1

Biomolecular Structure and Modeling: Historical Perspective

Chapter 1 Notation

SYMBOL	DEFINITION
Vectors	
$\mathbf{h}$	unit cell identifier (crystallography)
$\mathbf{r}$	position
$F_{\mathbf{h}}$	structure factor (crystallography)
$\phi_{\mathbf{h}}$	phase angle (crystallography)
Scalars	
d	distance between parallel planes in the crystal
$I_{\mathbf{h}}$	intensity, magnitude of structure factor (crystallography)
V	cell volume (crystallography)
θ	reflection angle (crystallography)
λ	wavelength of the X-ray beam (crystallography)

. . . physics, chemistry, and biology have been connected by a web of causal explanation organized by induction-based theories that telescope into one another. . . . Thus, quantum theory underlies atomic physics, which is the foundation of reagent chemistry and its specialized offshoot biochemistry, which interlock with molecular biology — essentially, the chemistry of organic macromolecules — and hence, through successively higher levels of organization, cellular,

T. Schlick, *Molecular Modeling and Simulation: An Interdisciplinary Guide*,
Interdisciplinary Applied Mathematics 21, DOI 10.1007/978-1-4419-6351-2_1,

> organismic, and evolutionary biology. ... Such is the unifying and highly productive understanding of the world that has evolved in the natural sciences.
>
> Edward O. Wilson: "Resuming the Enlightenment Quest", in *The Wilson Quarterly*, Winter 1998.

1.1 A Multidisciplinary Enterprise

1.1.1 Consilience

The exciting field of modeling molecular systems by computer has been steadily drawing increasing attention from scientists in varied disciplines. In particular, modeling large biological polymers — proteins, nucleic acids, and lipids — is a truly multidisciplinary enterprise. Biologists describe the cellular picture; chemists fill in the atomic and molecular details; physicists extend these views to the electronic level and the underlying forces; mathematicians analyze and formulate appropriate numerical models and algorithms; and computer scientists and engineers provide the crucial implementational support for running large computer programs on high-speed and extended-communication platforms. The many names for the field (and related disciplines) underscore its cross-disciplinary nature: computational biology, computational chemistry, *in silico* biology, computational structural biology, computational biophysics, theoretical biophysics, theoretical chemistry, and the list goes on.

As the pioneer of sociobiology Edward O. Wilson reflects in the opening quote, some scholars believe in a unifying knowledge for understanding our universe and ourselves, or *consilience*[1] that merges all disciplines in a biologically-grounded framework [1377]. Though this link is most striking between genetics and human behavior — through the neurobiological underpinnings of states of mind and mental activity, with shaping by the environment and lifestyle factors — such a unification that Wilson advocates might only be achieved by a close interaction among the varied scientists at many stages of study. The genomic era has such immense ramifications on every aspect of our lives — from health to technology to law — that it is not difficult to appreciate the effects of the biomolecular revolution on our 21st-century society. Undoubtedly, a more integrated synthesis of biological elements is needed to decode life [584].

In biomolecular modeling, a multidisciplinary approach is important not only because of the many aspects involved — from problem formulation to solution — but also since the best computational approach is often closely tailored to the

[1] *Consilience* was coined in 1840 by the theologian and polymath William Whewhell in his synthesis *The Philosophy of the Inductive Sciences*. It literally means the *alignment*, or *jumping together*, of knowledge from different disciplines. The sociobiologist Edward O. Wilson took this notion further recently by advocating in his 1998 book *Consilience* [1377] that the world is orderly and can be explained by a set of natural laws that are fundamentally rooted in biology.

biological problem. In the same spirit, close connections between theory and experiment are essential: computational models evolve as experimental data become available, and biological theories and new experiments are performed as a result of computational insights.[2]

Although few theoreticians in the field have expertise in experimental work as well, the classic example of Werner Heisenberg's genius in theoretical physics but naiveté in experimental physics is a case in point: Heisenberg required the resolving power of the microscope to derive the uncertainty relations. In fact, an error in the experimental interpretations was pointed out by Niels Bohr, and this eventually led to the 'Copenhagen interpretation of quantum mechanics'.

If Wilson's vision is correct, the interlocking web of scientific fields rooted in the biological sciences will succeed ultimately in explaining not only the functioning of a biomolecule and the workings of the brain, but also many aspects of modern society, through the connections between our biological makeup and human behavior.

1.1.2 What is Molecular Modeling?

Molecular modeling is the science and art of studying molecular structure and function through model building and computation. The model building can be as simple as plastic templates or metal rods, or as sophisticated as interactive, animated color stereographics and laser-made wooden sculptures. The computations encompass *ab initio* and semi-empirical quantum mechanics, empirical (molecular) mechanics, molecular dynamics, Monte Carlo, free energy and solvation methods, structure/activity relationships (SAR), chemical/biochemical information and databases, and many other established procedures. The refinement of experimental data, such as from nuclear magnetic resonance (NMR) or X-ray crystallography, is also a component of biomolecular modeling.

I often remind my students of Pablo Picasso's statement on art: *"Art is the lie that helps tell the truth"*. This view applies aptly to biomolecular modeling. Though our models represent a highly-simplified version of the complex cellular environment, systematic studies based on tractable quantitative tools can help discern patterns and add insights that are otherwise difficult to observe. *The key in modeling is to develop and apply models that are appropriate for the questions being examined with them.* Thus, the model's regime of applicability must be clearly defined and its predictability power demonstrated. A case in point is the use of limited historical data on home prices for extrapolative modeling of mortgage-backed securities and credit derivatives; the resulting mispricing of risk was a contributor to the U.S. subprime loan crisis that started in 2007.

The questions being addressed by computational approaches today are as intriguing and as complex as the biological systems themselves. They range

[2] See [176,362,395,396,948], for example, in connection to the characterization of protein folding mechanisms.

from understanding the equilibrium structure of a small biopolymer subunit, to the energetics of hydrogen-bond formation in proteins and nucleic acids, to the kinetics of protein folding, to the complex functioning of a supramolecular aggregate. As experimental triumphs are being reported in structure determination — from ion channel proteins, signaling receptor proteins (receptors), membrane transport proteins (transporters), ribosomes (see Figs. 1.1 and 1.2), various nucleosomes (see figures in Chapter 6), and non-coding RNAs — including new methodologies for their solution, such as advanced NMR, cryo-electron microscopy, and single-molecule biochemistry techniques, modeling approaches are needed to pursue many fundamental questions concerning their biological motions and functions. Modeling provides a way to systematically explore structural/dynamical/thermodynamic patterns, test and develop hypotheses, interpret and extend experimental data, and help better understand and extend basic laws that govern molecular structure, flexibility, and function. In tandem with experimental advances, algorithmic and computer technological advances, especially concerning distributed, loosely-coupled computer networks, have made problems and approaches that were insurmountable a few years ago now possible.

