Structural genomics of RNA

Jennifer A. Doudna

A detailed understanding of the functions and interactions of biological macromolecules requires knowledge of their molecular structures. Structural genomics, the systematic determination of all macromolecular structures represented in a genome, is focused at present exclusively on proteins. It is clear, however, that RNA molecules play a variety of significant roles in cells, including protein synthesis and targeting, many forms of RNA processing and splicing, RNA editing and modification, and chromosome end maintenance. To comprehensively understand the biology of a cell, it will ultimately be necessary to know the identity of all encoded RNAs, the molecules with which they interact and the molecular structures of these complexes. This report focuses on the feasibility of structural genomics of RNA, approaches to determining RNA structures and the potential usefulness of an RNA structural database for both predicting folds and deciphering biological functions of RNA molecules.

Over the past two decades it has become clear that a variety of RNA molecules have important or essential biological functions in cells, beyond the well-established roles of ribosomal, transfer and messenger RNAs in protein biosynthesis. A partial list of such molecules includes catalytic RNAs, small nuclear RNAs (snRNAs) that compose the pre-mRNA splicing machinery, guide RNAs involved in RNA editing, telomerase RNA required for chromosome end replication, signal recognition particle (SRP) RNA necessary for protein translocation and small nucleolar RNAs (snoRNAs) responsible for ribosomal RNA modification. Furthermore, the 5′ and 3′ untranslated regions (UTRs) of numerous messenger RNAs regulate gene expression through interactions with ribosomal subunits and cellular proteins. Each class of RNA is likely to have a unique fold that confers biochemical function. Indeed, RNA is proficient at forming complex and varied three-dimensional shapes, as revealed by high-resolution structures of a handful of large RNAs, RNA–protein complexes and the 50S and 30S ribosomal subunits. Large RNA structures remain scarce, however, due in part to difficulty in producing high quantities of conformationally and biochemically homogeneous RNA and a lack of focus of the structural biology community on this exciting area of research. Given the modest pace of RNA structural biology at present, does it make sense to launch a structural genomics effort for RNA? To address this question it is useful to consider the goals of high-throughput structure determination in light of the properties of RNA structure and the current state of knowledge about biologically important RNAs.

Structural genomics efforts have emerged in response to the fact that genome sequences encode many proteins of unknown function. Homologies between proteins are often undetectable from sequence comparison, and protein secondary and tertiary structures are highly coupled and difficult to predict accurately. Similarities of uncharacterized polypeptides to known proteins, revealed only at the level of high-resolution molecular structures, might suggest biological function. Furthermore, a database encompassing the complete set of protein folds that exist in biology might enable modeling of unknown protein structures by assembly of their component domains.

In RNA, sequence conservation among functional homologs is usually limited to short (<10 nucleotide) segments, making homology searching even more difficult than for proteins. In contrast to proteins, however, RNA secondary structures can be well defined by phylogenetic covariation analysis, giving RNA biologists an advantage in determining the functional family to which a molecule belongs. Frequently, sequence conservation within these RNA families becomes apparent in the context of the secondary structure. In group I introns, for example, the conserved positions of functionally critical residues within the RNA secondary structure revealed the location of the catalytic core (Fig. 1). Secondary structure analysis of telomerase RNA led to the surprising discovery of snoRNA motifs embedded within the molecule, suggesting an unanticipated function[1].

Ultimately, however, RNA tertiary structures are key to understanding biological activity, and these are much more difficult to model. The motifs that stabilize RNA three-dimensional folds are relatively small and often involve backbone functional groups, making them difficult or impossible to detect even within large families of secondary structures. Tetraloops and their receptors, U-turns, dinucleotide platforms, ribose zippers and S-turns all consist of 4–11 nucleotides and occur within a variety of sequence contexts[2]. In addition, non-canonical base pairs often create context-dependent helical geometries or surfaces used in RNA–RNA and RNA–protein recognition[3].

