

## Current Genomic Research: The Proteins Have It

LYNNE J. REGAN

*Department of Molecular Biophysics and  
Biochemistry and Department of Chemistry  
Yale University  
New Haven, Connecticut*

Proteins play crucial and essential roles in all aspects of cellular function in all organisms. They are composed of linear strings of discrete subunits called amino acids. There are 20 naturally occurring amino acids, which can be grouped into different classes based on the chemical nature of their side chains: acidic, basic, polar, aliphatic, aromatic. All the information necessary and sufficient to specify the final three-dimensional structure that is adopted by a protein is encoded by its amino acid sequence. The amino acid sequence of each protein is itself directly encoded in the chromosomal DNA, which is passed from generation to generation.

A single nucleotide base change in the chromosomal DNA that results in a single amino acid change in a single protein can have devastating, if not fatal, consequences for an individual organism and his or her offspring. Sickle cell anemia, in which the protein hemoglobin forms long aggregates that dramatically reduce its oxygen-carrying capacity and give rise to the characteristic sickling distortion of red blood cells, is a consequence of a mutation of a single amino acid from the negatively charged aspartic acid to the aliphatic valine. There are numerous other examples of diseases that can be linked directly to protein mutation.

Cloned human proteins produced in bacteria are of considerable therapeutic importance. Human growth hormone is administered to children for the treatment of dwarfism (and has been abused by athletes to promote muscle formation!); granulocyte colony-stimulating factor is administered to patients undergoing chemotherapy to promote white blood cell formation; and, of course, thousands of diabetics self-administer insulin daily. These proteins are targets of redesign to optimize their biological, physical, and pharmacokinetic properties for therapeutic purposes. Many human proteins are also potential drug targets.

To cite just a few examples, inhibition of liver phosphorylase would provide a route to the maintenance of adult onset diabetes, inhibition of tumor-promoting proteins provides a mechanism for novel cancer therapy, and inhibition of protein aggregation is a potential strategy by which to combat Alzheimer's disease.

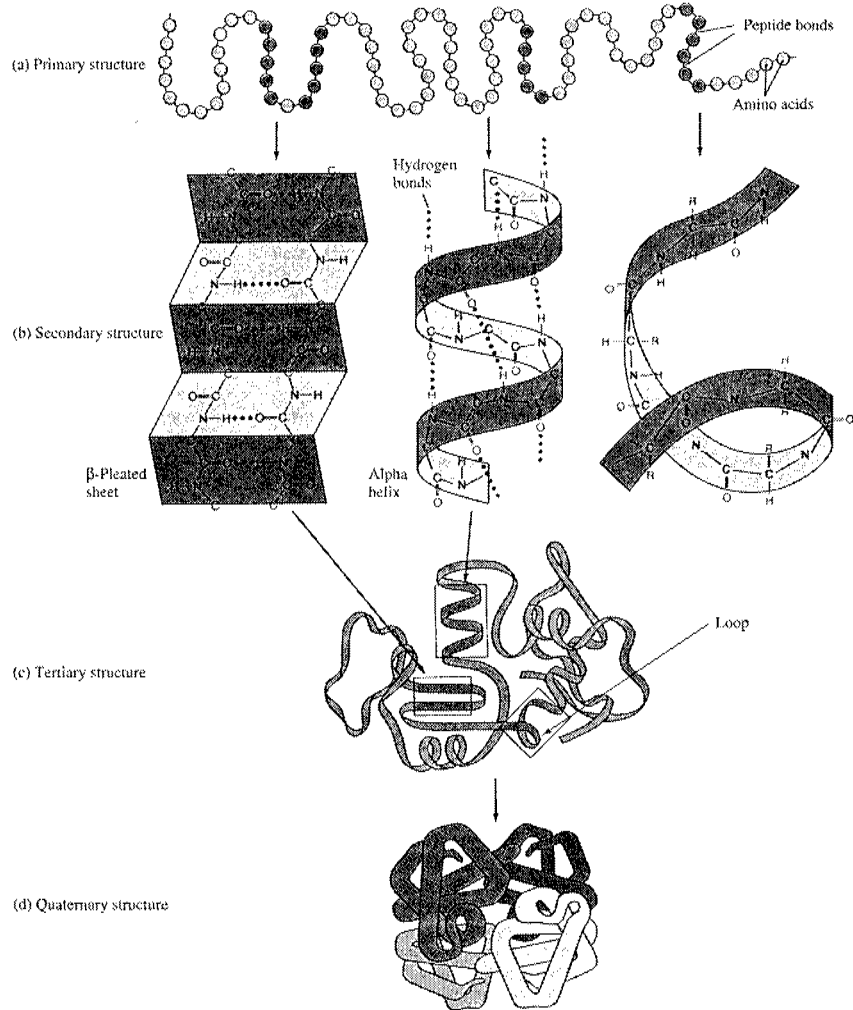
Proteins that are specific to an invading organism are targets for specific inhibition to counteract infection; examples range from the fairly recently developed anti-AIDS drugs, which target HIV's unique protease, to penicillin, which targets the enzymes involved in the bacteria-specific process of cell-wall synthesis.

