-strand, βE or βE-βF motif, has proven to be critical for protein/ligand recognition in pTyr signaling factors (10). However, the βF fragment is not always detectable or predictable, presumably due to the instability of this sequence in β-strand formation (15, 17). According to the distance between βC and βD motifs, the SH2 domain-carrying enzymes can be further divided into the long βC-βD loop and the short βC-βD loop groups (Fig. 1A). Src family members all belong to the long βC-βD group. In most of the adapters, the SH2 domains carry a putative short βC-βD loop. For transcription factor STAT, the length of the predicted βC-αB sequence varies with various family members and are overall shorter than those of signaling factors (Fig. 1A). But the most striking difference between Src-type and STAT-type SH2 domains is that STAT or STAT-like SH2 domains do not contain the βE or βE-βF motif. Instead, they all contain the αB′ or a nonsplit αB motif (Fig. 1A), which is considered as the critical region for STAT dimerization (16, 18). All these secondary structural features of STAT-type SH2 domains obtained from the secondary structural prediction are consistent with the findings obtained from the crystallization studies (16, 18). In Drosophila STAT, the SH2 domain carries the putative αB′ motif despite lacking the βD motif (Fig. 1A). Two putative STAT-like sequences (ce-STATa and ce-STATb) were previously found by homologous alignment in Caenorhabditis elegans genome data (19). Our secondary structural analysis predicts that these SH2 domains also contain αB′ motifs (Fig. 1A). Therefore, the level of detail obtained here can easily distinguish subtle domain differences between STAT SH2 and signaling factors.

We also analyzed the SH2 domains with divergent sequences. We investigated CBL and JAK family members that differences between STAT SH2 and signaling factors.
FIG. 1. Primary and secondary structural analysis of the SH2 domains. A, proteins that carry the SH2 domains are grouped into (i) signaling enzymes: SRC(K03218), BLK(S76617), FGR(M14333), HCK(M16591), LCK(X13529), LYK(M16038), YES(BC08960), YRK, BTK(X58957), ITK(D13720), TEC(D29767), IPP(1), SHP-1(M74903), SHP-2(L08807), BSK(Q13882), FRK(U00803), SYK(Z29630), ZAP-70(L29277), CSK(X74765), VAV1(AF030227), VAV2(S76992), SOCS1(U88326), SOCS2(AB004903), GRB10(U34355), RIN1(L36463), NSP1(AAQ88772), and NSP3(AAQ89948); (ii) signaling adaptors: DAPP(NP_055210), GRB2(M96995), GRAP(U52518), GRP2(Y18051), NCK1(X75767), NCK2(AF043119), SH2B(AF227967), SHC(U33778), SCK(AL360254), PI3K-p85(p110), PLCγ1(AB109343), PLCγ2(AB109343), VAV1(AF030227), VAV2(S76992), SOCS1(U88326), SOCS2(AB004903), GRB10(U34355), RIN1(L36463), NSP1(AAQ88772), and NSP3(AAQ89948); and (iii) self-signaling transcription factors (i.e. STAT): hs-STAT1: M97935, hs-STAT2: U18671, hs-STAT3: L29277, hs-STAT4: L78430, and two putative C. elegans STATs (ce-STATa: CEY51H4A and ce-STATb: AF163113) (19) were included for analysis. All the SH2 domains as indicated were submitted to secondary structure program analysis and aligned. The SH2 domains of the enzyme group are divided into long subgroups according to the distance between putative βC and αB motifs. The predicted β-strands are printed in yellow, and the predicted β-sheets are printed in cyan. The β-sheet motifs are shown in red. In the STAT SH2 domain, the βC motif is followed by a phenylalanine (F) residue, which is highlighted in blue. Hs-Src-Crys1 and Hs-Src-Crys2 represent the secondary structure obtained from two independent crystal studies of the Src (16, 17). B, the divergent amino acid sequence of the CBL-SH2 domain (U26170) was analyzed with secondary structure program and compared with the features obtained from crystallographic study (20). C, the suspected SH2-like sequences of human Jak1 (M64174, M35203), Jak2 (AF589825), Jak3 (U96807), Tyk2 (X54637), and Drosophila JAK (L26975) were submitted to secondary structure program analysis and aligned with each other.
are all known to contain a β-like motif flanked by ambiguous sequences. For CBL, the typical “αββα” topology obtained from secondary structural prediction (Fig. 1B) agreed with the conclusion drawn in a previous crystallography study (20). However, unlike Src-type or STAT-type SH2 domains, the CBL SH2 domain lacks the small βE or αβ motif. The immediate upstream sequence of the kinase-like domain in JAK has long been suspected as an SH2-like domain (21). Two-dimensional alignment clearly reveals that this sequence, though ambiguous, represents a typical Src-type SH2 domain (Fig. 1C). Among all the JAK members analyzed, Drosophila JAK contains the longest loop between βB and βC motifs. Utilizing other programs such as PHD (maple.bioc.columbia.edu/predictprotein) and JPRED (www.compbio.dundee.ac.uk) for prediction, similar results were obtained (data not shown). Thus, secondary structural prediction is particularly suited for, and reliable in, the detection of such fine structural difference of protein motifs with divergent sequences.