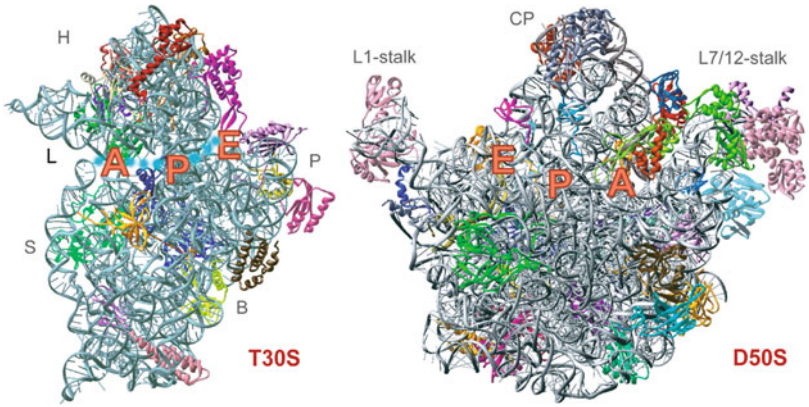

Figure 1.1. The inter-subunit interface of the two eubacterial ribosomal subunits at 3 Å resolution, showing their main architectural features. D50S is the large ribosomal subunit from *Deinococcus radiodurans* [516], and T30S is the small ribosomal subunit from *Thermus thermophilus* [1135], showing the head, platform, shoulder and latch (H,P,S,L, respectively). The cyan dots indicate the approximate mRNA channel; A, P, and E are the approximate positions of the anti-codon loops (on T30S) and the edges of the tRNAs acceptor stems (on D50S) of the three tRNA substrates: aminoacylated-tRNA (A), peptidyl-tRNA (P), and Exit tRNA (E). Image was kindly provided by Ada Yonath.

1.1.3 Need For Critical Assessment

The field of biomolecular modeling is relatively young, having started in the 1960s, and only gained momentum since the mid 1980s with the advent of supercomputers. Yet the field is developing with astonishing speed. Advances are driven by improvements in instrumentational resolution and genomic and structural databases, as well as in force fields, algorithms for conformational sampling and molecular dynamics, computer graphics, and the increased computer power and memory capabilities. These impressive technological and modeling advances are steadily establishing the field of theoretical modeling as a partner to experiment and a widely used tool for research and development.

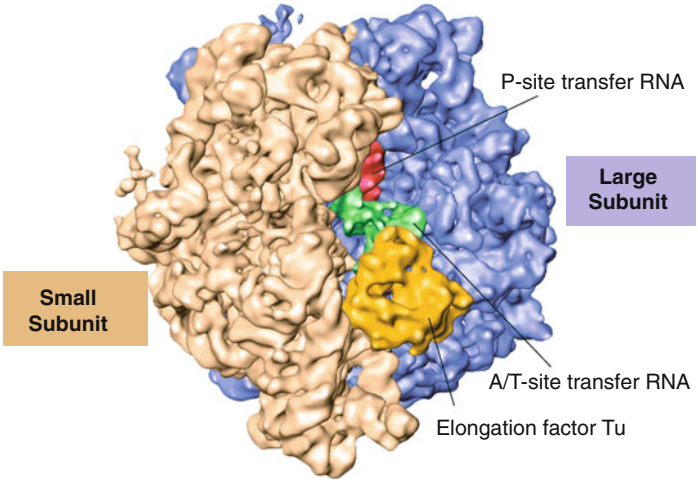

Figure 1.2. Cryo-EM view of of the 70S ribosome particle solved by J. Frank's group at 6.7 Å resolution in a complex with the EF-Tu-aa-tRNA ternary complex, GDP, and the antibiotic kirromycin [710]. Images were kindly provided by Michael Watters and Joachim Frank.

Yet as we witness the tantalizing progress, a cautionary usage of molecular modeling tools as well as a critical perspective of the field's strengths and limitations are warranted. This is because the current generation of users and application scientists in the industrial and academic sectors may not be familiar with some of the caveats and inherent approximations in biomolecular modeling and simulation approaches that the field pioneers clearly recognized. Indeed, the tools and programs developed by a handful of researchers several decades ago have now resulted in extensive profit-making software for genomic information, drug design, and every aspect of modeling. More than ever, a comprehensive background in the methodology framework is necessary for sound studies in the exciting era of computational biophysics that lies on the horizon.

1.1.4 Text Overview

This text aims to provide this critical perspective for field assessment while introducing the relevant techniques. Specifically, following an overview of biomolecular structure and modeling with a historical perspective and a description of current applications in this chapter and the next chapter, the elementary background for biomolecular modeling will be introduced in the chapters to follow: protein and nucleic-acid structure tutorials (Chapters 3–7), overview of theoretical approaches (Chapter 8), details of force field construction and evaluation (Chapters 9 and 10), energy minimization techniques (Chapter 11), Monte Carlo simulations (Chapter 12), molecular dynamics and related methods (Chapters 13 and 14), and similarity/diversity problems in chemical design (Chapter 15).

As emphasized in this book's Preface, given the enormously broad range of these topics, depth is often sacrificed at the expense of breadth. Thus, many specialized texts (e.g., in Monte Carlo, molecular dynamics, or statistical mechanics) are complementary, such as those listed in Appendix C; the representative articles used for the course (Appendix B) are important components. For introductory texts to biomolecular structure, biochemistry, and biophysical chemistry, see those listed in Appendix C, such as [163, 197, 275, 394, 1235]. For molecular simulations, a solid grounding in classical statistical mechanics, thermodynamic ensembles, time-correlation functions, and basic simulation protocols is important. Good introductory texts for these subjects, including biomolecular applications are [22, 165, 178, 428, 474, 494, 846, 853, 1038, 1067].

The remainder of this chapter and the next chapter provide a historical context for the field's development. Overall, this chapter focuses on a historical account of the field and the experimental progress that made biomolecular modeling possible. Chapter 2 introduces some of the field's challenges as well as practical applications of their solution.

Specifically, to appreciate the evolution of biomolecular modeling and simulation, we begin in the next section with an account of the milieu of growing experimental and technical developments. Following an introduction to the birth of molecular mechanics (Section 1.2), experimental progress in protein and nucleic-acid structure is described (Section 1.3). A selective reference chronology to structural biology is shown in Table 1.1.

The experimental section of this chapter discusses separately the early days of biomolecular instrumentation — as structures were emerging from X-ray crystallography — and the modern era of technological developments — stimulating the many sequencing projects and the rapid advances in biomolecular NMR and crystallography. Within this presentation, separate subsections are devoted to the techniques of X-ray crystallography and NMR and to the genome projects.

Chapter 2 continues this perspective by describing the computational challenges that naturally emerge from the overwhelming progress in genome projects and experimental techniques, namely deducing structure and function from sequence. Problems are exemplified by protein folding and misfolding. (Students unfamiliar with basic protein structure are urged to re-read Chapter 2 after the

protein minitutorial chapters). The sections that follow mention some of the exciting and important biomedical, industrial, and technological applications that lend enormous practical utility to the field. These applications represent a tangible outcome of the confluential experimental, theoretical, and technological advances.