Protein structure can be classified in a hierarchical fashion (Figure 1). The first level is the amino acid sequence or "primary structure," and the next is the "secondary structure," which is the regular repetition of backbone dihedral angles in a linear stretch of amino acids that gives rise to a common structural unit. The two dominant types of secondary structure in proteins are  $\alpha$ -helices and  $\beta$ -sheets. Elements of secondary structure are connected by loops or turns, which allow them to fold back on themselves and to associate to form a globular "tertiary structure." In some proteins the folded tertiary structure of a monomer is the active form, whereas in others monomers associate with other similar or dissimilar subunits to give specific higher-order complexes or "quaternary structure."

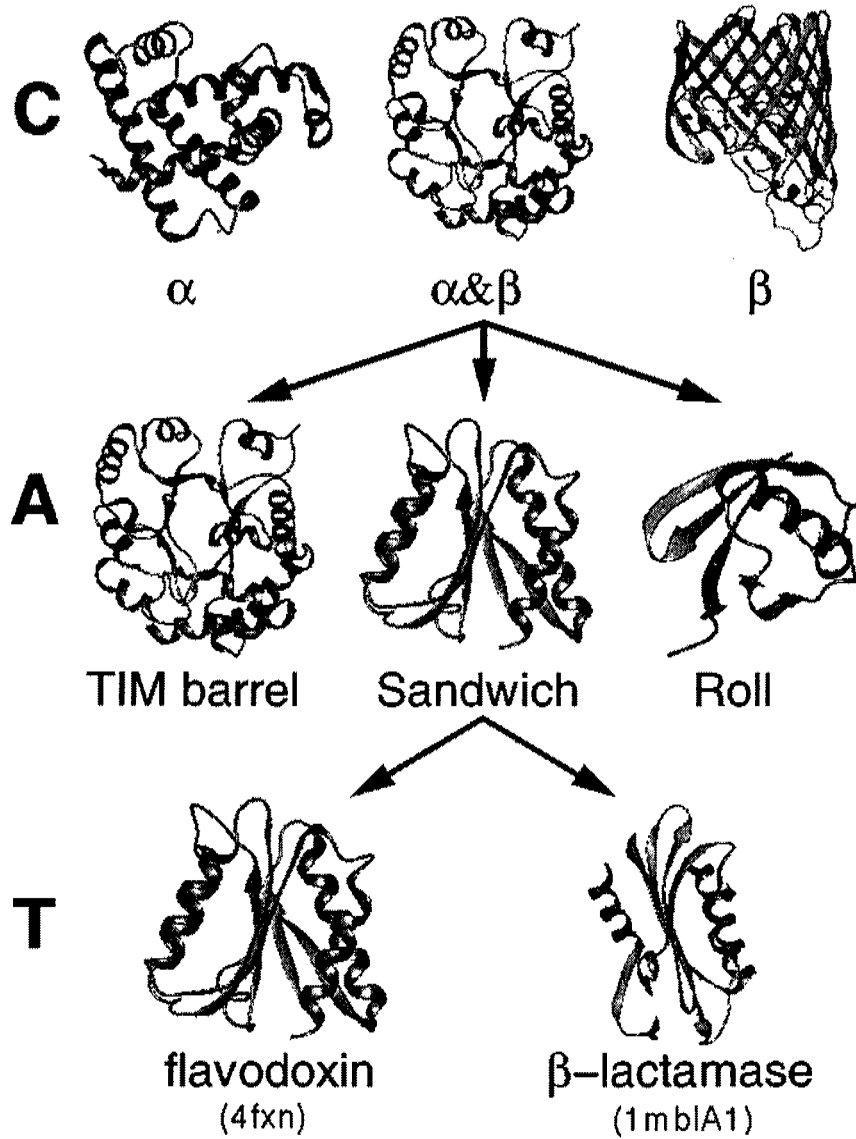
The different tertiary structures or "folds" that combinations of  $\alpha$ -helices and  $\beta$ -sheets give rise to are not a continuum but, rather, can be grouped into distinct structural classes. The tertiary folds are specified by the geometrical constraints that are imposed by optimally packing together the elements of secondary structure. A convenient classification scheme is the CATH scheme of Orengo and Thornton (Orengo et al., 1997), in which a protein structure is classified according to its class, architecture, topology, and homology. A sample CATH classification of a protein is shown in Figure 2. If we perform such a classification on all proteins for which structures are known at high resolution by X-ray crystallography or nuclear magnetic resonance, the distribution of structural classes can be illustrated on a CATHerine wheel (Figure 3). We see that approximately one-quarter of all folds are  $\alpha$ -helical, one-quarter are all  $\beta$ -sheet, and one-half are a mixture of  $\alpha$ -helix and  $\beta$ -sheet.