STATLs Are Novel Genes Carrying the STAT-like Linker-SH2 Domain in Plants—To locate putative SH2 domains in plants, we used Pat-match to screen the genomes of various eukaryotes for the βB core motif (see “Materials and Methods”). In the Arabidopsis Information Resource (TAIR; www.arabidopsis.org), we found 604 βB sequences in 583 putative genes, of which secondary structural analysis confirmed two putative genes containing the typical “αββα” structure of an SH2 domain [AC007651 (protein locus: AAD50031); AC007260 (protein locus: AAD30582)]. Subsequent cDNA cloning and DNA sequencing analysis indicate that these two genes are closely related to each other (65% identical at the amino acid level) (Fig. 2A). In both genes, the C termini are longer and divergent than those predicated by those published genomic sequences. The SH2 domains reside at the C-terminal regions and the predicted secondary structure match those found in STAT (Fig. 2B). Moreover, a sequence of 90 amino acids immediately N-terminal of the SH2 domain is also well conserved and resembles STAT’s linker domain (Fig. 2B) (16, 18, 22). We therefore named the two novel genes STAT-type linker-SH2 domain factor a and b (STATLa and STATLb). The linker-SH2 domain is well conserved in putative STATL genes identified in both monocot and dicot plants including soybean, sorghum, potato, tomato, medicago, wheat, and rice (Fig. 2C). Surprisingly, the STATL sequences were also retrieved in lower plants like the moss, Physcomitrella patens (pp-STATL) from NCBI translated BLAST searches, and the green algae, Chlamydomonas (cr-STATL) from the Chlamydomonas Resource Center (www.biology.duke.edu/chlamy_genome/crc.html), (Fig. 2C). The ubiquitous presence of STATL in plants suggests that the SH2 domain plays a fundamental role in plants and animals and rejects the possibility that this domain originated in plants through some accidental means (i.e., reverse horizontal gene transfer) or coevolved with tyrosine kinases or the SH3 domain (23). Amino acid sequence alignment indicates that the linker-SH2
domain identity is 29% between at-STATLa and hs-STAT3 and 33% between at-STATLb and hs-STAT3. About the same range of amino acid identity was obtained when the same domain of dd-STATTa (29%) or dd-STATTc (30%) was aligned with that of hs-STAT3. In the SH2 domains of all STAT proteins regardless of the species, the βB motif is exclusively followed by a phenylalanine. Most STAT members follow this rule without exception (Fig. 2C), whereas in the SH2 domains of signaling factors the βB motif has never been found to be followed immediately by phenylalanine. Another protein sequence feature is that, like most STAT members, the first residue of αA motif in STATL is also a lysine rather than an arginine, which coordinates with a phosphate group in metazoan STAT (16).

