The N-terminal domains of SOCS proteins: A conserved region in the disordered N-termini of SOCS4 and 5

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · MARCH 2012

Impact Factor: 2.63 · DOI: 10.1002/prot.23252 · Source: PubMed

6 AUTHORS, INCLUDING:

Zhi-Ping Feng
The Walter and Eliza Hall Institute of Medical …
34 PUBLICATIONS **1,076** CITATIONS
SEE PROFILE

Indu R Chandrashekaran
Monash University (Australia)
19 PUBLICATIONS **216** CITATIONS
SEE PROFILE

Andrew Low
CSL Limited
11 PUBLICATIONS **206** CITATIONS
SEE PROFILE

Raymond Norton
Monash University (Australia)
301 PUBLICATIONS **8,067** CITATIONS
SEE PROFILE

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

The N-terminal domains of SOCS proteins: A conserved region in the disordered N-termini of SOCS4 and 5

Zhi-Ping Feng,[1,2#] Indu R. Chandrashekaran,[3#] Andrew Low,[1,2] Terence P. Speed,[1,2] Sandra E. Nicholson,[1,2] and Raymond S. Norton[3*]

[1] The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria 3052, Australia
[2] The Department of Medical Biology, University of Melbourne, Parkville, Victoria 3052, Australia
[3] Medicinal Chemistry and Drug Action, Monash Institute of Pharmaceutical Sciences, Monash University, 381 Royal Parade, Parkville, Victoria 3052, Australia

ABSTRACT

Suppressors of cytokine signaling (SOCS) proteins function as negative regulators of cytokine signaling and are involved in fine tuning the immune response. The structure and role of the SH2 domains and C-terminal SOCS box motifs of the SOCS proteins are well characterized, but the long N-terminal domains of SOCS4–7 remain poorly understood. Here, we present bioinformatic analyses of the N-terminal domains of the mammalian SOCS proteins, which indicate that these domains of SOCS4, 5, 6, and 7 are largely disordered. We have also identified a conserved region of about 70 residues in the N-terminal domains of SOCS4 and 5 that is predicted to be more ordered than the surrounding sequence. The conservation of this region can be traced as far back as lower vertebrates. As conserved regions with increased structural propensity that are located within long disordered regions often contain molecular recognition motifs, we expressed the N-terminal conserved region of mouse SOCS4 for further analysis. This region, mSOCS4$_{86–155}$, has been characterized by circular dichroism and nuclear magnetic resonance spectroscopy, both of which indicate that it is predominantly unstructured in aqueous solution, although it becomes helical in the presence of trifluoroethanol. The high degree of sequence conservation of this region across different species and between SOCS4 and SOCS5 nonetheless implies that it has an important functional role, and presumably this region adopts a more ordered conformation in complex with its partners. The recombinant protein will be a valuable tool in identifying these partners and defining the structures of these complexes.

Proteins 2012; 80:946–957.

Key words: SOCS proteins; cytokine signaling; disordered protein; eukaryotic linear motifs; NMR; CD.

INTRODUCTION

Cytokines form an integral component of many biological responses, regulating fundamental processes such as haematopoiesis, embryonic development, and the innate and adaptive immune responses to pathogenic infection. They trigger cellular responses by binding to specific cell-surface receptors and activating intracellular signal transduction cascades such as the JAK-STAT pathway.[1] A family of cytokine-inducible proteins, termed suppressors of cytokine signaling (SOCS) proteins, function as negative regulators of cytokine signaling.[2–5] SOCS proteins inhibit components of the cytokine signaling cascade via direct binding to the signaling complex or by targeting signal transducers for proteosomal degradation.[6] Studies using gene-targeted mice have revealed important roles for SOCS proteins as key regulators of inflammation and adaptive immunity.[7,8]

The SOCS protein family consists of eight members in mammals (Fig. 1): cytokine-inducible SH2 domain-containing protein (CIS) and SOCS1–7.[2–5] They share a similar domain organization, with a central SH2 domain and a conserved C-terminal SOCS box.[9] In contrast, the N-terminal domains of SOCS proteins vary in length and amino acid sequence, and only SOCS1 and SOCS3 possess a kinase inhibitory region (KIR, 12 residues) immediately upstream of the central SH2 domain.[10]

Additional Supporting Information may be found in the online version of this article
Abbreviations: CD, circular dichroism; ELMs, eukaryotic linear motifs; HSQC, heteronuclear single quantum coherence; IUPs, intrinsically unstructured proteins; JAK, Janus kinase; KIR, kinase inhibitory region; MoRE, molecular recognition element; mSOCS4$_{86–155}$, N-terminal conserved region within mouse SOCS4; PPIs, protein–protein interactions; SH2, Src-homology 2; SOCS proteins, suppressors of cytokine signaling; STAT, signal transducer and activator of transcription; TCEP, tris(2-carboxyethyl)phosphine; TFE, trifluoroethanol.

[#]Zhi-Ping Feng and Indu R. Chandrashekaran contributed equally to this work.
Grant sponsor: National Health and Medical Research Council, Australia; Grant number: 461219; Grant sponsor: NHMRC IRIISS; Grant number: 361646; Grant sponsor: Victorian State Government OIS.
*Correspondence to: R. S. Norton, Medicinal Chemistry and Drug Action, Monash Institute of Pharmaceutical Sciences, Monash University (Parkville campus), 381 Royal Parade, Parkville, Victoria 3052, Australia. E-mail: ray.norton@monash.edu
Received 19 September 2011; Revised 6 November 2011; Accepted 9 November 2011
Published online 16 November 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.23252

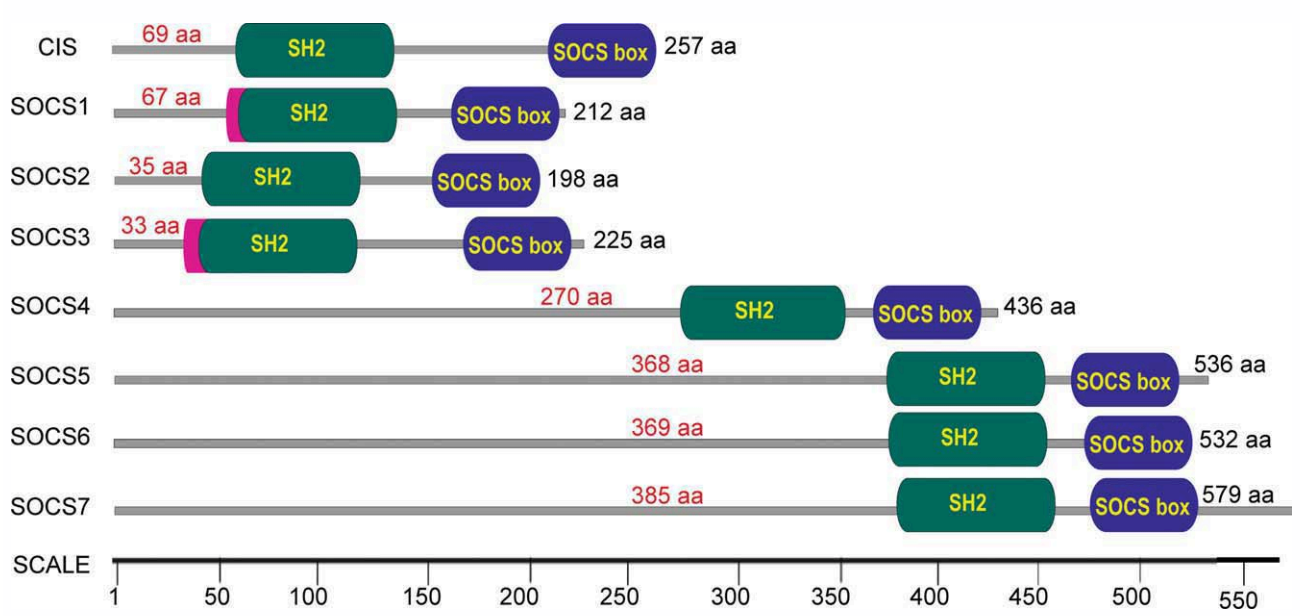

Figure 1

Domain structure of the SOCS protein family as exemplified by mouse proteins. The kinase inhibitory region of SOCS1 and SOCS3 is indicated by a purple bar; the SH2 domain and SOCS box are indicated in green and blue bars, respectively. The numbers of residues in the N-terminal domain and full-length protein are shown in red and black, respectively. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