But how can this information be combined with the results of the genome scale sequencing? The first thing to do when a novel gene is sequenced is to perform a sequence-alignment search against all known sequences from all organisms. If a match is found with a protein of known structure—one of the fold classes discussed above—a model for the structure of the novel protein can be built on the basis of its sequence homology with proteins of known fold. The greater the number of homologous sequences available, the better the alignment and the better the homology model that can be made.

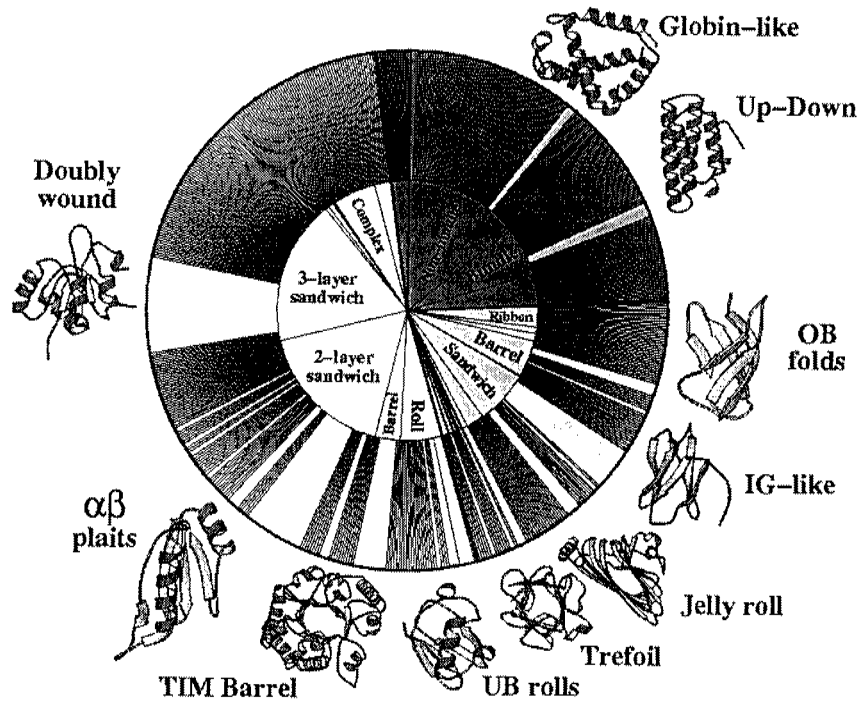
How far can we get with this approach? Is a homology model, rather than an actually determined three-dimensional structure, "good enough"? The answer to this question largely depends on the specific goal of a project. For some purposes, knowing the fold and having a homology model of the protein of



**FIGURE 1** The hierarchical nature of protein structure. SOURCE: Reprinted with permission of Wadsworth, an imprint of the Wadsworth Group, a division of Thomson Learning (Boyer, 1999).



