Whither structural biology?

Stephen C Harrison

We know the basic principles of protein, RNA and DNA structure, and we have atomic coordinates of many proteins and RNAs. Structural biology must now expand the range of length and timescales on which we can represent the molecular reality of a cell. Structural molecular biology and structural cell biology must merge into a single discipline, and we must establish a lively intellectual complementarity with the nascent 'systems biology' of the cell.

Biology rests on structural observation. Vesalius, Cajal, Spemann, Palade and Perutz all, in some sense, practiced 'structural biology.' What structural biology has come to mean during the past two decades—at least as it is used in the name of this journal—is a combination of the heritages of Perutz and Palade: structural molecular biology and structural cell biology. One short answer to the question in the title of this essay is that the next decade will (or should) see the fusion of these traditions into a unified discipline.

Structure has not been the principal theme in twentieth-century biology. 'Informational biology,' an intellectual thread that sprang from the discovery of genes and biochemical pathways and linked them to chemistry through classical molecular biology, has clearly dominated biological discourse. Its latest avatar, following the successes of genomics, is 'systems biology,' and the information transfer in question involves physiological signals rather than genetic specifications. Analysis of the transfer of genetic information—articulation of the 'central dogma'—was able to proceed with relatively little structural input, other than the crucial structure of DNA itself. Physiological signals are by their nature transient, contingent and complex. We are far less likely to understand them without structure, much less manage them or change them, than would be possible for the processes that determine sequences of proteins from those in DNA. Systems biology of the cell and structural

The author is at the Howard Hughes Medical Institute, Harvard Medical School and Children's Hospital, Boston, Massachusetts 02115, USA.
e-mail: harrison@crystal.harvard.edu