Table 1.1. Structural Biology Chronology.

1865	Genes discovered by Mendel
1910	Genes in chromosomes shown by Morgan's fruitfly mutations
1920s	Quantum mechanics theory develops
1926	Early reports of crystallized proteins
1930s	Reports of crystallized proteins continue and stimulate Pauling & Corey to compile bond lengths and angles of amino acids
1944	Avery proves genetic transformation via DNA (not protein)
1946	Molecular mechanics calculations reported (Westheimer, others)
1949	Sickle cell anemia identified as 'molecular disease' (Pauling)
1950	Chargaff determines near-unity A:T and G:C ratios in many species
1951	Pauling & Corey predict protein α-helices and β-sheets
1952	Hershey & Chase reinforce genetic role of DNA (phage experiments)
1952	Wilkins & Franklin deduce that DNA is a helix (X-ray fiber diffraction)
1953	**Watson & Crick report the structure of the DNA double helix**
1959	Myoglobin & hemoglobin deciphered by X-ray (Kendrew & Perutz)
1960s	Systematic force fields developed (Allinger, Lifson, Scheraga, others)
1960s	Genetic code deduced (Crick, Brenner, Nirenberg, Khorana, Holley, coworkers)
1969	Levinthal paradox on protein folding posed
1970s	Biomolecular dynamics simulations develop (Stillinger, Karplus, others)
1970s	Site-directed mutagenesis techniques developed by M. Smith; restriction enzymes discovered by Arber, Nathans, and H. Smith
1971	Protein Data Bank established
1974	t-RNA structure reported
1975	First simulation of protein folding
1975	Fifty solved biomolecular structures available in the PDB
1977	DNA genome of the virus ϕX174 (5.4 kb) sequenced; soon followed by human mitochondrial DNA (16.6 kb) and λ phage (48.5 kb)
1980s	Dazzling progress realized in automated sequencing, protein X-ray crystallography, NMR, recombinant DNA, and macromolecular synthesis
1985	PCR devised by Mullis; numerous applications follow
1985	NSF establishes five national supercomputer centers
1990	International Human Genome Project starts; spurs others
1994	RNA hammerhead ribozyme structure reported; other RNAs follow
1995	First non-viral genome completed (bacterium *H. influenzae*), 1.8 Mb
1996	Yeast genome (*Saccharomyces cerevisiae*) completed, 13 Mb
1997	Chromatin core particle structure reported; confirms earlier structure
1998	Roundworm genome (*C. elegans*) completed, 100 Mb
1998	Crystal structure of ion channel protein reported
1998	Private Human Genome initiative competes with international effort
1999	Fruitfly genome (*Drosophila melanogaster*) completed (Celera), 137 Mb

Table 1.1 (continued)

1999	Human chromosome 22 sequenced (public consortium)
1999	IBM announces petaflop computer to fold proteins by 2005
2000	**First draft of human genome sequence announced**, 3300 Mb
2000	Moderate-resolution structures of ribosomes reported
2000	ENCODE project consortium established to characterize the human genome
2001	First annotation of the human genome (February)
2002	First draft of rice genome sequence, 430 Mb (April)
2003	Human genome sequence completed (April)
2006	All human chromosomes sequenced
2007	Craig Venter's DNA solved and analyzed
2008	James Watson's DNA solved and analyzed in a triumph of sequencing technology
Ongoing	Many projects continue to interpret findings of the HGP, including ENCODE, 1000 genomes, HapMap, Cancer Atlas, Personal Genome Project, and the sequencing of many organisms to allow comparative studies

1.2 The Roots of Molecular Modeling in Molecular Mechanics

The roots of molecular modeling began with the notion that molecular geometry, energy, and various molecular properties can be calculated from mechanical-like models subject to basic physical forces. A molecule is represented as a mechanical system in which the *particles* — atoms — are connected by *springs* — the bonds. The molecule then rotates, vibrates, and translates to assume favored conformations in space as a collective response to the inter- and intramolecular forces acting upon it.

The forces are expressed as a sum of harmonic-like (from Hooke's law) terms for **bond-length** and **bond-angle** deviations from reference equilibrium values; trigonometric **torsional terms** to account for *internal rotation* (rotation of molecular subgroups about the bond connecting them); and **nonbonded van der Waals and electrostatic potentials**. See Chapter 9 for a detailed discussion of these terms, as well as of more intricate cross terms.

1.2.1 The Theoretical Pioneers

Molecular mechanics arose naturally from the concepts of molecular bonding and van der Waals forces. The Born-Oppenheimer approximation assuming fixed nuclei (see Chapter 8) followed in the footsteps of quantum theory developed in the 1920s. While the basic idea can be traced to 1930, the first attempts of molecular

Table 1.2. The evolution of molecular mechanics and dynamics.

Period	System and Size[a]	Trajectory Length[b] [ns]	CPU Time/Computer[c]
1973	Dinucleoside (GpC) in vacuum (8 flexible dihedral angles)	—	—
1977	BPTI, vacuum (58 residues, 885 atoms)	0.01	
1983	DNA, vacuum, 12/24 bp (754/1530 atoms)	0.09	several weeks each, Vax 780
1984	GnRH, vacuum (decapeptide, 161 atoms)	0.15	
1985	Myoglobin, vacuum (1423 atoms)	0.30	50 days, VAX 11/780
1985	DNA, 5 bp (2800 atoms)	0.50	20 hrs, Cray X-MP
1989	Phospholipid Micelle ($\approx$ 7,000 atoms)	0.10	
1992	HIV protease (25,000 atoms)	0.10	100 hrs.,Cray Y-MP
1997	Estrogen/DNA (36,000 atoms, multipoles)	0.10	22 days, HP-735 (8)
1998	DNA, 24 bp (21,000 atoms, PME)	0.50	1 year, SGI Challenge
1998	β-heptapeptide in methanol ($\approx$ 5000/9000 atoms)	200	8 months, SGI Challenge (3)
1998	Villin headpiece (36 residues, 12,000 atoms, cutoffs)	1000	4 months, 256-proc. Cray T3D/E
1999	bc_1 complex in phospholipid bilayer (91,061 atoms, cutoffs)	1	75 days, 64 450-MHz-proc. Cray T3E
2001	C-terminal β-hairpin of protein-G (177 atoms, implicit solvent)	38000[b]	$\sim$ 8 days, 5000 proc. Folding@home megacluster
2002	Channel protein in lipid membrane (106,189 atoms, PME)	5	30 hrs, 500 proc. LeMieux terascale system; 50 days, 32 proc. Linux (Athlon)
2006	Complete satellite tobacco mosaic virus (1 million atoms)	50	55 days ($\approx$1ns/day), 256 Altix nodes, NCSA Athlon 2600+, NAMD program
2007	B-DNA dodecamer in solvent, PME, AMBER parm98 (15,774 atoms)	1200	130 days, 32 PowerPC BladeCenter proc., MareNostrum Supercomputer, Barcelona
2007	Villin headpiece (9,684 atoms) AMBER-2003	1000	6 months, Folding@home X86 megacluster, GROMACS/MPI
2008	Ubiquitin protein, explicit solvent OPLS-AA/SPC forcefield, (19,471 atoms)	1200	14 days (87ns/day), 32 processors Operon cluster, Desmond program
2008	Fip35 protein, explicit solvent NAMD/CHARMM	10000	14 weeks, NCSA Abe cluster, NAMD program
2009	β_2AR protein mutants (50,000-99,000 atoms) CHARMM27 forcefield	2000	28 days, 32 (2.66 GHz) E5430 processors Desmond program

