VIEWPOINT

Protein Structure Prediction and Structural Genomics

David Baker[1] and Andrej Sali[2]

Genome sequencing projects are producing linear amino acid sequences, but full understanding of the biological role of these proteins will require knowledge of their structure and function. Although experimental structure determination methods are providing high-resolution structure information about a subset of the proteins, computational structure prediction methods will provide valuable information for the large fraction of sequences whose structures will not be determined experimentally. The first class of protein structure prediction methods, including threading and comparative modeling, rely on detectable similarity spanning most of the modeled sequence and at least one known structure. The second class of methods, de novo or ab initio methods, predict the structure from sequence alone, without relying on similarity at the fold level between the modeled sequence and any of the known structures. In this Viewpoint, we begin by describing the essential features of the methods, the accuracy of the models, and their application to the prediction and understanding of protein function, both for single proteins and on the scale of whole genomes. We then discuss the important role that protein structure prediction methods play in the growing worldwide effort in structural genomics.

Modeling of a sequence based on known structures consists of four steps: finding known structures related to the sequence to be modeled (i.e., templates), aligning the sequence with the templates, building a model, and assessing the model (*1*).

The templates for modeling may be found by sequence comparison methods, such as PSI-BLAST (*2*), or by sequence-structure threading methods (*3*) that can sometimes reveal more distant relationships than purely sequence-based methods. In the latter case, fold assignment and alignment are achieved by threading the sequence through each of the structures in a library of all known folds. Each sequence-structure alignment is assessed by the energy of a corresponding coarse model, not by sequence similarity as in sequence comparison methods.

Comparative structure prediction produces an all-atom model of a sequence, based on its alignment to one or more related protein structures. Comparative model building includes either sequential or simultaneous modeling of the core of the protein, loops, and side chains. In the original comparative approach, a model is constructed from a few template core regions and from loops and side chains obtained from either aligned or unrelated structures (*4–6*). Another family of comparative methods relies on approximate positions of conserved atoms from the templates to calculate the coordinates of other atoms (*7*). A third group of methods uses either distance geometry or optimization techniques to satisfy spatial restraints obtained from the sequence-template alignment (*8–10*). There are also many methods that specialize in the modeling of loops (*11*) and side chains (*12*) within the restrained environment provided by the rest of the structure.

[1]Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA. E-mail: dabaker@u.washington.edu. [2]Laboratory of Molecular Biophysics, Pels Family Center for Biochemistry and Structural Biology, The Rockefeller University, New York, NY 10021, USA. E-mail: sali@rockefeller.edu.

De novo Structure Prediction

Although comparative modeling is limited to protein families with at least one known structure, de novo structure prediction has no such limitation. De novo methods start from the assumption that the native state of a protein is at the global free energy minimum and carry out a large-scale search of conformational space for protein tertiary structures that are particularly low in free energy for the given amino acid sequence. The two key components of such methods are the procedure for efficiently carrying out the conformational search and the free energy function used for evaluating possible conformations. To allow rapid and efficient searching of conformational space, often only a subset of the atoms in the protein chain is represented explicitly; the potential functions must then include terms that reflect the averaged-out effects of the omitted atoms and solvent molecules.

Recently, there have been a number of promising advances in de novo structure prediction (*13–16*). A particularly successful method, called Rosetta, is based on a picture of protein folding in which short segments of the protein chain flicker between different local structures consistent with their local sequence, and folding to the native state occurs when these local segments are oriented such that low free energy interactions are made throughout the protein (*17*). In simulating this process, each short segment is allowed to sample the local structures adopted by the sequence segment in known protein structures, and a search is carried out through the combinations of these local structures for compact tertiary structures that bury the hydrophobic residues and pair the β-strands. This strategy resolves some of the problems with both the conformational search and the free energy function: The search is greatly accelerated because switching between different possible local structures can occur in a single step, and fewer demands are placed on the free energy function because the use of fragments of known structures ensures that the local interactions are close to optimal.

