Design and chance in the self-assembly of macromolecules

J.A.R. Worrall*, M. Górna*, X.Y. Pei*, D.R. Spring†, R.L. Nicholson† and B.F. Luisi*[1]

*Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1GA, U.K., and †Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1ER, U.K.

Abstract

The principles of self-assembly are described for naturally occurring macromolecules and for complex assemblies formed from simple synthetic constituents. Many biological molecules owe their function and specificity to their three-dimensional folds, and, in many cases, these folds are specified entirely by the sequence of the constituent amino acids or nucleic acids, and without the requirement for additional machinery to guide the formation of the structure. Thus sequence may often be sufficient to guide the assembly process, starting from denatured components having little or no folds, to the completion state with the stable, equilibrium fold that encompasses functional activity. Self-assembly of homopolymeric structures does not necessarily preserve symmetry, and some polymeric assemblies are organized so that their chemically identical subunits pack stably in geometrically non-equivalent ways. Self-assembly can also involve scaffolds that lack structure, as seen in the multi-enzyme assembly, the degradosome. The stable self-assembly of lipids into dynamic membraneous sheets is also described, and an example is shown in which a synthetic detergent can assemble into membrane layers.

The self-assembly of macromolecular complexes

Nearly 50 years ago, Anfisen and colleagues demonstrated that a denatured protein can reform into its native structure spontaneously. These observations were insightfully interpreted as indicating that the amino acid sequence contains sufficient information to guide the formation of the native structure through an astronomical number of possible conformations. It was shown subsequently that even massive assemblies have the capacity to be reconstituted from denatured components. One striking example is the reconstitution of the 2.6 MDa ribosome from RNA and protein constituents into an active assembly [1,2]. Thus it is possible to organize very complex and elaborate macromolecular structures spontaneously and without the requirement of auxiliary factors or the effort of exogenous work. In a cell, however, there are highly evolved machineries that accelerate assembly, or help to escape from kinetic traps, because life cannot afford the luxury of awaiting equilibrium. Furthermore, elaborate machinery exists to proofread the assembled molecules against misfolding, which is a potentially hazardous pathway implicated in severe cellular dysfunction [3].

The processes of life depend on intricate cellular machinery that is built upon the reversible association of proteins and nucleic acids into larger assemblies. These may be transient or stable on the scale of cellular lifetimes. Taking brewer's yeast (*Saccharomyces cerevisiae*) as a representative eukaryotic organism, it is estimated that there are roughly 800 different types of stable macromolecular complexes existing within the cell [4]. These complexes are composed of two to 25 proteins, and many exist only transiently. From the perspective of molecular construction, the cell might be viewed as having a modular nature, but it seems remarkable that the components recognize each other appropriately and discriminately from the vast potential of pairwise combinations, and avoid misassociations that can cause dysfunctional aggregation. This must be attributed to the selective pressures that constrain the molecular evolution [3,5].

The origins of stability of proteins and macromolecular complexes are generally well understood at the level of stereochemistry through extensive data from protein crystallography. From the smallest protein to the most intricate complex, the favourable process of folding arises mainly from the co-operative interplay of three effects: (i) the sequestration of non-polar side chains from the aqueous solvent, where they perturb the hydrogen-bonding pattern of the water [6,7]; (ii) the formation of hydrogen bonds within secondary structure, which replace comparable bonds made with the solvent in the unfolded state; and (iii) extensive van der Waals contacts between atoms of the protein. Analyses of multicomponent structures have identified several features that characterize a stable complex: complementarity of surface shape of components, so that the buried surface area is on the order of 10^3 Å^2 (1 Å = 0.1 nm), a match of non-polar patches, complementary hydrogen-bonding patterns (often supported by sequestered water molecules), and a match of surface-charge distribution [8,9]. Another important factor for recognition is the conformational plasticity of the interacting molecules, which can mutually accommodate to optimize

Key words: folding, degradosome, membrane, molecular recognition, protein–protein interaction, self-assembly.

Abbreviations used: RhlB, RNA helicase B; PNPase, polynucleotide phosphorylase.