It has been suggested that SOCS proteins attenuate cytokine signal transduction by binding through their SH2 domains to phosphorylated tyrosine residues on signaling intermediates, such as receptor subunits and JAKs. CIS and SOCS2, for instance, are thought to act by blocking STAT recruitment to phosphorylated tyrosines within the growth hormone receptor. SOCS1 and SOCS3 are induced in response to many cytokines and are thought to inhibit the catalytic activity of JAKs by binding to the kinase domain through both the KIR and SH2 domains. Binding of SOCS to JAK kinases therefore blocks further signaling in a negative feedback loop.[11]

In addition to the inhibitory mechanisms described above, the SOCS box binds to elongins B and C, cullin-5 and the ring protein Rbx-2 to form an E3 ubiquitin ligase complex. SOCS proteins therefore also act as adaptor proteins to recruit an E3 ubiquitin ligase that targets proteins bound to the SH2 domain for ubiquitination and proteasomal degradation.[12]

Structural studies on SOCS proteins have helped elucidate the molecular basis for their regulation of cytokine signaling. The central SH2 domain that determines target specificity and the C-terminal SOCS box that mediates proteosomal degradation of protein targets have been well-characterized structurally.[9,10,13–16] The structure of SOCS3 in complex with a phosphotyrosine-containing peptide from the interleukin-6 (IL-6) receptor signaling subunit gp130[10,13] and the structures of ternary complexes of SOCS2, 4, and 3 associated with elongin B and elongin C[15–17] have also been solved. By contrast, the structure and function of the N-terminal domains of the SOCS proteins remain poorly characterized. It has been suggested that SOCS proteins mediate the ubiquitination of substrate proteins bound to their N-terminal domains.[18] Furthermore, the N-terminal domains of SOCS4 and SOCS5 have been shown to play an important role in mediating interactions with the epidermal growth factor receptor[19,20] and interleukin-4 receptor.[21] However, in most of the structural studies to date, the N-terminal domains of SOCS proteins were removed in case they hampered protein expression or interfered with structure determination. As a result, structural studies on SOCS proteins have provided limited information on the N-terminal domains, even though in the case of SOCS4–7, these domains account for a large proportion of the protein.

It has been well documented that several important classes of proteins involved in cellular signaling and regulation are intrinsically unstructured proteins (IUPs).[22–25] As many as 40–50% of all eukaryotic genes are predicted to encode proteins containing lengthy disordered segments (>40 residues).[26–28] These proteins are usually involved in molecular recognition and assembly (as encountered in signaling pathways), protein modification (e.g., phosphorylation, acetylation, methylation), and entropic chain activities (e.g., linkers, springs and spacers).[29,30] Bioinformatic analyses have indicated that the majority of proteins involved in eukaryotic signal transduction are IUPs, and 79% of human cancer-associated proteins can be classified as IUPs, compared

with 47% of all eukaryotic proteins in the SWISS-PROT database.[31] A recent study showed that neurodegenerative diseases are also caused by aberrant protein–protein interactions (PPIs) involving IUPs.[32]

Although IUPs are disordered in isolation, they often undergo a disorder-to-order transition in which they adopt an ordered conformation upon binding to their biological partners.[23,25,33,34] IUPs thus represent a distinct molecular implementation of the principles of PPI and conformational promiscuity.[35,36] In contrast to globular proteins, IUPs usually use only single continuous segments designated molecular recognition elements (MoRE/MoRF) or linear motifs, whereas the binding sites of ordered proteins are more segmented.[37,38] Linear motifs are short, conserved elements embedded within larger protein segments that function as sites of regulation, and many are post-translationally modified. The significant differences in residue composition and physiochemical properties between IUPs and the linear motifs within them indicate that they are predictable.[37,39,40] As a result, primary sequence analysis with bioinformatics tools provides a viable means of studying such proteins and the PPIs in which they are involved.

In this study, we have analyzed the structural features of the N-terminal domains of the mammalian SOCS proteins. Our results indicate that the N-terminal domains are conserved across mammalian species and are mostly disordered. A more conserved region within the N-terminal domain of SOCS4 and SOCS5 was identified and could be traced back as far as lower vertebrates (although with less conservation than the SH2 domain and SOCS box). To understand the structural and functional roles of this region, a fragment of mouse SOCS4 (residues 86–155) has been expressed in bacteria and characterized by circular dichroism (CD) and nuclear magnetic resonance (NMR) spectroscopy. In addition, the possible functional linear motifs and partners of PPIs of the N-terminal domains of SOCS proteins have been analyzed. Our results provide a rational basis for the design of further experiments directed toward understanding the roles of the N-terminal domains in this important class of regulators of cytokine signaling.

MATERIALS AND METHODS

Bioinformatics

SOCS protein sequences from different species were extracted from the UniProtKB/Swiss-Prot,[41] UniProt Reference Clusters (UniRef),[42] and Ensembl databases. The accession IDs for these proteins are listed in Supporting Information Table S1.

Multiple sequence alignment was performed for the SOCS protein sequences from mammals (eutherians, marsupials, and monotreme) and their homologues from birds, amphibians, and fish with the ClustalW2 program[43] using default parameters. Multiple sequence alignment was also performed for these sequences using the T-Coffee program.[44] The conserved motif was searched for with MEME/MAST (version 3.5.4),[45] and sequence similarity searches were performed with BLAST and HMMER3.[46] The phylogenetic tree was built and bootstrap analyzed with the SeaView package[47] based on a maximal likelihood tree with a JTT distance model and displayed with Figtree [http://tree.bio.ed.ac.uk/software/figtree/]. Structural features were analyzed with VSL2,[48] VL3H,[49] PESTfind,[50] and Composition profiler.[51]

For retrieval of potential SOCS protein interacting partners, STRING (Search Tool for the Retrieval of Interacting Proteins (release7.0) (http://string.embl.de/)[52] was used. The hotspots or linear motifs of PPIs in SOCS proteins and their partners were analyzed with ISIS,[53] ANCHOR,[54] and ELM.[39]

Protein expression and purification

Codon-optimized DNA corresponding to the conserved region (amino acids 86–155) in the N-terminus of mouse SOCS4 (mSOCS4$_{86–155}$) was custom synthesized (Genscript). EcoR1 and thrombin sites were incorporated upstream of mSOCS4$_{86–155}$ gene and a HindIII site was incorporated downstream. This gene was ligated into the pET32a vector (Novagen) via EcoRI and HindIII sites and transformed into *E. coli* BL21-CodonPlus (DE3)-RP cells (Strategene).

mSOCS4$_{86–155}$ was expressed as a fusion protein with a thioredoxin tag in 1 L of Luria-Bertani medium. The cells were induced for 4 h at the logarithmic phase (OD_{600} 0.6–0.8) with 1 m*M* isopropyl β-D-thiogalactopyranoside. The fusion protein, which expressed as soluble protein, was purified initially using chelating Sepharose. One unit of thrombin (Roche) per 10 mg of fusion protein was used to cleave at 4°C for 12 h on a rotating mixer. The polypeptide corresponding to mSOCS4$_{86–155}$ was purified from the cleavage mixture containing thioredoxin and S-tag using cation-exchange chromatography. The purity of mSOCS4$_{86–155}$ was confirmed by SDS-PAGE and analytical RP-HPLC and the molecular mass was determined by LC-MS. $^{13}C/^{15}N$ labeled mSOCS4$_{86–155}$ was expressed in 1 L of M9 medium supplemented with 1 g/L of $^{15}NH_4Cl$ and 2 g/L of ^{13}C-glucose. Expression and purification conditions were as described for the unlabeled mSOCS4$_{86–155}$.