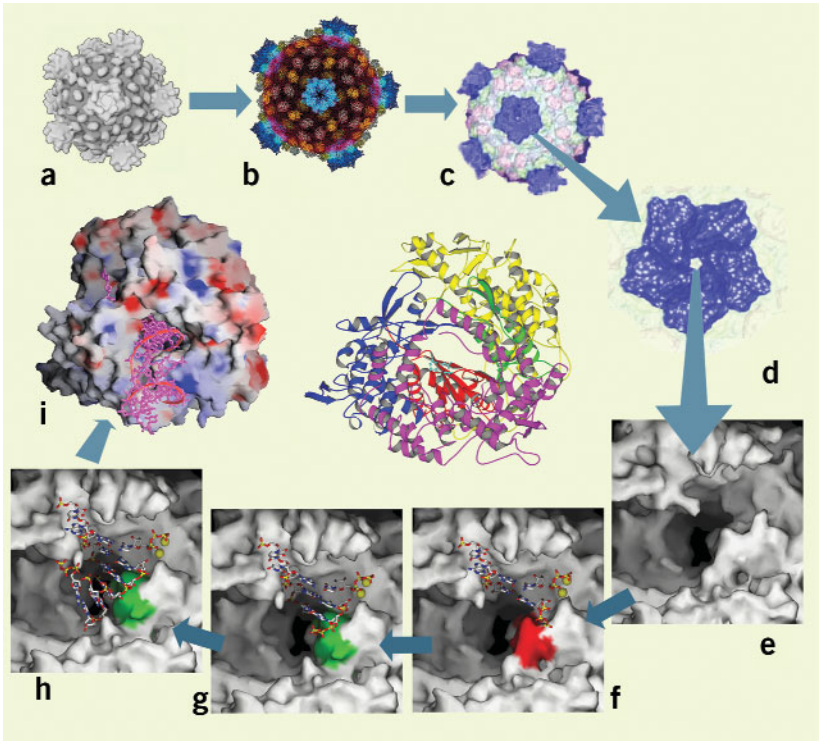

Figure 1 Reovirus cores as molecular machines. The reovirus core transcribes, caps, and exports mRNA, using segments of its dsRNA genome as templates. (**a**) Reconstruction of the reovirus core from cryoEM images, at ~25 Å resolution[21]. The diameter of the core is ~700 Å. (**b**) Crystal structure of the core, 3.8 Å resolution[22]. (**c**) Surface view, based on the crystal structure. (**d**) Each of the 5-fold turrets is a 'capping chamber', through which the nascent RNA passes en route to the cytoplasm. The subunits of the turret protein each have catalytic sites for guanyl transferase, *N*-methylase and *O*-methylase activities[22]. (**e**) Buried within the shell of the core, but tethered near each exit portal, are copies of the dsRNA-dependent, RNA polymerase (ribbon diagram in the center of the figure: ref. 23). The catalytic site of the polymerase lies within a cage-like superstructure. (**f**) An initiation complex, containing a template strand, a primer nucleotide, and a substrate nucleotide[23]. (**g**) The result of a first elongation step. (**h**) The catalytic complex after four elongation steps. Panels **e**–**h** are images derived directly from crystal structures[23]. Phosphodiester bond formation occurred in the polymerase crystals. (**i**) Surface representation of the polymerase, with a template strand entering and a double-strand product exiting. This figure is based on a model for the reaction in which the same polymerase generates a dsRNA genomic segment from a packaged ssRNA segment. A large series of structures, assembled by combining images derived from electron microscopy and X-ray crystallography, can yield the outlines of a 'molecular movie'.

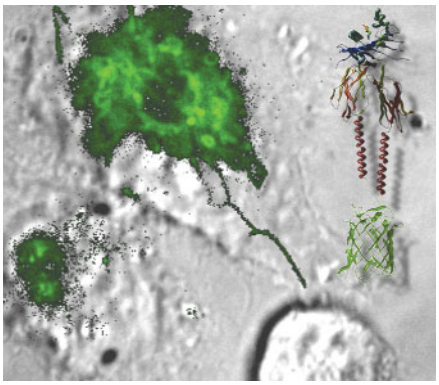

Figure 2 Dynamics of molecular transport *in vivo*, illustrated by a rapid, signal-induced redistribution of MHC class II molecules[24]. Tubules (green projections) emanate from the endosomal compartment (also green) of a dendritic cell, labeled with a chimeric protein in which EGFP has been fused to the cytoplasmic tail of an MHC class II protein (MHC-II–EGFP, for molecular representation, see inset). The dendritic cell, from a transgenic mouse that expresses MHC-II–EGFP, has been induced to display a particular peptide and then stimulated by contact with a T cell (lower right), from a hybridoma specific for the same peptide. The T cell induces tubulation of the endosome toward the point of contact.

biology as I describe it here are thus complementary ways to arrive at a complete description of how a cell works.

Structural biology in 2004

Where does structural biology stand in 2004? We understand most of the basic principles of protein, RNA and DNA structure. We have atomic structures for paradigms of many (perhaps most) of the important classes of proteins, of many of the key types of RNA and of a handful of larger assemblies[1,2]. Secondary structural elements form the architectural frameworks of protein domains, and the now ubiquitous ribbon diagrams succeed in communicating important three-dimensional relationships precisely for this reason. The modularity of proteins may be their most important characteristic for physiological signal transfer and response. Evolution can experiment with novel protein associations simply by recombining recognition modules. Interaction domains, switches and scaffolds have generated some of the most surprising and interesting structural discoveries of the last decade, and there seem to be more to find.

Membrane proteins, an almost unexplored frontier just ten years ago, now hold slightly fewer mysteries. Structures of ion channels[3], small-molecule transporters[4] and the secreted-protein translocation pore[5] demonstrate some of the devices that enable polar molecules to cross a lipid barrier. We have begun to see how membrane proteins can sense mechanical stress[6] and transmembrane potential[7].

From the point of view of technology, NMR spectroscopy, X-ray crystallography and electron microscopy now form a smooth continuum for visualizing macromolecular structures, from small proteins and RNA fragments to viruses, ribosomes and cytoskeletal filaments (**Fig. 1**). During the past few years, it has become clear that single-particle electron cryomicroscopy can indeed reach subnanometer resolution[8]. Atomic models can thus be inserted accurately into moderate-resolution image reconstructions of larger assemblies.

Confocal and deconvolution fluorescence microscopy and genetically targeted, expressible fluorophores (green fluorescent protein and its derivatives) have revolutionized structural cell biology. The static images of classical, thin-section electron microscopy can be linked to dynamic images in real time in living cells, with biochemically specific labels identifying the proteins observed (**Fig. 2**). Although optical microscopy does not offer the spatial resolution of the electron microscope, practical ways to breach the Rayleigh limit have been developed (reviewed in ref. 9), and we can look forward to a smoother link between electron and optical microscopy in the relatively near future.