[1]**To whom correspondence should be addressed (email bfl20@mole.bio.cam.ac.uk).**

Figure 1 | For legend see facing page

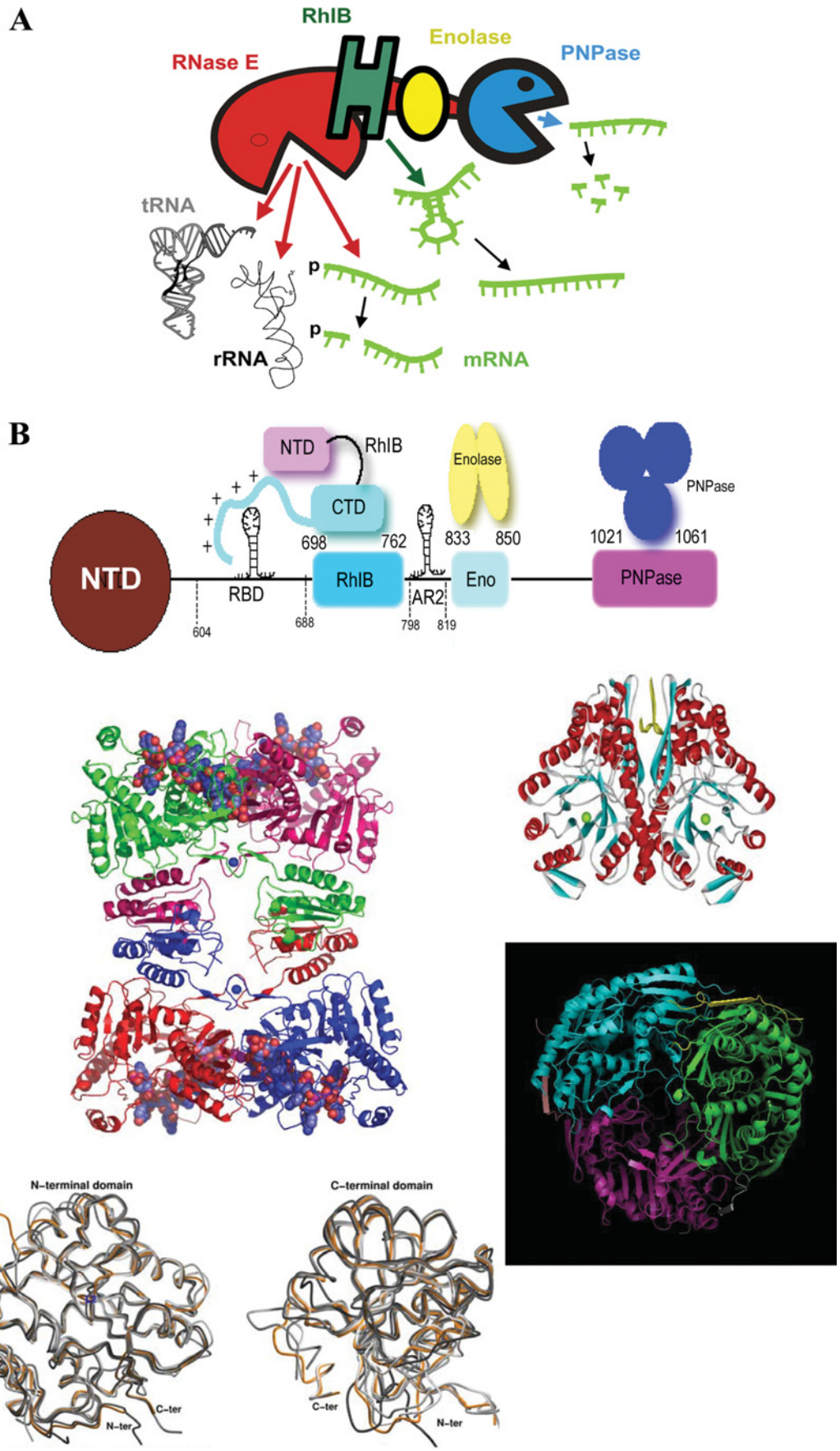


Figure 1 | Self-assembly of second order: the RNA degradosome, a complex assembly from the bacterium *Escherichia coli*

