2

Biomolecular Structure and Modeling: Problem and Application Perspective

All things come out of the one, and the one out of all things. Change, that is the only thing in the world which is unchanging.

Heraclitus of Ephesus (550–475 BC).

2.1 Computational Challenges in Structure and Function

2.1.1 *Analysis of the Amassing Biological Databases*

The experimental progress described in the previous chapter has been accompanied by an increasing desire to relate the complex three-dimensional (3D) shapes of biomolecules to their biological functions and interactions with other molecular systems. Structural biology, computational biology, genomics, proteomics, bioinformatics, chemoinformatics, and others are natural partner disciplines in such endeavors.

Structural biology focuses on obtaining a detailed structural resolution of macromolecules and relating structure to biological function.

Computational biology was first associated with the discipline of finding similarities among nucleotide strings in known genetic sequences, and relating these relationships to evolutionary commonalities; the term has grown, however, to encompass virtually all computational enterprises to molecular biology problems [126].

Comparative genomics — the search and comparison of sequences among species — is a natural outgrowth of the sequencing projects [116]. So are *structural and functional genomics*, the characterization of the 3D structure and biological function of these gene products [142, 20, 14, 111, 37].

In the fall of 2000, the U.S. National Institute of General Medical Sciences (NIGMS) launched a structural genomics initiative by funding seven research groups aiming to solve collectively the 3D structures of 10,000 proteins, each representing a protein family, over the next decade. This important step toward the assembly of a full protein library requires improvements in both structural biology's technology and methodology. (See the November 2000 supplement issue of *Nature Structural Biology*, volume 7, devoted to structural genomics and the progress report [37]). Ultimately, the functional identification of unknown proteins is likely to dramatically increase our understanding of human disease processes and the development of structure-based drug therapies.

Proteomics is another current buzzword defining the related discipline of protein structure and function (see [67] for an introduction), and even *cellomics* has been introduced.[1] Cellomics reflects the expanded interest of gene sequencers in integrated cellular structure and function. The *Human Proteomics Project* — a collaborative venture to churn out atomic structures using high-throughput and robotics-aided methods based on NMR spectroscopy and X-ray crystallography, rather than sequences — may well be on its way.

New instruments that have revolutionized genomics known as DNA microarrays, biochips, or gene expression chips (introduced in Chapter 1 and Box 1.5) allow researchers to determine which genes in the cell are active and to identify gene networks. A good introduction to these areas is volume 10 of the year 2000, pages 341–384, of *Current Opinion in Structural Biology* entitled "Sequences and Topology. Genome and Proteome Informatics", edited by P. Bork and D. Eisenberg.

The range of genomic sciences also continues [156] to the *metabolome*, the endeavor to define the complete set of metabolites (low-molecular cellular intermediates) in cells, tissues, and organs. Experimental techniques for performing these integrated studies are continuously being developed. For example, yeast geneticists have developed a clever technique for determining whether two proteins interact and, thereby by inference, participate in related cellular functions [210]. Such approaches to proteomics provide a powerful way to discover functions of newly identified proteins. DNA chip technology is also thought to hold the future of individualized health care now coined personalized medicine or *pharmacogenomics*; see Chapter 14 and Box 1.5.

[1] A glossary of biology disciplines coined with "ome" or "omic" terms can be found at http://www.genomicglossaries.com/content/omes.asp.

It has been said that current developments in these fields are *revolutionary rather than evolutionary*. This view reflects the clever exploitation of biomolecular databases with computing technology and the many disciplines contributing to the biomolecular sciences (biology, chemistry, physics, mathematics, statistics, and computer science). *Bioinformatics* is an all-embracing term for many of these exciting enterprises [99, 148] (structural bioinformatics is an important branch); *chemoinformatics* has also followed (see Chapter 14) [86]. Some genome-technology company names are indicative of the flurry of activity and grand expectations from our genomic era. Consider titles like Genetics Computer Group, Genetix Ltd., Genset, Protana, Protein Pathways, Inc., Pyrosequencing AB, Sigma-Genosys, or Transgenomic Incorporated.

This excitement in the field's developments and possibilities is echoed by the chief executive of the software giant Oracle Corp., Larry Ellison, who surrounds himself by molecular biologists — the scientists, board members, and fellows of his Ellison Medical Foundation; explaining to a *Wall Street Journal* reporter his preference of molecular biology over racing sailboats, Ellison said: "*The race is more interesting, the people in the race are more interesting and the prize is bigger.*" (*Wall Street Journal*, January 9, 2003).

Although the number of sequence databases has grown very rapidly and exceeds the amount of structural information,[2] the 1990s saw an exponential rise of structural databases as well. From only 50 solved 3D structures in the Protein Data Bank (PDB) in 1975, to 500 in 1988, another order of magnitude was reached in 1996 (around 5000 entries). In fact, the rate of growth of structural information is approaching the rate of increase of amino acid sequence information (see Figure 2.1 and Table 2.1). It is no longer a rare event to see a new crystal structure on the cover of *Nature* or *Science*. Now, on a weekly basis, groups compete to have their newly-solved structure on the cover of prominent journals, including the newer, colorful publications like *Nature Structural Biology* and *Structure*. The fraction of NMR-deduced structures deposited in the PDB is also rising steadily, reflecting roughly 15% by the end of 2001 (for updated information, see www.rcsb.org/pdb/holdings.html).

This trend, coupled with tremendous advances in genome sequencing projects [120], argues strongly for increased usage of computational tools for analysis of sequence/structure/function relationships and for structure prediction. Thus, besides genomics-based analyses and comparisons, important are accurate, reliable, and rapid theoretical tools for describing structural and functional aspects of gene products. (See [111], for example, for computational challenges in genomics).

[2] In 1991, it was pointed out [15] that the amount of 3D protein information is lagging by orders of magnitude behind the accessible sequence data [15].

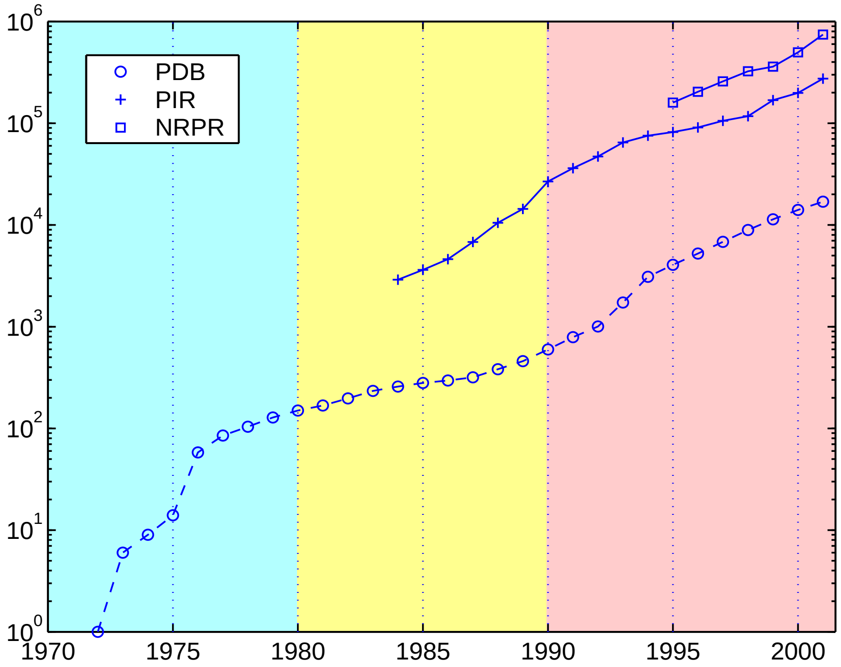


Figure 2.1. The growth of protein sequence databases (named PIR and NRPR) versus structural database of macromolecules (PDB). See Table 2.1 and www.dna.affrc.go.jp/htdocs/growth/P-history.html. The International Protein Information Resource (PIR) of protein sequences (pir.georgetown.edu) is a comprehensive, annotated public-domain database reflecting a collaboration among the United States (Georgetown University Medical Center in Washington, D.C.), Germany (Munich Information Center for Protein Sequences in Martinsried), and Japan (International Protein Sequence Database in Tsukuba). PIR is extensively cross referenced and linked to many molecular databases of genes, genomes, protein structures, structural classification, literature, and more. The NRPR database represents merged, non redundant protein database entries from several databases: PIR, SWISS-PROT, Genpept, and PDB.

2.1.2 Computing Structure From Sequence

One of the most successful approaches to date on structure prediction comes from *homology modeling* (also called comparative modeling) [3, 8].

