SAM breaks its stereotype

Traci M Tanaka Hall

Don't prejudge the function of a domain based on its sequence. Two recent studies drive home this message and add the sterile-α motif (SAM) to the growing list of domains that are able to interact with both proteins and nucleic acids.

The Pfam database describes the SAM domain as "a putative protein interaction module present in a wide variety of proteins[1]." But a revision in the annotation may be necessary, as recent papers by Green *et al.*[2] and Aviv *et al.*[3] have identified the SAM domain of *Drosophila* Smaug (Smg) protein as an RNA-binding domain and defined a new family of SAM RNA-binding proteins.

Post-transcriptional regulation of mRNA stability and translation is an important control point for gene expression during growth and development. This type of gene regulation is often modulated by *trans*-acting RNA-binding proteins that recognize specific sequences or structures in the 3′ untranslated region (UTR) of the mRNA. During the early patterning of *Drosophila* embryos, regulating expression of maternal mRNAs is crucial for the establishment of body axes. A textbook example is the localized expression of Nanos protein at the posterior end of the embryo[4]. This is accomplished by the repression of *nanos* mRNA translation in the bulk cytoplasm by the Smg protein and activation at the posterior pole[5–8]. Smg recognizes a stem-loop structure, termed the Smg recognition element (SRE), within the translational control element (TCE) in the 3′ UTR of the *nanos* mRNA[8–11]. A minimal RNA-binding fragment of Smg had been mapped to residues 584–763 (ref. 6). This fragment contains a SAM domain[7,12] and a C-terminal region of ~100 amino acids with no detectable sequence homology to proteins other than a Smg homolog from mosquito. This C-terminal region seemed a likely new RNA-binding motif, as the SAM domain was presumed to mediate protein-protein interactions, but the two recent studies prove otherwise.

In the June issue of *Molecular Cell*, Green *et al.*[2] report a crystal structure of the *Drosophila* Smg RNA-binding domain containing a SAM domain and a region similar in structure to HEAT repeats, christened the PHAT (pseudo-HEAT repeat analogous topology) domain. By using random mutagenesis of the *smg* gene and screening for mutants that retain RNA-binding activity in a yeast three-hybrid assay, they were able to map the locations of 45 residues not critical for RNA binding. The results clearly implicate one surface of the SAM domain as the RNA-binding surface (**Fig. 1**), and the PHAT domain appears not to be utilized. The surface is highly basic, consistent with a nucleic acid recognition surface. In addition, point mutations of His611 and Lys640 on the proposed RNA-binding surface reduce the RNA-binding activity of Smg.

In a complementary biochemical study published in the August issue of *Nature Structural Biology*, Aviv *et al.*[3] note a surface on the SAM domain that is highly conserved among Smg homologs and identify five residues (Lys606, Arg609, Lys612, Tyr613 and Ala642) that abolish or reduce Smg RNA-binding activity using a fluorescence polarization binding assay. Mapping the location of these residues on the Smg crystal structure reveals that they too are located on the surface proposed by Green *et al.*[2] as the RNA-binding