CD spectroscopy

Far-UV CD spectra were acquired on an Aviv 410SF CD spectrophotometer using a 1 mm path-length cell.

CD spectra were recorded with a step size of 0.5 nm, a bandwidth of 1 nm, and an averaging time of 2 s. mSOCS4$_{86–155}$ (10 μ*M*) was buffer exchanged into 10 m*M* phosphate buffer, pH 7.4, and CD measurements were performed. The ability of the polypeptide to adopt a helical structure was assessed by recording a CD spectrum in the presence of 50% trifluoroethanol (TFE) (v/v). CD spectra of phosphate buffer and 50% TFE solution were recorded and subtracted from the corresponding polypeptide spectra, which were fitted using databases of 48 proteins (SDP48) and 43 proteins (SP43), respectively, employing the CONTINLL algorithm included in the CDPro software package.[55]

NMR spectroscopy

NMR experiments were performed at 298 K on Avance-600 (Bruker Biospin) and Inova-600 (Varian) spectrometers equipped with a triple-resonance cryoprobe. All NMR samples were prepared by reconstituting the lyophilized mSOCS4$_{86–155}$ samples (2 mg/mL) in 20 m*M* sodium citrate buffer containing 3 m*M* TCEP, 90% H_2O, and 10% ^2H_2O, pH 4.0. Two-dimensional ^1H-^{15}N HSQC spectra were acquired with 1024 (^1H) and 256 (^{15}N) points and spectral widths of 14 ppm (^1H) and 34 ppm (^{15}N). Spectra were processed using TopSpin (version 3.0, Bruker Biospin) and analyzed using SPARKY (version 3.113).[56]

RESULTS

N-terminal domains of SOCS proteins have high conservation amongst mammalian species

Each of the SOCS protein sequences from mammals was aligned with ClustalW2. The sequence identity between human and mouse (~70 million years evolutionary distance) is >84% and between human and opossum (>170–190 million years evolutionary distance)[57] is >67%. Figure 2 shows a phylogram of SOCS proteins from different species. Clearly, they can be classified into two groups, one formed by CIS, SOCS1–3 and the other by SOCS4–7. In the first group, CIS and SOCS2 can be clustered, as can SOCS1 and SOCS3, whereas in the second group, SOCS4 and 5 cluster separately from SOCS6 and SOCS7.

In addition to the conserved SH2 domain and SOCS box, a longer conserved motif (~70 residues) in the N-terminal domains of SOCS4 and SOCS5 was detected with MEME/MAST,[45] and the alignment using ClustalW2 is shown in Figure 3. Multiple sequence alignments were also performed for SOCS4 and SOCS5 proteins using T-Coffee (Supporting Information Fig. S1). The alignments generated by this program are very similar to those generated by ClustalW2, with differences being limited to the nonconserved regions. The conserved motif in the N-terminal domain is unique in SOCS4, SOCS5, SOCS5a, and SOCS9 from fish[58] according to a search with a hidden Markov model built from the domain sequences of mammalian species against UniProtKB/Swiss-Prot database using the HMMER3 package with an *E*-value <0.006 and sequence bit score >13. This region is highly conserved among proteins from mammals to lower vertebrate species (birds, amphibians, fish, and lizards), except for a low complexity fragment insertion in the proteins from fish.

N-terminal domains of the SOCS proteins are largely disordered

Given the conservation of the N-terminal domains of SOCS proteins from mammalian organisms, the structural features can be exemplified by considering the mouse and human proteins. Sequence analysis with the disorder predictors VSL2,[48] VL3H[49] indicates that these domains are mostly disordered (Fig. 4). The amino acid compositions (Supporting Information Fig. S2) indicate that the N-terminal domains of the SOCS proteins are enriched in amino acids with higher flexible indices[59] and depleted in hydrophobic amino acids. The significant differences ($P \leq 0.05$, with 10,000 bootstrap iterations) in the N-terminal domain, relative to the SH2 domain and SOCS box, reflect the abundance of polar and charged residues and the dearth of hydrophobic residues. In addition, low complexity region(s) exist in the N-terminal domains of SOCS1 and SOCS4–7. These results all support the outcomes of the disordered predictions.

Although the N-terminal domains are mostly disordered, some short, ordered regions are predicted in both human and mouse SOCS4, 5, and 6 (Fig. 4). One of these predicted ordered segments occurs within the N-terminal conserved regions of SOCS4 and SOCS5; these conserved regions have very similar predicted scores and both contain a predicted structured region of 15–20 residues. Secondary structure prediction with Jpred[60] suggests that there are three short helices (mSOCS4: $_{88}$LKQKLQDAVG$_{97}$, $_{120}$HISELM$_{125}$, and $_{136}$DLAFRWHFIKR$_{146}$; mSOCS5: $_{177}$LRQRLQDTVG$_{186}$, $_{207}$KIHLSELML$_{215}$, and $_{225}$DLAQKWHLIKQ$_{235}$) within these conserved regions (Supporting Information Fig. S3). The possible ordered region in hSOCS6 is within a less conserved region around residue 200, whereas in mSOCS6, this region has a comparatively higher disordered score (Fig. 4).

Stretches of short linear motifs with ordered propensity within long disordered regions, such as the conserved region in the N-terminal domain of SOCS4 and SOCS5, often represent molecular recognition motifs (MoREs or ELMs) that are likely to be involved in PPIs or scaffolding functions.[25,61] We have therefore expressed this region of mouse SOCS4 (residues 86–155) in bacteria in order to characterize its conformation in solution by CD and NMR spectroscopy.