The secondary structure of STAT’s linker-SH2 domain (the five α-helices and the “αββαα” sandwich) (16, 18) is well conserved in at-STATLa proteins according to our prediction (Fig. 2B). When aligned with STAT, sequence gaps in at-STATLa and at-STATLb occur at neutral positions and do not interrupt the arrangement of α-helices and β-strands. For instance, a 9-aa stretch (ENMAGKFGS), absent in the linker domain at of at-STATLa or at-STATLb, forms an out loop between α9 and α10 in the linker domain, which does not interrupt the helicity between α9 and α10 (Fig. 2B) (16). For all STATL sequences obtained from different plants, the SH2 domains do not carry the βE motif. Instead, they exclusively carry the αB’ motif or a nonsplit αB’βB (Fig. 2C), which matches the secondary structural characteristics of the STAT SH2 domain. The lack of βD motif coupled with the presence of a large nonsplit αB in the SH2 domain of cr-STATL in Chlamydomonas may indicate a premature form of SH2 domain that arose during its development. The upstream of the linker-SH2 domain (Ser139-Pro232 in both at-STATLa and at-STATLb) is predicted to form a continuous β-sheet (Fig. 2D), which has some similarity to the DNA-binding domain of dd-STATTa or dd-STATTc but not that of human STAT (16, 18, 24). Therefore, the strong similarity between STAT and STATL within their linker-SH2 domains at both the amino acid sequence and secondary structure levels strongly indicate that these two domains share a common ancestor that evolved prior to the divergence of plants and animals.

Messenger RNAs of both genes were detected ubiquitously in different parts of Arabidopsis (Fig. 2E). Plants are not known to contain JAK-like nonreceptor tyrosine kinases (25). However, receptor protein kinases, serine/threonine plus tyrosine dual-function protein kinases, as well as protein phosphatases exist in plants (13, 26, 27). In Fig. 2F, tyrosine-phosphorylated proteins ranging in size from 60 to 120 kDa were detected in Arabidopsis treated with vanadate, a naturally occurring transition metal that can function as a nonspecific protein phosphatase inhibitor and trigger protein tyrosine phosphorylation in cells (28). Purified GST-STATLa-full length and GST-STATLb-SH2 domain proteins but not the GST control (Fig. 2G, right panel) were able to pull down the 120-kDa tyrosine phosphorylated protein in vanadate-treated samples but not in the samples without vanadate treatment (Fig. 2G, left panel). Therefore, the SH2 domain of STATL proteins might be involved in pTyr-dependent protein-protein interaction in plants.

SHK’s Linker-SH2 Domain Is Homologous to That of STAT in Dictyostelium—How does the STAT-type linker-SH2 domain relate phylogenetically to the tyrosine signaling- or Src-type SH2 domain, which quickly expanded in number in animal cells? To answer this question, we studied the database of Dictyostelium, a slime mold considered more closely related to fungi and animals than to plants (29). SH2-bearing genes cloned from Dictyostelium include two transcription factors (i.e., STATa and STATc) and one signaling factor (i.e., SHK1) (4–6). From the Dictyostelium discoideum Genome Project (www.sanger.ac.uk/Projects/D_discoideum), two additional putative STAT sequences, designated as dd-STATTa and dd-STATTd, and four additional putative SHK genes, designated as dd-SHK2, dd-SHK3, dd-SHK4, and dd-SHK5, were identified using βB core motif sequence as well as the
whole SH2 domain for BLAST searches. Like SHK1, the protein kinases of these novel putative SHK members are most closely related to the protein kinases found in plants (6). However, these same kinases in plants are not conjugated to any SH2 or SH2-like sequences. Using the kinase domain (C-region) of the putative SHK2 for a BLAST search, a large number of homologous expressed sequence tags (ESTs) were identified in the genomes of both Arabidopsis and Dictyostelium, but not in the databases of other eukaryotes (not shown). This suggests a close evolutionary relationship between plants and Dictyostelium.

