Solution Structure of the R3H Domain from Human Sμbp-2

Edvards Liepinsh¹, Ainars Leonchiks², Anatoly Sharipo²
Laurent Guignard¹ and Gottfried Otting¹,3*

1Department of Medical Biochemistry and Biophysics
Karolinska Institute, S-17177
Stockholm, Sweden
2Biomedical Research and Study Centre, University of Latvia, LV-1067 Riga, Latvia
3Research School of Chemistry
Australian National University
Canberra, ACT 0200, Australia

The R3H domain is a conserved sequence motif, identified in over 100 proteins, that is thought to be involved in polynucleotide-binding, including DNA, RNA and single-stranded DNA. In this work the 3D structure of the R3H domain from human Sμbp-2 was determined by NMR spectroscopy. It is the first 3D structure determination of an R3H domain. The fold presents a small motif, consisting of a three-stranded antiparallel β-sheet and two α-helices, which is related to the structures of the YhhP protein and the C-terminal domain of the translational initiation factor IF3. The similarities are non-trivial, as the amino acid identities are below 10%. Three conserved basic residues cluster on the same face of the R3H domain and could play a role in nucleic acid recognition. An extended hydrophobic area at a different site of the molecular surface could act as a protein-binding site. A strong correlation between conservation of hydrophobic amino acids and side-chain solvent protection indicates that the structure of the Sμbp-2 R3H domain is representative of R3H domains in general.

© 2003 Elsevier Science Ltd. All rights reserved

Keywords: R3H domain; Sμbp-2; NMR spectroscopy; C-terminal domain of IF3; YhhP

Introduction

The R3H domain was identified by Grishin as a conserved sequence motif in proteins from a diverse range of organisms, including Eubacteria, green plants, fungi and metazoans.¹ The Smart database² reports 145 R3H domains; yet, no 3D structure has been determined for any of them. The R3H domain was named after an invariant arginine residue and a highly conserved histidine residue, spaced in the sequence by three residues. On the basis of sequence similarities outside the R3H region, R3H domain-containing proteins were grouped into eight families. Thus, R3H domains are found in proteins together with ATPase domains, SF1 helicase domains, SF2 DEAH helicase domains, Cys-rich repeats, ring-type zinc fingers, and KH domains.¹ The functions of these domains indicate that the R3H domain might interact with nucleic acids. In particular, Sμbp-2 proteins from human, mouse and hamster were reported to bind single-stranded DNA with a 5′-phosphorylated guanine-rich sequence related to the immunoglobulin μ-chain switch region.³⁻⁵ A 150 residue region of the human Sμbp-2 protein has been identified to be essential for DNA binding.³ It contains an R3H domain. Similarly, an R3H domain is present in a 222 residue single-stranded DNA-binding segment of the Drosophila STC protein.⁶ On the basis of these data and the prediction of a fold similar to that of the C-terminal domain of the translational initiation factor IF3, it was suggested that the R3H domain might bind single-stranded DNA or RNA in a sequence-specific manner.¹,⁷ Here, we report the 3D solution structure of the R3H domain from human Sμbp-2. The structure is compared to that of other proteins, including the C-terminal domain of IF3 (IF3-C), and functional implications are discussed.

Results and Discussion

Structure determination of the R3H domain

The gene of the R3H domain was obtained from a cDNA library (see Materials and Methods). The initial construct was based on the domain
boundaries suggested by Grishin, supplemented by an N-terminal His-tag (denoted R3H(2) in Figure 1). Determination of the 3D structure revealed a fold similar to IF3-C, but devoid of the N-terminal β-strand present in IF3-C. In order to verify the absence of a homologous β-strand from the R3H domain and to avoid possible interference by the His-tag, a second construct was made with an N-terminal extension and the His-tag placed at the C terminus (Figure 1). The resulting structures were indistinguishable. In the following, the results are presented for the longer construct with the C-terminal His-tag.