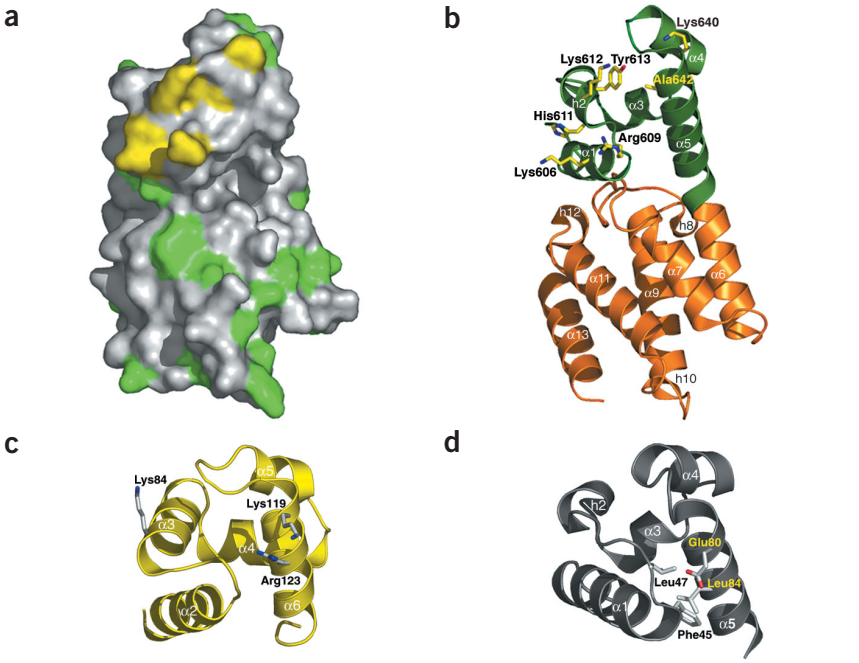

Figure 1 Mutational analysis maps a RNA-binding surface of Smg protein. (**a**) Surface representation of the crystal structure of the Smg RNA-binding domain (PDB entry 1OXJ). Random mutants that did not affect RNA binding in a yeast three-hybrid assay (green) are located on all surfaces of the PHAT domain and on one surface of the SAM domain, suggesting that the surface comprising α1, h2 and α5 forms the RNA-binding surface. Point mutations that reduce or abolish RNA binding (yellow) are on the proposed RNA-binding surface (adapted from ref. 2). (**b**) Ribbon diagram of the Smg RNA-binding domain. The orientation of Smg is the same as in **a**. SAM domain, green; PHAT domain, orange; side chains important for RNA binding, yellow. All figures were created with PyMOL (www.pymol.org). (**c**) Ribbon diagram of domain II of RuvA (PDB entry 1C7Y)[27], shown in the same orientation as the Smg SAM domain in **b**. Side chains that contact Holliday junction DNA are shown[27,28]. (**d**) Ribbon diagram of the TEL SAM domain (PDB entry 1JI7)[15], shown in the same orientation as the Smg SAM domain in **b**. Side chains involved in the oligomeric protein-protein interface are shown.

The author is at the National Institute of Environmental Health Sciences, Laboratory of Structural Biology, National Institutes of Health, Research Triangle Park, North Carolina 27709, USA.
e-mail: hall4@niehs.nih.gov