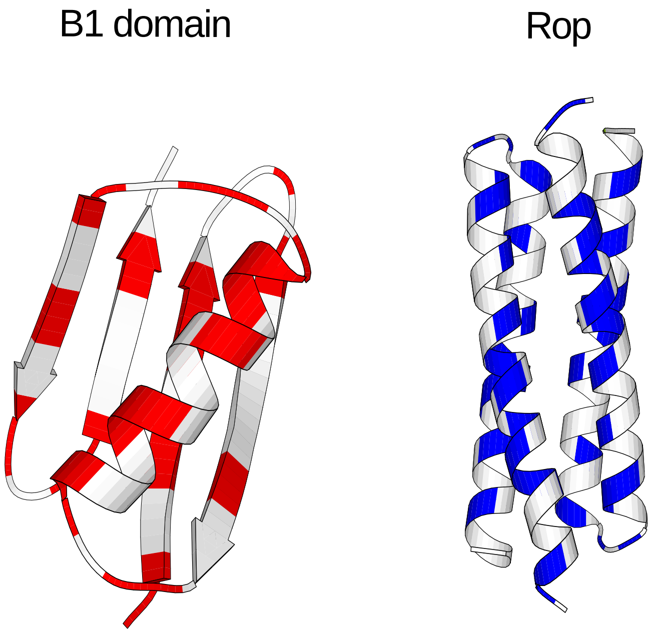
In general, a large degree of sequence similarity often suggests similarity in 3D structure. It has been reported, for example, that a sequence identity of greater than 40% usually implies more than 90% 3D-structure overlap (defined as percentage of C^α atoms of the proteins that are within 3.5 Å of each other in a rigid-body alignment; see definitions in Chapter 3) [184]. Thus, sequence similarity of at least 50% suggests that the associated structures are similar overall. Conversely, small sequence similarity generally implies structural diversity.

Table 2.1. Growth of protein sequence databases.

Year	PDB[a]	PIR[b]	NRPR[c]
1972	1		
1973	6		
1974	9		
1975	14		
1976	58		
1977	85		
1978	104		
1979	128		
1980	150		
1981	168		
1982	197		
1983	234		
1984	258	2898	
1985	280	3615	
1986	296	4612	
1987	318	6796	
1988	382	10527	
1989	459	14372	
1990	597	26798	
1991	790	36150	
1992	1006	47234	
1993	1727	64760	
1994	3091	75511	
1995	4056	82066	159808
1996	5240	91006	204123
1997	6833	105741	258272
1998	8942	117482	324237
1999	11364	168808	360674
2000	14063	198801	497787
2001	16973	274514	744991

[a] From the Protein Data Bank (PDB), www.rcsb.org/pdb/holdings.html.
[b] From the Protein International Resource (PIR), www-nbrf.georgetown.edu/cgi-bin/dbinfo.
[c] From the 'Non Redundant Proteins database merged Regular release', www.dna.affrc.go.jp/htdocs/growth/P-history.html.

There are many exceptions, however, as demonstrated humorously in a contest presented to the protein folding community (see Box 2.1). The *myoglobin* and *hemoglobin* pair is a classic example where large structural, as well as evolutionary, similarity occurs despite little sequence similarity (20%). Other exceptional



```
B1     MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE
Janus  MTKKAILALNTAKFLRTQAAVLAAKLEKLGAQEANDNAVDLEDTADDLYKTLLVLA
Rop    GTKQEKTALNMARFIRSQTLTLLEKLNELDADEQADICESLHDHADELYRSCLARF
```

Figure 2.2. Ribbon representations of the B1 domain of IgG-binding protein G and the Rop monomer (first 56 residues), which Janus resembles [46], with corresponding sequences. Half of the protein G β domain (B1) residues were changed to produce Janus in response to the Paracelsus challenge (see Box 2.1 and [177]). The origin of the residues is indicated by the following color schemes: residues from B1: red; residues from Rop: blue; residues in both: green; residues in neither: **black**. While experimental coordinates of protein G and Rop are known, the structure of Janus was deduced by modeling. The single-letter amino acid acronyms are detailed in Table **??**, Chapter 3.

examples and various sequence/structure relationships are discussed separately in Chapter 3, as well as Homework 6 and [79] for example.

More general than prediction by sequence similarity is structure prediction *de novo* [8], a *Grand Challenge* of the field, as described next.

2.2 Protein Folding – An Enigma

2.2.1 *'Old' and 'New' Views*

There has been much progress on the protein folding challenge since Cyrus Levinthal first posed the well-known "paradox" named after him; see [97] for a historical perspective. Levinthal suggested that well defined folding pathways might exist [123] since real proteins cannot fold by exhaustively traversing through their entire possible conformational space [124]. (See [213, 51, 214] for a recent related discussion of whether the number of protein conformers depends exponentially or nonexponentially on chain length). Levinthal's paradox led to the development of two views of folding — the 'old' and the 'new' — which have been merging of late [49, 50, 122].

The old view accents the existence of a specific folding pathway characterized by well-defined intermediates. The new version emphasizes the rugged, heterogeneous multidimensional energy landscape governing protein folding, with many competing folding pathways [228]. Yet, the boundary between the two views is pliant and the intersection substantial [97]. This integration has resulted from a variety of information sources: theories on funnel-shaped energy landscape (e.g., [158, 50, 25]); folding and unfolding simulations of simplified models (e.g., [77, 87, 112, 141, 190, 71]), at high temperatures or low pH concentrations (e.g., [122, 240]); NMR spectroscopic experiments that monitor protein folding intermediates (e.g., [60, 157]); predictions of secondary and/or tertiary structure on the basis of evolutionary information [183]; and statistical mechanical theories.

Such studies suggest that while wide variations in folding pathways may occur, 'fast folders' possess a unifying pattern for the evolution of native-structure contacts. In particular, the pathway and other kinetic aspects governing the folding of a particular protein depend on the free energy landscape, which is temperature-dependent. Namely, the folding ensemble is sensitive to shallow energy traps at lower temperatures. Thus, changes in temperature can substantially change the folding kinetics.

2.2.2 *Folding Challenges*

The great progress in the field can also be seen by evaluations of biannual prediction exercises held in Asilomar, California (termed CASP for Critical Assessment of Techniques for Protein Structure Prediction) [58, 114, 209] (see Prediction Center.llnl.gov, and special issues of the journal *Proteins*: vol. 23, 1995; Suppl. 1, 1997; Suppl. 3, 1999, and Suppl. 5, 2001).

The CASP organizers assign specific proteins for theoretical prediction that protein crystallographers and NMR spectroscopists expect to complete by the next CASP meeting. Prediction assessors then consider several categories of structural prediction tools, for example: comparative (homology) modeling, fold recognition (i.e., based on a library of known protein folds), and *ab initio* prediction (i.e., using first principles). The quality of the prediction has been characterized in terms of C^α root-mean-square (RMS) deviations, the best of which is in the range of 2–6 Å; the lowest values have been obtained from the best comparative modeling approaches.

To the fourth CASP meeting in December 2000, two additional competitive experiments were added, dealing with protein structure prediction and rational drug design: CAFASP2 (assessing automatic methods for predicting protein structure), and CATFEE (evaluating methods for protein ligand docking); see predictioncenter.llnl.gov/casp4/Casp4.html). These additions reflect the rapid developments of many genome efforts and the rise of computational biology as an important discipline of biology, medicine, and biotechnology.

Results of the CASP4 meeting indicated considerable progress in fully-automated approaches for structure prediction and hence promise that computer modeling might become an alternative to experimental structure determination. See [146] for an overview with articles collected on CASP4 in that special issue of *Proteins* (Suppl. 5, 2001), as well as a related discussion of field progress and prospects [13] and of the limitations and challenges in the comparative modeling section of CASP4 [134, 147, 182].

Results of the CASP5 meeting [209] revealed the evolving importance of the CASP initative for motivating progress in protein prediction and helping related global initatives, like the “Ten Most Wanted” proteins of unknown structure that the community aims to solve because of suggested biological importance (see tmw.llnl.gov/). In particular, the meeting demonstrated that comparative modeling approaches can produce reasonably good structural models but that it is still difficult to predict the structure of regions that are substantially different from the target. Recognizing novel folds remains a challenge, as well as predicting secondary structures and long-range contacts.

Modeling work in the field is invaluable because it teaches us to ask, and seek answers to, systematic questions about sequence/structure/function relationships and about the underlying forces that stabilize biomolecular structures. Still, given the rapid improvements in the experimental arena, the pace at which modeling predictions improve must be expeditious to make a significant contribution to protein structure prediction from sequence. To this end, computational biologists are soliciting candidates for the “Ten Most Wanted” proteins.

The annual Johns Hopkins Coolfront Protein Folding meetings also reflect the great progress made in this exciting field. As gleaned from reports of the fourth [160] and fifth [220] meetings, progress is rapid on many theoretical and experimental fronts. Particularly exciting are emerging areas of study dealing with protein folding in membranes and protein folding diseases (e.g., from misfolding or chaperone interference; see below). Furthermore, progress on both experimental ultrafast methods for studying protein folding kinetics and theoretical predictions based on sequence and structural genomics is impressive.

Box 2.1: Paracelsus Challenge

In 1994, George Rose and Trevor Creamer posed a challenge, named after a 16th-century alchemist: change the sequence of a protein by 50% or less to create an entirely different 3D global folding pattern [177]. Though this challenge might sound not particularly difficult, imagine altering at most 50% of the ingredients for a chocolate cake recipe so as to produce bouillabaisse instead! Rose and Creamer offered a reward of $1000 to entice entrants.

The transmutation was accomplished four years later by Lynne Regan and coworkers [46], who converted the four-stranded β-sheet B1 domain of protein G — which has the β sheets packed against a single helix — into a four-helix bundle of two associating helices called Janus (see Figure 2.2). These contestants achieved this wizardry by replacing residues in a β-sheet-encoding domain (i.e., those with high β-sheet-forming propensities) with those corresponding to the four-helix-bundle protein Rop (repressor of primer). Other modifications were guided by features necessary for Rop stability (i.e., internal salt bridge), and the combined design was guided by energy minimization and secondary-structure prediction algorithms.