Primary and predicted secondary structure alignment indicates that the SHK SH2 domains carry some features of the STAT SH2 domains in Dictyostelium. Phenylalanine (F) has been found to follow the ββ motif in the SH2 domains of SHK1, SHK2, and SHK3 (Fig. 3A). However, secondary structural modifications have been noted in SHK’s SH2 domain that contrast STAT. In the region between the predicted ββ and αβ motifs, the αβ’ motif has been replaced by the βF or βE-βF motif. The introduction of βE in SHK5 lengthened the distance between αA and αβ motifs (Fig. 3A). This seems to reflect the trend of SH2 maturation because this distance between αA and αβ motifs have become even longer in the Src SH2 domain in metazoans (Fig. 1A). In this region, both

Fig. 3. The SH2 domain coevolves with the linker domain but not with the SH3 domain. A, the SH2 domains of SHK and STAT members from D. discoideum were analyzed with secondary structure program and aligned. SHK family members include dd-SHK1, putative dd-SHK2 (Contig13319, Sanger Center), putative dd-SHK3 (AX4a11Blb2.r1), putative dd-SHK4 (JC2d33h04.s1), and putative dd-SHK5 (Contig12179, Sanger Center). STAT family members include dd-STATa, dd-STATc, putative dd-STATb (Contig17339, Sanger Center), and putative dd-STATd (JC1b225h10.r1). In dd-STATd, an S/T-rich sequence (KDSLSKSSNDKLQPTTTTT TTTS) between βC and βD was omitted. B, the linker domains of dd-SHK and dd-STAT members were analyzed with secondary structure program and aligned. C, the SH3 domains of human Src (hs-Src, P12931), yeast NAP1-binding protein (sc-NBP, YDR162C), a putative gene of Dictyostelium (dd-SH3a, C94356), and three putative genes of A. thaliana (at-SH3a, AAG5264; at-SH3b, AAL32440; and at-SH3c, AAL32439) were analyzed with secondary structure program and compared with the structural features of the SH3 domain obtained from crystallization (15).
Fig. 4. SPT6 carries a putative atypical linker-SH2 domain within the C terminus. The C-terminal regions of SPT6 proteins were analyzed with secondary structure program and aligned with the SH2 domains of hs-STAT1 and at-STATLb. SPT6 or putative SPT6 genes included for this study are: sc-SPT6 (S. cerevisiae, ACP23615), sp-SPT6 (Schizosaccharomyces pombe, ACQ00915), at-SPT6a (A. thaliana, AC010795), at-SPT6b (A. thaliana, AC022355), dd-SPT6 (D. discoideum, Contig17121, Sanger Center), hs-SPT (Homo sapiens, HSU46691), dm-SPT6 (Drosophila melanogaster, AF104400), ce-SPT6 (C. elegans, D14635), xi-SPT6 (Xenopus laevis, BJ098459), lm-SPT6 (Leishmania major, ALAI822), tb-SPT6 (Trypanosoma brucei, AC003868), dr-SPT6 (Danio rerio, AF421378), pc-SPT6 (Pneumocystis carinii f. sp. Canhii, AW334882), lJ-SPT6 (Lotus japonicus, BG662187), gx-SPT6 (Glycine max, BI424708), mt-SPT6 (Medicago truncatula, BG645814), and sb-SPT6 (Sorghum bicolor, Br139864).

Sequence similarities and gaps were observed, suggesting that the accumulation of favorable mutations, insertions, and deletions might all contribute to the α-helix/β-sheet switch (Fig. 3A). Thus, this αβ'/β-βF-containing region in the SH2 domain serves as an evolutionarily active region (EAR) within an otherwise conserved domain essential for its function. When STATs’s linker domain was used for a BLAST search, the sequence between the protein kinase domain and the SH2 domain (the linker) of SHK was recovered, suggesting a close relationship among these molecules within this region. SHK’s linker domain is predicted to contain a α-helix repeat composed of α7 to α10 motifs, which is indeed homologous to that of STAT (Fig. 3B). The C-terminal three α-helices (α9, α10, and α11) formed a large α-helix immediately upstream of the SH2 domain. While the linker domains of most SHK members are relatively similar in size, SHK5’s linker domain is apparently much longer due to homopolymerism, a characteristic of many genes found in Arabidopsis (4, 7). Comparing with STAT or STATL, the predicted helical characteristic of the linker domain is degenerating in SHK (Fig. 3B), perhaps due to a functional regression of this domain in tyrosine signaling.