Structure of the R3H domain

The overall fold of the R3H domain was predicted accurately by Grishin. It consists of a triple-stranded antiparallel β-sheet, against which two helices pack from one side (Figure 2). Transpeptide bonds were found throughout. The only residue with a positive φ angle is Lys740. The ordered part of the structure is composed of only 60 residues, while the residues N-terminal of Gly725 are mobile. Mobility of the N-terminal 18 residues was documented by narrow linewidths in the 1H NMR spectrum, resulting in zero-quantum cross-peaks and weak nuclear Overhauser effects (NOEs) in NOESY spectra. The 3J(Hα, Hα) coupling constants measured for this segment were likewise characteristic of random-coil conformations. Due to rapid amide-proton exchange, the amide-proton resonances of glycine, serine and histidine residues of the first ten and last eight residues of the construct (Figure 1) were very broad or unobservable.

A strong correlation between the conservation of hydrophobic amino acid side-chains and protection from solvent access suggests that the NMR structure of the R3H domain of human Ssubp-2 is representative of R3H domains in general. Thus, uncharged residues with buried side-chains in the Ssubp-2 R3H domain (< 5% solvent accessibility) are consistently uncharged in all 24 full-length R3H domains shown in the alignment by Grishin, except for Leu750, which is the only buried residue in a loop region (Figure 1). Remarkably, the reverse holds too, i.e. all residues that are consistently uncharged in Grishin’s alignment are buried in the Ssubp-2 R3H domain with less than 5% side-chain solvent accessibility, with a single exception. The exception is Ser771, with a solvent accessibility of more than 35% (Figure 3). This residue is always uncharged in the alignment, but almost invariably hydrophilic, so that its solvent exposure is not unexpected. Finally, structural conservation between R3H domains is further indicated by the absence of insertions and deletions in regions of regular secondary structure (Figure 4).

Comparison to YhhP and the C-terminal domain of IF3

A DALI search of the protein data bank (PDB) revealed about 30 protein structures with a Z-score of 3.0 and higher. In almost all of these proteins, the fold of the R3H domain is embedded in a much larger structure and it is not clear
whether the motif would be stable in isolation. Two proteins, however, YhhP and the C-terminal domain of IF3 (IF3-C), are nearly as small as the R3H domain and structurally very similar, except for the presence of a fourth N-terminal β-strand in the sheet (Figs. 1 and 2). YhhP is an *Escherichia coli* protein of unknown function. IF3-C from *Bacillus stearothermophilus* yielded a Z-score of 5.2 and an rmsd of 2.5 Å for 56 aligned residues. IF3 is a translational initiation factor consisting of two structurally independent domains, both of which bind to ribosomal RNA of the 30 S subunit.

The structural similarity between the Sμbp-2 R3H domain and YhhP is reflected by the fact that all residues with buried side-chains in the R3H domain (Figure 3) have less than 15% solvent accessibility in the YhhP structure. The reverse
Resolution Structure of the R3H Domain from S\textsubscript{bp}-2

Figure 4. Multiple alignment of representative members of eight families of R3H-domain sequences.\textsuperscript{1} The locations of α-helices and β-strands found in the R3H domain from S\textsubscript{bp}-2 are marked by open and filled bars, respectively. Small characters identify residues outside the R3H domain. Arrows mark the conserved arginine and histidine residues, spaced by three residues, that gave the R3H domain its name, and a third, highly conserved, basic residue.\textsuperscript{1}

Small characters identify residues outside the R3H domain. Arrows mark the conserved arginine and histidine residues, spaced by three residues, that gave the R3H domain its name, and a third, highly conserved, basic residue.\textsuperscript{1} Residues with \(<5\%\) side-chain solvent accessibility in the NMR structure of the S\textsubscript{bp}-2 R3H domain are identified by bold characters. Names and gene identification numbers in the NCBI/GenBank protein sequence database: J\textsubscript{AG}\_BACSU: JAG protein from \(B.\) \textit{subtilis} (400078); YC45\_PORPU: hypothetical protein YCF45 from \textit{Porphyra purpurea} chloroplast (1723347); ITBA4: human ITBA4 protein (2274972); KIAA0029: human KIAA0029 protein (2104218); SMB2\_HUMAN: human S\textsubscript{bp}-2 protein (730751); helicase A: putative RNA helicase A from \textit{Arabidopsis thaliana} (1353239).