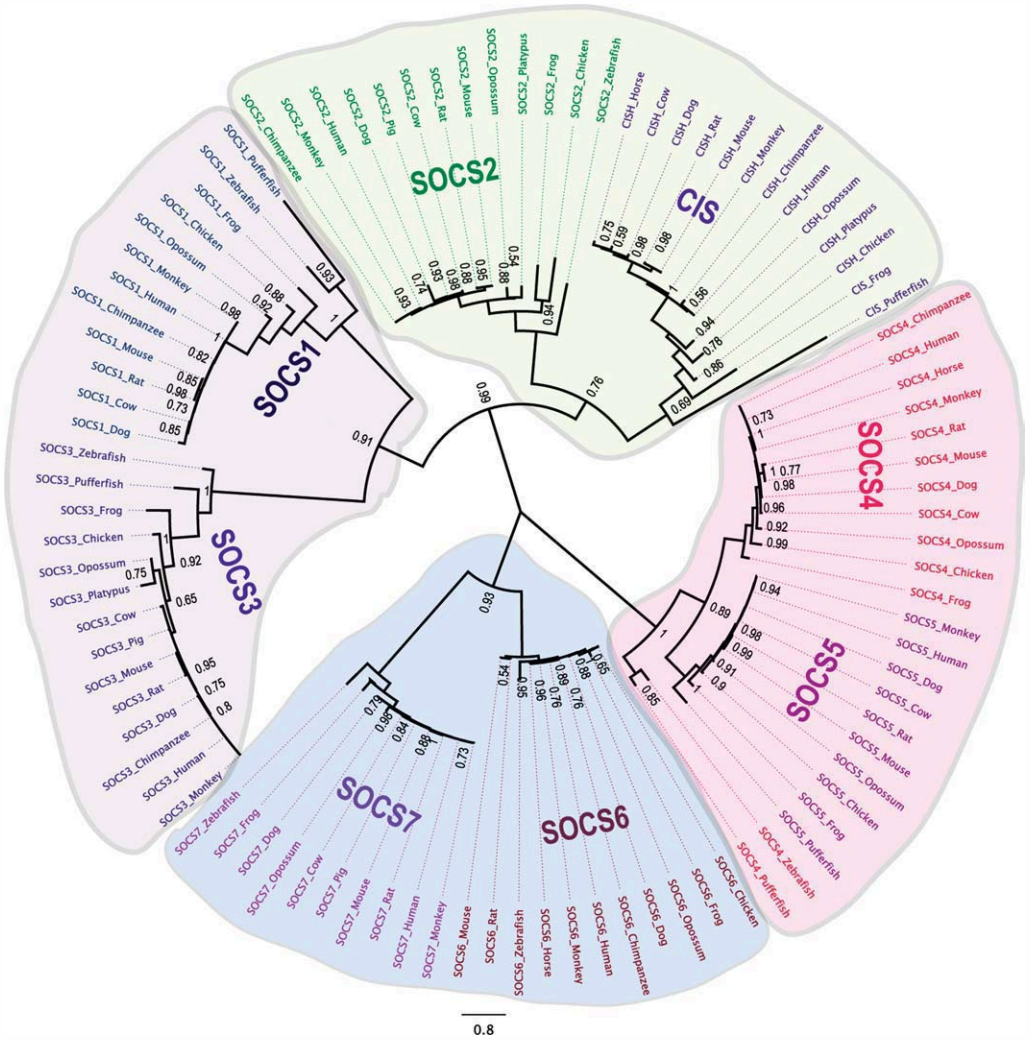

Figure 2

Phylogeny of SOCS proteins derived from a multiple protein sequence alignment of 96 SOCS protein sequences with ClustalW2.[43] The maximal likelihood tree was built with a JTT distance model, analyzed using the SeaView package[47] and displayed with FigTree software (http://tree.bio.ed.ac.uk/software/figtree/). The numbers in the figure indicate percentage support in bootstrap analyses (1000 replicates); only numbers >50% are shown. The edge length is proportional to the distance between the sequences, which is a reflection of substitution of residues over the course of evolution. The sequences in a cluster with very short edges have high sequence similarity scores in multiple sequence alignment, such as in the case of SOCS 4 and 5. The protein IDs of the SOCS proteins are listed in Supporting Information Table S1. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Expression and purification of the N-terminal conserved region in mSOCS4

mSOCS4$_{86–155}$ was expressed as a thioredoxin fusion protein in *E. coli* in rich medium with a yield of around 0.1 g/L of bacterial culture. The thioredoxin fusion protein was purified from the cell lysate using an initial purification step with chelating Sepharose, followed by cleavage with thrombin. mSOCS4$_{86–155}$ was purified from the thrombin cleavage mixture by cation-exchange chromatography (Supporting Information Fig. S4). After thrombin cleavage and ion-exchange chromatography, around 10 mg of purified mSOCS4$_{86–155}$ was isolated. After cleavage with thrombin, two additional residues (Gly and Ser) remained at the N-terminal end of the sequence, resulting in a total length of 72 residues. The purified mSOCS4$_{86–155}$ migrates as a single band of apparent molecular mass ~8 kDa under reducing and nonreducing conditions on an SDS-PAGE gel (Supporting Information Fig. S5). The purity of the polypeptide was also analyzed by analytical RP-HPLC and mass spectrometry (Supporting Information Fig. S6). mSOCS4$_{86–155}$ has a molecular mass of 8195.7 Da as determined by ESI mass spectrometry, corresponding closely to its theoretical molecular mass of 8196.5 Da. mSOCS4$_{86–155}$ was found to be monomeric in solution on analysis by size-exclusion chromatography (data not shown).

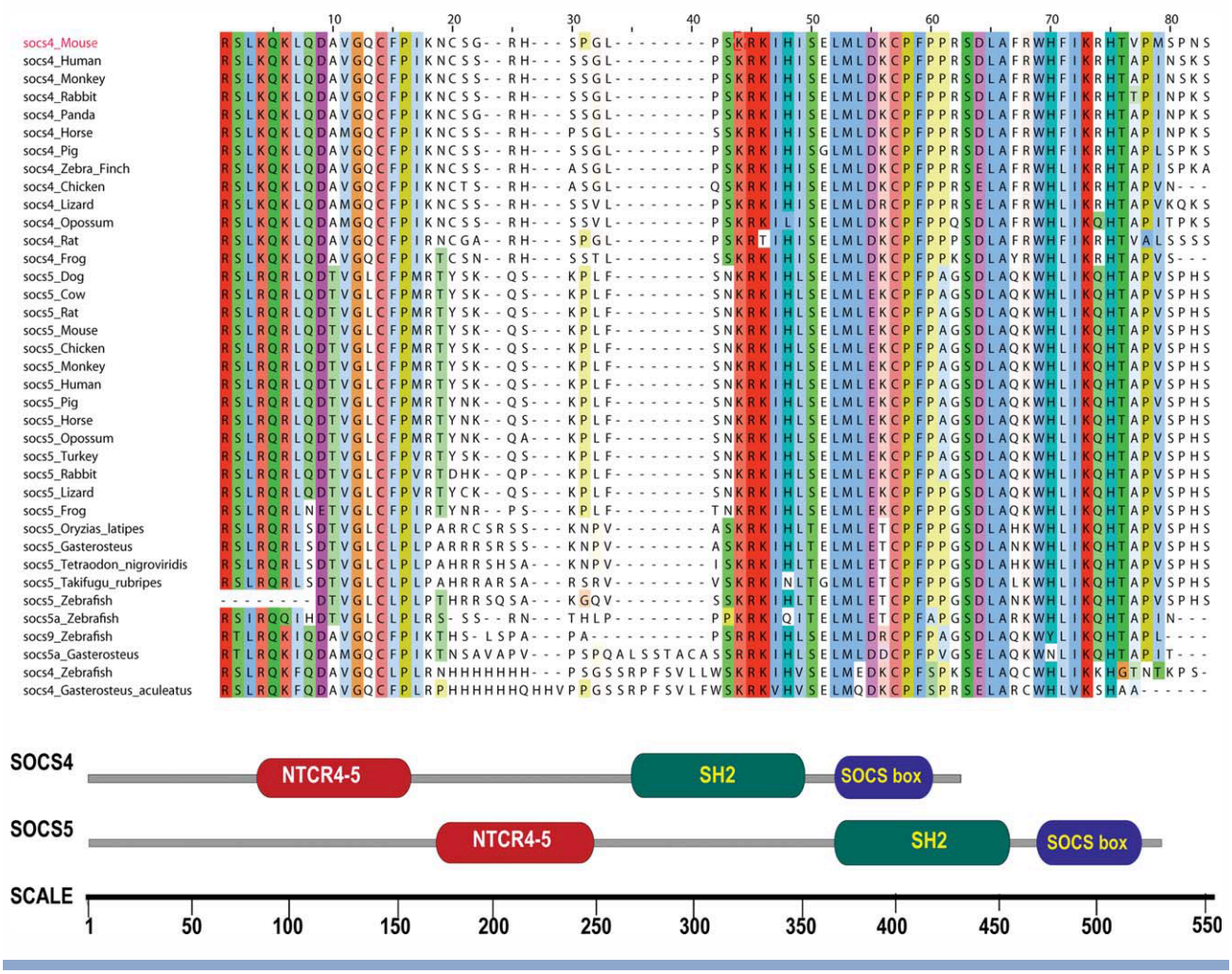

Figure 3

Domain structure of SOCS4 and SOCS5. The N-terminal conserved region (labeled NTCR4-5) is indicated by a red bar, and the sequence alignment is shown. The percentage identities in this conserved region are as follows: hSOCS4 versus hSOCS5, 58%; mSOCS4 versus mSOCS5, 61%; mSOCS4 versus hSOCS4, 90%; and mSOCS5 versus hSOCS5, 100%. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

For NMR characterization, $^{13}C/^{15}N$-labeled $mSOCS4_{86–155}$ was expressed in minimal medium supplemented with 1 g of $^{15}NH_4Cl$ and 2 g of ^{13}C-glucose per litre of culture. Protein purification was performed as described for the unlabeled protein, although only 20 mg of $^{13}C/^{15}N$-labeled thioredoxin fusion protein was obtained from 1 L of bacterial culture. Two milligrams of purified $^{13}C/^{15}N$ $mSOCS4_{86–155}$ was isolated following thrombin cleavage and ion-exchange chromatography.