Attempts to understand large RNA tertiary structures by studying isolated secondary structural domains have met with mixed success. These constructs are readily prepared and are amenable to structure determination. The resulting structures reveal non-canonical base pairings, helical geometry and potential sites of ligand binding. Deriving biological function is more difficult, however, due to the dynamics of small RNAs. In the case of signal recognition particle RNA, for example, NMR and crystal structures determined for the isolated molecule differed from the structure of the RNA bound to its protein partner (Fig. 2)[4–6]. Similarly, the structures of 5S ribosomal RNA and a hairpin RNA derived from the core of the hepatitis delta virus

Howard Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA.
email: doudna@csb.yale.edu

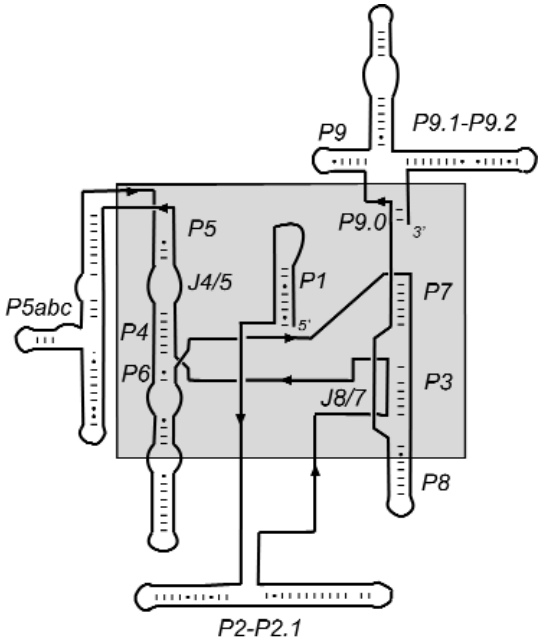

Fig. 1 Secondary structure of the group I class of self-splicing introns. The representation shows features typical of RNA secondary structures, including base-paired segments (P) connected by joining regions (J). The boxed region is the catalytic core, identified by conservation of functionally critical residues.

ribozyme differ from the structures of these RNAs in the context of the 50S ribosomal subunit and the intact HDV ribozyme, respectively (refs 7,8 and refs therein). In some cases such conformational differences may provide clues to functionally relevant structural dynamics within an RNA, as proposed for the tetraloop receptor[9,10].

These examples suggest that an initiative to systematically determine RNA structures would be valuable if the structural targets are biologically functional RNAs and their protein partners. Such an effort, coupled with approaches for identifying new RNA genes and a comprehensive database for storing and disseminating the information, would provide the tools for RNA research in the post-genomic era.

Ribonomics: the RNA analog of proteomics

Our understanding of RNA in biology is currently limited in part by a lack of structural data, but perhaps more profoundly by limited knowledge of the cast of characters. It is not yet clear how many structured RNAs are expressed in different cell types, what biochemical pathways they participate in and what proteins they bind. Structural genomics of RNA will be most interesting when integrated with experimental and computational methods for identifying novel RNA genes and determining their biological relevance: an approach defined previously as 'ribonomics'[11]. Such an effort would have at least three essential goals: (i) to develop and implement methodologies for identifying and characterizing novel RNA genes; (ii) to develop techniques for high-throughput determination of RNA and RNA–protein structures; and (iii) to create and maintain a centralized database of RNA structures, sequences, functional data and modeling tools.

How are RNA-encoding genes to be located? As mentioned above, RNAs often share little sequence homology within families, making it difficult to identify them by homology searches. Instead, secondary structure conservation has been used successfully to identify new members of the snoRNA family[12,13], an approach that could be applied to other RNA classes. This method is useful for uncovering members of known RNA families, but for novel RNA genes, different computational algorithms will be required. One possibility is to search for sequences that contain higher than average proportions of purines, tetraloop sequences and short complementary regions likely to characterize structured RNAs. The resulting set of potential RNA-encoding genes could be tested for expression *in vivo* using microarray analysis. Secondary structure prediction and homology searches could be performed with other genome sequences, on the assumption that functionally important RNAs are likely to be conserved. Since many of the identified RNAs are likely to function within ribonucleoprotein (RNP) complexes, it will be essential to determine their protein ligands, perhaps genetically using suppressor screens, or by direct binding assays or cross-linking.