The challenge proposers, though delighted at the achievement they stimulated, concluded that in the future only tee-shirt prizes should be offered rather than cash!

2.2.3 *Folding by Dynamics Simulations?*

While molecular dynamics simulations are beginning to approach the timescales necessary to fold small peptides [47] or small proteins [55], we are far from finding the Holy Grail, if there is one [12]. Indeed, IBM's ambitious announcement in December 1999 of building a 'Blue Gene' *petaflop* computer (i.e., capable of 10^{15} floating-point operations per second) for folding proteins by 2005 depends on the computational models guiding this ubiquitous cellular process.[3] For example, some proteins require active escorts to assist in their folding *in vivo*. These *chaperone* molecules assist in the folding and rescue misbehaved polymers. Though many details are not known about the mechanisms of chaperone assistance (see below), we recognize that chaperones help by guiding structure assembly and preventing aggregation of misfolded proteins. For an overview of chaperones, see

[3] This $100 million program to build the world's most powerful supercomputer depends on a million processors running in parallel with the PIM (processor in memory) design. This architecture puts the memory and processor on the same chip: a total of 40,000 chips with 25 or more processors per chip. Besides the crucial role of software in the success leading to protein folding 'in silico', it is a challenge to eliminate bottlenecks in this enormous parallel system that might result between memory and processors when sequential (rather than parallel) computations must be performed.

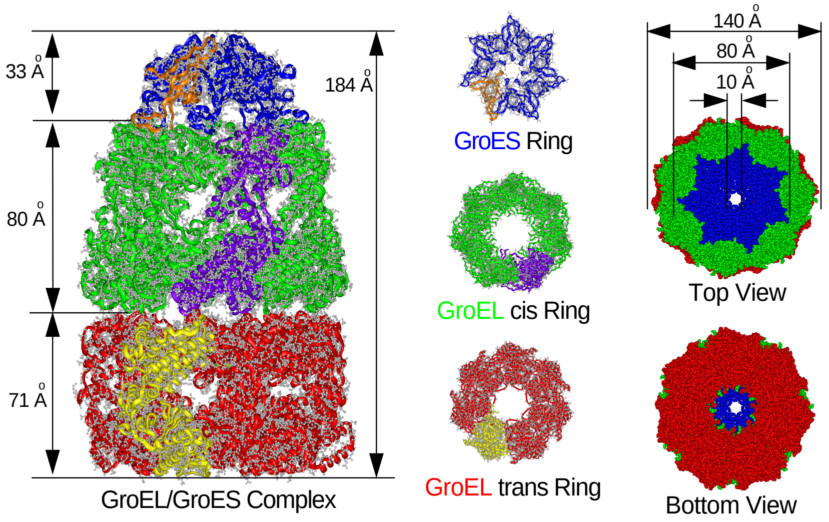


Figure 2.3. The bullet-shaped architecture of the GroEL/GroES chaperonin/co-chaperonin complex sequence. Overall assembly and dimensions are shown from a side view (left). The top ring is the GroES 'cap', and the other layers are GroEL rings. Sidechains are shown in grey. As seen from the top and bottom views (right), a central channel forms in the interior, conducive to protein folding. The protein is organized as three rings that share a 7-fold rotational axis of symmetry (middle), where GroEL contains 14 identical proteins subunits assembled in two heptameric rings, and GroES contains 7 smaller identical subunits in its heptamer ring.

[90, 217], for example, and the 1 August 2001 issue of the *Journal of Structural Biology*, volume 135, devoted to chaperones.

2.2.4 Folding Assistants

Current studies on chaperone-assisted folding, especially of the archetypal chaperone duo, the *E. Coli* bacterial chaperonin GroEL and its cofactor GroES, are providing insights into the process of protein folding [68, 205] (see Figure 2.3 and Box 2.2). The rescue acts of chaperones depend on the subclass of these escorts and the nature of the protein being aided. Some chaperones can assist a large family of protein substrates, while others are more restrictive (see Box 2.2); detailed structural explanations remain unclear. Many families of chaperones are also known, varying in size from small monomers (e.g., 40 or 70 kDa for DnaJ and DnaK of Hsp70) to large protein assemblies (e.g., 810 kDa for GroEL or 880 kDa for the GroEL/GroES complex).

The small assistants bind to short runs of hydrophobic residues[4] to delay premature folding and prevent aggregation. Larger chaperones are likely needed to prevent aggregation of folded compact intermediates in the cell termed ‘molten globules’, requiring a complex trap-like mechanism involving co-chaperones (see also Box 2.2).

Recent work also suggests that small chaperones in the PapD-like superfamily — which directs the folding of various surface organelles — facilitate folding by providing direct *steric* information to their substrates, in addition to capping their interactive surfaces [154, 10]. In certain assemblies (pilus subunits called pilins), the chaperone donates a β-strand that completes the fold of the pilin to promote correct folding and binding to neighboring subunits.

Since macromolecular crowding affects aggregation and diffusion properties of proteins, uncertainties remain regarding the interpretation of *in vitro* experiments on chaperone-assisted folding. Different models for chaperone activity have also been proposed: a ‘pathway’ route for direct folding, and a ‘network’ model, involving iterative folding. They differ by the extent to which unfolded proteins are released to the cellular medium while they remain vulnerable to aggregation [64, 63].

2.2.5 Unstructured Proteins

Though our discussion has focused on the concept of native folds, not all proteins are intrinsically structured [61]. The intrinsic lack of structure can be advantageous, for example in binding versatility to different targets or in ability to adapt different conformations. Unfolded or non-globular structures are recognized in connection with regulatory functions, such as binding of protein domains to specific cellular targets [61, 230]. Examples include DNA and RNA-binding regions of certain protein complexes (e.g., basic region of leucine zipper protein GCN4, DNA-binding domain of NFATC1, RNA recognition regions of the HIV-1 Rev protein). Here, the unstructured regions become organized only upon binding to the DNA or RNA target. This folding flexibility offers an evolutionary advantage, which might be more fully appreciated in the future, as more gene sequences that code for unstructured proteins are discovered and analyzed.

[4] The terms *hydrophobic* (‘water-hating’) and *hydrophilic* (‘water-loving’) characterize water-soluble and water-insoluble molecular groups, respectively.

Box 2.2: Studies on Protein Escorts

The archetypal chaperone GroEL is a member of a chaperone class termed *chaperonins*; hsp60 of mitochondria and chloroplasts is another member of this class. These chaperones bind to partially-folded peptide chains and assist in the folding with the consumption of ATP. The solved crystal structure of GroEL/GroES [233] suggests beautifully, in broad terms, how the large central channel inside a barrel-shaped chaperone might guide protein compaction in its container and monitor incorrect folding by diminishing aggregation (see Figure 2.3). The two-ringed GroEL chaperonin (middle and bottom levels in Figure 2.3) attaches to its partner chaperonin GroES (top ring) upon ATP binding, causing a major conformational change; the size of GroEL nearly doubles, and it assumes a cage shape, with GroES capping over it. This capping prevents the diffusion of partially folded compact intermediates termed ‘molten globules’ and offers them another chance at folding correctly.

Experiments that track hydrogen exchange in unfolded rubisco protein by radioactive tritium (a hydrogen isotope) labeling suggest how misfolded proteins fall into this cavity and are released: a mechanical stretching force triggered by ATP binding partially or totally unfolds the misfolded proteins, eventually releasing the captive protein [192].

These results also support an *iterative annealing* (or *network* model) for chaperone-guided folding, in which the process of forceful unfolding of misfolded molecules, their trapping in the cavity, and their subsequent release is iterated upon until proper folding.

The identification of preferential substrates for GroEL *in vivo* [98], namely multidomain *E. Coli* proteins with complex $\alpha\beta$ folds, further explains the high-affinity interactions formed between the misfolded or partially folded proteins and binding domains of GroEL. These proteins require the assistance of a chaperone because assembly of β-sheet domains requires long-range coordination of specific contacts (not the case for formation of α-helices). Natural substrates for other chaperones, like Eukaryotic type II chaperonin CCT, also appear selective, favoring assistance to proteins like actin [129].

However, such insights into folding kinetics are only the tip of the iceberg. Chaperone types and mechanisms vary greatly, and the effects of macromolecular crowding (not modeled by *in vitro* experiments) complicate interpretations of folding mechanisms *in vivo*. Unlike chaperonins, members of another class of chaperones that includes the heat-shock protein Hsp70 bind to exposed hydrophobic regions of newly-synthesized proteins and short linear peptides, reducing the likelihood of aggregation or denaturation. These are classified as ‘stress proteins’ since their amount increases as environmental stresses increase (e.g., elevated temperatures). Other chaperones are known to assist in protein translocation across membranes.

2.3 Protein Misfolding – A Conundrum

2.3.1 Prions and Mad Cows

Further clues into the protein folding enigma are also emerging from another puzzling discovery involving certain proteins termed *prions*. These misfolded proteins — triggered by a conformational change rather than a sequence mutation — appear to be the source of infectious agent in fatal neurodegenerative diseases like bovine spongiform encephalopathy (BSE) or 'mad cow disease' (identified in the mid 1980s in Britain), and the human equivalent Creutzfeld-Jacob disease (CJD).[5] The precise mechanism of protein-misfolding induced diseases is not known, but connections to neurodegenerative diseases, which include Alzheimer's, are growing and stimulating much interest in protein misfolding [42, 52].