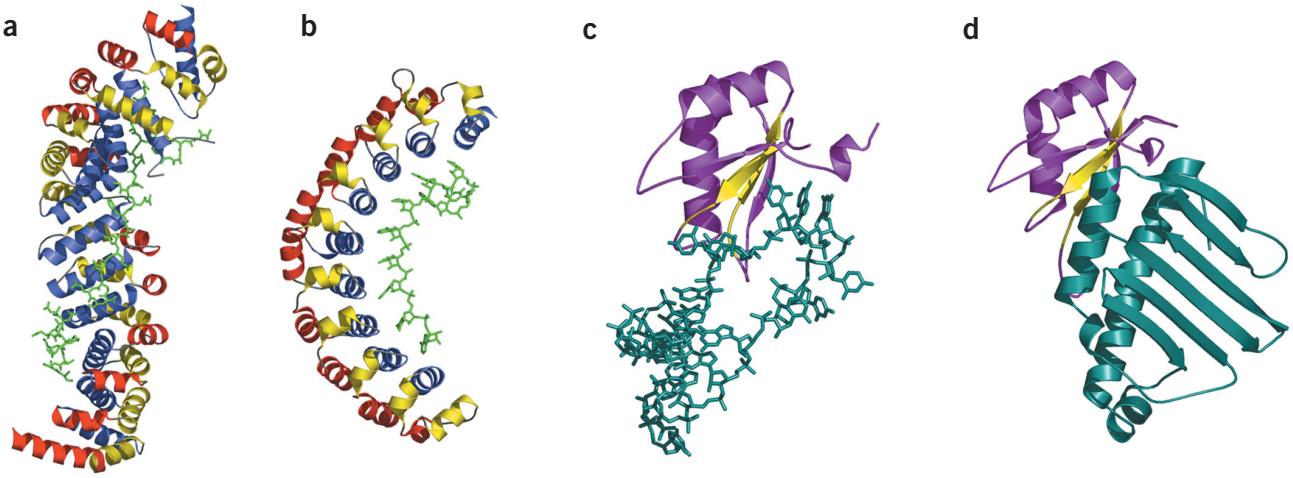

Figure 2 Protein modules that function as either RNA-binding or protein-interaction domains. (**a**) The Armadillo repeat protein β-catenin binds peptides (green) using the surface formed by the H3 helices (blue; PDB entry 1G3J)[20]. (**b**) Pumilio binds RNA (green) using the surface formed by α2 helices (blue), equivalent to the H3 surface in Armadillo proteins (PDB entry 1M8Y)[21]. (**c**) U1A (magenta) binds an RNA hairpin (cyan) using conserved RNP1 and RNP2 motifs (yellow; PDB entry 1URN)[29]. (**d**) Y14 (magenta) binds to Mago (cyan) using its RNP1 and RNP2 motifs (yellow; PDB entry 1OOO)[26]. Adapted from ref. 26.

surface of the SAM domain (**Fig. 1**). This surface is structurally equivalent to the DNA-binding surface in domain II of the Holliday junction–binding protein, RuvA[2] (**Fig. 1c**), which shares the SAM domain–like protein fold, as classified in the SCOP database, and represents one of a number of SAM domain–like proteins that bind DNA[13]. If indeed the Smg SAM domain interacts with RNA, the SAM domain–like fold can be used to interact with protein, DNA or RNA.

To confirm that the SAM domain of Smg mediates interaction with RNA, Aviv *et al.*[3] expressed the SAM domain of the yeast Smg homolog, Vts1, and analyzed its binding characteristics. The SAM domain of *Drosophila* Smg is not stable on its own as it shares an extensive interface with the PHAT domain and so could not be studied as an independent domain. However, 29% of the amino acid positions are identical in the Smg and Vts1 SAM domains. The Vts1 SAM domain binds a model SRE stem-loop RNA with the same affinity as full-length Vts1 protein. Point mutants of conserved residues in the proposed RNA-binding surface of Vts1 reduce or eliminate RNA binding in both the full-length and SAM-domain proteins.

An analysis of the consensus binding sequence for Smg and Vts1 demonstrated that both proteins require a stem-loop structure with a four- or five-base loop of the sequence CNGG or CNGGN (where N at the second and fifth positions is any base). Thus the SAM domains of Smg and Vts1 have similar, possibly identical binding specificities. But does this similarity in binding specificity indicate a similarity in cellular function? In a bold experiment, Aviv *et al.*[3] asked whether Vts1 could alter the expression of a target mRNA containing SRE sequences in its 3′ UTR. They found that endogenous Vts1 in yeast can induce the degradation of the reporter mRNA and that this degradation is dependent on the cytoplasmic deadenylase CCR4. Consistent with this, Green *et al.*[2] found that frog and mouse SAM domain proteins could bind to SRE RNA hairpins in a yeast three-hybrid assay with substitutions permitted at the second and fifth positions in the loop sequence. The independent SAM domains of fly, frog and mouse proteins bound relatively weakly in their assays, perhaps owing to instability of the independent domains. Together Green *et al.*[2] and Aviv *et al.*[3] identify SAM domains from yeast, fungi, insects, worms and vertebrates that contain several conserved basic residues on the Smg RNA-binding surface. Aviv *et al.*[3] note that Smg and Vts1 are two of the most divergent sequences, suggesting that these may represent a new family of SAM domains that bind stem-loop RNA structures similar to those recognized by Smg and that function in post-transcriptional gene regulation.

Thus, the SAM domain–like fold is represented in at least three functional families: protein interaction, DNA binding and RNA binding. And in some cases, the equivalent surface is used for either protein, DNA or RNA recognition (**Fig. 1a–c**). For example, the transcriptional repressor TEL forms a helical head-to-tail structure resulting from interaction of EH (end helix) and ML (mid-loop) surfaces of the SAM monomer[14,15]. The EH surface of TEL, involving residues in the first and last α-helices of the SAM domain, overlaps with the RNA-binding surface of Smg and the DNA-binding surface of RuvA.