Although the linker domain in STAT may play a role in transcription (22), it was either lost or replaced by the SH3 domain in SH2-bearing signaling proteins of animal cells. For a long time, the linker domain of metazoan STAT was confused as the SH3 domain based upon amino acid sequence alignment (2). The SH3 domain consists of five β-strands arranged as two tightly packed anti-parallel β-sheets (14, 15). Our secondary structural analysis revealed the typical SH3 domain structure in the C-terminal regions of three putative Arabidopsis genes among 14 putative SH3-carrying genes under both plant and prokaryote categories given by the SMART protein database (Fig. 3C). The SH3 domains of these three putative genes showed either a α-helix or ω/β-hybridized motif in the region of the βC motif (Fig. 3C), suggesting that this region is a putative EAR. Although a weak sequence homology exists between the STAT linker domain and the SH3 domain (2), the presence of well-developed SH3 domains as well as the fully developed linker-SH2 domain in independent genes in plants indicates the SH3 domain is unlikely evolved from the linker domain. While SH3 domain carrying genes exist in plants and quickly multiplied in its number in lower eukaryotes such as Dictyostelium and yeast, SH3-SH2 domain conjugation has not been discovered in these organisms.

SPT6 Gene Carries a Putative Immature Linker-SH2-like Domain—We next studied the yeast genome in which no typical SH2 domains were identified (30). Using the same approach, we identified 89 β sequences from 86 putative genes in the Saccaromyces genome database (genome-www.stanford.edu/Saccharomyces). The suppressor of Ty 6 gene (SPT6) attracted our attention after secondary structural analysis of all these sequences. SPT6 was previously reported in yeast and animal cells and is involved in transcriptional initiation and DNA/RNA binding (31, 32). Using the yeast SPT6 protein sequence for a BLAST search, putative SPT6 genes were identified in plants, Dictyostelium, and other eukaryotes (NCBI BLAST searches), suggesting that SPT6 is also an ancient gene that existed prior to the divergence of plants and animals. The conserved third residue should be either Phe or Tyr in the standard βB sequence GXF/YBBR (Fig. 4); however, sc-SPT6 of yeast is the only one to follow this rule (Fig. 4). Nevertheless, the putative αββα-like structure, albeit less typical, is well maintained in all SPT6 proteins according to our secondary structural prediction, supporting the suspicion of a degenerate SH2 domain in SPT6 genes (33, 34). The αA, βB, and βC motifs that compose the evolutionary inactive
region in the SH2 domains are conserved in SPT6s regardless of the origin. The putative α′β helical structure in SPT6 does not split as STATα’s SH2 domain does. In the suspected EARs of most SPT6s, a short β-strand and a α-helix are predicted, and the βD motif was not fully extended according to the prediction analysis. Such a poorly developed putative βD motif may not efficiently form an anti-parallel β-sheet with βB and βC, and may eventually hamper its function in protein-protein interactions. The large Lys/Glu-rich helical structure in the linker domain of SPT6 resembles the ancient STATL-like α9-α10-α11 continuing α-helices that became three discrete helices in the linker domain of human STAT or was discontinued in SHK (Figs. 2A and 3B). Moreover, in the region upstream of the linker domain, SPT6 contains a putative 6-β-strand repeat resembling the DNA-binding domain in both STAT and at-STATL (not shown). Interestingly, our secondary structure prediction analysis indicates that SPT6 and cr-STATL of the unicellular plant Chlamydomonas share some critical features at secondary structural level. Both are predicted to bear a large nonsplit α′β′α′β motif and a poorly developed βD motif (Figs. 2B and 4). Therefore, it is possible that SPT6 and STAT are evolutionarily related.