Structural studies on the N-terminal conserved region of SOCS4

The CD spectrum of $mSOCS4_{86–155}$ in phosphate buffer was indicative of a largely random coil structure, with an ellipticity minimum between 200 and 210 nm (Fig. 5). However, a broad but weak negative ellipticity was observed between 215 and 235 nm, indicative of a small population of residual helical or β-strand structure. Nonlinear least-squares analyses yielded good fits (RMSD < 0.22) to ~6% helix, 14% β-strand, 13% turn, and 67% disordered structure. Given the indication of some residual structure being present in $mSOCS4_{86–155}$, we tested the ability of TFE to induce a helical structure. The CD spectrum in the presence of 50% TFE clearly indicates that the polypeptide adopts a helical structure under these conditions (Fig. 5). Nonlinear least-squares analyses of the data yielded good fits (RMSD < 0.26) to ~30% helix, 22% β-strand, 24% turn, and 24% disordered structure.

One-dimensional NMR spectra of $mSOCS4_{86–155}$ displayed limited chemical shift dispersion and considerable peak overlap, confirming that this region of SOCS4 is predominantly unstructured. NMR spectra were recorded at different temperatures and pH values in the range

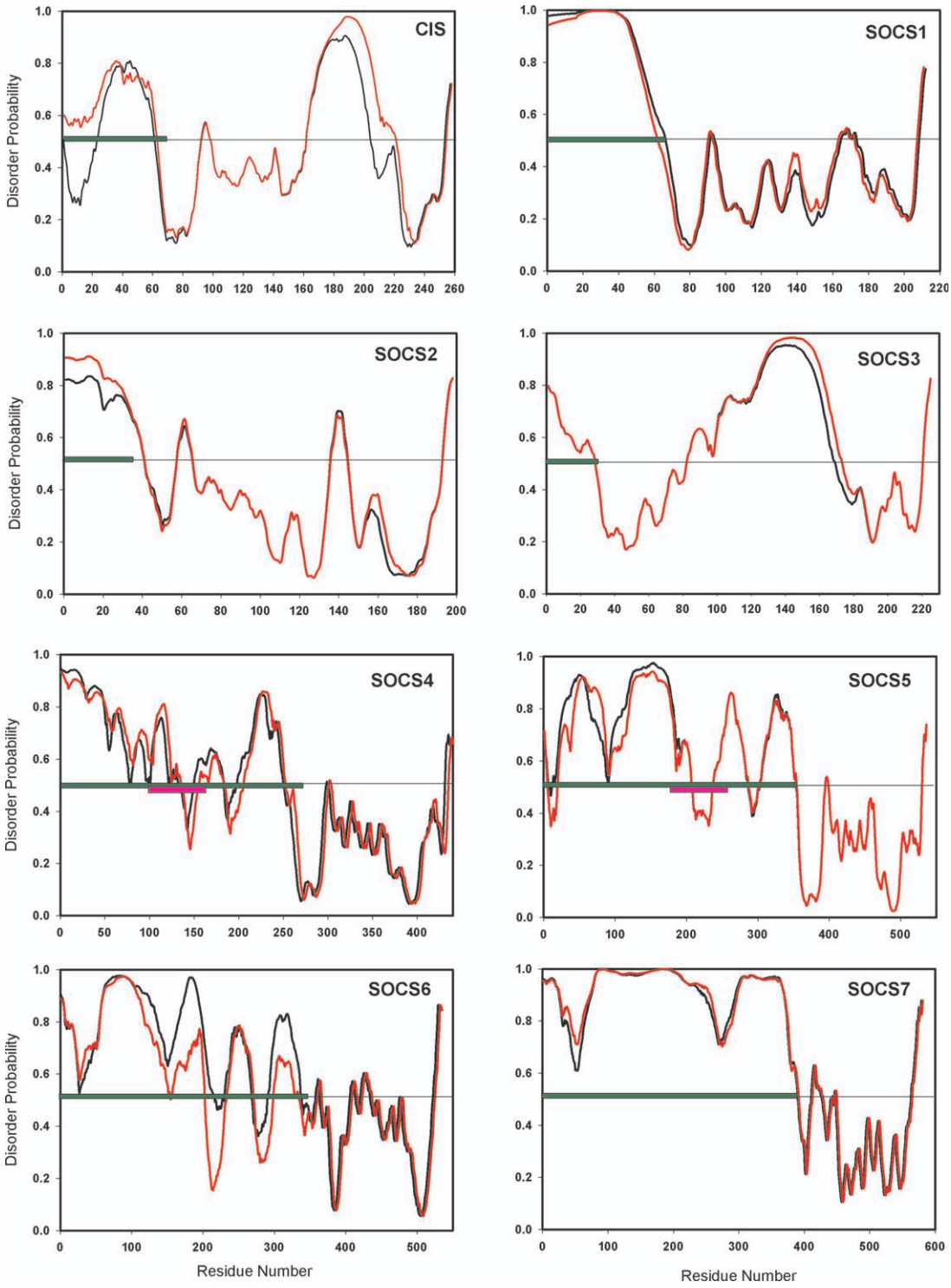

Figure 4

Disordered scores predicted with VSL2 for SOCS proteins from mouse (black) and human (red). The N-terminal domain of each protein is highlighted with a green horizontal line at the threshold 0.5 and the conserved region in SOCS4 and 5 is highlighted with a pink horizontal line at the same threshold.

3.6-6.5 in an effort to improve spectral dispersion. pH 4.0 was chosen for the NMR experiments to minimize the exchange of labile protons that occurs at higher pH. A ^{1}H-^{15}N HSQC spectrum of uniformly $^{13}C/^{15}N$-labeled $mSOCS4_{86–155}$ at 25°C and pH 4.0 is shown in Figure 6. Amide ^{1}H chemical shifts were dispersed over a narrow

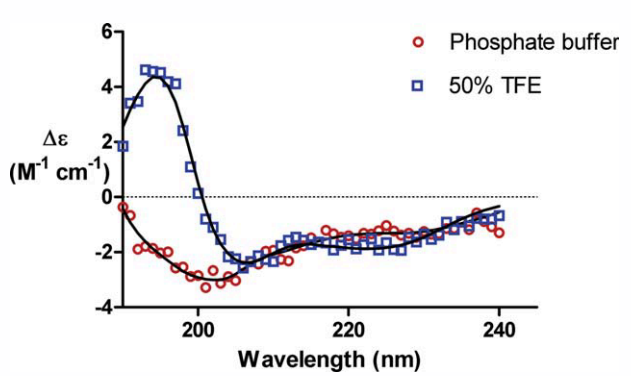

Figure 5

Far-UV CD spectra of mSOCS4$_{86–155}$ in aqueous solution (○) and 50% TFE (□). Change in ellipticity (M^{-1} cm^{-1}) is plotted as a function of wavelength (nm). The nonlinear least-squares best fits employing the CDPro software package with SDP48 and SP43 reference sets and CONTINLL algorithm are shown as continuous lines. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

range, as expected for a largely disordered polypeptide. Most of the peaks were relatively sharp and intense, which again reflects the flexible nature of this polypeptide in solution. However, the HSQC spectrum displays more peaks than anticipated for a 72-residue protein, several of which were of lower intensity. It is possible that some of these additional peaks arise from minor conformations of the protein in slow chemical exchange on the NMR time scale. As this region contains several Pro residues (Fig. 3), conformers with *cis* peptide bonds preceding one or more Pro residues presumably account for some of these minor peaks.