(**A**) Cartoon representation of the components of the degradosome. The endoribonuclease RNase E, composed of 1061 amino acids, forms the scaffold for protein–protein and protein–RNA interactions; RhlB is an ATP-dependent RNA helicase with a widely occurring DEAD sequence motif; enolase is an enzyme from glycolysis; the exoribonuclease PNPase uses inorganic phosphate to cleave the terminal phosphodiester linkage of the RNA substrate. The arrows indicate the substrates cleaved by the RNase E catalytic domain, unwound by helicase or cut by PNPase. (**B**) A cartoon representation of the binding domains of components of the degradosome, and a gallery of structural models of the degradosome components and their interactions. The schematic diagram shows RNA hairpins associated with the RNA-binding regions (RBD and AR2). Middle left: the homotetrameric catalytic N-terminal domain (NTD) of RNase E. Middle right: the homodimer enolase in complex with its recognition peptide from RNase E (yellow). Lower right: homotrimeric PNPase with recognition peptide from RNase E (protomers are blue, purple and green, and RNase E segments are the interprotomer single strands, coloured yellow, brown and grey). Lower left: overlays of homology models of the RhlB helicase with two RecA-like domains that comprise an internal repeat in the structure. RhlB is similar, but contains loop extensions (L1).

the surface match [10,11]. These deformations may be energetically demanding in some cases, and so contribute to the discrimination of cognate from non-cognate interactions.

When considering these biological macromolecules and assemblies with an eye for application in nanoscale engineering, one must keep in mind that they are optimized for organism fitness, and not necessarily for desired engineering properties, such as thermodynamic or mechanical stability. Most folded proteins are stabilized weakly by only small energy differences from the unfolded state, and are on the threshold of stability. Although counterintuitive, in fact it seems that overstabilization is likely to be avoided in the course of protein evolution [12]. Similarly, overstabilization of multicomponent assemblies is likely to be avoided. There may be benefit in having many weak interactions, because they can contribute co-operatively to the assembly process, and this may provide kinetic benefits for optimal rates while maintaining accuracy [13]. While mechanical and thermodynamic stability might be desired objectives for engineering, the weaker interactions that are associated with assembly of plastic components offers the possibility of designing co-operative effects [14], and so can be useful in applications where non-linear responses are the aim.

Nonetheless, Nature does provide many examples of assemblies having mechanical strength and thermodynamic stability, as, for example, in the adaptations necessary for life to exist at extremes of pH or temperatures above the boiling point of water and pressures in excess of an atmosphere. For example, the four-helix bundles that lie on the surface of thermophilic bacteria have a pattern of residues that allow association into a stable configuration involving the right-handed association of α-helices, in contrast with the left-handed form found in many helical assemblies that do not face the evolutionary pressure for maintaining thermal stability. In this case, Nature might provide some useful hints about forming mechanically stable assemblies.

The armour plating on the outermost surface of many prokaryotes, known as S-layers, is a salient example of a robust assembly. These layers are composed of identical protein subunits that vary in mass between 40 and 200 kDa, depending on species, and can form arrays of near-perfect crystalline regularity. The recombinant proteins of S-layers can self-assemble into sheets, cylinders and on the surface of liposomes (see below), and have already proved very useful in various applications to organize metals in crystalline sheets [15,16]. Similarly, filamentous viral coats with engineered metal-binding sites can self-assemble into structures with magnetic and semi-conducting properties [17].

Nature also provides examples of recognition specificity that can be used for engineering. The precision of base-pairing in duplex DNA has proved to be a very effective guide for engineering nanoscale assemblies [18,19]. The stability of duplex DNA arises from the co-operation of many numbers of complementary hydrogen bonds and from the considerable favourable energy of base stacking, which pack the atoms of duplex DNA with densities exceeding those for equivalent atoms in crystalline lattices of related small molecules. It might be possible to take the planar assemblies that have been engineered using DNA and to bring them into three-dimensional shapes, or to form dynamic and co-operative switches. This might be achieved using the stability of guanine tetraplexes in DNA and RNA, and their ability to co-ordinate dehydrated metals, such as sodium, potassium or thallium, or of certain recurring motifs for RNA secondary structure [20]. The use of modified bases that fluoresce upon pairing [21,22], or the binary engagement of sequence-specific binding proteins might permit the engineering of co-operativity of ligand binding or non-linear optical and conductive properties.