Stanley Prusiner, a neurology professor at the University of California at San Francisco, coined the term prion to emphasize the infectious source as the protein ('proteinaceous'), apparently in contradiction to the general notion that nucleic acids must be transferred to reproduce infectious agents. Prusiner won the 1998 Nobel Prize in Physiology or Medicine for this *"pioneering discovery of an entirely new genre of disease-causing agents and the elucidation of the underlying principles of their mode of action"*.

Prions add a new symmetry to the traditional roles long delegated to nucleic acids and proteins! Since the finding in the 1980s that nucleic acids (catalytic RNAs) can *catalyze* reactions — a function traditionally attributed to proteins only — the possibility that certain proteins, prions, *carry genetic instructions* — a role traditionally attributed to nucleic acids — completes the duality of functions to both classes of macromolecules.

2.3.2 Infectious Protein?

Is it possible for an ailment to be transmitted by 'infectious proteins' rather than viruses or other traditional infectious agents? The prion interpretation for the infection mechanism remains controversial for lack of clear molecular explanation. In fact, one editorial article stated that "*whenever prions are involved, more open questions than answers are available*" [1]. Yet the theory is winning more converts with laboratory evidence that an infectious protein that causes mad cow disease also causes a CJD variant in mice [187]. These results are somewhat frightening because they suggest that the spread of this illness from one species to another is easier than has been observed for other diseases.

The proteinaceous theory suggests that the prion protein (see Figure 2.4) in the most studied neurodegenerative prion affliction, *scrapie* (long known in sheep and goats), becomes a pathologic agent upon conversion of one or more of its α-helical regions into β-regions (e.g., parallel β-helix [223]); once this confor-

[5] See information from the UK Department of Health on www.doh.gov.uk, the UK CJD Surveillance Unit at www.cjd.ed.ac.uk, and the CJD Disease Foundation at cjdfoundation.org.

mational change occurs, the conversion of other cellular neighbors proceeds by a domino-like mechanism, resulting in many abnormally-folded molecules which eventually reap havoc in the mammal. This protein-only hypothesis was first formulated by J.S. Griffith in 1967, but Prusiner first purified the hypothetical abnormal protein thought to cause BSE. New clues are rapidly being added to this intriguing phenomenon (see Box 2.3).

Both the BSE and CJD anomalies implicated with prions have been linked to unusual deposits of protein aggregates in the brain. (Recent studies on mice also open the possibility that aberrant proteins might also accumulate in muscle tissue). It is believed that a variant of CJD has caused the death of dozens of people in Britain (and a handful in other parts of the world) since 1995 who ate meat infected with BSE, some only teenagers. Recent studies also suggest that deaths from the human form of mad cow disease could be rising significantly and spreading within Europe as well as to other continents.

Since the incubation period of the infection is not known — one victim became a vegetarian 20 years before dying of the disease — scientists worry about the extent of the epidemic in the years to come. The consequences of these deaths have been disastrous to the British beef industry and have led indirectly to other problems (e.g., the 2001 outbreak of foot-and-mouth disease, a highly infectious disease of most farm animals except horses). The panic has not subsided, as uncertainties appear to remain regarding the safety of various beef parts, as well as sheep meat, and the possible spread of the disease to other parts of the world.

2.3.3 Other Possibilities

Many details of this intriguing prion hypothesis and its associated diseases are yet to be discovered and related to normal protein folding. Some scientists believe that a lurking virus or virino (small nonprotein-encoding virus) may be involved in the process, perhaps stimulating the conformational change of the prion protein, but no such evidence has yet been found. Only creation of an infection *de novo* in the test tube is likely to convince the skeptics, but the highly unusual molecular transformation implicated with prion infection is very difficult to reproduce in the test tube.

Box 2.3: Prion: Structural Evidence

The detailed structural picture associated with the prion conformational change is only beginning to emerge as new data appear [2]. In 1997, Kurt Wütrich and colleagues at the Swiss Federal Institute of Technology in Zurich reported the first NMR solution structure of the 208-amino acid glycoprotein “prion protein cellular” $\mathrm{PrP}^{\mathrm{C}}$ anchored to nerve cell membranes. The structure reveals a flexibly disordered assembly of helices and sheets (see Fig. 2.4). This organization of the harmless protein might help explain the conversion process to its evil isoform $\mathrm{PrP}^{\mathrm{Sc}}$. It has been suggested that chaperone molecules may bind to $\mathrm{PrP}^{\mathrm{C}}$ and drive its conversion to $\mathrm{PrP}^{\mathrm{Sc}}$ and that certain membrane proteins may also be involved in the transformation.

In early 1998, a team from the University of California at San Francisco discovered a type of prion, different from that associated with mad cow disease, that attaches to a major structure in neuron cells and causes cells to die by transmitting an abnormal signal. This behavior was observed in laboratory rats who quickly died when a mutated type of prion was placed into the brains of newborn animals; their brains revealed the abnormal prions stuck within an internal membrane of neuron cells. The researchers believe that this mechanism is the heart of some prion diseases. They have also found such abnormal prions in the brain tissue of patients who died from a rare brain disorder called Gerstmann-Straussler-Scheinker disease (GSS) — similar to Creutzfeld-Jacob disease (CJD) — that destroys the brain.

Important clues to the structural conversion process associated with prion diseases were further offered in 1999, when a related team at UCSF, reported the NMR structure of the core segment of a prion protein rPrP that is associated with the scrapie prion protein $\mathrm{PrP}^{\mathrm{Sc}}$ [104, 128]. The researchers found that part of the prion protein exhibits multiple conformations. Specifically, an intramolecular hydrogen bond linking crucial parts of the protein can be disrupted by a single amino acid mutation, leading to different conformations. This compelling evidence on how the molecule is changed to become infectious might suggest how to produce scrapie-resistant or BSE-resistant species by animal cloning.

Prion views from several organisms (including human and cow) have been obtained [239], allowing analyses of species variations, folding, and misfolding relationships; see [223], for example. This high degree of similarity across species is shown in Figure 2.4.

Still, until prions are demonstrated to be infectious *in vivo*, the proteinaceous hypothesis warrants reservation. Clues into how prions work may emerge from parallel work on yeast prions, which unlike their mammalian counterparts do not kill the organism but produce transmitted heritable changes in phenotype; many biochemical and engineering studies are underway to explore the underlying mechanism of prion inheritance.

2.3.4 Other Misfolding Processes

There are other examples of protein misfolding diseases (e.g., references cited in [85, 52]). The family of amyloid diseases includes Alzheimer's, Parkinson's, and type II (late-onset) diabetes. For example, familial amyloid polyneuropathy is a heritable condition caused by the misfolding of the protein transthyretin. The amyloid deposits that result interfere with normal nerve and muscle function.

Dobson [52] intriguingly suggests that understanding the evolution of proteins holds the key to protein misfolding diseases. Namely, he argues that since evolutionary processes have selected sequences of amino acids that form close-packed, globular proteins, the effectively irreversible formation of amyloid fibrils reflects a conversion of proteins to their 'primordial' rather than evolved states, possibly from aging-induced mutations that destabilize native proteins.

As in mad cow disease, a molecular understanding of the misfolding process may lead to treatments of the disorders. In the case of familial amyloid polyneuropathy, research has shown that incorporating certain mutant monomers in the tetramer protein transthyretin reduces considerably the formation of amyloid deposits (amyloid fibrils); moreover, incorporating additional mutant monomers can prevent misfolding entirely [85]. These findings suggest potential therapeutic strategies for amyloid and related misfolding disorders. See also [164] for a new pharmacological approach for treating human amyloid diseases by using a small-molecule drug that targets a protein present in amyloid deposits; the drug links two pentamers of that protein and leads to its rapid clearance by the liver.

Recent studies also suggest that misfolded proteins generated in the pathway of protein folding can be dangerous to the cell and cause harm (whether or not they convert normal chains into misfolded structures, as in prion diseases) [27, 216]. The cellular mechanisms associated with such misfolded forms and aggregates are actively being pursued.

2.3.5 Deducing Function From Structure

Having the sequence and also the 3D structure at atomic resolution, while extremely valuable, is only the beginning of understanding biological function. How does a complex biomolecule accommodate its varied functions and interactions with other molecular systems? How sensitive is the 3D architecture of a biopolymer to its constituents?

Despite the fact that in many situations protein *structures* are remarkably stable to tinkering (mutations), their *functional* properties can be quite fragile. In other words, while a protein often finds ways to accommodate substitutions of a few amino-acids so as not to form an entirely different overall folding motif [33], even the most minute sequence changes can alter biological activity significantly. Mutations can also influence the *kinetics* of the folding pathway.

An example of functional sensitivity to sequence is the altered transcriptional activity of various protein/DNA complexes that involve single base changes in the TATA-box recognition element and/or single protein mutations in TBP (TATA-Box binding protein) [161]. For example, changing just a single residue in the common nucleotide sequence of TATA-box element, TA**T**AAAAG, to TA**A**AAAAG impairs binding to TBP and hence disables transcriptional activity.