In terms of amino acid sequence, the SH2 domains of SPT6, STAT, and JAK are among the most divergent. Unlike JAK family members, both SPT6 and STAT are transcription factors that contain special secondary structural characteristics, especially in the EAR sequence as seen above. Based on the phylogenetic alignment, SH2 domains can be grouped into two categories, i.e., STAT-type and Src-type (Fig. 5). SHK family members fall in between STAT-type and Src-type but are closer to the STAT-type (Fig. 5). This strongly indicates a close relationship between SHK and STAT families in their SH2 domains and further supports the notion that SHK’s linker-SH2 domain evolved from STAT or STATL. In SHK, STAT, and SPT6, the linker-SH2 domains all reside exclusively in the C-terminal regions. The appearance of SHK gene, which bears

Fig. 5. Different types of SH2 domains are revealed by phylogenetic analysis. The amino acid sequences of SH2 domains of different genes are aligned with ClustalW (v1.4) program and grouped into two categories: Src-type and STAT-type. The alignment parameters are gap distance = 8; similarity matrix = blosum; open gap penalty = 10.0; and extend gap penalty = 0.1.
both the linker-SH2 domain and the kinase domain, in Dictyo-
stellum started the new era toward pTyr signal transduction.

The combination of amino acid sequence screening with
secondary structural analysis improves the accuracy in pro-
tein motif prediction. The βB core motif-like and its surround-
ing sequences in the Rht-B1/D1 gene of wheat have been
considered as the first SH2 domain discovered in plant (35).
Moreover, GAI, RGA, and SCR, all members of a putative
transcription factor family termed GRAS, were suspected as
STAT-like factors in plant and carry the SH2 domains due to
βB-like core sequences found in their C-terminal regions (11,
35). However, according to our secondary structural predic-
tion “βααα” rather than “αβββαα” topology is present in all those
SH2-like sequences. Therefore, a perfect βB sequence does
not necessarily reveal an SH2 domain structure. In contrast,
as long as the “αβββαα” secondary structure maintains, the
amino acid sequence can be very divergent. The balance
between evolution and conservation in the SH2 domain de-
velopment reflects evolution at the primary structural level but
conservation at secondary structural level.

The discovery of STAT linker-SH2 domain-bearing genes in
plants underscores the proposal that SH2 domain develop-
ment was an early step in the evolution of multicellularity. The
SH2 domain formed most likely in a common eukaryote an-
cestor prior to divergence of any of the major eukaryote taxa.
Hence, the linker-SH2 domain of the transcription factor STAT
was an early step in the evolution of multicellularity. The
conservation at secondary structural level.

Acknowledgments—We are grateful to J. M. Wang for critical sug-
gestions in performing secondary structural analysis and D. Chatter-
jee for careful reading the manuscript.

‡ This work was supported in part by National Institutes of Health
Grant RO1 CA82549 (to E. Y. C.) and by Grant RR-15578 from the
National Center for Research Resources to Brown University. The
Center of Biomedical Research Excellence (COBRE) Program of the
Grant RO1 CA82549 (to E. Y. C.) and by Grant RR-15578 from the
National Institutes of Health.