[a]The examples for each period are representative. The first five systems are modeled in vacuum and the others in solvent. Except for the dinucleoside, simulations refer to molecular dynamics (MD). The two system sizes for the β-heptapeptide [285] reflect two (temperature-dependent) simulations. See text for definitions of abbreviations and further entry information.

[b]The 38 μs β-hairpin simulation in 2001 represents an ensemble (or aggregate) dynamics simulation, as accumulated over several short runs, rather than one long simulation [1428].

[c]The computational time is given where possible; estimates for the vacuum DNA, heptapeptide, β-hairpin, and channel protein simulations [285, 746, 1247, 1428] were kindly provided by M. Levitt, W. van Gunsteren, V. Pande, and K. Schulten, respectively.

mechanics calculations were recorded in 1946. Frank Westheimer's calculation of the relative racemization rates of biphenyl derivatives illustrated the success of such an approach. However, computers were not available at that time, so it took several more years for the field to gather momentum.

In the early 1960s, pioneering work on development of systematic force fields — based on spectroscopic information, heats of formation, structures of small compounds sharing the basic chemical groups, other experimental data, and quantum-mechanical information — began independently in the laboratories of the late Shneior Lifson at the Weizmann Institute of Science (Rehovot, Israel) [747], Harold Scheraga at Cornell University (Ithaca, New York), and Norman Allinger at Wayne State University (Detroit, Michigan) and then the University of Georgia (Athens). These researchers and their talented coworkers (notably Warshel and Levitt) began to develop force field parameters for families of chemical compounds by testing calculation results against experimental observations regarding structure and energetics. In 1969, following the pioneering Cartesian coordinate treatment described by Lifson and Warshel a year earlier [766], Levitt and Lifson reported the first energy calculation on entire protein molecules (myoglobin and lysozyme), in which molecular potentials and experimental constraints defined the target energy function minimized in Cartesian coordinates by the steepest descent method to refine low-resolution experimental coordinates [748]. Such formulations in Cartesian coordinates paved the way for all subsequent energy minimization and molecular dynamics calculations of biomolecules. In fact, Warshel's recognition in the mid 1960s that programming molecular force fields in Cartesian coordinates rather than internal coordinates [766] led to efficient evaluation of the functions along with analytic first and second derivatives and to program segments in many current macromolecular modeling programs [747].

In the early 1970s, Rahman and Stillinger reported the first molecular dynamics work of a polar molecule, liquid water [1034, 1035]; results offered insights into the structural and dynamic properties of this life sustaining molecule. Rahman and Stillinger built upon the simulation technique described much earlier (1959) by Alder and Wainwright but applied to hard spheres [19].

In the late 1970s, the idea of using molecular mechanics force fields with energy minimization as a tool for refinement of crystal structures was presented [598] and developed [670]. This led to the modern versions employing simulated annealing and related methods [248, 676].

It took a few more years, however, for the field to gain some 'legitimacy'.[3] In fact, these pioneers did not receive much general support at first, partly because their work could not easily be classified as a traditional discipline of chemistry (e.g., physical chemistry, organic chemistry). In particular, spectroscopists criticized the notion of transferability of the force constants, though at the same time

[3]Personal experiences shared by Norman L. Allinger on those early days of the field form the basis for the comments in this paragraph. I am grateful for him sharing these experiences with me.

they were quite curious about the predictions that molecular mechanics could make. In time, it indeed became evident that force constants are not generally transferable; still, the molecular mechanics approach was sound since nonbonded interactions are included, terms that spectroscopists omitted.

Ten to fifteen more years followed until the first generation of biomolecular force fields was established. The revitalized idea of molecular dynamics in the late 1970s propagated by Martin Karplus and colleagues at Harvard University sparked a flame of excitement that continues with full force today with the fuel of supercomputers. Most programs and force fields today, for both small and large molecules, are based on the works of the pioneers cited above (Allinger, Lifson, and Scheraga) and their coworkers. The water force fields developed in the late 1970s and early 1980s by Berendsen and coworkers (e.g., [1079]) and by Jorgensen and coworkers [617] (SPC and TIP3P/TIP4P, respectively) laid the groundwork for biomolecular simulations in solution. Important concepts in protein electrostatics and enzyme/substrate complexes in solution laid by Warshel and colleagues [1344, 1346] paved the way to quantitative modeling of enzymatic reactions and hybrid quantum/molecular mechanics methods [1342].

Peter Kollman's legacy is the development and application of force field methodology and computer simulation to important biomolecular, as well as medicinal, problems such as enzyme catalysis and protein/ligand design [1335]; his group's free energy methods and combined quantum/molecular mechanics approaches have opened many new doors of applications. With Kollman's untimely death in May 2001, the community mourned the loss of a great leader and innovator.

Modern versions of these and other molecular simulation packages have led to competition for "better, bigger, and faster" program design. For example, free software at the University of Illinois at Urbana-Champaign by Klaus Schulten and coworkers called NAMD for nanoscale MD (see the NAMD homepage) can be run on hundreds of parallel microprocessors with various force fields [996]. David Shaw group's program Desmond is fast even on a small number of processors [156]. GROMACS [124, 770] is also widely used for long MD simulations (e.g., [364]). And Anton, a specialized computer hard-wired for long MD simulations, has been launched [1169]. With these excellent teams of software and hardware engineers and biomolecular scientists, the average MD user can look forward to improved and faster applications.

1.2.2 Biomolecular Simulation Perspective

Table 1.2 and Figures 1.3 and 1.4 provide a perspective of biomolecular simulations. Specifically, the selected examples illustrate the growth in time of system complexity (size and model resolution) and simulation length. The three-dimensional (3D) rendering in Figure 1.3 shows 'blocks' with heights

proportional to system size. Figure 1.4 offers molecular views of the simulation subjects and extrapolations for long-time simulations of proteins and cells based on [339].

Representative Progress

Starting from the first entry in the table, **dinucleoside GpC** (guanosine-3′, 5′-cytidine monophosphate) posed a challenge in the early 1970s for finding all minima by potential energy calculations and model building [1222]. Still, clever search strategies and constraints found a correct conformation (dihedral angles in the range of helical RNA and sugar in C3′-endo form) as the lowest energy minimum. *Global optimization remains a difficult problem!* (See Chapter 11).