Symmetry and non-equivalence in assemblies

The cases shown so far are of complexes comprising many different components that have no apparent higher symmetry. Self-assembly is perhaps most intuitively apparent in polymers made of one or a few different types of subunits, such as the shells of viruses or the helical bundles that are used as propellers in certain bacteria. It would seem that perfect symmetry favours self-assembly, since the protomers are in equivalent environments and would be expected to lie at an energy minimum. However, biological assemblies may often demand geometrical imperfection in their construction. For instance, the apparent platonic ideality of icosahedral shells

of certain viruses can mask an underlying non-equivalence that is required to pack the subunits. Sixty identical subunits can indeed fit with perfect geometry on a shell with icosahedral symmetry and are related by one of the 2-fold, 3-fold or 5-fold elements of the icosahedral point group [23]. However, the capsids of many viruses have multiples of 60 chemically identical protomers. Most integer multiples of 60 can be accommodated through sub-triangulation of the surface, as described by Caspar and Klug [24]. The protomers in a sub-triangulated capsid shell encounter quasi-equivalent environments, and they meet the imperfection in part by making several different types of contact as well as through modest structural distortions [25]. The packing of protomers by quasi-equivalence is an economical means of packaging larger viral genomes.

Helices are another visually appealing application of symmetry, and, in Nature, they occur frequently in a range of mechanical structures. Salient amongst these are the flagella, which are whip-like organelles that are found on the surface of some bacteria and which propel the organism. The flagellum can be viewed as being composed of 11 co-axial protofilaments that are made of one type of protein: flagellin. In a resting state where all the protofilaments are perfectly aligned, the flagellum is straight. However, if one of the protofilaments shortens along its length, then it can occupy the innermost line of a corkscrew trajectory, with the consequence that the flagellum supercoils. The superhelical state can be modulated by the number of adjacent protofilaments that shortened [26], and details of the process have been explored experimentally and computationally by Keiichi Namba and colleagues [27]. The change in superhelical state originates from the mechanical force (torque) generated when the rotating motor at the base of the flagellum abruptly reverses direction. Thus we see that the flagellum represent a case of polymorphism where the protomers are capable of undergoing conformational adjustments so that the chemically identical subunits pack stably in geometrically non-equivalent ways.

The helical polymorphism of flagella and the quasi-equivalence of viral capsids both illustrate the role of conformational adjustment in protein–protein assembly. 'Conformability' also occurs in the formation of protein–nucleic acid complexes, and is a general feature of macromolecular recognition of biological molecules.

Association through an unstructured scaffold

Cells contain many regulatory assemblies that control the expression of genetic information, and which illustrate some of the principles of assembly that we have described above. One such system that we are studying is the multi-enzyme RNA degradosome from *Escherichia coli* (shown schematically in Figure 1A). The components of this large complex include RNase E, which is a ribonuclease that cleaves RNA substrates internally and PNPase (polynucleotide phosphorylase), a second type of ribonuclease that cleaves RNA at the 3′ end of the polymer [28]. RNase E cleaves single-stranded regions of RNA by a hydrolytic mechanism that is activated by an allosteric switch [29]. PNPase uses phosphate to cleave the 3′-terminal phosphodiester linkage in the backbone of the RNA substrate to release sequentially nucleotide diphosphates [30]. Another component of the assembly is the helicase RhlB (RNA helicase B), which can unwind secondary structure in RNA so that it becomes a suitable substrate for the nucleolytic components of the degradosome [31]. RhlB uses the free energy of ATP binding and hydrolysis to do the mechanical work of unwinding and translocating the folded RNA. The fourth main component of the *E. coli* degradosome is the glycolytic enzyme enolase, whose function is still yet to be determined. *In vivo* studies indicate that enolase may have a role in regulating gene expression in conjunction with small regulatory RNAs [32]. Molecular genetic studies show that the degradosome contributes to global regulation of mRNA levels, and disruption of the assembly affects many transcripts as well as organism fitness.

Based on our analysis of the protein–protein interactions between the components of the degradosome, we estimate that its cumulative mass may be in excess of 2.5 MDa [28]. In many stable complexes in the cell that are made of multiple components, the subunits associate through stable extensive interfaces. However, in the *E. coli* degradosome, the interactions of the four types of components appear to be mediated instead by small segments of polypeptide that may have little propensity for forming stable three-dimensional folds [33]. Our crystallographic data confirm that recognition of PNPase and enolase in the degradosome is mediated through small segments of the RNase E C-terminal domain, and that these segments do not form a globular fold (Figure 1B) ([28,34], and S. Nurmohamed and B.F. Luisi, unpublished work).