Comparison with Pfam and Smart alignments

Since June 2002, the Pfam\textsuperscript{13} and Smart\textsuperscript{2} databases have provided sequence alignments of R3H domains with the accession codes PF01424 and SM0393, respectively. Since the sequence alignment by Grishin,\textsuperscript{12} many more R3H domains have been identified, and Grishin’s original alignment as well as the domain boundaries were readjusted in these databases. Both databases provide a most plausible sequence alignment for the C-terminal two-thirds of the domain, consistently aligning uncharged residues at positions where side-chains are buried in the NMR structure. By the same criterion, the N-terminal third of the domain is less well aligned. The Smart database includes about 15 additional N-terminal residues as part of the R3H domain that were found to be mobile in the R3H domain of S\textsubscript{bp}-2. In contrast, the Pfam database entry truncates the structurally well-defined part of the domain (residues 720-786) by about four residues on either end. The domain boundaries and sequence alignments originally proposed\textsuperscript{4} are in best agreement with the NMR structure of the S\textsubscript{bp}–2 R3H domain.

Functional implications

R3H domains are found in proteins in association with other typical DNA-binding or RNA-binding domains. The most extensive biological data are available for S\textsubscript{bp}-2. On the basis of its ability to bind target DNA, it was named differently in the literature: immunoglobulin S-μ binding protein-2 (S\textsubscript{bp}-2)\textsuperscript{3,4} or IGHMBP2\textsuperscript{14} glial factor 1 (Gf1)\textsuperscript{15,16} rat insulin enhancer binding protein 1 (Rip1),\textsuperscript{7} and cardiac transcription factor 1 (Catf1).\textsuperscript{17} S\textsubscript{bp}-2 was originally proposed to be involved in immunoglobulin class switching.\textsuperscript{3} Other studies suggest that it is a transcription...
factor. Therefore, S\textsubscript{ubp}-2 has been shown to repress the Epstein–Barr virus lytic switch promoter. S\textsubscript{ubp}-2 and its truncated form, Gf1, transactivate the JC virus early and late promoters in glial cells. The rat homologue, Catf1, presumably transactivates the atrial natriuretic factor promoter. The hamster homologue Rip1 binds to a tissue-specific element from the rat insulin II gene promoter. Diseases such as spinal motor neuron degeneration and spinal muscular atrophy with respiratory distress type I have been linked to mutations in S\textsubscript{ubp}-2, but these mutations affected the helicase domain rather than the R3H domain.

By N and C-terminal deletion mutations, Fukita and co-workers identified a segment in S\textsubscript{ubp}-2 that specifically binds 5'-phosphorylated, guanine-rich, single-stranded DNA (ssDNA) sequences. This segment spans residues 638–786, including the R3H domain of S\textsubscript{ubp}-2 at its C-terminal end. ssDNA-binding specificity similar to that of the full-length protein was demonstrated. Secondary structure prediction indicates the presence of an irregular structure in the segment N-terminal of the R3H domain, except for one short helix separated from the R3H domain by about 30 residues in “random coil” conformation. Yet, significantly reduced ssDNA-binding affinities were observed for two constructs starting in and after that putative helix, respectively, suggesting that the R3H domain alone is not sufficient for high-affinity binding.

As predicted by Grishin, the side-chains of Arg755, His759 and Arg778 are spatially close in the R3H structure and could present a binding motif for phosphoester backbones of DNA and RNA (Figure 2(d)). These three residues are located at a position of the domain different from that of the RNA-contacting residues of IF3-C; in the complex of IF3-C with the small ribosomal subunit, IF3-C straddles the phosphoester backbones of different rRNA segments with its N-terminal α-helix (residues 96–111). Interestingly, a different face of the R3H domain presents a sizeable surface area composed of hydrophobic amino acid residues (Phe729, Met732, Leu744, Phe746, Pro747 and Leu750) and devoid of any charged or hydrophilic residues (Figure 2(e)). This area could be involved in yet unknown protein–protein interactions. The corresponding residues are consistently uncharged in families III and VII of R3H domains (Figure 4). As the concave shape of this hydrophobic area prohibits symmetric self-association of the domain, hetero-association would be more likely than formation of a homodimer. Experimentally, the monomeric state of the R3H domain is indicated by the narrow line shapes observed in the 1H NMR spectrum (data not shown).