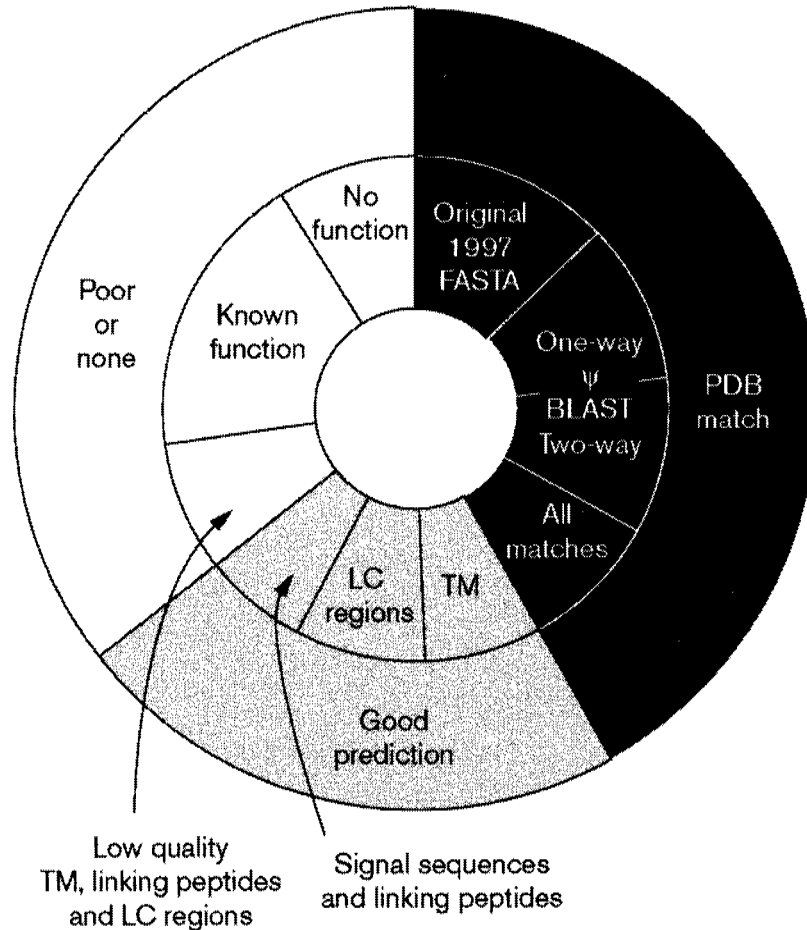
**FIGURE 2** The class, architecture, topology, homology (CATH) classification of a protein. SOURCE: Department of Biochemistry and Molecular Biology, University College, London. This figure can be viewed in color at <[http://www.biochem.ucl.ac.uk/bsm/cath\\_new/cath\\_info.html](http://www.biochem.ucl.ac.uk/bsm/cath_new/cath_info.html)>.



**FIGURE 3** CATherine wheel illustration of the distribution of fold classes in the database of known protein structures. SOURCE: Department of Biochemistry and Molecular Biology, University College, London. This figure can be viewed in color at <[http://www.biochem.ucl.ac.uk/bsm/cath/lex/CATherine\\_orig.html](http://www.biochem.ucl.ac.uk/bsm/cath/lex/CATherine_orig.html)>.

interest will be sufficient. Will a homology model provide a good enough structural model for rational drug design? The answer to this question is subject to debate. One relevant piece of information is that the goodness of the model depends upon the percent sequence identity of the unknown protein and the target fold. It was first pointed out by Chothia and Lesk (1986) that there is a direct and predictable relationship between the percent sequence identity of two proteins and the backbone root mean square deviation of the two structures.