In principle, theoretical approaches should be able to explain these relations between sequence and structure from elementary physical laws and knowledge of basic chemical interactions. In practice, we are encountering immense difficulty pinpointing what Nature does so well. After all, the notorious "*protein folding*" problem is a challenge to us, not to Nature.

Much work continues on this active front.

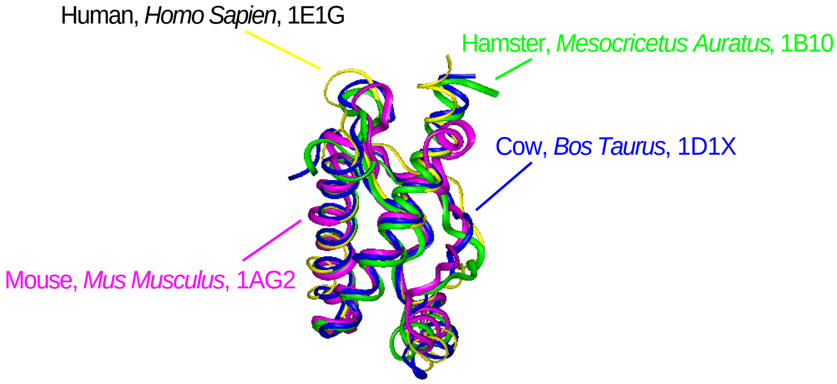

Human, *Homo Sapien*, 1E1G
Hamster, *Mesocricetus Auratus*, 1B10
Cow, *Bos Taurus*, 1D1X
Mouse, *Mus Musculus*, 1AG2

Figure 2.4. Structure of the prion protein.

2.4 From Basic to Applied Research

An introductory chapter on biomolecular structure and modeling is aptly concluded with a description of the many important practical applications of the field, from food chemistry to material science to drug design. A historical perspective on drug design is given in Chapter 14. Here, we focus on the current status of drug development as well as other applied research areas that depend strongly on progress in molecular modeling. Namely, as biological structures and functions are being resolved, natural disease targets that affect the course of disease can be proposed. Other biological and polymer targets, such as the ripening genes of vegetables and fruit or strong materials, can also be manipulated to yield benefits to health, technology, and industry.

2.4.1 Rational Drug Design: Overview

The concept of systematic drug design, rather than synthesis of compounds that mimic certain desired properties, is only about 50 years old (see Chapter 14). Gertrude Elion and George Hitchings of Burroughs Wellcome, who won the 1988 Nobel Prize in Physiology or Medicine, pioneered the field by creating analogues of the natural DNA bases in an attempt to disrupt normal DNA synthesis. Their strategies eventually led to a series of drugs based on modified nucleic-acid bases targeted to cancer cells. Today, huge compound libraries are available for systematic screening by various combinatorial techniques, robotics, other automated technologies, and various modeling and simulation protocols (see Chapter 14).

Rational pharmaceutical design has now become a lucrative enterprise. The sales volume for the world's best seller prescription drug in 1999, *prilosec* (for ulcer and heartburn), exceeded six billion dollars. A vivid description of the climate in the pharmaceutical industry and on Wall Street can be found in *The Billion-Dollar Molecule: One Company's Quest for the Perfect Drug* [222]. This thriller describes the racy story of a new biotech firm for drugs to suppress the immune system, specifically the discovery of an alternative treatment to *cyclosporin*, medication given to transplant patients. Since many patients cannot tolerate cyclosporin, an alternative drug is often needed.

Tremendous successes in 1998, like Pfizer's anti-impotence drug *viagra* and Entre-Med's drugs that reportedly eradicated tumors in mice, have generated much excitement and driven sales and earnings growth for drug producers. A glance at the names of biotechnology firms is an amusing indicator of the hope and prospects of drug research: Biogen, Cor Therapeutics, Genetech, Genzyme, Immunex, Interneuron Pharmaceuticals, Liposome Co., Millennium Pharmaceuticals, Myriad Genetics, NeXstar Pharmaceuticals, Regeneron Pharmaceuticals, to name a few. Yet, both the monetary cost and development time required for each successful drug remains very high [16].

2.4.2 *A Classic Success Story: AIDS Therapy*

HIV Enzymes

A spectacular example of drugs made famous through molecular modeling successes are inhibitors of the two viral enzymes *HIV protease* (HIV: human immunodeficiency virus) and *reverse transcriptase* for treating AIDS, acquired immune deficiency syndrome.

This world's most deadly infectious disease is caused by an insidious retrovirus. Such a virus can convert its RNA genome into DNA, incorporate this DNA into the host cell genome, and then spread from cell to cell. To invade the host, the viral membrane of HIV must attach and fuse with the victim's cell membrane; once entered, the viral enzymes reverse transcriptase and integrase transform HIV's RNA into DNA and integrate the DNA into that of the host [92].

Current drugs inhibit enzymes that are key to the life cycle of the AIDS virus (see Figure 2.5). **Protease inhibitors** like indinavir block the activity of proteases, protein-cutting enzymes that help a virus mature, reproduce, and become infectious [39]. **Reverse transcriptase** (RT) inhibitors block the action of an enzyme required by HIV to make DNA from its RNA [173].

AIDS Drug Development

A typical drug cocktail is the triplet drug combination of the protease inhibitor indinavir with the two nucleoside analogue RT inhibitors AZT (zidovudine, or 3′-azido-3′-deoxythymidine) and 3TC. More than one drug is needed because mutations in the HIV enzymes can confer drug resistance; thus, acting on different sites as well as on different HIV proteins increases effectiveness of the therapy.

Two types of RT blockers are *nucleoside analogues* and *non-nucleoside* inhibitors. Members of the former group like AZT interfere with the HIV activity by replacing a building block used to make DNA from the HIV RNA virus with an inactive analog and thereby prevent accurate decoding of the viral RNA. Non-nucleoside RT inhibitors (e.g., *nevirapine*, *calanolide* molecules, and *sustiva*

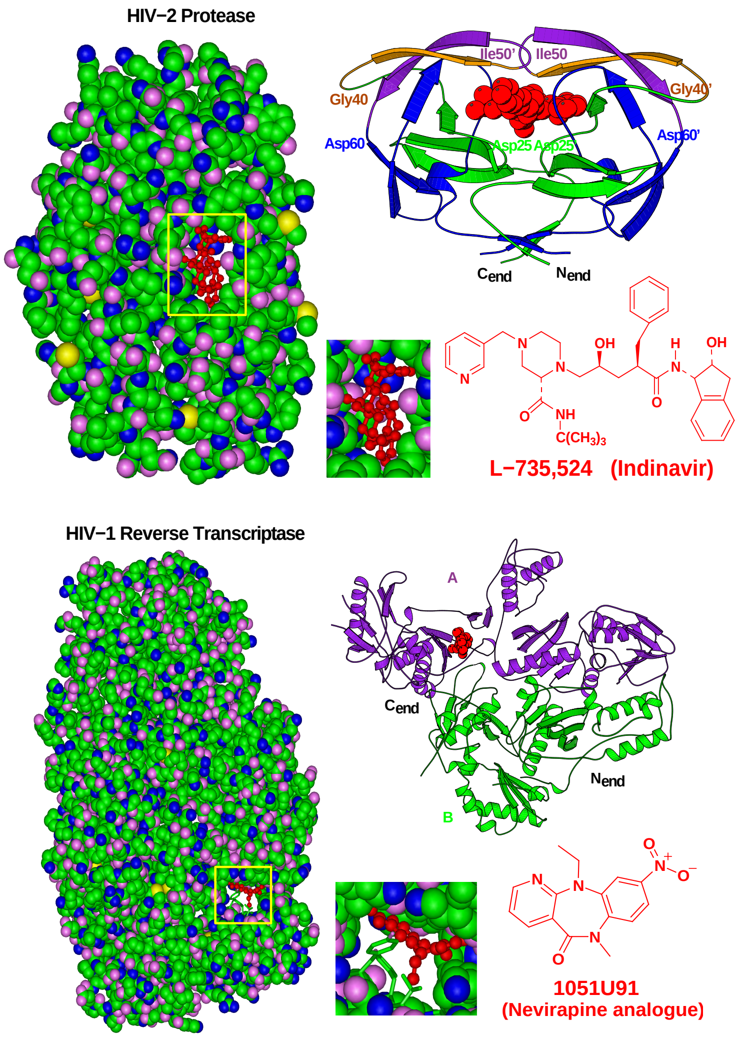


Figure 2.5. Examples of AIDS drug targets — the HIV protease inhibitor and reverse transcriptase (RT) — with corresponding designed drugs. The protease inhibitor indinavir (crixivan) binds tightly to a critical area of the dimer protease enzyme (HIV-2, 198 residues total shown here [39]), near the flaps (residues 40 to 60 of each monomer), inducing a conformational change (flap closing) that hinders enzyme replication; intimate interactions between the ligand and enzyme are observed in residues 25 and 50 in each protease monomer. The non-nucleoside RT inhibitor 1051U91 (a nevirapine analogue), approved for use in combination with nucleoside analogue anti-HIV drugs like AZT, binds to a location near the active site of RT that does not directly compete with the oligonucleotide substrate. The large RT protein of 1000 residues contains two subdomains (A and B).

under development) are designed to bind with high affinity to the active site of reverse transcriptase and therefore physically interfere with the enzyme's action.