Availability of the 3D structure of the R3H domain paves the way for the informed design of site-directed mutagenesis experiments for a more detailed functional characterization.

Materials and Methods

Protein cloning, expression and purification

The 207 bp and 235 bp DNA fragments encoding the S\textsubscript{ubp}-2 R3H domain were PCR amplified from human leukocyte cDNA library (Clontech). Primers contained BamHI and HindIII sites for the 207 bp fragment or NcoI and BamHI sites for the 235 bp variant, respectively.

PCR cycle conditions were: 20 seconds of denaturation at 95 °C, 30 seconds of annealing at 58 °C, and two minutes of polymerization at 68 °C. The first extension steps used two minutes predenaturation at 95 °C followed by 30 PCR cycles. The reaction mixture contained 10 pmol of each primer, 100 pg of the template, 0.25 mM each dNTP, and 2.5 units of Pfu DNA polymerase (Stratagene) in 100 µl of 100 mM KCl, 100 mM (NH₄)₂SO₄, 200 mM Tris–HCl (pH 8.75), 20 mM MgSO₄, 1% (v/v) Triton X-100, and 1 mg/ml of bovine serum albumin. The 207 bp BamHI-HindIII fragment bearing the R3H domain gene was inserted between the same two sites of pQE30 (Qiagen) resulting in an amino-terminally His-tagged open reading frame (ORF), while the 235 bp long NcoI-BamHI fragment was cloned into pQE60 with six carboxy-terminal His residues.

Recombinant proteins were expressed in E. coli M15 (Qiagen) for three to four hours after IPTG-induction and purified by affinity chromatography on Ni-NTA Sepharose (Qiagen) under native conditions according to the manufacturer’s protocol.

For both constructs, the final yields of soluble protein were about 10 mg per litre of bacterial suspension.

NMR spectroscopy and collection of structural restraints

NMR spectra were recorded using ~0.4 mM solutions of R3H(2) and R3H at 25 °C and 600 MHz 1H NMR frequency on a Bruker DMX 600 NMR spectrometer. Sequential resonance assignments were first obtained for the R3H(2) sample at pH 6.1 using a NOESY spectrum recorded with 100 ms mixing time and a TOCSY spectrum. In addition, a NOESY spectrum with 40 ms mixing time was recorded at pH 6.9 for the collection of structural restraints. The resonances of the R3H sample were assigned from TOCSY and NOESY spectra and compared to the spectra of the R3H(2) sample. Distance restraints for the structure calculation were collected from a NOESY spectrum recorded with 40 ms mixing time, using \( t_{max} = 64.3 \) ms, \( t_{min} = 210 \) ms, and a total recording time of about 72 hours. Dihedral angle restraints were derived from \( \gamma_{HNHA} \) and \( \gamma_{HAHY} \) coupling constants and NOEs. The \( \gamma_{HNHA} \) couplings were measured from the line-splittings of NOE cross-peaks. The \( \gamma_{H-H} \) couplings were estimated as 11.0 and 4.0 Hz (± 3.0 Hz), respectively, when TOCSY and NOESY cross-peaks indicated the presence of large and small couplings, respectively, together with staggered conformations around the C\(^{\alpha}-\)C\(^{\beta}\) bond.

Structure calculations

The cross-peaks in the NOESY spectra were assigned and integrated using the program XEASY. The NMR structure was calculated using the program DYANA. Eighty random conformers were annealed in 40,000 steps using torsion-angle dynamics. The 20 conformers...
with the lowest residual restraint violations were energy-
minimized in a shell of water using the program OPAL
with standard parameters.22

Table 1 shows an overview of the restraints used and
structural statistics. The Ramachandran plot was
analyzed using PROCHECK-NMR.23 No residue was
found in disallowed regions.

Secondary structure elements and rmsd values were
calculated using the program MOLMOL,24 which was
used to generate Figure 2.

Data Bank accession codes

The coordinates of the 20 energy-refined DYANA con-
formers of the R3H domain have been deposited in the
PDB with the accession code 1MSZ. The NMR chemical
shifts have been deposited at the BioMagResBank
(BMRB) under accession code 5535.