It seems increasingly common to find proteins with the same fold that use equivalent surfaces to bind either proteins or RNA. For example, the helical repeats of β-catenin and karyopherin-α bind linear peptide epitopes[16–20] whereas the topologically equivalent repeats of Pumilio proteins bind single-stranded RNA[21–23] (**Fig. 2a,b**). Although Armadillo and Pumilio repeats pack differently to form superhelical versus curved structures, the equivalent and most highly conserved surface is used for either peptide or RNA recognition depending on the protein. Even more surprising was the recent finding that the RNA recognition motif (RRM) of the exon-exon junction complex protein Y14 uses the typical RRM RNA-binding surface to interact instead with the Mago protein[24–26] (**Fig. 2c,d**). We may need to reconsider whether the presence of an RRM, long the hallmark of an RNA-binding protein, actually denotes this, although in this case it has been suggested that Y14 may bind to protein and RNA using the same surface[25]. Similarly, we may need to consider whether characteristic protein interaction motifs can serve alternate functions. What will be next—an SH2 domain that binds RNA?

1. Bateman, A. *et al. Nucleic Acids Res.* **30**, 276–280 (2002).
2. Green, J.B., Gardner, C.D., Wharton, R.P. & Aggarwal,

A.K. *Mol. Cell* **11**, 1537–1548 (2003).
3. Aviv, T. *et al. Nat. Struct. Biol.* **10**, 614–621 (2003).
4. Wang, C. & Lehmann, R. *Cell* **66**, 637–647 (1991).
5. Bergsten, S.E. & Gavis, E.R. *Development* **126**, 659–669 (1999).
6. Dahanukar, A., Walker, J.A. & Wharton, R.P. *Mol. Cell* **4**, 209–218 (1999).
7. Smibert, C.A., Lie, Y.S., Shillinglaw, W., Henzel, W.J. & Macdonald, P.M. *RNA* **5**, 1535–1547 (1999).
8. Smibert, C.A., Wilson, J.E., Kerr, K. & Macdonald, P.M. *Genes Dev.* **10**, 2600–2609 (1996).
9. Gavis, E.R., Lunsford, L., Bergsten, S.E. & Lehmann, R. *Development* **122**, 2791–2800 (1996).
10. Dahanukar, A. & Wharton, R.P. *Genes Dev.* **10**, 2610–2620 (1996).
11. Crucs, S., Chatterjee, S. & Gavis, E.R. *Mol. Cell* **5**, 457–467 (2000).
12. Schultz, J., Ponting, C.P., Hofmann, K. & Bork, P. *Protein Sci.* **6**, 249–253 (1997).
13. Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. *J. Mol. Biol.* **247**, 536–540 (1995).
14. Kim, C.A., Gingery, M., Pilpa, R.M. & Bowie, J.U. *Nat. Struct. Biol.* **9**, 453–457 (2002).
15. Kim, C.A. *et al. EMBO J.* **20**, 4173–4182 (2001).
16. Conti, E. & Kuriyan, J. *Structure Fold. Des.* **8**, 329–338 (2000).
17. Conti, E., Uy, M., Leighton, L., Blobel, G. & Kuriyan, J. *Cell* **94**, 193–204 (1998).
18. Eklof Spink, K., Fridman, S.G. & Weis, W.I. *EMBO J.* **20**, 6203–6212 (2001).
19. Huber, A.H. & Weis, W.I. *Cell* **105**, 391–402 (2001).
20. Graham, T.A., Weaver, C., Mao, F., Kimelman, D. & Xu, W. *Cell* **103**, 885–896. (2000).
21. Wang, X., McLachlan, J., Zamore, P.D. & Hall, T.M. *Cell* **110**, 501–512 (2002).
22. Wang, X., Zamore, P.D. & Hall, T.M.T. *Mol. Cell* **7**, 855–865 (2001).
23. Edwards, T.A., Pyle, S.E., Wharton, R.P. & Aggarwal, A.K. *Cell* **105**, 281–289 (2001).
24. Fribourg, S., Gatfield, D., Izaurralde, E. & Conti, E. *Nat. Struct. Biol.* **10**, 433–439 (2003).
25. Lau, C.K., Diem, M.D., Dreyfuss, G. & Van Duyne, G.D. *Curr. Biol.* **13**, 933–941 (2003).
26. Shi, H. & Xu, R.M. *Genes Dev.* **17**, 971–976 (2003).
27. Ariyoshi, M., Nishino, T., Iwasaki, H., Shinagawa, H. & Morikawa, K. *Proc. Natl. Acad. Sci. USA* **97**, 8257–8262 (2000).
28. Roe, S.M. *et al. Mol. Cell* **2**, 361–372 (1998).
29. Oubridge, C., Ito, N., Evans, P.R., Teo, C.H. & Nagai, K. *Nature* **372**, 432–438 (1994).