Following the first MD simulation of a biological process of duration 100 fs [1344], the small protein **BPTI** (Bovine Pancreatic Trypsin Inhibitor) was simulated **1977** [845], showing substantial atomic fluctuations on the picosecond timescale.

The 12 and 24-base-pair (bp) **DNA** simulations in **1983** [746] were performed in vacuum without electrostatics, and that of the DNA pentamer system in 1985, with 830 water molecules and 8 sodium ions and full electrostatics [1158]. Stability problems for nucleic acids emerged in the early days — unfortunately, in some cases the strands untwisted and separated [746]. Stability became possible with the introduction of scaled phosphate charges in other pioneering nucleic-acid simulations [523, 1013, 1260] and the introduction a decade later of more advanced treatments for solvation and electrostatics; see, for example, [220], for a discussion.

The linear **decapeptide** GnRH (gonadotropin-releasing hormone) was studied in **1984** for its pharmaceutical potential, as it triggers LH and FSH hormones [1234].

The 300 ps dynamics simulation of the protein **myoglobin** in **1985** [752] was considered three times longer than the longest previous MD simulation of a protein. The results indicated a slow convergence of many thermodynamic properties.

The large-scale **phospholipid** aggregate simulations in **1989** [1361] was an ambitious undertaking: it incorporated a hydrated micelle (i.e., a spherical aggregate of phospholipid molecules) containing 85 LPE molecules (lysophosphatiadylethanolamine) and 1591 water molecules.

The **HIV protease** system simulated in solution in **1992** [518] captured an interesting flap motion at the active site. See also Figure 2.5 and a discussion of this motion in the context of protease inhibitor design.

The **1997 estrogen/DNA simulation** [679] sought to understand the mechanism underlying DNA sequence recognition by the protein. It used the multipole electrostatic treatment, crucial for simulation stability, and also parallel processing for speedup [1133].

The **1998 DNA** simulation [1417] used the alternative, Particle Mesh Ewald (PME) treatment for consideration of long-range electrostatics (see Chapter 10) and uncovered interesting properties of A-tract sequences.

The **1998 peptide** simulation in methanol used periodic boundary conditions (defined in Chapter 10) and captured reversible, temperature-dependent folding [285]; the 200 ns time reflects four 50 ns simulations at various temperatures.

The **1998** 1 μs **villin-headpiece** simulation (using periodic boundary conditions) [338] was considered longer by three orders of magnitude than prior simulations. A folded structure close to the native state was approached; see also [340].

The solvated protein $\boldsymbol{bc}_1$ **embedded in a phospholipid bilayer** [597] was simulated in **1999** for over 1 ns by a 'steered molecular dynamics' algorithm (45,131 flexible atoms) to suggest a pathway for proton conduction through a water channel. As in villin, the Coulomb forces were truncated.

In **2002**, an aquaporin membrane channel protein in the glycerol conducting subclass (*E. coli* **glycerol channel, GlpF**) in a lipid membrane (106,189 total atoms) was simulated for 5 ns (as well as a mutant) with all nonbonded interactions considered, using the PME approach [1247]. The simulations suggested details of a selective mechanism by which water transport is controlled; see also [606] for simulations examining the glycerol transport mechanism.

By early 2002, the longest simulation published of 38μs reflected aggregate (or ensemble) dynamics — usage of many short trajectories to simulate the microsecond timescale — for the C-terminal β-**hairpin from protein G** (16 residues) in **2001** [1428]. Whereas the continuous 1 μs villin simulation required months of dedicated supercomputing, the β-hairpin simulation (177 atoms, using implicit solvation and Langevin dynamics) was performed to analyze folding kinetics on a new distributed computing paradigm which employs personal computers from around the world (see Folding@home at folding.stanford.edu and [1177]). About 5000 processors were employed and, with the effective production rate of 1 day per nanosecond per processor, about 8 days were required to simulate the 38 μs aggregate time. See also [1205] for a later set of simulations (reviewed in [177]) and a large-scale molecular dynamics study of a variant of the villin headpiece that consisted of hundreds of 1μs simulations [364].

Several years later, longer and larger simulations, though not yet routine, are clearly possible using specialized programs that exploit high-speed multiple-processor systems, like NAMD, GROMACS, and/or specialized computing resources like Anton [1169].

While trends continue to simulate larger biomolecular systems (e.g., entire **satellite mosaic virus** in 2006 with one million atoms) [425] and longer time frames (e.g., **B-DNA dodecamer** [987], **ubiquitin** protein [827], **Fip35 protein** [424], and β_2**AR protein receptor** [337] for over one microsecond, and small proteins for milliseconds [1462]) with specialized MD programs and dedicated supercomputers, coarse-grained models and combinations of enhanced sampling methods are emerging as the way to go for simulating complex biomolecular systems (recently reviewed in [655, 729, 1116, 1117]). This is because computer power alone is not likely to solve the folding problem in general. For example, the 10 μs simulation of Fip35 [424] did not provide the anticipated folded conformation nor the folding trajectory from the extended state, as expected from

experimental measurements; this long simulation also raised force field and algorithmic stability questions, which were explored later [426]. Still, for other proteins, folding simulations can be very successful (e.g., [337, 427, 637]).

Trends

Note from the table and figure the transition from simulations in vacuum (first five entries) to simulations in solvent (remaining items). Observe also the steady increase in simulated system size, with a leap increase in simulation lengths made more recently.

Large system sizes or long simulation times can generally be achieved by sacrificing other simulation aspects. For example, truncating long-range electrostatic interactions makes possible the study of large systems over short times [597], or small systems over long times [338]. Using implicit solvent and cutoffs for electrostatic interactions also allows the simulation of relatively small systems over long times [1428]. And simplified, coarse-grained models with effective potentials also allow simulations over longer time frames, with the correct physical behavior. (These topics are discussed in later chapters). In fact, with the increased awareness of the sampling problem in dynamic simulation (see Chapter 13), long single simulations are often replaced by several trajectories, leading to overall better sampling statistics, and coarse graining is being applied to biological systems and problems of greater complexity [655].

Duan *et al.* make an interesting 'fanciful' projection on the computational capabilities of modeling in the coming decades [339]: they suggest the feasibility, in 20 years, of simulating a second in the life-time of medium-sized proteins and, in 50–60 years, of following the entire life cycle of an *E. Coli* cell (1000 seconds or 20 minutes, for 30 billion atoms). This estimate was extrapolated on the basis of two data points — the 1977 BPTI simulation [845] and the 1998 villin simulation [338, 340] discussed above — and relied on the assumption that computational power increases by a factor of 10 every 3–4 years (Even better progress was actually realized). These projections are displayed by entries for the years 2020 and 2055 in Figure 1.4.

1.3 Emergence of Biomodeling from Experimental Progress in Proteins and Nucleic Acids

At the same time that molecular mechanics developed, tremendous progress on the experimental front also began to trigger further interest in the theoretical approach to structure determination.