What is in store for the next ten years? In the following sections I suggest that the answer to this question has three linked parts: more complete structural pictures of the molecular machines that execute a cell's principal activities, detailed analyses of the dynamics of these assemblies, and clear connections between their properties studied *in vitro* and their behavior in the context of a living cell.

Molecular machines

A few of the essential molecular machines of the cell, such as the F1 ATP synthase[10] and the ribosome[11], are relatively invariant entities (although even the ribosome has 'factors' that come on and off, and it dissociates into two major subunits when it has finished synthesizing a polypeptide chain). But many are transient, with a variety of states and a succession of structures. DNA replication complexes or transcription initiation complexes are a succession of devices in which signal input or progression from one state to another corresponds to a substantial reconstruction of the machine. Thus, the first key challenge for structural biology in the coming years is to visualize these stages of reconstruction for important subcellular molecular assemblies.

As in other areas of cell biology, viruses have often been the simplified tools of basic discovery. Conformational transformations of viral proteins and viral-protein assemblies have yielded some of the most striking images of the workings of molecular machines. The Ca^{2+}- and pH-regulated expansion of small RNA plant viruses, the reorganization of influenza-virus hemagglutinin, and the transformations and activities of the reovirus particle are all useful examples (reviewed in ref. 12). Analysis of bacteriophage assembly, through which many of the basic notions of regulated protein association were first adduced, has returned at the atomic-resolution level—for example, through a combination of X-ray crystallographic, electron microscopic and single-molecule biophysical experiments on phage ϕ29 (ref. 13). The extent to which we understand—or will understand—reovirus or ϕ29 organization and reorganization is a standard for what we can do with assemblies such as replication origins, spliceosomes and nuclear pores.

From molecular machines to molecular movies

X-ray crystallography and electron microscopy provide snapshots of molecules or molecular assemblies at defined stages of their activity. Stringing these snapshots together into a movie is not always straightforward, both because some of the intermediates are inaccessible for direct structural studies and because directed mutagenesis, the best link between structure and function in many circumstances, has proved to be a blunter instrument of dissection than we might have wished. Structural biologists want to think in terms of individual molecules and individual structures rather than ensembles of molecules and ensembles of structures. Moreover, cells have only one or two copies of each gene, and biological processes are inherently phenomena carried out by a relatively small number of molecular effectors. The advent of single-molecule methods for studying a wide variety of protein functions is thus a major event.

The relevant timescales in cell biology range from 10^{-6} s for some diffusional events to 10^{2} s or longer for some intracellular processes. Individual molecular events happen rapidly, however, and relatively high time resolution will probably be required to determine mechanism. The successes and limitations of single-molecule studies of protein unfolding[14] and molecular-motor activity[15] probably illustrate the current state of experimental methods. We can certainly anticipate being able to look at individual nascent polypeptides folding (or being trapped by chaperones) as they emerge from the ribosome, individual polypeptide chains translocating across an endoplasmic reticulum or mitochondrial membrane and individual RNA molecules passing through a

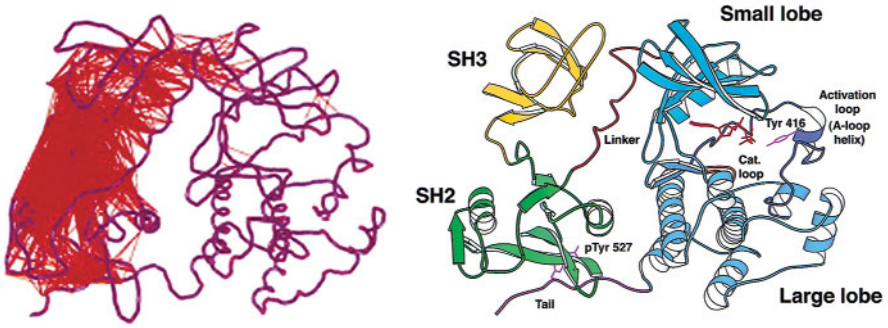

Figure 3 Activation dynamics of c-Src tyrosine kinase. The autoregulated state of Src-family kinases is an assembled arrangement of three domains (the Src-homology 3 or SH3 domain, the Src-homology 2 or SH2 domain, and the kinase domain)[25]. The SH3-SH2 regulatory module is held in place by a 'latch' formed by association of a phosphotyrosine-containing 'tail' with the specificity pocket of the SH2 domain. How does this arrangement, which does not occlude access to the catalytic site, inhibit the kinase? Molecular dynamics calculations and mutational data suggest that the regulatory module clamps the two lobes of the kinase domain firmly together[19]. Red lines in the diagram on the left show correlated motions of individual residues during a simulation that mimics initial steps in activation. The SH3 and SH2 domains move as a concerted unit. Displacement of this regulatory module will release constraints on the kinase.