REFERENCES
1. Pawson, T., Gish, G. D., and Nash, P. (2001) SH2 domains, interaction
modules and cellular wiring. Trends Cell Biol. 11, 504–511
2. Fu, X. Y. (1992) A transcription factor with SH2 and SH3 domains is directly
activated by an interferon-α-induced cytoplasmic protein tyrosine ki-
rase(s). Cell 70, 323–335
3. Darnell, J. E., Jr. (1997) STATs and gene regulation. Science 277,
1630–1635
4. Kawata, T., Shevchenko, A., Fukuzawa, M., Jermn, K. A., Totty, N. F.,
Zhukovskaya, N. V., Sterling, A. E., Mann, M., and Williams, J. G. (1997)
SH2 signaling in a lower eukaryote: A STAT protein that regulates stakl
cell differentiation in dictyostelium. Cell 89, 909–916
5. Fukuzawa, M., Araki, T., Adrian, I., and Williams, J. G. (2001) Tyrosine
phosphorylation-independent nuclear translocation of a dictyostelium
STAT in response to DIF signaling. Mol. Cell 7, 779–788
6. Moniaux, J., Funamoto, S., Fukuzawa, M., Meisenhelder, J., Araki, T., Abe,
Meili, R., Hunter, T., Williams, J., and Firtel, R. A. (2001) An SH2-
domain-containing kinase negatively regulates the phosphatidylinosi-
 tol-3 kinase pathway. Genes Dev. 15, 687–698
7. Kimmel, A. R., and Firtel, R. A. (1985) Sequence organization and de-
veloptment expression of an interspersed, repetitive element and associ-
ated single-copy DNA sequences in Dictyostelium discoideum. Mol. Cell
Biol. 5, 2123–2130
8. Kay, R. R. (1997) Development at the edge of multi-cellularity: Dictyostelium
discoideum. Curr. Biol. 7, R723–725
9. Pawson, T., and Gish, G. D. (1992) SH2 and SH3 domains: From structure
to function. Cell 71, 359–362
10. Waksman, G., Kominos, D., Robertson, S. C., Pant N., Baltimore, D., Birge,
R. B., Cowburn, D., Hanafusa, H., Mayer, B. J., Overduin, M., et al. (1992)
Crystal structure of the phosphotyrosine recognition domain SH2 of
v-src complexed with tyrosine-phosphorylated peptides. Nature
358, 646–653
11. Richards, D. E., Peng, J., and Harberd, N. P. (2000) Plant GRAS and
metazon STATs: one family? BioEssays 22, 575–577
12. Wang, Y., Wu, T. R., Cai, S., Welte, T., and Chin, Y. E. (2000) Stat1 as a
component of tumor necrosis factor α receptor 1-TRADD signaling com-
pex to inhibit NF-κB activation. Mol. Cell Biol. 20, 4505–4512
13. Hof, P., Pluskey, S., Dhe-Paganon, S., Eck, M. J., and Shoelson, S. E.
(1998) Crystal structure of the tyrosine phosphate SHP-2. Cell 92,
441–450
14. Silver, F., Moore, I., and Kuriyan, J. (1997) Crystal structure of the Src
family tyrosine kinase Hck. Nature 385, 602–609
15. Xu, W., Harrison, S. C., and Eck, M. J. (1997) Three-dimensional structure
of the tyrosine kinase c-Src. Nature 385, 595–602
16. Chen, X., Vinkemeier, U., Zhao, Y., Jeruzalmi, D., Darnell, J. E., Jr., and
Kuriyan, J. (1998) Crystal structure of a tyrosine phosphorylated STAT-1
dimer bound to DNA. Cell 93, 827–839
17. Charlton, P. S., Shewchuk, L. M., Rocque, W., Hummel, C. W., Jordan,
S. R., Mohr, C., Pacofsky, G. J., Peel, M. R., Rodriguez, M., Sternbach,
D. D., and Consler, T. G. (1997) Peptide ligands of ppp6(c-src) SH2
domains: A thermodynamic and structural study. Biochemistry 36,
6283–6293
18. Becker, S., Groner, B., and Müller, C. W. (1998) Three-dimensional struc-
ture of the STAT3J homodimer bound to DNA. Nature 394, 145–151
19. Liu, X. D., Quinn, M., Chin, Y. E., and Fu, X. Y. (1999) STAT genes in C.
ellegans. Science 285, 1695–1698