N-terminal domains of the SOCS proteins contain numerous eukaryotic linear motifs

Possible linear motifs or functional binding sites were searched for in a known Eukaryotic Linear Motif (ELM) database using a regular expression search followed by globular domain filtering, structural filtering, and context filtering.[39] As short patterns applied to proteins are usually not statistically significant, most matches shown are more likely to be false positives than true matches. Nonetheless, they are useful as guides to future experimentation. The results for human and mouse SOCS4 and 5 are shown in Supporting Information Figure S7. The N-terminal domains contain the majority of the ELMs in these proteins, except for SOCS3, where most of them reside in the PEST motif, which is rich in proline, glutamic

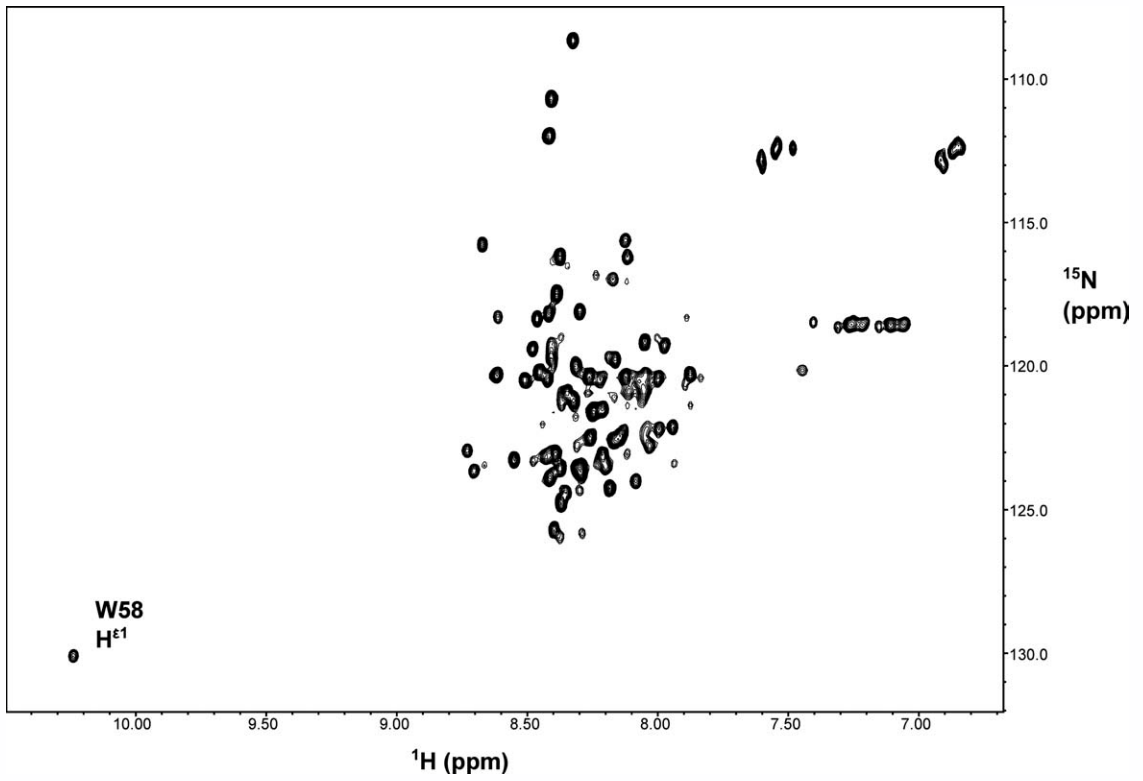

Figure 6

^{1}H-^{15}N HSQC NMR spectrum of mSOCS4$_{86–155}$. The spectrum was recorded in 20 m*M* citrate buffer, pH 4.0, containing 3 m*M* TCEP at 25°C on a 600 MHz spectrometer.

acid, serine, and threonine, between the SH2 domain and SOCS box. This is also supported by the results of a hotspot search with ISIS and ANCHOR; among the 22 ELMs found in mSOCS4, 20 localize to the N-terminal domain, whilst in hSOCS4, 20 of the 23 ELMs localize to the N-terminal domain (Supporting Information Fig. S7).

The number and type of the ELMs can change with a single point mutation in the sequence. For example, the FHA phosphothreonine binding motifs (LIG_FHA_1)[62] in hSOCS4 ($_{149}$RHTAPIN$_{155}$ and $_{209}$PMTGSVM$_{215}$) are generated by the single point mutations M154I and M214V in mSOCS4, and the STAT3 SH2 domain binding motif (LIG_SH2_STAT3)[63] found in hSOCS4 ($_{264}$YHTQ$_{267}$) arises from the F264Y mutation (Supporting Information Fig. S7). The single point mutation F264Y also creates a motif ($_{261}$PPKY$_{264}$, LIG_WW_1) that can be recognized by group I WW domains. Therefore, although most SOCS proteins, including their N-terminal domains, are conserved between mouse and human (>86% similarity), the number and type of ELMs found from the current ELM database are different, and in most cases, hSOCS proteins contain a greater number.

DISCUSSION

Disordered N-terminal domains of SOCS proteins and their function

Little is known experimentally about the 3D structures of the N-terminal domains of the SOCS proteins, with the exception of a short extension of mSOCS3 (residues 15–28 including residues 22–28 of the KIR), which has been shown to be disordered in the context of the SH2 domain bound to the gp130 phosphopeptide.[13] Sequence analyses indicate that the N-terminal domains of all SOCS proteins are likely to be disordered in isolation. This is further supported by evidence from our spectroscopic studies on the conserved region found in the N-terminal domain of SOCS4. NMR spectra of mSOCS4$_{86–155}$ displayed limited chemical shift dispersion and considerable peak overlap, showing that this region of SOCS4 is predominantly unstructured in solution. However, this region may contain some residual structure, and results from our CD studies in the presence of 50% TFE point to its helical propensity. Further characterization of the conformational preferences of this polypeptide in solution will require more detailed NMR studies, including a full set of sequence-specific resonance assignments.

Unlike globular proteins, where the interaction interfaces are often composed of conserved polar or charged residues that provide critical anchoring interactions, surrounded and sealed from hydration by more variable shielding residues, the hydrophobic amino acids in disordered regions counter the tendency to collapse into some structure by their special amino acid composition, keeping them exposed for interaction with their partners.[40,64] Among the ELMs found in the N-terminal domains of SOCS proteins, Arg is common in all the motifs, Leu and Pro are enriched in many ligand-binding motifs, and Cys, Ser, and Thr are more abundant in post-translational modification motifs. ELMs in the N-terminal domains of SOCS proteins appear to be mostly depleted of hydrophobic residues, with the exception of Leu, Cys, Phe, and Ile, which appear in a putative substrate-recognition motif that interacts with cyclin and increases phosphorylation by cyclin-dependent kinases (LIG_CYCLN1: [RK]L [FYLIVMP]).[65] This motif is found in mouse CIS, SOCS2, SOCS4, SOCS5, SOCS6, and SOCS7 (Supporting Information Fig. S7). It has been established that IUPs are abundant in all important cellular processes, including transport, transcription, translation metabolism, and signaling,[66,67] where multiprotein complexes are usually key players. A recent study further indicates that structural disorder promotes assembly of protein complexes, and, in the case of proteins regulated by phosphorylation, nearby disordered regions could help display the phosphorylation sites to both kinases and phosphatases.[67] These common disordered properties may also apply to the N-terminal domains of SOCS proteins and contribute to their regulatory roles by providing a large interaction surface and enabling rapid association and dissociation rates.

Roles of N-terminal domains in SOCS subcellular localization

While those SOCS proteins with short N-terminal domains (CIS, SOCS1, SOCS2, and SOCS3) are localized primarily in the cytoplasm, they were recently reported to also localize to the nucleus, with localization regulated by the cellular conditions.[68]

SOCS6 has been found in both the cytoplasm and the nucleus, and the N-terminal domain of SOCS6 (the first 210 residues) is capable of transporting proteins into the nucleus.[69] Our ELM search indicates that SOCS6 contains a tyrosine-based sorting signal ($_{355}$YDSV$_{358}$) responsible for interaction with the μ subunit of adaptor protein complex. Other sorting signals may exist in the long disordered N-terminal domain of SOCS6.