Accuracy and Applications of Models

The accuracy of a comparative model is related to the percentage sequence identity on which it is based, correlating with the relationship between the structural and sequence similarity of two proteins (Fig. 1) (*1*, *18*, *19*). High-accuracy comparative models are based on more than 50% sequence identity to their templates. They tend to have about 1 Å root mean square (RMS) error for the main-chain atoms, which is comparable to the accuracy of a medium-resolution nuclear magnetic resonance (NMR) structure or a low-resolution x-ray structure. The errors are mostly mistakes in side-chain packing, small shifts or distortions of the core main-chain regions, and occasionally larger errors in loops. Medium-accuracy comparative models are based on 30 to 50% sequence identity. They tend to have about 90% of the main-chain modeled with 1.5 Å RMS error. There are more frequent side-chain packing, core distortion, and loop modeling errors, and there are occasional alignment mistakes (*18*). Finally, low-accuracy comparative models are based on less than 30% sequence identity. The alignment errors increase rapidly below 30% sequence identity and become the most substantial origin of errors in comparative models. In addition, when a model is based on an almost insignificant alignment to a known structure, it may also have an entirely incorrect fold. Accuracies of the best model building methods are relatively similar when used optimal-

ly (*19*, *20*). Other factors such as template selection and alignment accuracy usually have a larger impact on the model accuracy, especially for models based on less than 40% sequence identity to the templates.

There is a wide range of applications of protein structure models (Figs. 1 and 2). For example, high- and medium-accuracy comparative models frequently are helpful in refining functional predictions that have been based on a sequence match alone because ligand binding is more directly determined by the structure of the binding site than by its sequence. It is often possible to correctly predict features of the target protein that do not occur in the template structure. The size of a ligand may be predicted from the volume of the binding site cleft (Fig. 2A). For example, the complex between docosahexaenoic fatty acid and brain lipid-binding protein was modeled on the basis of its 62% sequence identity to the crystallographic structure of adipocyte lipid-binding protein (PDB code 1ADL) (*21*). A number of fatty acids were ranked for their affinity to brain lipid-binding protein consistently with site-directed mutagenesis and affinity chromatography experiments, even though the ligand specificity profile of this protein is different from that of the template structure. Another example is prediction of a binding site for a charged ligand based on a cluster of charged residues on the protein, as was done for mouse mast cell protease 7 (Fig. 2B) (*22*). The prediction of a proteoglycan binding patch was confirmed by site-directed mutagenesis and heparin-affinity chromatography experiments. Fortunately, errors in the functionally important regions in comparative models are many times relatively low because the functional regions, such as active sites, tend to be more conserved in evolution than the rest of the fold. The utility of low-accuracy comparative models can be illustrated by a molecular model of the whole yeast ribosome, whose construction was facilitated by fitting comparative models of many ribosomal proteins into the electron microscopy map of the ribosomal particle (*23*). This example also suggests that structural genomics of single proteins or their domains, combined with protein structure prediction, may contribute substantially to efficient structural characterization of large macromolecular assemblies.

The accuracy and reliability of models produced by de novo methods is much lower than that of comparative models based on alignments with more than 30% sequence identity, but the basic topology of a protein or domain can in some cases be predicted reasonably well (Fig. 1, D and E). For roughly 40% of proteins shorter than 150 amino acids that have been examined, one of the five most commonly recurring models generated by Rosetta has sufficient global similarity to the true structure to recognize it in a search of the protein structure database. Reasonable models can in some cases be produced for domains of even very large proteins by using multiple sequence alignments to identify domain boundaries (Fig. 1D).