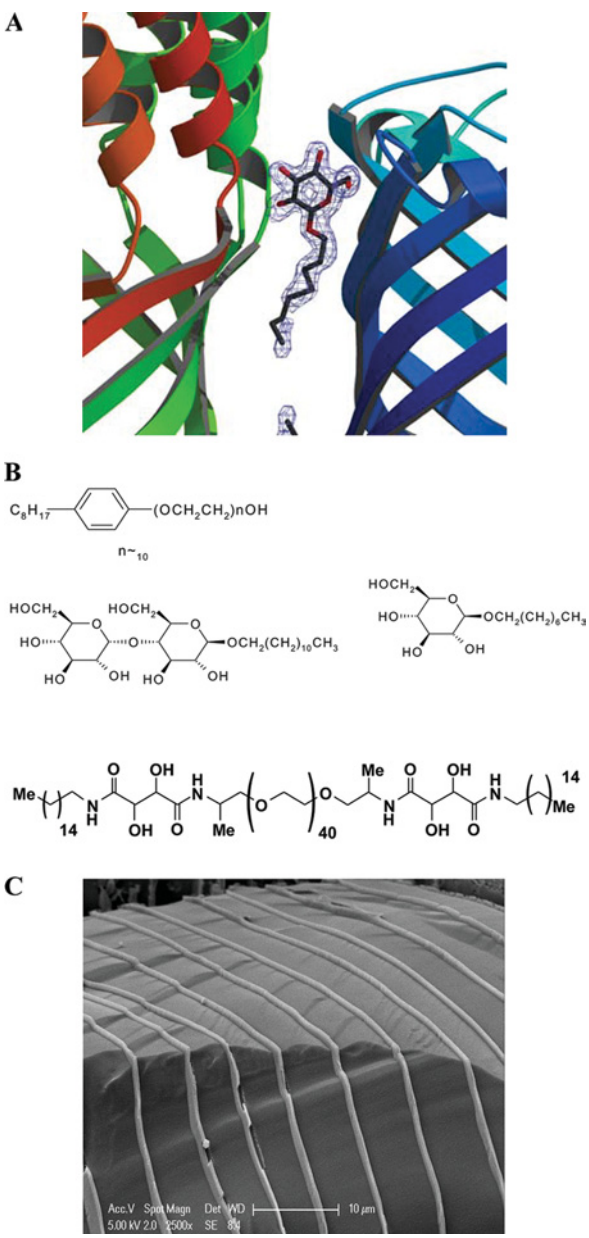
The use of small, unstructured motifs is a common theme in molecular recognition in all domains of life. It is a characteristic feature of weak interactions, and might provide some kinetic advantages in systems, such as signalling assemblies, where multiple components work together co-operatively to rapidly build and re-dissociate to affect a rapid binary switching between an inactive resting state and an activated state [35].

Self-assembly in natural and designed membranes

Membranes are another biological assembly that may offer some useful design principles for application. Cells are enveloped by a bimolecular layer of self-associating amphiphilic molecules that have non-polar and polar constituents; these form the internal structure and the solvent-exposed surface respectively (Figure 2A). In natural lipids, the non-polar component is a pair of acyl chains of 12–16 carbons and, in a typical bilayer, spans roughly 34 Å. Membranes are in dynamic equilibrium, and the layer has the property of a two-dimensional fluid that maintains a stable boundary with the bulk solvent. Some proteins can bind to the surfaces

Figure 2 | Self assembly of multiple order: organization of lipids and synthetic detergents in aqueous solution

(**A**) Natural phospholipids have a polar phosphate group and an apolar diglycerol chain. Detergents mimic this amphipathic nature. The detergent *n*-octyl-β-D-glucopyranoside is Commonly used to isolate and solubilize membrane proteins in crystallographic and biophysical studies. Shown are the interactions of *n*-octyl-β-D-glucopyranoside with the outer membrane protein VceC from *Vibrio cholerae* [38]. The electron density for the detergent is shown at the hydrophobic interface between two adjacent VceC molecules. Adapted from [37]. © 2006 Royal Society of Chemistry. (**B**) Chemical structures of popular detergents. Top: Triton X-100; left middle: *n*-dodecyl-β-D-maltoside; right middle; *n*-octyl-β-D-glucopyranoside. Bottom: a new type of detergent with a hydrophilic spacer. (**C**) Freeze–etch electron micrograph of meso-structures formed by the linked detergent shown in the bottom panel of (**B**). The sheets extend for many microns. The detailed structure within the sheets is not yet known.



of membranes and induce curvature, and so nucleate the formation of elongated tubes [36].