nuclear pore. It will likewise soon be possible to visualize formation of individual membrane fusion pores with fluorescently tagged fusion proteins.

Biologists studying ion channels have known about the advantages of single-molecule analysis, in the form of the patch clamp, for over two decades. Structural biologists interested in molecular motors were the first to use single-molecule methods in a structural context, because they could use an entire filament (of actin or tubulin) as their reporter and because the structures of myosin and kinesin became available not too long after the initial single-filament experiments. The level at which we now understand the mechanism and dynamics of muscle contraction (from a series of experiments spanning more than four decades by Huxley, Taylor, Holmes, Spudich and many others (recently reviewed in ref. 16)) is a standard for how we should analyze dynamic processes generally.

The living cell

Some structures and some dynamic processes, such as clathrin coats and processive motion of kinesin along a microtubule, can be studied well with purified proteins *in vitro*. Others, such as patterns of membrane traffic, require an intact cell (or in favorable cases a permeabilized or broken cell, or perhaps an extract). It is here, I would argue, that we find the most challenging, but also the most exciting, possibilities for the fusion of the molecular and cellular traditions in structural biology. How can we integrate isolated pieces of structural information about actin, Arp2/3, N-WASP, Cdc42 and other components into a coherent picture of the behavior of the actin cytoskeleton? How do we relate the sequence of events we see during formation, budding and targeting of a clathrin-coated vesicle to the underlying structures and interactions of clathrin, adaptors, Eps15, dynamin, auxilin, Hsc70, SNAREs and other proteins, not to mention membrane lipids? How do we relate the behavior of the mitotic spindle to the structure and properties of tubulin, MAPs, motors, kinetochore proteins and so on? What is the molecular mechanism of chromosome condensation? The answers to questions like these will come from a combination of the directions outlined in the preceding two sections, coupled with innovations in live-cell imaging and developments in tomography by electron cryomicroscopy. More powerful live-cell imaging will probably require three sorts of methodological advances: super-resolution microscopes (for example, as developed by Hell, Gustafsson and others[9]), new image-analysis software[17] and even more powerful and flexible fluorescent tags[18]. Indications of how each of these may proceed are already in the literature.

What is the role of structural genomics?

The activity known as structural genomics, application of high-throughput organization and method to structure determination, has been advertised as a principal new direction in structural biology. Within the context of the broader research program outlined above, structural genomics will have at best an ancillary rather than a central function. Large libraries of structures will certainly be useful, and they may even make homology modeling somewhat more successful. But interesting biological function is likely to reside in the most variable and least predictable features of a protein, even when its framework can be derived confidently and accurately from those of homologous proteins. Indeed, experience so far suggests that even quite dense libraries of structures of particular folds fail to allow useful prediction of critical functional elements. It may be a jump from this critique to the conclusion that National Institutes of Health dollars are being wasted on high-throughput structural biology, but I have no doubt that what we really need for major advances in biological understanding is better integration of information on multiple length and timescales, as suggested by the kinds of studies shown in **Figures 1** and **2**.

There is, of course, a sense in which we are all 'structural genomicists.' The information derived from genome sequences informs everything we do in biology, including our choice of structural problems. Moreover, a structural approach to any complex molecular machine will almost always require determining a catalog of structures, compiled from a list of the components. Again, complex viruses are good examples. The way we have gone about understanding the reovirus infectious cycle (**Fig. 1**) and the way that other groups have studied adenovirus show that determining a complete set of substructures is an essential first step toward characterizing dynamic processes such as viral entry or regulated viral assembly. The international (and still unfinished) effort to determine structures for all the HIV-specific and HIV-related proteins illustrates that structural genomics, in the sense of generating a structural catalog, can have a substantial impact on therapeutics as well as on basic mechanistic understanding. But the time, talent and biological insight devoted to that international effort have been exceptional, and no high-throughput approach could possibly have done as well.