Having discussed how we can make use of sequence information to predict structure, based on our understanding of sequence-structure relationships and homology modeling to proteins of known fold, we must now consider what fraction of all proteins in an organism either have known structure or can be modeled reasonably satisfactorily by homology. Figure 4 shows the complete distribution of folds and sequences in the genome of *Mycoplasma genitalium* (MG), which is, to date, the smallest known self-replicating organism. Less than



**FIGURE 4** Illustration of the distribution of protein sequences in the genome of *Mycoplasma genitalium*. SOURCE: Reprinted with permission from Elsevier Science (Teichmann et al., 1999).

half of the protein structures have been determined experimentally, about a third have structures that can be predicted “well” by homology modeling, and about half are completely unknown. This distribution, which is similar in other organisms, provides an illustration of the scale of missing “fold space.” Filling in these missing structures via a number of different strategies is a goal of current genomics research. This information is important not only because it provides fundamental insights into how different protein folds can be used for the same or different functions but also because it is of practical importance in drug design.

There are two complementary approaches to the characterization of unknown proteins in structural genomics projects. First is the “low hanging fruit” approach, in which proteins are cloned, expressed, purified and structures solved, but, initially, at each step only the proteins that behave well and are easy to work with are carried forward to the next step. Second is the rational preselection of representatives of predicted new fold classes, or proteins that may have unusual physical properties, and a concentrated focus on characterizing these proteins, regardless of how easy or difficult it is to work with each protein. Each approach has a number of advantages and disadvantages, and at this stage in structural genomics research, there is clearly a need for both.

As an example, let’s consider our work on MG (Balasubramanian et al., 2000). MG has the smallest genome known for any self-replicating organism, encoding approximately 483 proteins (the exact number of predicted proteins varies slightly, depending on the identification method used). Of these, 202 are structurally uncharacterized, 70 are both functionally and structurally uncharacterized, and 25 are completely structurally and functionally uncharacterized over their entire length. Of these 25, 15 are unique to MG, and 10 have homologues in related organisms. We have expressed, purified, and characterized 12 of these 25 proteins. Seven behave like “normal” proteins, display substantial secondary structure, and likely represent novel folds. These are candidates for high-resolution structure determination. One protein is unstructured and may require a partner molecule (either another protein subunit or a nucleic acid or other cellular component) in order to fold, and two display unusual thermodynamic properties: they are highly helical and extremely resistant to thermal denaturation. These latter two proteins are highly conserved from MG to man.

What stages in this work provide engineering challenges and opportunities? Currently, the most common means of production of the large amounts of protein required for biophysical and structural characterization is expression in live bacteria. A better high-throughput method would speed the process. Purification of large quantities of protein to a reasonable level of purity can be accomplished quite readily by attaching a universal “tag,” which allows all proteins to be purified by that same method, regardless of their individual chemical nature. Development of an additional universally applicable purification step, which would easily allow proteins to be purified to the high levels required for crystallization, is important. At the moment, a more significant problem than these is that at least half of all proteins, when expressed in bacteria, do not partition into the soluble phase but, rather, aggregate and partition as unfolded protein into the insoluble fraction. This means that they must be refolded before characterization can be begun, or bacterial growth conditions must be manipulated to “coax” as much protein as possible into the soluble phase. In our work with the uncharacterized proteins of MG, obtaining reasonable quantities of pure folded protein certainly was the rate-limiting step.

In summary, this brief overview is intended to provide a sampling of the ways in which studying protein structure and function in the “genomic era” furnishes new challenges and opportunities and is likely to give rise to a host of unexpected and exciting discoveries.

### ACKNOWLEDGMENTS

I thank all members of the North East Structural Genomics Consortium, but especially my collaborator at Yale, Mark Gerstein, for his insights and advice on illustrations for my presentation. I acknowledge Christine Orenga and Janet Thornton for the CATH protein classification system and for the illustrations in this article that are taken from their work.

### REFERENCES

- Balasubramanian, S., T. Schneider, M. Gerstein, and L. Regan. 2000. Proteomics of *Mycoplasma genitalium*: Identification and characterization of unannotated and atypical proteins in a small model genome. *Nucleic Acids Research* 28:3075–3082 and references therein.
- Boyer, R. 1999. *Concepts in Biochemistry*. Pacific Grove, Calif.: Brooks/Cole Publishing Co.
- Chothia C., and A. M. Leske. 1986. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal* 5(4):823–826.
- Orengo, C. A., A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. 1997. CATH: A hierarchic classification of protein domain structures. *Structure* 5(8):1093–1108.
- Teichmann, S. A., C. Chothia, and M. Gerstein. 1999. Advances in structural genomics. *Current Opinion in Structural Biology* 9:390–399.