Design of such drugs was made possible in part by molecular modeling due to the structure determination of the HIV protease by X-ray crystallography in 1989 and RT a few years later [197]. Figure 2.5 shows molecular views of these HIV enzymes complexed with drugs.

Besides the HIV protease and reverse transcriptase, a third target is the HIV integrase, which catalyzes the integration of a DNA copy of the viral genome into the host cell chromosomes. Scientists at Merck have identified 1,3-diketo acid integrase inhibitors that block strand transfer, one of the two specific catalytic functions of HIV-1 integrase [94]; this function has not been affected by previous inhibitors. This finding paves the way for developing effective integrase inhibitors.

AIDS Drug Limitations

Much progress has been made in this area since the first report of the rational design of such inhibitors in 1990 [176] (see [16] for a review). In fact, the dramatic decline of AIDS-related deaths by such drug cocktails can be attributed in large part to these new generation of designer drugs (see Box 2.4) since the first introduction of protease inhibitors in 1996. Indeed, the available drug cocktails of protease inhibitors and nucleoside analogues RT inhibitors have been shown to virtually suppress HIV, making AIDS a manageable disease.

However, the cocktails are not a cure. The virus returns once patients stop the treatment. The mechanisms of drug resistant mutations and the interactions among them are not clearly understood [173]. In addition, few countires in the developing world, like Africa, can afford the virus suppressing drugs; the drug-cocktail regimen is complex, requiring many daily pills taken at multiple times and separated from eating, most likely for life; serious side effects also occur. In certain parts of the world, the situation is profoundly distressing: the life expectancy of sexually-active Ugandans, for example, has fallen from 64 years before the epidemic to 42 today, and the number continues to drop.

Lurking Virus

As mentioned, even available treatments cannot restore the damage to the patient's immune system; the number of T-cell (white blood cells), which that HIV attaches itself to, is still lower than normal (which lowers the body's defenses against infections), and there remain infected immune cells that the drugs cannot reach because of integration. Thus, new drugs are being sought to interrupt the first step in the viral life cycle — binding to a co-receptor on the cell surface to rid the body of the cell's latent reservoirs of the HIV virus, to chase the virus out of cells where it hides for subsequent treatment, or to drastically reduce the HIV reservoir so that the natural immune defenses can be effective. New structural and mechanistic targets are currently being explored (see Box 2.4).

A better understanding of the immune-system mechanism associated with AIDS, for example, may help explain how to prime the immune system to recognize an invading AIDS virus. Unlike traditional AIDS drug cocktails which inhibit division of already infected cells, fusion (or entry) inhibitors define another class of drugs that seek to prevent HIV from entering the cell membrane. This entry, called fusion, releases the virus's genetic material and allows it to replicate. The promising drug T-20 or enfuvirtide (which must be injected into the skin) is a member of fusion-inhibitor drugs that, when added to a combination of standard drugs, can significantly reduce HIV levels in the blood. As manifested by its complex components of invasion that include the fusion apparatus, the AIDS virus has developed a complex, tricky, and multicomponent-protection infection machinery, as well as drug-resistant defense.

Besides integrase and fusion inhibitors, among the newer drugs to fight AIDS being developed are immune stimulators and antisense drugs. The former stimulate the body's natural immune response, and the latter mimic the HIV genetic code and prevent the virus from functioning.

Vaccine?

Still, many believe that only an AIDS vaccine offers true hope against this deadly disease. Yet the research on vaccines trails behind the development of drugs, which offer much greater financial incentives and lower risks The vaccine AIDSVAX by the California-base company VaxGen (ready by 2004 or 2005) could protect in part people from HIV, but the early 2003 results from the first large-scale trial were not as encouraging as had been hoped. This vaccine is a genetically amplified version of a single protein from the outer shell of the AIDS virus; because the shell changes rapidly, the vaccine may offer only limited protection.

Another vaccine under development by an Oxford team (part of the International AIDS vaccine Initiative) is exploiting for vaccine development the immunological data gleaned from Nairobi women who have remained unaffected by AIDS despite many years of high-risk sexual behavior. These women's T-cells were found to fight off the disease by attacking two particular proteins produced by the AIDS virus. The DNA sequences making those proteins were subsequently identified and used to create a vaccine specific to viral infections in East Africa; besides the DNA component associated with the relevant genes, the vaccine was amplified with a benign virus copy with same DNA sequences inserted.

Other vaccines are being developed (e.g., Harvard, Merck) but, like those further along, response is far from ideal. Unfortuntately, experience suggests that a constant level of exposure (e.g., booster shots) is needed to yield immunity, and this defeats the main vaccine advantage of convenience and low cost. Observations also suggest that more than one vaccine may be needed, since the HIV virus mutates and replicates quickly.[6] Still, it is hoped that therapeutic vaccination in

[6] For example, there is an enormous variation in the HIV-1 envelope protein. It has also been found that nearly all of non-nucleoside reverse transcriptase inhibitors can be defeated by site-directed mutation of tyrosine 181 to cysteine in reverse transcriptase. For this reason, the derivatives of *calano-*

combination with anti-HIV-1 drug treatment, even if it fails to eradicate infection, will suppress AIDS infection and the rate of transmission, and ultimately decrease the number of AIDS deaths substantially.

For a comprehensive overview of the biology of AIDS, the HIV life cycle, current status of the AIDS pandemic, and efforts for treating AIDS, see *Nature Insight* in the 19 April 2001 issue of *Nature* (Volume 410). This review was written at the 20 years after the first hints of the disease were reported in the summer of 1981, in clusters of gay men in large American cities; these groups exhibited severe symptoms of infection by certain pneumonias combined with those from Kaposi's sarcoma (KS) cancer.

Box 2.4: Fighting AIDS

AIDS drugs attributed to the success of molecular modeling include *AZT* (zidovudine) sold by Bristol-Myers Squibb, and the newer drugs *viracept* (nelfinavir) made by Agouron Pharmaceuticals, *crixivan* (indinavir) by Merck & Company, and *amprenavir* discovered at Vertex Pharmaceuticals Inc. and manufactured by Glaxo Wellcome. Amprenavir, in particular, approved by the U.S. Government in April 1999, is thought to cross the 'blood-brain barrier' so that it can attack viruses that lurk in the brain, where the virus can hide. This general class of inhibitors has advanced so rapidly that drug-resistant AIDS viruses have been observed.

Current structural investigations are probing the structural basis for the resistance mechanisms, which remain mysterious, particularly in the case of nucleoside analogue RT inhibitors like AZT [121]. The solved complex of HIV-1 reverse transcriptase [100] offers intriguing insights into the conformational changes associated with the altered viruses that influence the binding or reactivity of inhibitors like AZT and also suggests how to construct drug analogues that might impede viral resistance.

Basic research on the virus's process of invading host cells — by latching onto receptors (e.g., the CD4 glycoprotein, which interacts with the viral envelope glycoprotein, gp120, and the transmembrane component glycoprotein, gp41), and co-receptors (e.g., CCR5 and CXCR4) — may also offer treatments, since developments of disease intervention and vaccination are strongly aided by an understanding of the complex entry of HIV into cells; see [119] for example.

The HIV virus uses a spear-like agent on the virus' protein coat to puncture the membrane of the cells which it invades; vaccines might be designed to shut the chemical mechanism or stimuli that activate this invading harpoon of the surface protein. The solved structure of a subunit of gp41, for example, has been exploited to design peptide inhibitors that disrupt the ability of gp41 to contact the cell membrane [69]. A correlation has been noted, for example, between co-receptor adaptation and disease progression.

lide A under current development are attractive drug targets because they appear more robust against mutation [113].

Novel techniques for gene therapy for HIV infections are also under development, such as internal antibodies (*intrabodies*) against the Tat protein, a vehicle for HIV infection of the immune cells; it is hoped that altered T-cells that produce their own anti-Tat intrabody would lengthen the survival time of infected cells or serve as an HIV ‘dead-end’.

Other clues to AIDS treatments may come from the finding that HIV-1 originally came from a subspecies of chimpanzees [80]. Since chimps have likely carried the virus for hundreds of thousands of years but not become ill from it, understanding this observation might help fight HIV-pathogeny in humans. Help may also come from the interesting finding that a subset of humans have a genetic mutation (32 bases deleted from the 393 of gene CCR5) that creates deficient T-cell receptor; this mutation intriguingly slows the onset of AIDS. Additionally, a small subset of people is endowed with a huge number of helper (CD4) T-cells which can coordinate an attack on HIV and thus keep the AIDS virus under exquisite control for many years; such people may not even be aware of the infection for years.

2.4.3 Other Drugs

Another example of drug successes based on molecular modeling is the design of potent *thrombin inhibitors*. Thrombin is a key enzyme player in blood coagulation, and its repressors are being used to treat a variety of blood coagulation and clotting-related diseases. Merck scientists reported [17] how they built upon crystallographic views of a known thrombin inhibitor to develop a variety of inhibitor analogues. In these analogues, a certain region of the known thrombin inhibitor was substituted by hydrophobic ligands so as to bind better to a certain enzyme pocket that emerged crucial for the fit. Further modeling helped select a subset of these ligands that showed extremely compact thrombin/enzyme structures; this compactness helps oral absorption of the drug. The most potent inhibitor that emerged from these modeling studies has demonstrated good efficacy on animal models [17].