Acknowledgements

G.O. thanks the Australian Research Council for a
Federation Fellowship. L.G. acknowledges an
EU training fellowship (project HPRN-CT-2000-
00092). Financial support by the Swedish Research
Council is gratefully acknowledged.

Table 1. Structural statistics for the NMR conformers

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of assigned NOE cross-peaks</td>
<td>1094</td>
</tr>
<tr>
<td>Number of non-redundant NOE upper-distance limits</td>
<td>663</td>
</tr>
<tr>
<td>Number of scalar coupling constants(^a)</td>
<td>123</td>
</tr>
<tr>
<td>Number of dihedral-angle restraints</td>
<td>170</td>
</tr>
<tr>
<td>Intra-protein AMBER energy (kcal/mol)</td>
<td>(-2413 \pm 44)</td>
</tr>
<tr>
<td>Maximum NOE-restraint violations (Å)</td>
<td>0.1 ± 0.0</td>
</tr>
<tr>
<td>Maximum dihedral-angle restraint violations (deg.)</td>
<td>2.4 ± 0.1</td>
</tr>
<tr>
<td>rmsd to the mean for N, C(^\alpha), and C(^\beta) (Å)(^b)</td>
<td>0.44 + / – 0.09</td>
</tr>
<tr>
<td>rmsd to the mean for all heavy-atoms (Å)(^c)</td>
<td>0.87 + / – 0.08</td>
</tr>
</tbody>
</table>

Ramachandran plot appearance\(^\text{abc}\)

Most favoured regions (%) | 81.1 |
Additionally allowed regions (%) | 18.9 |
Generously allowed regions (%) | 0.0 |
Disallowed regions (%) | 0.0 |

\(^a\) 50 \(^3\)(H\(^\alpha\)–H\(^\gamma\)), 73 \(^3\)(H\(^\alpha\)–H\(^\delta\)).
\(^b\) For residues 725–786.
\(^c\) From PROCHECK NMR.23

References

2. Letunic, I., Goodstadt, L., Dickens, N. J., Doerkis, T.,
Schultz, J., Mott, R. et al. (2002). Recent improvements
to the SMART domain-based sequence
3. Fukita, Y., Mizuta, T. R., Shirozu, M., Ozawa, K.,
Chem. 268, 17463–17470.
Honjo, T. (1993). Isolation of cDNA encoding a
binding protein specific to 5'-phosphorylated single-
stranded DNA with G-rich sequences. *Nucl. Acids
Res.* 21, 1761–1766.
Molecular characterization of the rat insulin
enhancer-binding complex 3Kr. *J. Biol. Chem.* 270,
21503–21508.
homolog of human transcription factor NF-X1
encoded by the *Drosophila shuttle craft* gene is
required in the embryonic central nervous system.
7. Anantharaman, V., Koonin, E. V. & Aravind, L.
(2002). Comparative genomics and evolution of
proteins involved in RNA metabolism. *Nucl. Acids
Res.* 30, 1427–1446.
comparison by alignment of distance matrices. *J. Mol.
Biol.* 233, 123–136.
9. Katoh, E., Hatta, T., Shindo, H., Ishii, Y., Yamada, H.,
NMR structure of YhP, a novel *Escherichia coli*
protein implicated in cell division. *J. Mol. Biol.* 304,
219–229.
crystallography shows that translational initiation
factor IF3 consists of two compact \(\alpha/\beta\) domains
11. Pioletti, M., Schlünzen, F., Harms, J., Zarivach, R.,
Glühmann, M., Avila, H. et al. (2001). Crystal structures
of complexes of the small ribosomal subunit
with tetracycline, edeine and IF3. *EMBO J.* 20,
1829–1839.
12. Palm, G. J., Billy, E., Filipowicz, W. & Wlodawer, A.
(1999). Crystal structure of RNA 3'-terminal
phosphate cyclase, a ubiquitous enzyme with unusual
13. Bateman, A., Birney, E., Cerruti, L., Durbin, R.,


Edited by P. Wright

(Received 20 September 2002; received in revised form 19 November 2002; accepted 26 November 2002)