The CTD code

Stephen Buratowski

How does the C-terminal domain (CTD) of RNA polymerase II interact specifically with multiple targets? A recent paper describing the structure of this domain with a mRNA capping enzyme guanylyltransferase suggests that the CTD is a contortionist that, upon post-translational modification, adopts different configurations specifically recognized by its partners.

The gene expression field has been experiencing a period of remarkable integration. Eukaryotic RNA polymerase II (RNAPII) produces mRNA, but that is only its most basic function. While transcribing, RNAPII also scans for DNA damage and modifies the surrounding chromatin. Through protein-protein interactions, RNAPII also acts as a platform for several mRNA processing factors that modify the mRNA as it is being synthesized. One particularly important component for these interactions is the C-terminal domain (CTD) of the RNAPII largest subunit. The CTD couples transcription with histone methylation, mRNA splicing, and polyadenylation, but its best-characterized direct interaction is with the mRNA capping enzyme. A recent report in *Molecular Cell* presents the crystal structure of the CTD bound to the capping enzyme guanylyltransferase (Cgt1), extending our understanding of this interaction to the atomic level[1]. This and other studies lend insight into how transcription by RNAPII is linked to so many other processes.

The CTD is a simple repetition (27–52 repeats, depending upon the organism) of the heptapeptide sequence Tyr-Ser-Pro-Thr-Ser-Pro-Ser. No analogous domain exists on the related RNAPI and RNAPIII enzymes, and the CTD is completely dispensable for RNA polymerization. The CTD is highly phosphorylated *in vivo*, and many proteins are thought to bind to this domain. Interacting partners include the Mediator complex that regulates transcription initiation, several histone methyltransferases, the capping enzyme that modifies the 5′ end of mRNA, and the polyadenylation factors that modify the 3′ end (for reviews, see refs. 2,3). How does such a simple sequence interact with so many targets? Rather than carrying all these factors throughout the transcription cycle, the CTD interacts dynamically with each factor at the appropriate time. A series of different phosphorylations and conformation changes generates configurations specific for binding of particular factors. In essence, there is a CTD 'code' that specifies the position of RNAPII in the transcription cycle.

The two major CTD phosphorylations occur at distinct points in the transcription cycle. The serine in the fifth position (Serine 5) is phosphorylated by the basal transcription factor TFIIH near the promoter, and genetic and biochemical data show that capping enzyme is recruited by this modification[4–7]. The structure of the *Candida albicans* guanylyltransferase (Cgt1)-CTD phosphopeptide complex illustrates how a CTD code can be read. The peptide used contains four heptad repeats, each phosphorylated at serine 5, but only seventeen residues (two repeats) are visible in the structure. The phosphopeptide binds in a cleft on the nucleotidyl transferase domain, with an extended β-like conformation containing one turn at proline 6. The phosphates on two serine 5 residues from adjacent repeats bind in positively charged pockets and act as electrostatic anchors to either end of the binding cleft. In addition to serine 5, the tyrosine and two prolines within each repeat also make specific contacts with Cgt1. These interactions are consistent with mutagenesis data reported by Fabrega *et al.*[1], as well as with previous biochemical and genetic studies.

Serine 2 is phosphorylated during elongation by a different kinase. There are suggestions that polyadenylation factors may interact specifically with the serine 2 phosphorylated form of the CTD. Therefore, the two phosphorylations help distinguish early and late phases of transcription[7].

In addition to phosphorylation, a CTD code probably also includes *cis-trans* isomerization at the two prolines that follow the phosphorylated serines. The proline isomerase Pin1/Ess1 acts at prolines preceded by a phosphorylated residue and has been implicated in mRNA 3′ end formation (ref. 8 and references therein). It is informative to compare the CTD (serine 5-P)–capping enzyme structure with that of Pin1 bound to the CTD phosphorylated at both serine 2 and serine 5 (ref. 9). Whereas the capping enzyme-bound CTD has a β-like configuration, the Pin1-bound CTD is more like a type II polyproline helix. In both structures, the prolines are in the *trans* configuration. Pin1 binds to the CTD (at least in part) via its WW domain, a motif found in several other CTD-binding

The author is in the Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts, USA.
e-mail: steveb@hms.harvard.edu