The accuracy of de novo models is too low for problems requiring high-resolution structure information. Instead, the low-resolution models produced by these methods can reveal structural and functional relationships between proteins not apparent from their amino acid sequences and provide a framework for analyzing spatial relationships between evolutionarily conserved residues or between residues shown experimentally to be functionally important. These applications are illustrated by examples from the recent CASP4 blind protein structure prediction experiment (*24*, *25*). The predicted structure of a protein involved in cell lysis (*26*) was found to be structurally related to a protein with a similar function but no significant sequence similarity (Fig. 2B). The predicted structure of a domain of the mismatch repair protein MutS (27) (Fig. 1D) has structural similarity to proteins with related functions (*28*). Functionally important residues of the signaling protein Frizzled (*29*) were clustered in the predicted structure in a surface patch likely to be involved in a key protein-protein interaction (Fig. 2C). Thus, in favorable cases de novo predictions can provide some of the most important functional insights obtainable from experimentally determined structures.

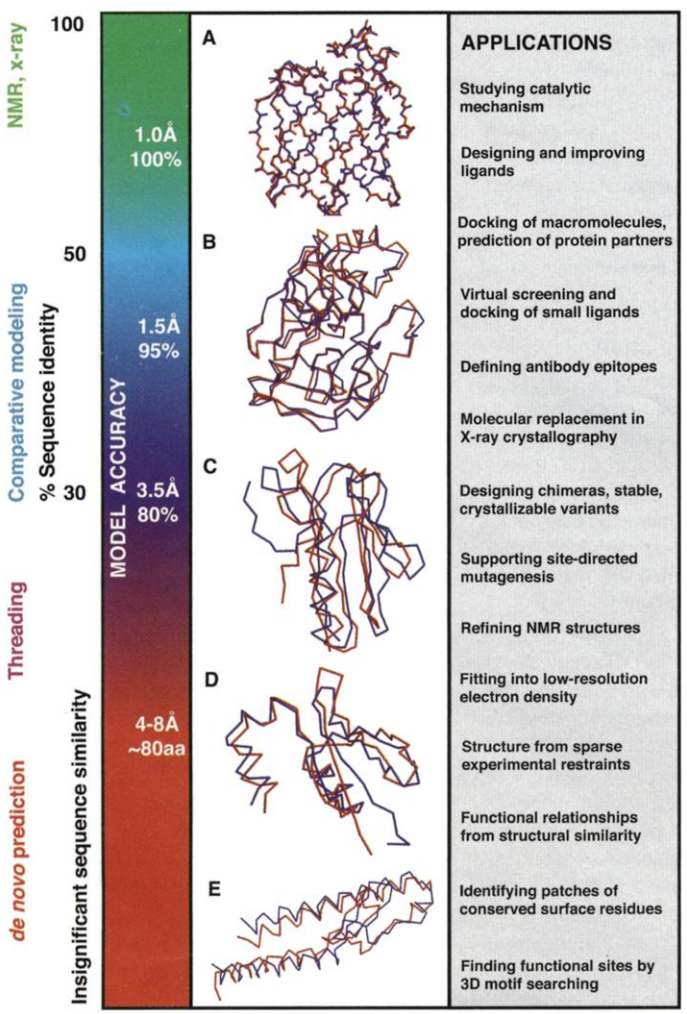

Fig. 1. Accuracy and application of protein structure models. Shown are the different ranges of applicability of comparative protein structure modeling, threading, and de novo structure prediction; the corresponding accuracy of protein structure models; and their sample applications. (**A** through **C**). Sample comparative models based on about 60% (A), 40% (B), and 30% (C) sequence identity to their template structure. (**D** and **E**) Examples of Rosetta de novo structure predictions for the CASP4 structure prediction experiment. Predicted structures are in red, and actual structures are in blue. The accuracy of the models decrease significantly in going from (A) to (E), but the overall structure is still roughly correct. (D) A domain from the 811-residue MUtS protein which was recognized as an autonomous unit from an alignment of homologous sequences; such parsing of large proteins into domains can make structure prediction more tractable.