Other examples of drugs developed in large part by computational techniques include the *antibacterial agent* norfloxacin of Kyorin Pharmaceuticals (noroxin is one of its brand names), *glaucoma treatment* dorzolamide (“trusopt”/Merck), *Alzheimer’s disease treatment* donepezil (“aricept”/ Eisai), and *migraine medicine* zolmitriatan (“zomig”) discovered by Wellcome and marketed by Zeneca [16]. The headline-generating drug that combats impotence (*viagra*) was also found by a rational drug approach. It was interestingly an accidental finding: the compound had been originally developed as a drug for hypertension and then angina.

There are also notable examples of *herbicides and fungicides* that were successfully developed by statistical techniques based on linear and nonlinear regression and classical multivariate analysis (or QSAR, see Chapter 14): the herbicide metamitron — a bestseller in 1990 in Europe for protecting sugar beet crops — was discovered by Bayer AG in Germany.

2.4.4 A Long Way To Go

With an annual yield of about 50 new approved pharmaceutical agents that has become accepted in the last couple years, we are enjoying improved treatments for cancer, AIDS, heart disease, Alzheimer and Parkinson's disease, migraine, arthritis, and many more ailments. Yet the average cost of $500 million and time of 12–15 years required to develop a single drug remains extremely high. It can now be hoped that through the new fields of knowledge-based biological information, like *bioinformatics* [99, 148] and *chemoinformatics* [86], computers will reduce drastically these costs. Perhaps such revolutionary advances in drug development, expected in the next decade, will also alleviate the industry's political problems, associated with inadequate availability of drugs to the world's poor population.

Improved modeling and library-based techniques, coupled with robotics and high-speed screening, are also likely to increase the demand for faster and larger-memory computers.

"In a marriage of biotech and high tech," writes the New York Times reporter Andrew Pollack on 10 November 1998, *"computers are beginning to transform the way drugs are developed, from the earliest stage of drug discovery to the late stage of testing the drugs in people"*. Chapter 14 in this text points to some of these computational challenges.

2.4.5 Better Genes

Looking beyond drugs, gene therapy is another approach that is benefiting from key advances in biomolecular structure/function studies. Gene therapy attempts to compensate for defective or missing genes that give rise to various ailments — like hemophilia, the severe combined immune deficiency SCID, sickle-cell anemia, cystic fibrosis, and Crigler-Najjar (CN) syndrome — by trying to coerce the body to make new, normal genes. This regeneration is attempted by inserting replacement genes into viruses or other vectors and delivering those agents to the DNA of a patient (e.g., intravenously). However, delivery control, biological reliability, as well as possible unwelcome responses by the body against the foreign invader remain technical hurdles. See Box 2.5 for examples of gene therapy.

The first death in the fall of 1999 of a gene-therapy patient treated with the common fast-acting weakened cold virus adenovirus led to a barrage of negative publicity for gene therapy.[7] However, the first true success of gene therapy was reported four months later: the lives of most infants who would have died of the severe immune disorder SCID (and until then lived in airtight bubbles to avoid

[7] The patient of the University of Pennsylvania study was an 18-year old boy who suffered from ornithine transcarbamylase (OTC) deficiency, a chronic disorder stemming from a missing enzyme that breaks down dietary protein, leading to accumulation of toxic ammonia in the liver and eventually brain and kidney failure. The teenager suffered a fatal reaction to the adenovirus vector used to deliver healthy DNA rapidly. Autopsy suggests that the boy might have been infected with a second cold virus, parvovirus, which could have triggered serious disorders and organ malfunction that ultimately led to brain death.

the risk of infection) were not only saved, but able to live normal lives following gene therapy treatments that restore the ability of a gene essential to make T cells [35]. Unfortunately, complications arose is several of the treated infants by late 2002 (see Box 2.5).

Though such medical advances appear just short of a miracle, it remains to be seen how effective gene therapy will be on a wide variety of diseases and over a long period. Still, given that gene therapy is a young science in a state of continuous flux, results to date indicate a promising future for the field [6].

A related technique for designing better genes is another relatively new technique known as *directed molecular evolution*. Unlike protein engineering, in which natural proteins are improved by making specific changes to them, directed evolution involves mutating genes in a test tube and screening the resulting ('fittest') proteins for enhanced properties. Companies specializing in this new Darwinian mimicking (e.g., Maxigen, Diversa, and Applied Molecular Evolution) are applying such strategies in an attempt to improve the potency or reduce the cost of existing drugs, or improve the stain-removing ability of bacterial enzymes in laundry detergents. Beyond proteins, such ideas might also be extended to evolve better viruses to carry genes into the body for gene therapy or evolve metabolic pathways to use less energy and produce desired nutrients (e.g., carotenoid-producing bacteria).

Box 2.5: Gene Therapy Examples

A prototype disease model for gene therapy is hemophilia, whose sufferers lack key blood-clotting protein factors. Specifically, Factor VIII is missing in hemophilia A patients (the common form of the disease); the much-smaller Factor IX is missing in hemophilia B patients (roughly 20% of hemophiliacs in the United States).

Early signs of success in treatment of hemophilia B using adeno-associated virus (a vector not related to adenovirus, which is slower acting and more suitable for maintenance and prevention) were reported in December 1999. However, introducing the much larger gene needed for Factor VIII, as required by the majority of hemophiliacs, is more challenging. Here, the most successful treatments to date only increase marginally this protein's level. Yet even those minute amounts are reducing the need for standard hemophilia treatment (injections of Factor IX) in these patients.

Larger vectors to stimulate the patient's own cells to repair the defective gene are thus sought, such as retroviruses (e.g., lentiviruses, the HIV-containing subclass), or non-virus particles, like chimeraplasts (oligonucleotides containing a DNA/RNA blend), which can in theory correct point mutations by initiating the cell's DNA mismatch repair machinery.

An interesting current project involving chimeraplasts is being tested in children of Amish and Mennonite communities to treat the debilitating Crigler-Najjar (CN) syndrome. Sufferers of this disease lack a key enzyme which break down the toxic waste product bilirubin, which in the enzyme's absence accumulates in the body and causes jaundice and overall toxicity. Children with CN must spend up to 18 hours a day under a blue light to clear bilirubin and seldom reach adulthood, unless they are fortunate to receive and respond to a liver transplant. Chimeraplasty offers these children hope, and might reveal to be safer than the adenovirus approach, but the research is preliminary and the immune response is complex and mysterious.

Recent success was reported for treating children suffering from the severe immune disorder SCID type XI [35]. Gene therapy involves removing the bone marrow from infants, isolating their stem-cells, inserting the normal genes to replace the defective genes via retroviruses, and then re-infusing the stem cells into the blood stream. As hoped, the inserted stem cells were able to generate the cells needed for proper immune functioning in the patients, allowing the babies to live normal lives. Though successful for 2–3 years for most infants, complications arose when several infants developed leukemia-like conditions. Scientists believe that the retrovirus vectors lodged near a cancer-causing gene and activated it. Of course, it remains to be seen whether the overall benefits outweigh the risks, and how in the long term children's new immune systems will behave (i.e., deteriorate over time or continue to function properly).

Though clearly many bumps in the road are expected when new therapies are developed, scientists remain hopeful. Indeed, success in any such gene therapy endeavors would lead to enormous progress in treating inherited diseases caused by point mutations.

2.4.6 Designed Compounds and Foods

From our farms to medicine cabinets to supermarket aisles, designer foods are big business.

As examples of these practical applications, consider the transgenic organisms designed to manufacture medically-important compounds: bacteria that produce *human insulin*, goats whose milk contains *proteins to make silk* for use in surgical thread or bulletproof clothing, silkworms that produce *mammalian-type collagen and silk* for use in tissue engineering and other medical applications, and the food product *chymosin to make cheese*, a substitute for the natural rennet enzyme traditionally extracted from cows' stomachs. Genetically modified bacteria, more generally, hold promise for administering drugs and vaccines more directly to the body (e.g., the gut) without the severe side effects of conventional therapies. For example, a strain of the harmless bacteria *Lactococcus lactis* modified to secrete the powerful anti-inflammatory protein interleukin-10 (IL-10) has shown to reduce bowel inflammation in mice afflicted with inflammatory bowel disease (IBD), a group of debilitating ailments that includes Crohn's disease and ulcerative colitis.

The production of drugs in genetically-altered plants — "biopharming" or "molecular pharming" — represents a growing trend in agricultural biotechnology. The goal is to alter gene structure of plants so that medicines can be grown on the farm, such as to yield an edible vaccine from a potato plant against hepatitis B, or a useful antibody to be extracted from a tobacco plant. As in bioengineered foods, many obstacles must be overcome to make such technologies effective as medicines, environmentally safe, and economically profitable. Proponents of molecular pharming hope eventually for far cheaper and higher yielding drugs.

Genetically-engineered crops are also helping farmers and consumers by improving the taste and nutritional value of food, protecting crops from pests, and enhancing yields. Examples include the roughly one-half of the soybean and one-third of the corn grown in the United States, sturdier salad tomatoes,[8] corn pollen that might damage monarch butterflies, papaya plants designed to withstand the papaya ringspot virus, and caffeine-free plants (missing the caffeine gene) that produce decaffeinated cups of java.