Proteins can be incorporated into membranes with stable globular folds. The fluctuating hydrocarbons provide a complementary, but dynamic, match to the non-polar surface of the protein and effectively exerts a type of lateral force on the protein. The stability of the protein fold is a delicate balance between the optimal packing of side chains on the protein interior, the accommodation of the transmembrane protein surface by the lipid hydrocarbons, and the polar interaction of the extramembraneous portions of the protein with the polar headgroup and the bulk solvent.

One difficulty faced in investigating membrane proteins is isolating them from the densely packed hydrophobic environment of the membrane into conditions where the proteins are monodispersed. This can be achieved by detergents which, like natural lipids, are amphipathic molecules with polar and non-polar domains, and, like natural lipids, they can self-assemble into complex organizations (Figure 2B). These can form smaller organizations, known as small spherical bodies called micelles, which have a very extensive and complex phase space and can partition into different structures, such as lamellar sheets or cubic lattices [37]. Detergents can form small highly curved surfaces that match the non-polar portion of the solubilized membrane protein, and can be visualized in some crystal structures of membrane proteins (Figure 2A) [38]. One shortcoming of detergents is that they only approximate the hydrocarbon density. Lipopeptides may overcome this limitation by mimicking the planar organization of the membrane, and they have proved useful for solubilizing and stabilizing membrane proteins [39,40].

We have been exploring the use of linked detergents on the formation of lamellar sheets of membranes. Preliminary analyses of the assemblies formed by a mixture of the bis-detergent and conventional detergent indicate that it can form extensive two-dimensional sheets that extend for hundreds of micrometers (Figure 2C). It might be possible to combine this material with the specificity of nucleic acid base-pairing to form elaborate three dimensional lattices for many different applications.

Our work is supported by the Wellcome Trust.

References

1 Nomura, M., Traub, P. and Bechmann, H. (1968) Nature **219**, 793–799
2 Nierhaus, K.H. and Dohme, F. (1974) Proc. Natl. Acad. Sci. U.S.A. **71**, 4713–4717
3 Dobson, C.M. (2003) Nature **426**, 884–890
4 Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B. et al. (2006) Nature **440**, 631–636
5 Wright, C.F., Teichmann, S.A., Clarke, J. and Dobson, C.M. (2005) Nature **438**, 878–881
6 Chandler, D. (2002) Nature **417**, 491
7 Chandler, D. (2005) Nature **437**, 640–647
8 Lo Conte, L., Chothia, C. and Janin, J. (1999) J. Mol. Biol. **285**, 2177–2198
9 Chakrabarti, P. and Janin, J. (2002) Proteins Struct. Funct. Genet. **47**, 334–343