20. Meng, W., Sawadikosol, S., Burakoff, S. J., and Eck, M. J. (1999) Structure
of the amino-terminl domain of Cbl complexed to its binding site on
ZAP-70 kinase. Nature 399, 84–90
21. Zhou, Y. J., Chen, M., Cusack, N. A., Kimmel, L. H., Magnuson, K. S., Boyd,
J. G., Lin, W., Roberts, J. L., Lengi, A., Buckley, R. H., Geahlen, R. L.,
Candotti, F., Gadina, M., Changelian, P. S., and O’Shea, J. J. (2001)
Unexpected effects of ERM domain mutations on catalytic activity of
Jak3: Structural implication for Janus kinases. Mol. Cell 8, 959–969
22. Yang, E., Wen, Z., Haspel, R. L., Zhang, J. J., and Darnell, J. E., Jr. (1999)
The linker domain of Stat1 is required for γ interferon-driven transcrip-
tion. Mol. Cell Biol. 19, 5106–5112
23. Nars, M., and Vihinen, M. (2001) Coevolution of the domains of cytoplasmic
 tyrosine kinases. Mol. Biol. Evol. 18, 312–321
24. Horvath, C. M., Wen, Z., and Darnell, J. E., Jr. (1998) A STAT protein domain
that determines DNA sequence recognition suggests a novel DNA-bind-
ing domain. Genes Dev. 9, 984–994
25. The Arabidopsis Genome Initiative. (2000) Nature 408, 796–815
26. Ali, N., Haffter, U., and Chu, N. H. (1994) Cloning and biochemical char-
terization of a plant protein kinase that phosphorylates serine, threo-
ine, and tyrosine. J. Biol. Chem. 269, 31626–31629
27. Stone, J. M., Collinge, M. A., Smith, R. D., Horn, M. A., and Walker, J. C.
(1994) Interaction of a protein phosphatase with an Arabidopsis serine-
threonine receptor kinase. Science 266, 793–795
28. Sweitzer, S. M., Calvo, S., Kraus, M. H., Finbloom, D. S., and Larner, A. C. (1995) Characterization of a Stat-like DNA binding activity in Drosophila melanogaster. J. Biol. Chem. 270, 16510–16513
29. Baldauf, S., and Doolittle, W. F. (1997) Origin and evolution of the slime molds (Mycetozoa). Proc. Natl. Acad. Sci. U. S. A. 94, 12007–12012
30. Hunter, T., and Plowman, G. D. (1997) The protein kinases of budding yeast: Six score and more. Trends Biochem. Sci. 22, 18–22
31. Clark-Adams, C. D., and Winston, F. (1987) The SPT6 gene is essential for growth and is required for δ-mediated transcription in Saccharomyces cerevisiae. Mol. Cell Biol. 7, 679–686
32. Swanson, M. S., Carlson, M., and Winston, F. (1990) SPT6, an essential gene that affects transcription in Saccharomyces cerevisiae, encodes a nuclear protein with an extremely acidic amino terminus. Cell Biol. 10, 4935–4941
33. Maclennan, A. J., and Shaw, G. (1993) A yeast SH2 domain. Trends Biochem. 18, 464–465
34. Chiang, P. W., Wang, S. Q., Smithivas, P., Song, W. J., Crombez, E., Akhtar, A., Im, R., Greenfield, J., Ramamoorthy, S., Van Keuren, M., Blackburn, C. C., Tsai, C. H., and Kumit, D. M. (1996) Isolation and characterization of the human and mouse homologues (SUPT4H and Sup4h) of the yeast SPT4 gene. Genomics 34, 328–333
35. Peng, J., Richards, D. E., Hartley, N. M., Murphy, G. P., Devos, K. M., Flintham, J. E., Beales, J., Fish, L. J., Worland, A. J., Pelica, F., Sudhakar, D., Christou, P., Snape, J. W., Gale, M. D., and Harberd, N. P. (1999) “Green revolution” genes encode mutant gibberellin response modulators. Nature 400, 256–261
36. Darnell, J. E., Jr. (1997) Phosphotyrosine signaling and the single cell: Metazoa boundary. Proc. Natl. Acad. Sci. U. S. A. 94, 11767–11769
37. Kuriyan, J., and Darnell, J. E., Jr. (1999) An SH2 domain in disguise. Nature 398, 22–23