1.3.1 Protein Crystallography

The first records of crystallized polypeptides or proteins date back to the late 1920s / early 1930s (1926: urease, James Sumner; 1934: pepsin, J. D. Bernal and

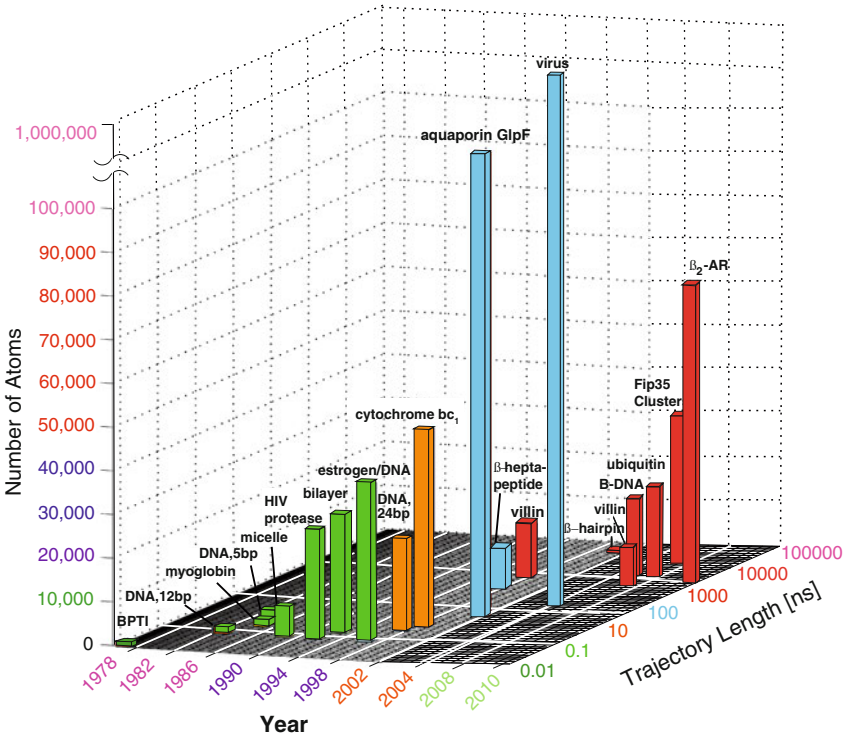

Figure 1.3. The evolution of molecular dynamics simulations with respect to system sizes and simulation lengths (see also Table 1.2).

Dorothy Crowfoot-Hodgkin; 1935: insulin, Crowfoot-Hodgkin). However, only in the late 1950s did John Kendrew (Perutz' first doctoral student) and Max Perutz succeed in deciphering the X-ray diffraction pattern from the crystal structure of the protein (1958: myoglobin, Kendrew; 1959: hemoglobin, Perutz). This was possible by Perutz' crucial demonstration (around 1954) that structures of proteins can be solved by comparing the X-ray diffraction patterns of a crystal of a native protein to those associated with the protein bound to heavy atoms like mercury (i.e., by 'isomorphous replacement'). The era of modern structural biology began with this landmark development.

As glimpses of the first X-ray crystal structures of proteins came into view, Linus Pauling and Robert Corey began in the mid 1930s to catalogue bond lengths and angles in amino acids. By the early 1950s, they had predicted the two basic structures of amino acid polymers on the basis of hydrogen bonding patterns: α helices and β sheets [974, 976]. As of 1960, about 75 proteins had been crystallized, and immense interest began on relating the sequence content to catalytic activity of these enzymes.

By then, the exciting new field of molecular biology was well underway. Perutz, who founded the Medical Research Council Unit for Molecular Biology at the

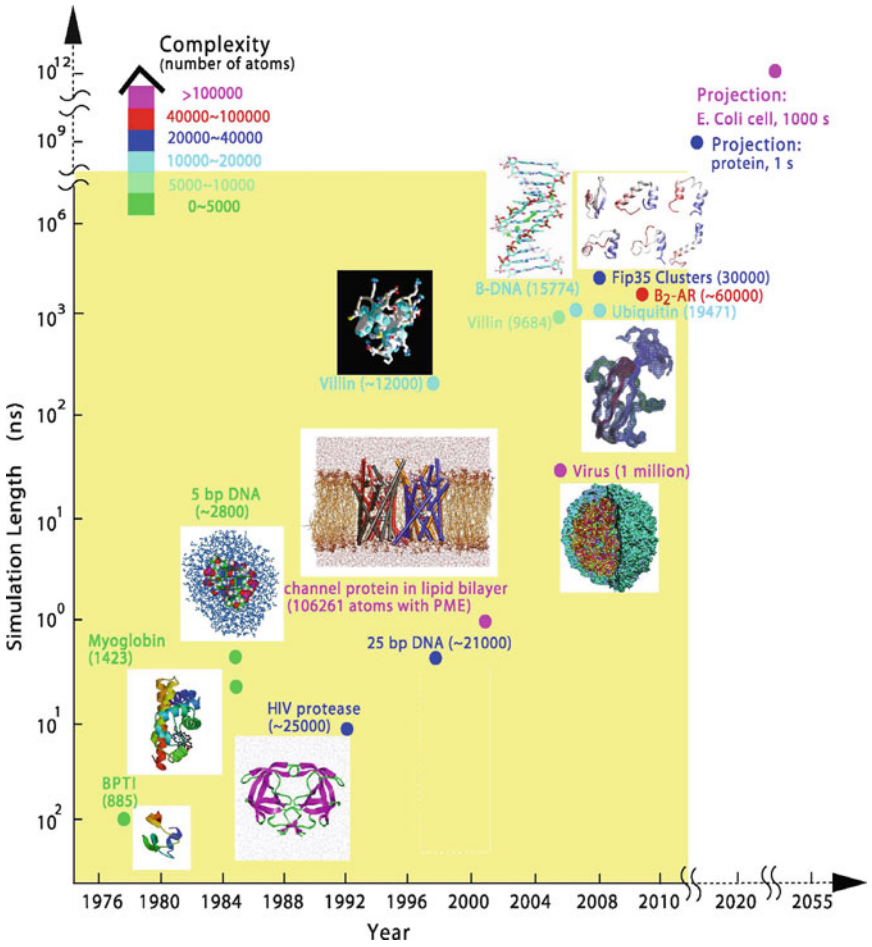

Figure 1.4. The evolution of molecular dynamics simulations with respect to simulation lengths (see also Table 1.2 and Figure 1.3). The data points for 2020 and 2055 represent extrapolations from the 1977 BPTI [845] and 1998 villin [338, 340] simulations, assuming a computational power increase by a factor of 10 every 3–4 years, as reported in [339].