10 Jones, S. and Thornton, J.M. (1996) Proc. Natl. Acad. Sci. U.S.A. **93**, 13–20
11 Russell, R.B., Alber, F., Aloy, P., Davis, F.P., Korkin, D., Pichaud, M., Topf, M. and Sali, A. (2004) Curr. Opin. Struct. Biol. **14**, 313–324
12 DePristo, M.A., Weineich, D.M. and Hartl, D.L. (2005) Nat. Rev. Genet. **6**, 678–687
13 Hofmann, K.P., Spahn, C.M.T., Heinrich, R. and Heinemann, U. (2006) Trends Biochem. Sci. **31**, 497–508
14 Dwyer, M.A. and Hellinga, H.W. (2004) Curr. Opinion Struct. Biol. **14**, 495–504
15 Sleytr, U.B., Egelseer, E.M., Ilk, N., Pum, D. and Schuster, B. (2007) FEBS J. **274**, 323–334
16 Mark, S.S., Bergkvist, M., Yang, X., Angert, E.R. and Batt, C.A. (2006) Biomacromolecules **7**, 1884–1897
17 Mao, C., Solis, D.J., Reiss, B.D., Kottmann, S.T., Sweeney, R.Y., Hayhurst, A., Georgiou, G., Iverson, B. and Belcher, A.M. (2004) Science **303**, 213–217
18 Rothermund, P.W.K. (2006) Nature **440**, 297–302
19 Seeman, N.C. (2005) Q. Rev. Biophys. **38**, 363–371
20 Nissen, P., Ippolito, J.A., Ban, N., Moore, P.B. and Steitz, T.A. (2001) Proc. Natl. Acad. Sci. U.S.A. **98**, 4899–4903
21 Kool, E., Lu, H., Kim, S.J., Tan, S., Wilson, J.N., Gao, J. and Liu, H. (2006) Nucleic Acids Symp. Ser. **50**, 1–16
22 Lynch, S.R., Liu, H., Gao, J. and Kool, E.T. (2006) J. Am. Chem. Soc. **128**, 14704–14711
23 Crick, F.H. and Watson, J.D. (1956) Nature **177**, 473–475
24 Caspar, D.L.D. and Klug, A. (1962) Cold Spring Harbor Symp. Quant. Biol. **27**, 1–24
25 Harrison, S.C. (2001) Curr. Opin. Struct. Biol. **11**, 195–199
26 Calladine, C.R. (1975) Nature **255**, 121–124
27 Kitao, A., Yonekura, K., Maki-Yonekura, S., Samatey, F.A., Imada, K., Namba, K. and Go, N. (2006) Proc. Natl. Acad. Sci. U.S.A. **103**, 4894–4899
28 Marcaida, M.J., DePristo, M.A., Cahndran, V., Carpousis, A.J. and Luisi, B.F. (2006) Trends Biol. Sci. **31**, 359–365
29 Callaghan, A.J., Marcaida, M.J., Stead, J.A., McDowall, K.J., Scott, W.G. and Luisi, B.F. (2005) Nature **437**, 1187–1191
30 Symmons, M.F., Jones, G.H. and Luisi, B.F. (2000) Structure **8**, 1215–1226
31 Chandran, V., Poljak, L., Vanzo, N.F., Leroy, A., Miguel, R.N., Fernandez-Recio, J., Parkinson, J., Burns, C., Carpousis, A.J. and Luisi, B.F. (2007) J. Mol. Biol. **367**, 113–132
32 Morita, T., Kawamoto, H., Mizota, T., Inada, T. and Aiba, H. (2004) Mol. Microbiol. **54**, 1063–1075
33 Callaghan, A.J., Aurikko, J.P., Ilag, L.L., Grossmann, J.G., Chandran, V., Kuhnel, K., Poljak, L., Carpousis, A.J., Robinson, C.V., Symmons, M.F. and Luisi, B.F. (2004) J. Mol. Biol. **340**, 965–979
34 Chandran, V. and Luisi, B.F. (2006) J. Mol. Biol. **358**, 8–15
35 Neduva, V., Linding, R., Su-Angrand, I., Stark, A., de Masi, F., Gibson, T.J., Lewis, J., Serrano, L. and Russell, R.B. (2005) PloS Biol. **3**, e405
36 Gallop, J.L., Jao, C.C., Kent, H.M., Butler, P.J., Evans, P.R., Langen, R. and McMahon, H.T. (2006) EMBO J. **25**, 2898–2910
37 Walas, F., Matsumura, H. and Luisi, B. (2006) in Structural Biology of Membrane Proteins (Grisshammer, R. and Buchanan, S.K., eds.), RSC Publishing, London
38 Federici, L., Du, D., Walas, F., Matsumura, H., Fernandez-Recio, J., McKeegan, K.S., Borges-Walmsley, M.I., Luisi, B.F. and Walmsley, A.R. (2005) J. Biol. Chem. **280**, 15307–15314
39 Kelly, E., Prive, G.G. and Tieleman, D.P. (2005) J. Am. Chem. Soc. **127**, 13446–13447
40 Kiley, P., Zhao, X., Vaughn, M., Baldo, M.A., Bruce, B.D. and Zhang, S. (2005) PloS Biol. **3**, e230

Received 9 January 2007