Modeling on a Genomic Scale

Threading and comparative modeling methods have already been applied on a genomic scale (*18*, *30*, *31*). In total, domains in 58% of all 600,000 known protein sequences were modeled with ModPipe (*18*) and MODELLER (*9*) and deposited into a comprehensive database of comparative models, ModBase (*32–34*). The Web interface to the database allows flexible querying for fold assignments, sequence-structure alignments, models, and model assessments of interest. An integrated sequence/structure viewer, ModView, allows inspection and analysis of the query results. ModBase will be increasingly interlinked with other applications

and databases such that structures and other types of information can be easily used for functional annotation. Although the current number of modeled proteins may look impressive given the early stage of structural genomics, usually only one domain per protein is modeled (on the average, proteins have slightly more than two domains), and two-thirds of the models are based on less than 30% sequence identity to the closest template.

Automation and large-scale modeling with de novo methods have lagged behind those of comparative modeling methods, because of the relatively poor quality of the models produced and the relatively large amount of computer time required. However, inspired by the potential for functional insights, large-scale modeling calculations have been initiated with Rosetta. In the first such project, models for representatives of all PFAM families with less than 150 amino acids and currently not linked to proteins of known structure have been produced. Strong structural similarities of these models to structures of previously determined proteins can indicate previously unidentified relationships that may provide functional insights. It should soon be possible to extend these large-scale calculations to cover most of the domains not represented in ModBase.

The Role of Protein Structure Prediction in Structural Genomics

Structural genomics aims to structurally characterize most protein sequences by an efficient combination of experiment and prediction (*35*–*37*). This aim will be achieved by careful selection of target proteins and their structure determination by x-ray crystallography or NMR spectroscopy. There are a variety of target selection schemes (*38*), ranging from focusing on only novel folds to selecting all proteins in a model genome. A model-centric view requires that targets be selected such that most of the remaining sequences can be modeled with useful accuracy by comparative modeling. Even with structural genomics, the structure of most of the proteins will be modeled, not determined by experiment. As discussed above, the accuracy of comparative models and correspondingly the variety of their applications decrease sharply below the 30% sequence identity cutoff, mainly as a result of a rapid increase in alignment errors. Thus, we will need to determine protein structures so that most of the remaining sequences are related to at least one known structure at higher than 30% sequence identity (*36*, *37*). It was recently estimated that this cutoff requires a minimum of 16,000 targets to cover 90% of all protein domain families, including those of membrane proteins (*36*). These 16,000 structures will allow the modeling of a very much larger number of proteins. For example, New York Structural Genomics Research Consortium measured the impact of its structures by documenting the number and quality of the corresponding models for detectably related proteins in the nonredundant sequence database. For each new structure, on average, ~100 protein sequences without any prior structural characterization could be modeled at least at the fold level (*39*). This large leverage of structure determination by protein structure modeling illustrates and justifies the premise of structural genomics.

De novo structure prediction will contribute to structural genomics in several ways. Large-scale de novo prediction can guide target selection by focusing experimental structure determination on proteins likely to adopt novel folds. De novo methods should also be useful in complementing comparative modeling methods by building portions of proteins not present in template structures. In addition, de novo methods supplemented by restraints from cross linking or other experiments can provide models for proteins not readily amenable to x-ray crystallographic or NMR analysis. Finally, large-scale de novo modeling may allow coarse structure-based insights into protein function of a large number of proteins well in advance of experimentally determined structures.

Conclusions

Improvement in the accuracy of models produced by both de novo and comparative modeling approaches will require methods that finely sample protein conformational space using a free energy or scoring function that has sufficient accuracy to distinguish the native structure from the nonnative conformations. Despite many years of development of molecular simulation methods, attempts to refine models that are already relatively close to the native structure have met with relatively little success. This failure is likely to be due to inaccuracies in the potential functions used in the simulations, particularly in the treatment of electrostatics and solvation effects. Improvements in sampling strategies may also be necessary, given the relatively long time scale of protein folding (milliseconds to seconds). Combination of physical chemistry with the vast amount of information in known protein structures may provide a route to development of improved potential