Cavendish Laboratory in Cambridge in 1947, also created the Laboratory of Molecular Biology there in 1962. Perutz and Kendrew received the Nobel Prize in Chemistry for their accomplishments in 1962.[4]

[4]See the formidable electronic museum of science and technology, with related lectures and books that emerged from Nobel-awarded research, on the website of the Nobel Foundation (nobelprize.org). This virtual museum was recently constructed to mark the 100th anniversary in 2001 of Alfred B. Nobel's legacy. See also a marvelous account in [1258] of Perutz (who died at the age of 87 in 2002) as scientist and person, including his relationship with Kendrew.

1.3.2 DNA Structure

Momentum at that time came in large part from parallel experimental work that began in 1944 in the nucleic acid world and presaged the discovery of the DNA double helix.

Inspired by the 1928 work of the British medical officer Fred Griffith, Oswald Avery and coworkers Colin MacLeod and Maclyn McCarty studied pneumonia infections. Griffith's intriguing experiments showed that mice became fatally ill upon infection from a live but harmless (coatless) strain of pneumonia-causing bacteria mixed with the DNA from heat-killed pathogenic bacteria; thus, the DNA from heat-killed pathogenic bacteria transformed live harmless into live pathogenic bacteria. Avery and coworkers mixed DNA from virulent strains of pneumococci with harmless strains and used enzymes that digest DNA but not proteins. Their results led to the cautious announcement that the 'transforming agent' of traits is made exclusively of DNA.[5]

Their finding was held with skepticism until the breakthrough, Nobel prize-winning phage experiments of Alfred Hershey and Martha Chase eight years later, which demonstrated that only the nucleic acid of the phage entered the bacterium upon infection, whereas the phage protein remained outside.[6]

Much credit for the transforming agent evidence is due to the German theoretical physicist and Nobel laureate Max Delbrück, who brilliantly suggested to use bacterial viruses as the model system for the genome demonstration principle. Delbrück shared the Nobel Prize in Physiology or Medicine in 1969 with Hershey and Salvador Luria for their pioneering work that established bacteriophage as the premier model system for molecular genetics.

In 1950, Erwin Chargaff demonstrated that the ratios of adenine-to-thymine and guanine-to-cytosine bases are close to unity, with the relative amount of each kind of pair depending on the DNA source.[7] These crucial data, together with the X-ray fiber diffraction photographs of hydrated DNA taken by Rosalind Franklin[8] and Raymond Gosling (both affiliated with Maurice Wilkins who was engaged in related research [622]), led directly to Watson and Crick's ingenious proposal of the structure of DNA in 1953. The photographs were crucial as they suggested a helical arrangement.

---

[5] Interested readers can visit the virtual gallery of Profiles in Science at www.profiles.nlm.nih.gov/ for a profile on Avery.

[6] A wonderful introduction to the rather recluse Hershey, who died at the age of 88 in 1997, can be enjoyed in a volume edited by Franklin W. Stahl titled *We can sleep later: Alfred D. Hershey and the origins of molecular biology* (Cold Spring Harbor Press, New York, 2000). The title quotes Hershey from his letter to contributors of a volume on *bacteriophage* λ which he edited in 1971, urging them to complete and submit their manuscripts!

[7] Chargaff died in June 2002 at the age of 96. Sadly, he was a sardonic man who did not easily fit into the sharply focused world of most scientists; he further isolated himself when he denounced the molecular biology community in the late 1950s.

[8] See an outstanding study on "the dark lady of DNA" in a recent biography [810] and [358].

Although connecting these puzzle pieces may seem straightforward to us now that the DNA double helix is a household word, these two ambitious young Cambridge scientists deduced from the fiber diffraction data and other evidence that the observed base-pairing specificity, together with steric restrictions, can be reconciled in an *anti-parallel* double-helical form with a sugar-phosphate backbone and nitrogenous-bases interior. Their model also required a key piece of information from the organic chemist Jerry Donahue regarding the *tautomeric* states of the bases.[9] Though many other DNA forms besides the classic Crick and Watson (B-DNA) form are now recognized, including triplexes and quadruplexes, the B-form is still the most prevalent under physiological conditions. Indeed, the 50th anniversary in April 2003 of Watson and Crick's seminal paper was celebrated with much fanfare throughout the world.

RNA crystallography is at an earlier stage, but has recently made quantum leaps with the solution of several ribosomes (see Fig. 1.1), other significant RNA molecules, and newly-identified novel roles for RNA, including, most prominently, non-coding RNAs, aptamers and riboswitches, with many potential benefits to biomedicine and nanotechnology (see Chapter 7). These developments followed the exciting discoveries in the 1980s that established that RNA, like protein, can act as an enzyme in living cells, and discoveries in recent years — that RNA has numerous regulatory roles in biological processes — which have transformed our understanding of RNA's functional repertoire. Sidney Altman and Thomas Cech received the 1989 Nobel Prize in Chemistry for their discovery of RNA biocatalysts, *ribozymes*, and twice in 2006 were RNA discoveries recognized by Nobel Prizes, including for uncovering gene silencing, which can be exploited to selectively knock out protein functions for disease analysis. RNA made headlines again in 2009 when the Nobel Prize in Chemistry was aptly awarded to three scientists who independently uncovered the atomic-level detail of the magnificent RNA/protein machine that makes up the ribosome: Ada Yonath, Venkatraman Ramakrishnan, and Thomas Steitz. Their X-ray crystallographic views and functional analyses have also led to antibiotic design.

The next two subsections elaborate upon the key techniques for solving biomolecular structures: X-ray crystallography and NMR. We end this section on experimental progress with a description of modern technological advances and the genome sequencing projects they inspired.

1.3.3 The Technique of X-ray Crystallography

Much of the early crystallographic work was accomplished without computers and was inherently very slow. *Imagine calculating the Fourier series by hand!*

[9] Proton migrations within the bases can produce a *tautomer*. These alternative forms depend on the dielectric constant of the solvent and the pH of the environment. In the bases, the common *amino* group (–N–H_2) can tautomerize to an *imino* form (=N–H), and the common *keto* group (–C=O) can adopt the *enol* state (=C–O–H); the fraction of bases in the rare imino and enol tautomers is only about 0.01% under regular conditions.

Only in the 1950s were direct methods for the phase problem developed, with a dramatic increase in the speed of structure determination occurring about a decade later.

Structure determination by X-ray crystallography involves analysis of the X-ray diffraction pattern produced when a beam of X-rays is directed onto a well-ordered crystal. Crystals form by vapor diffusion from purified protein solutions under optimal conditions. See [163, 930] for overviews.

The diffraction pattern can be interpreted as a reflection of the primary beam source from sets of parallel planes in the crystal. The diffracted spots are recorded on a detector (electronic device or X-ray film), scanned by a computer, and analyzed on the basis of Bragg's law[10] to determine the unit cell parameters.

Each such recorded diffraction spot has an associated amplitude, wavelength, and phase; all three properties must be known to deduce atomic positions. Since the phase is lost in the X-ray experiments, it must be computed from the other data. This central obstacle in crystal structure analysis is called the *phase problem* (see Box 1.1). Together, the amplitudes and phases of the diffraction data are used to calculate the electron density map; the greater the resolution of the diffraction data, the higher the resolution of this map and hence the atomic detail derived from it.