Once a comprehensive set of RNAs has been defined for a given organism, structural analysis would go hand-in-hand with biochemical approaches to determining function. Secondary structure prediction and testing would be critical for defining fold families and functional centers, as discussed above. Approaches to determining RNA tertiary structures include chemical probing and modification interference mapping, nuclear magnetic resonance (NMR), X-ray crystallography and cryo-electron microscopy. Chemical probing and interference mapping, such as RNA modification with alkylating agents and phosphorothioates or cleavage with free radicals, has been central to RNA tertiary structure prediction for RNAs whose secondary structure is well defined phylogenetically[14–17]. Since most structured RNAs require magnesium ions to fold, RNA modified before and after addition of magnesium salts reveals those nucleotides whose solvent accessibility changes upon structure formation. These approaches have been useful for modeling RNAs including tRNA, ribosomal RNA, group I and group II introns, Ribonuclease P and internal ribosome entry site (IRES) RNA. NMR structure determination has been particularly useful for small RNAs and RNA–protein complexes (<30 kDa), and for elucidating dynamic motions of RNA in solution. The availabili-

Fig. 2 Solution and crystal structures of the 50-nucleotide domain IV of signal recognition particle 4.5S RNA. The left and middle images are structures of the RNA determined in the absence of the Ffh M domain protein, while the image on the right is the structure of the complex[4–6]. A large conformational change occurs within an asymmetric loop of the RNA upon protein binding.

ty of methods for crystallizing and derivatizing RNA molecules has recently led to several RNA structures determined by X-ray crystallography. Cryo-electron microscopy is particularly useful for visualizing large macromolecular complexes[18].

Each of the approaches described above involves painstaking preparation of RNA samples and often months of work to obtain and interpret structural information. Can the sorts of high-throughput approaches being developed for protein structural genomics speed up the RNA structure determination pipeline? The bottlenecks for RNA structure determination are typically identification of stable, well-behaved constructs and preparation of highly pure samples. These issues are currently addressed empirically for each RNA to be studied, through preparation and screening of dozens of samples. It would thus be enormously beneficial to develop faster, parallel methods of screening large numbers of constructs for amenability to NMR, X-ray and other structure determination techniques. For example, a series of plasmids could be designed for *in vitro* transcription in which ribozymes are positioned flanking the RNA sequence of interest, to enable production of chemically homogeneous samples[19]. Assays for conformational homogeneity, such as native gel electrophoresis or size exclusion chromatography, could be carried out using robotics on many samples at once. For crystallization experiments, rational approaches including designing reagent screens and engineering RNA–RNA and RNA–protein interactions into molecules of interest have been useful[20–24]. Nanoliter sample sizes and automation of crystal analysis will reduce the amount of material and time required for such screening. The protein components of RNA–protein complexes can be expressed, purified and analyzed using some of the same approaches being developed for high-throughput protein structure determination[25].

One goal of determining large numbers of macromolecular structures is to provide a database of information that will guide prediction and accurate modeling of unknown folds. It is difficult to predict whether such modeling will ultimately be easier for RNA than for proteins. On the one hand, there may be fewer and smaller tertiary structural motifs for RNA than for proteins, as discussed above. On the other hand, the limited size of these motifs is not very information-rich. Nonetheless, knowledge of the complete 'parts list' of RNA tertiary structural motifs, together with secondary structure prediction and recognition programs, is likely to enable more rapid and accurate modeling of RNA three-dimensional folds. A classic success story in the RNA modeling world is that of group I self-splicing introns, for which a largely correct three-dimensional model was constructed based on sequence covariation analysis and biochemical data[26,27]. However, in other cases, such as the hepatitis delta virus ribozyme, modeling has been less successful due to the lack of sufficient sequences for the identification of structural interactions such as pseudoknots[8,28]. The availability of extensive sequence databases and development of appropriate programs for identifying RNA family members based on secondary structure conservation will undoubtedly continue to be essential for accurate RNA three-dimensional modeling.