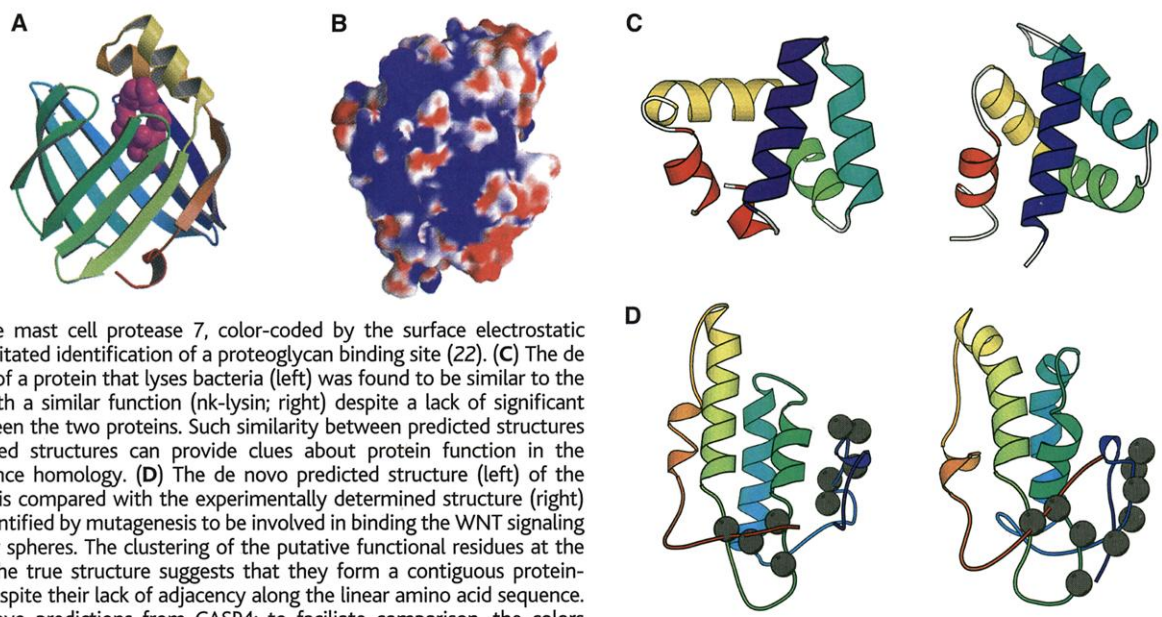

Fig. 2. Sample applications of protein structure models. (**A**) A comparative model of a complex between docosahexaenoic fatty acid (violet) and brain lipid-binding protein. Such models for a number of fatty acid ligands were used to rank their binding affinities (*21*). (**B**) A comparative model of mouse mast cell protease 7, color-coded by the surface electrostatic potential. This model facilitated identification of a proteoglycan binding site (*22*). (**C**) The de novo predicted structure of a protein that lyses bacteria (left) was found to be similar to the structure of a protein with a similar function (nk-lysin; right) despite a lack of significant sequence similarity between the two proteins. Such similarity between predicted structures and previously determined structures can provide clues about protein function in the absence of strong sequence homology. (**D**) The de novo predicted structure (left) of the signaling protein Frizzled is compared with the experimentally determined structure (right) (*26*), with the residues identified by mutagenesis to be involved in binding the WNT signaling protein, indicated by gray spheres. The clustering of the putative functional residues at the right in the model and the true structure suggests that they form a contiguous protein-protein interaction site despite their lack of adjacency along the linear amino acid sequence. (C and D) Rosetta de novo predictions from CASP4; to faciliate comparison, the colors indicate position along the the chain from the NH_2 terminus (blue) to the COOH terminus (red).

functions. The refinement of de novo and comparative models provides a good test and application of the molecular dynamics methods widely used to simulate biological macromolecules (*40*).