SOCS7 has been found in the cytoplasm and nucleus and at the cell membrane.[70] The N-terminal domain of SOCS7 contains a putative nuclear export signal (NES, $_{56}$LEAQLAALGL$_{65}$) and two sorting signals ($_{284}$ETVSLV$_{289}$, $_{312}$RSLSLL$_{317}$), which are also found in the cytoplasmic juxta-membrane region of type I transmembrane proteins, and can interact with adaptor protein complexes.[71] The N-terminal domain of SOCS7 has been found to bind a multifunctional adaptor protein NCK and transport NCK into the nucleus;[72] nuclear

accumulation of NCK is necessary for activation of the tumour suppressor p53.[72]

The ELM search results indicate that both human and mouse SOCS4 contain C-extended and N-extended versions of the monopartite variant of the classical basic nuclear localization signal, but human and mouse SOCS5 only contain a bipartite variant of this signal. At present, there is no experimental evidence on the subcellular localization of SOCS4 and SOCS5.

CONCLUSIONS

Bioinformatic analysis indicates that the N-terminal domains of the mammalian SOCS proteins are conserved and disordered. These disordered domains are likely to play important roles in the recognition and regulation of multiprotein complexes and to promote assembly of protein complexes in the JAK-STAT signaling pathway. The highly interconnected roles of these disordered domains in networks of PPIs presumably contribute to their sequence conservation during evolution, especially for the conserved region in SOCS4 and 5 identified here.

Several studies have highlighted the importance of disorder in proteins involved in various diseases, such as cancer,[31] cardiovascular diseases,[73] and conformational diseases such as Alzheimer's disease.[74] A detailed understanding of the conformation and interactions of the disordered N-terminal domains of the SOCS proteins will provide opportunities to develop inhibitors against their PPIs, which at the very least should prove to be valuable probes of the biological functions of these domains.

The high degree of sequence conservation in the N-terminal conserved region of SOCS4 and 5 across all species suggests that this region is likely to play an important functional role in the regulation of cytokine signaling by these proteins. Although this sequence is unstructured in solution, the availability of pure, recombinant material now provides a basis for further biological studies on its role in the regulation of cytokine signaling by SOCS4 and 5.

ACKNOWLEDGMENTS

The authors thank Drs. Jeff Babon and Chris MacRaild for helpful discussions. S.E.N. and R.S.N. acknowledge fellowship support from the NHMRC.

REFERENCES

1. Ihle JN, Kerr IM. Jaks and Stats in signaling by the cytokine receptor superfamily. Trends Genet 1995;11:69–74.
2. Starr R, Willson TA, Viney EM, Murray LJ, Rayner JR, Jenkins BJ, Gonda TJ, Alexander WS, Metcalf D, Nicola NA, Hilton DJ. A family of cytokine-inducible inhibitors of signalling. Nature 1997;387:917–921.
3. Endo TA, Masuhara M, Yokouchi M, Suzuki R, Sakamoto H, Mitsui K, Matsumoto A, Tanimura S, Ohtsubo M, Misawa H, Miyazaki T, Leonor N, Taniguchi T, Fujita T, Kanakura Y, Komiya S, Yoshimura A. A new protein containing an SH2 domain that inhibits JAK kinases. Nature 1997;387:921–924.
4. Ilangumaran S, Ramanathan S, Rottapel R. Regulation of the immune system by SOCS family adaptor proteins. Semin Immunol 2004;16:351–365.
5. Nicholson SE, Hilton DJ. The SOCS proteins: a new family of negative regulators of signal transduction. J Leukoc Biol 1998;63:665–668.
6. Alexander WS, Hilton DJ. The role of suppressors of cytokine signaling (SOCS) proteins in regulation of the immune response. Annu Rev Immunol 2004;22:503–529.
7. Alexander WS, Starr R, Fenner JE, Scott CL, Handman E, Sprigg NS, Corbin JE, Cornish AL, Darwiche R, Owczarek CM, Kay TW, Nicola NA, Hertzog PJ, Metcalf D, Hilton DJ. SOCS1 is a critical inhibitor of interferon-γ signaling and prevents the potentially fatal neonatal actions of this cytokine. Cell 1999;98:597–608.
8. Croker BA, Krebs DL, Zhang JG, Wormald S, Willson TA, Stanley EG, Robb L, Greenhalgh CJ, Forster I, Clausen BE, Nicola NA, Metcalf D, Hilton DJ, Roberts AW, Alexander WS. SOCS3 negatively regulates IL-6 signaling in vivo. Nat Immunol 2003;4:540–545.
9. Hilton DJ, Richardson RT, Alexander WS, Viney EM, Willson TA, Sprigg NS, Starr R, Nicholson SE, Metcalf D, Nicola NA. Twenty proteins containing a C-terminal SOCS box form five structural classes. Proc Natl Acad Sci USA 1998;95:114–119.
10. Babon JJ, McManus EJ, Yao S, DeSouza DP, Mielke LA, Sprigg NS, Willson TA, Hilton DJ, Nicola NA, Baca M, Nicholson SE, Norton RS. The structure of SOCS3 reveals the basis of the extended SH2 domain function and identifies an unstructured insertion that regulates stability. Mol Cell 2006;22:205–216.
11. Elliott J, Johnston JA. SOCS: role in inflammation, allergy and homeostasis. Trends Immunol 2004;25:434–440.
12. Babon JJ, Sabo JK, Zhang JG, Nicola NA, Norton RS. The SOCS box encodes a hierarchy of affinities for Cullin5: implications for ubiquitin ligase formation and cytokine signalling suppression. J Mol Biol 2009;387:162–174.
13. Bergamin E, Wu J, Hubbard SR. Structural basis for phosphotyrosine recognition by suppressor of cytokine signaling-3. Structure 2006;14:1285–1292.
14. Nicholson SE, Willson TA, Farley A, Starr R, Zhang JG, Baca M, Alexander WS, Metcalf D, Hilton DJ, Nicola NA. Mutational analyses of the SOCS proteins suggest a dual domain requirement but distinct mechanisms for inhibition of LIF and IL-6 signal transduction. EMBO J 1999;18:375–385.
15. Bullock AN, Debreczeni JE, Edwards AM, Sundstrom M, Knapp S. Crystal structure of the SOCS2-elongin C-elongin B complex defines a prototypical SOCS box ubiquitin ligase. Proc Natl Acad Sci USA 2006;103:7637–7642.
16. Bullock AN, Rodriguez MC, Debreczeni JE, Songyang Z, Knapp S. Structure of the SOCS4-ElonginB/C complex reveals a distinct SOCS box interface and the molecular basis for SOCS-dependent EGFR degradation. Structure 2007;15:1493–1504.
17. Babon JJ, Sabo JK, Soetopo A, Yao S, Bailey MF, Zhang JG, Nicola NA, Norton RS. The SOCS box domain of SOCS3: structure and interaction with the elonginBC-cullin5 ubiquitin ligase. J Mol Biol 2008;381:928–940.
18. Yoshimura A, Naka T, Kubo M. SOCS proteins, cytokine signalling and immune regulation. Nat Rev Immunol 2007;7:454–465.
19. Nicholson SE, Metcalf D, Sprigg NS, Columbus R, Walker F, Silva A, Cary D, Willson TA, Zhang JG, Hilton DJ, Alexander WS, Nicola NA. Suppressor of cytokine signaling (SOCS)-5 is a potential negative regulator of epidermal growth factor signaling. Proc Natl Acad Sci USA 2005;102:2328–2333.
20. Segatto O, Anastasi S, Alema S. Regulation of epidermal growth factor receptor signalling by inducible feedback inhibitors. J Cell Sci 2011;124:1785–1793.
21. Seki Y, Hayashi K, Matsumoto A, Seki N, Tsukada J, Ransom J, Naka T, Kishimoto T, Yoshimura A, Kubo M. Expression of the suppressor of cytokine signaling-5 (SOCS5) negatively regulates IL-