What is the role of computational modeling?

Even if homology modeling and structure prediction still have a long way to go, it would clearly be silly to throw up our hands and stop trying. But the time and distance scales of special interest in creating 'molecular movies' of events in a cell suggest other kinds of efforts. Some of the interesting transformations of protein assemblies (expansion of viral capsids and processive transcriptional elongation) are like 'displacive' solid-state phase changes, involving relatively continuous and concerted molecular motions. Others (the fusion-promoting conformational change in

influenza hemagglutinin and the activation of N-WASP by Cdc42) are 'reconstructive,' involving unfolding and refolding. The former are probably good initial targets for simulation. What sorts of mean-field approximations will be necessary to model these reasonably continuous transformations, which tend to occur on a distance scale of 10–100 Å and a timescale of 10^{-2} to 10^{1} s? Conventional molecular dynamics can begin to nibble at the short-distance and rapid-timescale end of this range. My favorite illustration is a series of computations carried out by Kuriyan's group[19] on activation of Src-family kinases (**Fig. 3**). They suggest a concerted displacement of the SH3 and SH2 modules and thereby lead to a mechanical description of how their correlated arrangement inhibits the kinase. Experiments on a mutated kinase are consistent with the model hinted at by the simulations.

On coarser time and distance scales, can we derive a computational scheme for simulating more complex events in a cell—for example, the back-and-forth jumps along a filament that a processive motor might make, or the translocation of a polypeptide chain across a pore in a membrane, or the movement and targeting of transport vesicles? Each level of modeling will require a specific physical and mathematical formalism and a specific computational framework, and may require a specific programming language.

Specific goals for the coming decade

In view of the thoughts outlined above, what kinds of experimental results might we hope to have ten years from now? At the purely architectural level, we would certainly like to see the details of the mitotic spindle, the organization of a condensed chromosome, and the atomic structures of a nuclear pore, a coated vesicle and an assembled enhanceosome. At a mechanistic level, we would like to understand—at 1 ms or faster and at 2.5 Å or better—myosin, kinesin, dynein, AAA+ and ABC ATPases, topoisomerases, helicases, polymerases, chromatin remodelers and ribosomes. Single-molecule biophysical experiments will be essential for adding the time dimension. In membrane biology, we still need structural answers to the following questions. What are the principles by which protein assemblies are organized on membranes? What is the significance of the large variety of lipids in a membrane, only a few of which (phosphoinositides, for example) have well defined specific functions? How do the bilayer rearrangements in membrane fusion and fission really proceed? And of course we need a complete understanding of how channels and pumps really work.

In this consciously one-sided account, I have slighted another important structural biology interface—that with chemistry. Genuinely atomic-resolution protein structures are resolving old mysteries and in some cases exposing fundamental surprises. The high-resolution structure of a bacterial potassium channel showed positions of water molecules and ions in the pore, answering some critical questions about the mechanism of rapid yet specific ion transport. A recent structure of the MoFe protein of bacterial nitrogenase at a resolution of 1.16 Å (from Rees and co-workers[20]) revealed a completely unexpected light atom (probably a nitrogen) within the metal cluster that catalyzes one of the most crucial chemical reactions in the biosphere. One hundred years after Haber, we still do not know exactly how nitrogen is reduced by living organisms.

What about biomedical applications? Structure-based development of therapeutics is now well integrated into the pharmaceutical industry, but primarily in the context of drugs directed at enzyme active sites. There are daring and important problems for which academic foundations still need to be created. Some of them involve precisely the kinds of goals listed in the preceding paragraph. Surely there is a better way to inhibit AAA+ ATPases, for example, than by targeting the ATP-binding site. In the specific realm of infectious diseases, there are splendid problems in 'grand' architecture that are also important for the mechanism of bacterial pathogenesis: type III secretion systems and the multidomain enzymes that produce polypeptide and polyketide antibiotics are obvious examples. From a broader perspective, there is a serious need to base design and development of vaccines on a molecular-structural foundation. Such a change will come about only if we educate a generation of vaccine developers in the principles of structural biology, as we educated an earlier generation of drug developers.