Automated methods for deducing function from structure will be critical to obtaining functional insights from both predicted and experimentally determined structures. Considerable insight can be gained from structural comparison of a given structure with all other known protein structures using methods such as Dali (*41*), which can frequently detect structural relationships with functional significance that are not evident from sequence comparisons. Also promising are methods that match a structure against a library of structural motifs associated with different functions (*42–44*). For higher resolution models produced by comparative modeling methods, functional sites on proteins can potentially be identified and characterized by explicit ligand docking calculations. Finally, large-scale protein-protein docking calculations in years to come may contribute to the identification and characterization of protein interaction networks.

References

1. M. A. Marti-Renom *et al.*, *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291 (2000).
2. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
3. A. E. Torda, *Curr. Opin. Struct. Biol.* **7**, 200 (1997).
4. W. J. Browne, A. C. North, D. C. Phillips, *J. Mol. Biol.* **42**, 65 (1969).
5. J. Greer, *J. Mol. Biol.* **153**, 1027 (1981).
6. T. L. Blundell, B. L. Sibanda, M. J. Sternberg, J. M. Thornton, *Nature* **326**, 347 (1987).
7. M. Levitt, *J. Mol. Biol.* **226**, 507 (1992).
8. T. F. Havel, M. E. Snow, *J. Mol. Biol.* **217**, 1 (1991).
9. A. Sali, T. L. Blundell, *J. Mol. Biol.* **234**, 779 (1993).
10. A. Kolinski, M. R. Betancourt, D. Kihara, P. Rotkiewicz, J. Skolnick, *Proteins* **44**, 133 (2001).
11. A. Fiser, R. K. G. Do, A. Sali, *Protein Sci.* **9**, 1753 (2000).
12. M. J. Bower, F. E. Cohen, R. L. Dunbrack Jr., *J. Mol. Biol.* **267**, 1268 (1997).
13. R. Samudrala, Y. Xia, E. Huang, M. Levitt, *Proteins* **3** (suppl.), 197 (1999).
14. A. R. Ortiz, A. Kolinski, P. Rotkiewicz, B. Ilkowski, J. Skolnick, *Proteins* **37**, 177 (1999).
15. J. Pillardy *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 2329 (2001).
16. D. T. Jones, *Proteins* **1** (suppl.), 185 (1997).
17. K. Simons, C. Strauss, D.Baker, *J. Mol. Biol.* **306**, 1191 (2000).
18. R. Sanchez, A. Sali, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 13597 (1998).
19. P. Koehl, M. Levitt, *Nature Struct. Biol.* **6**, 108 (1999).
20. M. A. Marti-Renom, M. S. Madhusudhan, A. Fiser, B. Rost, *Structure*, in press.
21. L. Z. Xu, R. Sanchez, A. Sali, N. Heintz, *J. Biol. Chem.* **271**, 24711 (1996).
22. R. Matsumoto, A. Sali, N. Ghildyal, M. Karplus, R. L. Stevens, *J. Biol. Chem.* **270**, 19524 (1995).
23. C. M. T. Spahn *et al.*, *Cell*, in press.
24. See http://predictioncenter.llnl.gov/casp4.
25. R. Bonneau *et al.*, *Proteins: Structure, Function and Genetics*, in press.
26. A. Gonzalez *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11221 (2000).
27. G. Obmolova *et al.*, *Nature* **407**, 703 (2000).
28. R. Bonneau, J. Tsai, I. Ruczinski, D. Baker, *J. Struct. Biol.* **134**, 186 (2001).
29. A. E. Dann *et al.*, *Nature* **412**, 86 (2001).
30. D. Fischer, D. Eisenberg, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 11929 (1997).
31. N. Guex, A. Diemand, M. C. Peitsch, *Trends. Biochem. Sci.* **24**, 364 (1999).
32. See http://guitar.rockefeller.edu/modbase/.
33. R. Sanchez *et al.*, *Nature Struct. Biol.* **7** (suppl.), 986 (2000).
34. R. Sánchez *et al.*, *Nucleic Acids Res.* **28**, 250 (2000).
35. S. K. Burley *et al.*, *Nature Genet.* **23**, 151 (1999).
36. D. Vitkup, E. Melamud, J. Moult, C. Sander, *Nature Struct. Biol.* **8**, 559 (2001).
37. A. Sali, *Nature Struct. Biol.* **5**, 1029 (1998).
38. S. E. Brenner, *Nature Struct. Biol.* **7** (suppl.), 967 (2000).
39. See http://nysgrc.org/.
40. A. L. Brooks, M. Karplus, B. M. Pettit (Wiley, New York, 1988).
41. L. Holm, C. Sander, *Trends Biochem. Sci.* **20**, 478 (1995).
42. A. Zhang *et al.*, *Protein Sci.* **8**, 1104 (1999).
43. A. C. Wallace, N. Borkakoti, J. M. Thornton, *Protein Sci.* **6**, 2308 (1997).
44. P. Aloy, E. Querol, F. X. Aviles, M. J. Sternberg, *J. Mol. Biol.* **311**, 395 (2001).
45. R. Bonneau, D. Baker, *Annu. Rev. Biophys. Biomol. Struct.* **30**, 173 (2001).