4-dependent STAT6 activation and Th2 differentiation. Proc Natl Acad Sci USA 2002;99:13003–13008.
22. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. Flexible nets. The roles of intrinsic disorder in protein interaction networks. FEBS J 2005;272:5129–5148.
23. Dyson H, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6:197–208.
24. Tompa P, Szasz C, Buday L. Structural disorder throws new light on moonlighting. Trends Biochem Sci 2005;30:484–489.
25. Uversky VN, Oldfield CJ, Dunker AK. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. J Mol Recognit 2005;18:343–384.
26. Feng ZP, Zhang X, Han P, Arora N, Anders RF, Norton RS. Abundance of intrinsically unstructured proteins in *P. falciparum* and other apicomplexan parasite proteomes. Mol Biochem Parasitol 2006;150:256–267.
27. Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. Structure 2003;11:1453–1459.
28. Liu J, Tan H, Rost B. Loopy proteins appear conserved in evolution. J Mol Biol 2002;322:53–64.
29. Mohan A, Sullivan WJ, Jr, Radivojac P, Dunker AK, Uversky VN. Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes. Mol Biosyst 2008;4:328–340.
30. Schreiber G, Keating AE. Protein binding specificity versus promiscuity. Curr Opin Struct Biol 2011;21:50–61.
31. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol 2002;323:573–584.
32. Das S, Mukhopadhyay D. Intrinsically unstructured proteins and neurodegenerative diseases: conformational promiscuity at its best. IUBMB Life 2011;63:478–488.
33. Sugase K, Dyson HJ, Wright PE. Mechanism of coupled folding and binding of an intrinsically disordered protein. Nature 2007;447:1021–1025.
34. Uversky VN, Dunker AK. Controlled chaos. Science 2008;322:1340–1341.
35. Dosztanyi Z, Sandor M, Tompa P, Simon I. Prediction of protein disorder at the domain level. Curr Protein Pept Sci 2007;8:161–171.
36. Fuxreiter M, Tompa P, Simon I. Local structural disorder imparts plasticity on linear motifs. Bioinformatics 2007;23:950–956.
37. Vacic V, Oldfield CJ, Mohan A, Radivojac P, Cortese MS, Uversky VN, Dunker AK. Characterization of molecular recognition features, MoRFs, and their binding partners. J Proteome Res 2007;6:2351–2366.
38. Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, Dunker AK. Coupled folding and binding with α-helix-forming molecular recognition elements. Biochemistry 2005;44:12454–12470.
39. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, Ferre F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Kuster B, Helmer-Citterich M, Hunter WN, Aasland R, Gibson TJ. ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 2003;31:3625–3630.
40. Meszaros B, Simon I, Dosztanyi Z. Prediction of protein binding regions in disordered proteins. PLoS Comput Biol 2009;5:e1000376.
41. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31:365–370.
42. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 2007;23:1282–1288.
43. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and Clustal X version 2.0. Bioinformatics 2007;23:2947–2948.
44. Di Tommaso P, Moretti S, Xenarios I, Orobitg M, Montanyola A, Chang JM, Taly JF, Notredame C. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res 2011;39:W13–W17.
45. Bailey TL, Gribskov M. Methods and statistics for combining motif match scores. J Comput Biol 1998;5:211–221.
46. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14:755–763.
47. Gouy M, Guindon S, Gascuel O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 2010;27:221–224.
48. Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK. Exploiting heterogeneous sequence properties improves prediction of protein disorder. Proteins 2005;61 Suppl 7:176–182.
49. Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown C, Dunker AK. Predicting intrinsic disorder from amino acid sequence. Proteins 2003;53:566–572.
50. Rogers S, Wells R, Rechsteiner M. Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. Science 1986;234:364–368.
51. Vacic V, Uversky VN, Dunker AK, Lonardi S. Composition Profiler: a tool for discovery and visualization of amino acid composition differences. BMC Bioinformatics 2007;8:211.
52. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P. STRING 7-recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 2007;35:D358–D362.
53. Ofran Y, Rost B. ISIS: interaction sites identified from sequence. Bioinformatics 2007;23:e13–e16.
54. Dosztanyi Z, Meszaros B, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics 2009;25:2745–2746.
55. Sreerama N, Woody RW. Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. Anal Biochem 2000;287:252–260.
56. Goddard TD, Kneller DG. SPARKY 3. San Francisco: University of California.
57. Gentles AJ, Wakefield MJ, Kohany O, Gu W, Batzer MA, Pollock DD, Jurka J. Evolutionary dynamics of transposable elements in the short-tailed opossum *Monodelphis domestica*. Genome Res 2007;17:992–1004.
58. Jin HJ, Shao JZ, Xiang LX, Wang H, Sun LL. Global identification and comparative analysis of SOCS genes in fish: insights into the molecular evolution of SOCS family. Mol Immunol 2008;45:1258–1268.
59. Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions. Proteins 1994;19:141–149.
60. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 2000;40:502–511.
61. Cortese MS, Uversky VN, Keith Dunker A. Intrinsic disorder in scaffold proteins: getting more from less. Prog Biophys Mol Biol 2008;98:85–106.
62. Hofmann K, Bucher P. The FHA domain: a putative nuclear signalling domain found in protein kinases and transcription factors. Trends Biochem Sci 1995;20:347–349.
63. Stahl N, Farruggella TJ, Boulton TG, Zhong Z, Darnell JE, Jr, Yancopoulos GD. Choice of STATs and other substrates specified by modular tyrosine-based motifs in cytokine receptors. Science 1995;267:1349–1353.
64. Meszaros B, Tompa P, Simon I, Dosztanyi Z. Molecular principles of the interactions of disordered proteins. J Mol Biol 2007;372:549–561.
65. Takeda DY, Wohlschlegel JA, Dutta A. A bipartite substrate recognition motif for cyclin-dependent kinases. J Biol Chem 2001;276:1993–1997.
66. Devos D, Russell RB. A more complete, complexed and structured interactome. Curr Opin Struct Biol 2007;17:370–377.

67. Hegyi H, Schad E, Tompa P. Structural disorder promotes assembly of protein complexes. BMC Struct Biol 2007;7:65.
68. Lee KH, Moon KJ, Kim HS, Yoo BC, Park S, Lee H, Kwon S, Lee ES, Yoon S. Increased cytoplasmic levels of CIS, SOCS1, SOCS2, or SOCS3 are required for nuclear translocation. FEBS Lett 2008;582:2319–2324.
69. Hwang MN, Min CH, Kim HS, Lee H, Yoon KA, Park SY, Lee ES, Yoon S. The nuclear localization of SOCS6 requires the N-terminal region and negatively regulates Stat3 protein levels. Biochem Biophys Res Commun 2007;360:333–338.
70. Martens N, Wery M, Wang P, Braet F, Gertler A, Hooghe R, Vandenhaute J, Hooghe-Peters EL. The suppressor of cytokine signaling (SOCS)-7 interacts with the actin cytoskeleton through vinexin. Exp Cell Res 2004;298:239–248.
71. Craig HM, Pandori MW, Guatelli JC. Interaction of HIV-1 Nef with the cellular dileucine-based sorting pathway is required for CD4 down-regulation and optimal viral infectivity. Proc Natl Acad Sci USA 1998;95:11229–11234.
72. Kremer BE, Adang LA, Macara IG. Septins regulate actin organization and cell-cycle arrest through nuclear accumulation of NCK mediated by SOCS7. Cell 2007;130:837–850.
73. Cheng Y, LeGall T, Oldfield CJ, Dunker AK, Uversky VN. Abundance of intrinsic disorder in protein associated with cardiovascular disease. Biochemistry 2006;45:10448–10460.
74. Csizmok V, Dosztanyi Z, Simon I, Tompa P. Towards proteomic approaches for the identification of structural disorder. Curr Protein Pept Sci 2007;8:173–179.