From structures to systems

The complexity that makes a eukaryotic cell a really interesting 'system' comes not primarily from the molecular activities it contains but rather from the intricacy with which they are regulated. The logic of intracellular signaling involves scaffolds, switches and sequential states in ways that are only beginning to become clear. There are hints that specific kinds of control logic are embodied in specific kinds of molecular architecture—a conclusion that makes sense in terms of the evolution of regulation but that may or may not hold out as we analyze signaling further, from both the structural end and the informational end (see, for example, **Fig. 3**). A fundamental understanding of the structural foundations of intracellular regulation will be essential for rational experimental analysis and for rational intervention when dysregulation leads to disease; it will probably be equally essential for understanding the nature of the information in any particular regulatory network. Thus, structural biology must seek to understand information transfer in terms of its underlying molecular agents by analyzing the molecular hardware that executes the information-transfer software. Unlike most man-made computers, the hardware and software of physiological regulation co-evolved. The possibilities for storage, retrieval, transfer and destruction of information are not independent of the molecular devices that execute these functions. The architectural principles of the cell's control systems and the dynamics of their operation are no less proper studies of structural biology than are the organizational and dynamical properties of the molecular machines that execute the regulated commands.

1. Branden, C. & Tooze, J. *Introduction to Protein Structure* (Garland, New York, 1999).
2. Petsko, G. & Ringe, D. *Protein Structure and Function* (Sinauer Associates, Sunderland, Massachusetts, USA, 2003).
3. Doyle, D.A. *et al. Science* **280**, 69–77 (1998).
4. Huang, Y., Lemieux, M.J., Song, J., Auer, M. & Wang, D.N. *Science* **301**, 616–620 (2003).
5. van den Berg, B. *et al. Nature* advance online publication, 3 December 2003 (doi:10.1038/nature 02218).
6. Chang, G., Spencer, R.H., Lee, A.T., Barclay, M.T. & Rees, D.C. *Science* **282**, 2220–2226 (1998).
7. Jiang, Y. *et al. Nature* **423**, 33–41 (2003).
8. Bottcher, B., Wynne, S.A. & Crowther, R.A. *Nature* **386**, 88–91 (1997).
9. Gustafsson, M.G. *Curr. Opin. Struct. Biol.* **9**, 627–634 (1999).
10. Abrahams, J.P., Leslie, A.G., Lutter, R. & Walker, J.E. *Nature* **370**, 621–628 (1994).
11. Ramakrishnan, V. *Cell* **108**, 557–572 (2002).
12. Harrison, S.C. In *Fields Virology* 4th edn (eds. Knipe, D.M. & Howley, P.M.) 53–85 (Lippincott Williams and Wilkins, Philadelphia, 2002).
13. Tao, Y. *et al. Cell* **95**, 431–437 (1998).
14. Oberhauser, A.F., Hansma, P.K., Carrion-Vazquez, M. & Fernandez, J.M. *Proc. Natl. Acad. Sci. USA* **98**, 468–472 (2001).
15. Yildiz, A *et al. Science* **300**, 2061–2065 (2003).
16. Geeves, M.A. & Holmes, K.C. *Annu. Rev. Biochem.* **68**, 687–728 (1999).
17. Swedlow, J.R., Goldberg, I., Brauner, E. & Sorger, P.K. *Science* **300**, 100–102 (2003).
18. Zhang, J., Campbell, R.E., Ting, A.Y. & Tsien, R.Y. *Nat. Rev. Mol. Cell Biol.* **3**, 906–918 (2002).
19. Young, M.A., Gonfloni, S., Superti-Furga, G., Roux, B. & Kuriyan, J. *Cell* **105**, 115–126 (2001).
20. Einsle, O. *et al. Science* **297**, 1696–1700 (2002).
21. Dryden K.A. *et al. J. Cell Biol.* **122**, 1023–1041 (1993).
22. Reinisch, K.M., Nibert, M.L. & Harrison, S.C. *Nature* **404**, 960–967 (2000).
23. Tao, Y., Farsetta, D.L., Nibert, M.L. & Harrison, S.C. *Cell* **111**, 733–745 (2002).
24. Boes, M. *et al. Nature* **418**, 983–988 (2002).
25. Xu, W., Doshi, A., Lei, M., Eck, M.J. & Harrison, S.C. *Mol. Cell* **3**, 629–638 (1999).