46. We are grateful to J. Frank and R. Beckmann for the picture of the ribosomal particle; M. A. Marti-Renom, R. Bonneau, and N. Eswar for help in preparing the figures; and members of our groups for many discussions about protein structure prediction. Supported by NIH/GM 54762 (A.S.), the Mathers Foundation (A.S.), a Merck Genome Research Award (A.S.), and the Howard Hughes Medical Institute (D.B.). A.S. is an Irma T. Hirschl Trust Career Scientist.

REVIEW

Making Sense of Eukaryotic DNA Replication Origins

David M. Gilbert

DNA replication is the process by which cells make one complete copy of their genetic information before cell division. In bacteria, readily identifiable DNA sequences constitute the start sites or origins of DNA replication. In eukaryotes, replication origins have been difficult to identify. In some systems, any DNA sequence can promote replication, but other systems require specific DNA sequences. Despite these disparities, the proteins that regulate replication are highly conserved from yeast to humans. The resolution may lie in a current model for once-per-cell-cycle regulation of eukaryotic replication that does not require defined origin sequences. This model implies that the specification of precise origins is a response to selective pressures that transcend those of once-per-cell-cycle replication, such as the coordination of replication with other chromosomal functions. Viewed in this context, the locations of origins may be an integral part of the functional organization of eukaryotic chromosomes.

Transmission of genetic information from one cell generation to the next requires the accurate and complete duplication of each DNA strand exactly once before each cell division. Typically, this process begins with the binding of an "initiator" protein to a specific DNA sequence or "replicator." In response to the appropriate cellular signals, the initiator directs a local unwinding of the DNA double helix and recruits additional factors to initiate the process of DNA replication. This paradigm describes most of the currently tractable replication systems and, although derived from prokaryotic and viral systems, there is no compelling reason to doubt that it will apply to all eukaryotic organisms. In fact, the proteins that regulate replication are highly conserved from yeast to humans, including the origin recognition complex (ORC), which binds directly to replication origin sequences in budding yeast (*1*, *2*). However, in several eukaryotic replication systems, it appears that any DNA sequence can function as a replicator. Those outside the field are often perplexed as to how investigators of different eukaryotic systems can work with assumptions that range from very specific to completely random origin sequence recognition, yet all agree on the basic mechanism regulating DNA replication. This review summarizes our current understanding of eukaryotic replication origins and then presents some simple guidelines to help demystify these seemingly disparate observations, providing a framework for understanding eukaryotic origins that includes all existing data.

Department of Biochemistry and Molecular Biology, SUNY Upstate Medical University, 750 East Adams Street, Syracuse, NY 13210, USA. E-mail: gilbertd@mail.upstate.edu