Closer to the supermarket, one of the fastest growing category of foods today is *nutraceuticals* (a.k.a. functional foods or pharmaceuticals), no longer relegated only to health-food stores. These foods are designed to improve our overall nutrition as well as to help ward off disease, from cancer prevention to improved brain function. See Box 2.6 for examples. Related are diet ingredients and supplements customized to genetic variations based on gene/diet connections (*nutritional genomics* or *nutrigenomics*), such as diets low in certain proteins (for patients with phenylketonuria) or high in liver, broccoli, and other folic-acid rich foods (for people with a genetic variation that produces a less efficient enzyme involved in processing folic acid).

The general public (first in Europe and now in the United States) has resisted genetically-modified or biotech crops, and this was followed by several blockades of such foods by leading companies, as well as global biosafety accords to protect the environment. Protesters have painted these products as unnatural, hazardous, evil, and environmentally dangerous (‘Frankenfoods’).[9]

With the exception of transferred allergic sensitivities — as in Brazil nut allergies realized in soybeans that contained a gene from Brazil nuts — most negative reactions concerning *food safety* are not scientifically well-grounded in this writer’s opinion. In fact, not only do we abundantly use various sprays and chemicals to kill flies, bacteria, and other organisms in our surroundings and on the farm; each person consumes around 500,000 kilometers of DNA on an

[8] The *Flavr Savr* tomato that made headlines when introduced in 1993 contained a gene that reduces the level of the ripening enzyme polygalacturonase. However, consumers were largely disappointed: though beautiful, the genetically engineered fruit lacked taste. This is because our understanding of fruit ripening is still limited; a complex, coordinated series of biochemical steps is involved — modifying cell wall structure, improving texture, inducing softening, and producing compounds in the fruit that transform flavor, aroma, and pigmentation. Strawberries and other fruit are known to suffer similarly from the limitations of our understanding of genetic regulation of ripening and, perhaps, also from the complexity of human senses! See [215], for example, for a recent finding that a tomato plant whose fruit cannot ripen carries a mutation in a gene encoding a transcription factor.

[9] Amusing Opinion/Art ads that appeared in The New York Times on 8 May 2000 include provocative illustrations with text lines like “GRANDMA’S MINI-MUFFINS are made with 100% NATURAL irradiated grain and other ingredients”; “TOTALLY ORGANIC Biomatter made with Nucleotide Resequencing”; “The Shady Glen Farms Promise: Our Food is fresh from the research labs buried deep under an abandoned farm”. [Note: the font size and form differences here are intentional, mimicking the actual ads].

average day! Furthermore, there are many potential benefits from genetically-engineered foods, like higher nutrients and less dependency on pesticides, and these considerations might win in the long run. Still, environmental effects must be carefully monitored so that genetically-altered food will succeed in the long run (see Box 2.6 for possible problems).

Perhaps to counter fear of introduced allergens, bioengineering is also being used to reduce or remove compounds that cause allergic reactions in people. Though at a relatively early stage, various companies worldwide are using genetic engineering to try to reduce allergies from foods like wheat, rice, soybean, ryegrass, and peanuts. Genes responsible for producing allergenic proteins can be removed (i.e., *knocked out*), as done for soybeans, or the associated proteins redesigned, as in peanuts, so that allergenicity is lost but other nut characteristics are retained. As above, care must be taken to retain flavor, freshness, and looks of the original product, and not to introduce other possible allergens.

In addition to tampering with plants to remove allergens, such biotech companies are also expanding effort on the removal of genes associated with natural toxins. For example, companies (with support of national security organizations) are attempting to remove the toxin *ricin* — one of the deadliest substances known — from castor plants. Castor beans have been cultivated for centuries, and the plant's natural oils (which lack toxicity) are widely used as laxatives and as component in brake fluid, dyes, soaps, and cosmetics. However, the toxic protein ricin can also be extracted from the castor plant, and has been associated with terrorist groups like Al Qaeda, with production of weapons "for mass destruction" in Iraq, and with a famous killing of the spy Georgi Markov on a London sidewalk in 1978 by Bulgarian agents who injected ricin from an umbrella tip into the defector's leg. Once removed, ricin-free castor plants can become more attractive to growers.

Box 2.6: Nutraceuticals Examples

The concept of fortified food is not new. Vitamin-D supplemented milk has eradicated rickets, and fortified breakfast cereals have saved many poor diets. In fact, classic bioengineering has been used for a long time to manipulate genes through conventional plant and animal inter-breeding. But the new claims — relying on our increased understanding of our body's enzymes and many associated vital processes — have been making headlines. ("Stressed Out? Bad Knee? Try a Sip of These Juices.", J.E. Barnes and G. Winter, *New York Times*, Business, 27 May 2001). Tea brews containing sedative roots like kava promise to tame tension and ease stress. Fruit-flavored tonics with added glucosamine (building block of cartilage) and calcium are claimed to soothe stiff knees of aging bodies. (See Chapter 3 on the fibrous protein collagen). Fiber-rich grains are now touted as heart-disease reducers. Herb-coated snacks, like potato corn munchies coated with ginkgo biloba, are advertised as memory and alertness boosters.

With this growing trend of designer foods, the effect of these manipulations on our environment demands vigilant watch. This is because it is possible to create 'super-resistant weeds' or genetically-improved fish that win others in food or mate competitions. This potential danger emerges since, unlike conventional cross-breeding (e.g., producing a tangelo from a tangerine and grapefruit), genetic engineering can overcome the species barrier — by inserting nut genes in soybeans or fish genes in tomatoes, for example. This newer type of tinkering can have unexpected results in terms of toxins or allergens which, once released to the environment, cannot be stopped easily. For example, the first genetically-modified animal to reach American dinner plates is likely to be a genetically-altered salmon endowed with fortified genes that produce growth hormones, making the fish grow twice as fast as normal salmon. The effect of these endowed fish on the environment is yet unknown.

Popular examples of fortified food products with added vitamins and minerals (e.g., calcium and vitamin E) that also help protect against osteoporosis are orange juice, specialty eggs, and some vegetarian burritos. Other designer disease-fighting foods include drinks enriched with echinacea to combat colds; juices filled with amino acids and herbs claimed to boost muscle and brain function; margarines containing plant stanol esters (from soybean or pine trees) to fight heart disease and cancer (by blocking cholesterol absorption from the digestive tract), as well as green teas enriched with ginseng and other herbs; superyogurts to enhance the immune system; and tofu and yams to combat hot flashes. Such functional foods are also touted to lower cholesterol, provide energy, fight off depression, or to protect against salmonella and E. coli poisoning (e.g., yogurt fortified with certain bacteria). Many other enriched food products are under design, for example fruit with increased vitamin C levels using a recently-isolated gene in strawberries (GalUR) that plays an important role in the production of vitamin C.

Will Ginkgo Biloba chips, Tension Tamer cocktails, or Quantum Punch juice become part of our daily diet (and medicine cabinet) in this millennium?

2.4.7 Designer Materials

New specialty materials are also being developed in industry with the needed thermochemistry, stereochemistry (e.g., compounds that bind to one chemical but not its mirror image), and kinetic properties. Examples are enzymes for manufacturing detergents, adhesives and coatings, photography film, or biosensors for explosives [130]. Fullerene nanotubes (giant linear fullerene chains that can sustain enormous elastic deformations [234]), formed from condensed carbon vapor, have many potential applications. These range from architectural components of bridges and buildings, cars, and airplanes to heavy-duty shock absorbers, to components of computer processors, scanning microscopes, and semiconductors.

Long buckyball nanotubes have even been proposed as elements of 'elevators' to space in the new millennium [234]. These applications arise from their small size (their thickness is five orders of magnitude smaller than human hair), amazing electronic properties, and enormous mechanical strength of these polymers. In particular, these miniscule carbon molecules conduct heat much faster than silicon, and could therefore replace the silicon-based devices used in microelectronics, possibly overcoming current limitations of computer memory and speed.

2.4.8 Cosmeceuticals

Cosmeceutical companies are also rising — companies that specialize in design of cosmetics with bioactive ingredients (such as designer proteins and enzymes), including cosmetics that are individually customized (by *pharmacogenomics*) based on genetic markers, such as single nucleotide polymorphisms (SNPs). Most popular are products for sun or age-damaged skin containing alpha hydroxy acids (mainly glycolic and lactic acid), beta hydroxy acids (e.g., salicylic acid), and various derivatives of vitamin A or retinol (e.g., the tretinoin-containing *Retin-A* and *Renova* topical prescriptions). Besides reducing solar scars and wrinkling, products can also aid combat various skin diseases. Many of these compounds work by changing the metabolism of the epidermis, for example by increasing the rate of cell turnover, thereby enhancing exfoliation and the growth of new cells. New cosmeceuticals contain other antioxidants, analogues of various vitamins (A, D, and E), and antifungal agents.

The recent information gleaned from the Human Genome Project can help recognize changes that age and wrinkle skin tissue, or make hair or teeth gray. This in turn can lead to the application of functional genomics technology to develop agents that might help rejuvenate the skin, or color only target gray hair or tooth enamel. Computational methods have an important role in such developments by screening and optimizing designer peptides or proteins. Such biotechnology research to produce products for personal care will likely rise sharply in the coming years.