Ultimately, of course, researchers in structural genomics hope to decipher functional properties of biologically interesting molecules through determining their molecular structures. Will this be possible for RNA? The atomic resolution structures of both ribosomal subunits will certainly be instructive in this regard[7,29–31]. Perhaps the vast expanses of structured RNA and RNA–protein interactions in the ribosome will reveal unsuspected functional roles or sites of ligand contacts based purely on structural features. This depends largely upon whether RNA structural modules correspond to identifiable functional units. In the case of small motifs this seems unlikely. For example, the U-turns observed in crystal structures of tRNA and the hammerhead ribozyme both occur at helical junctions, but in tRNA the turn is purely structural while in the hammerhead it creates a catalytic pocket opposite the self-cleavage site. Whether collections of motifs within larger structural units will correlate with distinct recurring biochemical functions remains to be seen.

1. Mitchell, J.R., Cheng, J. & Collins, K. *Mol. Cell. Biol.* **19**, 567–576 (1999).
2. Moore, P.B. *Annu. Rev. Biochem.* **68**, 287–300 (1999).
3. Leontis, N.B. & Westhof, E. *Quart. Rev. Biophys.* **31**, 399–455 (1998).
4. Schmitz, U. *et al. RNA* **5**, 1419–1429 (1999).
5. Batey, R.T., Rambo, R.P., Lucast, L., Rha, B. & Doudna, J.A. *Science* **287**, 1232–1239 (2000).
6. Jovine, L. *et al. Structure* **8**, 527–540 (2000).
7. Ban, N., Nissen, P., Hansen, J., Moore, P.B. & Steitz, T.A. *Science* **289**, 905–920 (2000).
8. Ferre–D'Amare, A.R., Zhou, K. & Doudna, J.A. *Nature* **395**, 567–574 (1998).
9. Cate, J.H. *et al. Science* **273**, 1678–1685 (1996).
10. Butcher, S.E., Dieckmann, T. & Feigon, J. *EMBO J.* **16**, 7490–7299 (1997).
11. Bourdeau, V., Ferbeyre, G., Pageau, M., Paquin, B. & Cedergren R. *Nucleic Acids. Res.* **27**, 4457–4467 (1999).
12. Lowe, T.M. & Eddy, S.R. *Science* **283**, 1168–1171 (1999).
13. Omer, A.D. *et al. Science* **288**, 517–522 (2000).
14. Latham, J.A. & Cech, T.R. *Science* **245**, 276–282 (1989).
15. Christian, E.L. & Yarus, M. *J. Mol. Biol.* **228**, 743–758 (1992).
16. Harris, M.E., Kazantsev, A.V., Chen, J.L. & Pace, N.R. *RNA* **3**, 561–576 (1997).
17. Ryder, S.P., Ortoleva-Donnelly, L., Kosek, A.B. & Strobel, S.A. *Methods Enzymol.* **317**, 92–109 (2000).
18. Stowell, M.H., Miyazawa, A. & Unwin, N. *Curr. Opin. Struct. Biol.* **8**, 595–600 (1998).
19. Ferre-D'Amare, A.R. & Doudna, J.A. *Nucleic Acids Res.* **24**, 977–978 (1996).
20. Doudna, J.A., Grosshans, C., Gooding, A. & Kundrot, C.E. *Proc. Natl. Acad. Sci. USA* **90**, 7829–7833 (1993).
21. Scott, W.G. *et al. J. Mol. Biol.* **250**, 327–332 (1995).
22. Golden, B.L., Podell, E.R., Gooding, A.R. & Cech, T.R. *J. Mol. Biol.* **270**, 711–723 (1997).
23. Ferre-D'Amare, A.R., Zhou, K., Doudna, J.A. *J. Mol. Biol.* **279**, 621–631 (1998).
24. Ferre-D'Amare, A.R. & Doudna, J.A. *J. Mol. Biol.* **295**, 541–556 (2000).
25. Burley, S.K. *et al. Nature Genet.* **23**, 151–157 (1999).
26. Michel, F. & Westhof, E. *J. Mol. Biol.* **216**, 585–610 (1990).
27. Golden, B.L., Gooding, A.R., Podell, E.R. & Cech, T.R. *Science* **282**, 259–264 (1998).
28. Tanner, N.K. *et al. Curr. Biol.* **4**, 488–498 (1994).
29. Schluenzen, F. *et al. Cell* **102**, 615 (2000).
30. Wimberly, B.T. *et al. Nature* **407**, 327–339 (2000).
31. Carter, A.P. *et al. Nature* **407**, 340–348 (2000).