Both the laborious crystallization process [852] and the necessary mathematical analysis of the diffraction data limit the amount of accurate biomolecular data available. Well-ordered crystals of biological macromolecules are difficult to grow, in part because of the disorder and mobility in certain regions. Crystallization experiments must therefore screen and optimize various parameters that influence crystal formation, such as temperature, pH, solvent type, and added ions or ligands.

The phase problem was solved by direct methods for small molecules (roughly ≤ 100 atoms) by Jerome Karle and Herbert Hauptman in the late 1940s and early 1950s; they were recognized for this feat with the 1985 Nobel Prize in Chemistry. For larger molecules, biomolecular crystallographers have relied on the method pioneered by Perutz, Kendrew and their coworkers termed *multiple isomorphous replacement* (MIR).

MIR introduces new X-ray scatters from complexes of the biomolecule with heavy elements such as selenium or heavy metals like osmium, mercury, or uranium. The combination of diffraction patterns for the biomolecule, heavy elements or elements or metals, and biomolecule/heavy-metal complex offers more information for estimating the desired phases. The differences in diffracted

[10]The Braggs (father William-Henry and son Sir William-Lawrence) observed that if two waves of electromagnetic radiation arrive at the same point in phase and produce a maximal intensity, the difference between the distances they traveled is an integral multiple of their wavelengths. From this they derived what is now known as *Bragg's law*, specifying the conditions for diffraction and the relation among three key quantities: d (distance between parallel planes in the crystal), λ (the wavelength of the X-ray beam), and θ (the reflection angle). Bragg's condition requires that the difference in distance traveled by the X-rays reflected from adjacent planes is equal to the wavelength λ. The associated relationship is $\lambda = 2d \sin \theta$.

intensities between the native and derivative crystals are used to pinpoint the heavy atoms, whose waves serve as references in the phase determination for the native system.

To date, advances in the experimental, technological, and theoretical fronts have dramatically improved the ease of crystal preparation and the quality of the obtained three-dimensional (3D) biomolecular models [163, last chapter]. Techniques besides MIR to facilitate the phase determination process — by analyzing patterns of heavy-metal derivatives using *multi-wavelength anomalous diffraction* (MAD) or by molecular replacement (deriving the phase of the target crystal on the basis of a solved related molecular system) [540, 541] have been developed. Very strong X-ray sources from synchrotron radiation (e.g., with light intensity that can be 10,000 times greater than conventional beams generated in a laboratory) have become available. New techniques have made it possible to visualize short-lived intermediates in enzyme-catalyzed reactions at atomic resolution by time-resolved crystallography [433, 870, 994]. And improved methods for model refinement and phase determination are continuously being reported [1287]. Such advances are leading to highly refined biomolecular structures[11] (resolution ≤ 2 Å) at much greater numbers [110], even for nucleic acids [894].

Box 1.1: The Phase Problem

The mathematical phase problem in crystallography [531,634] involves resolving the phase angles $\phi_{\mathbf{h}}$ associated with the structure factors $F_{\mathbf{h}}$ when only the intensities (squares of the amplitudes) of the scattered X-ray pattern, $I_{\mathbf{h}} = |F_{\mathbf{h}}|$, are known. The structure factors $F_{\mathbf{h}}$, defined as

$$F_{\mathbf{h}} = |F_{\mathbf{h}}| \ \exp(i\phi_{\mathbf{h}}) \,, \tag{1.1}$$

describe the scattering pattern of the crystal in the Fourier series of the electron density distribution:

$$\rho(\mathbf{r}) = \frac{1}{V} \sum_{\mathbf{h}} F_{\mathbf{h}} \ \exp(-2\pi i \mathbf{h} \cdot \mathbf{r}). \tag{1.2}$$

Here $\mathbf{r}$ denotes position, $\mathbf{h}$ identifies the three defining planes of the unit cell (e.g., h, k, l), V is the cell volume, and $\cdot$ denotes a vector product. See [1047], for example, for details.

1.3.4 The Technique of NMR Spectroscopy

The introduction of NMR as a technique for protein structure determination came much later (early 1960s), but since 1984 both X-ray diffraction and NMR have been valuable tools for determining protein structure at atomic resolution. Kurt

[11] The resolution value is similar to the quantity associated with a microscope: objects (atoms) can be distinguished if they are separated by more than the resolution value. Hence, the lower the resolution value the more molecular architectural detail that can be discerned.

Wütrich was awarded the 2002 Nobel Prize in Chemistry[12] for his pioneering efforts in developing and applying NMR to biological macromolecules.

Nuclear magnetic resonance is a versatile technique for obtaining structural and dynamic information on molecules in solution. The resulting 3D views from NMR are not as detailed as those that can result from X-ray crystallography, but the NMR information is not static and incorporates effects due to thermal motions in solution.

In NMR, powerful magnetic fields and high-frequency radiation waves are applied to probe the magnetic environment of the nuclei. The local environment of the nucleus determines the frequency of the resonance absorption. The resulting NMR spectrum contains information on the interactions and localized motion of the molecules containing those resonant nuclei.

The absorption frequency of particular groups can be distinguished from one another when high-frequency NMR devices are used (*"high resolution NMR"*). Until recently, this requirement for nonoverlapping signals to produce a clear picture has limited the protein sizes that can be studied by NMR to systems with masses in the range of 50 to 100 kDa. However, dramatic increases (such as a tenfold increase) have been possible with novel strategies for isotopic labeling of proteins [1423] and detection of signals from disordered residues with fast internal motions by cross correlated relaxation-enhanced polarization transfer [397]. For example, the Horwich and Wütrich labs collaborated in 2002 to produce a high resolution solution NMR structure of the chaperonin/cochaperonin GroEL/GroES complex ($\sim$900 kDa) [397]. Advances in solid-state NMR techniques may be particularly valuable for structure analysis of membrane proteins.

As in X-ray crystallography, advanced computers are required to interpret the data systematically. NMR spectroscopy yields a wealth of information: a network of distances involving pairs of spatially-proximate hydrogen atoms. The distances are derived from Nuclear Overhauser Effects (NOEs) between neighboring hydrogen atoms in the biomolecule, that is, for atom pairs separated by less than 5–6 Å.

To calculate the 3D structure of the macromolecule, these NMR distances are used as conformational restraints in combination with various supplementary information: primary sequence, reference geometries for bond lengths and bond angles, chirality, steric constraints, spectra, and so on. A suitable energy function must be formulated and then minimized, or surveyed by various techniques, to find the coordinates that are most compatible with the experimental data. See [248] for an overview. Such deduced models are used to back calculate the spectra inferred from the distances, from which iterative improvements of the model are pursued to improve the matching of the spectra. Indeed, the difficulty of using

[12]The other half of the 2002 Chemistry prize was split between John B. Fenn and Koichi Tanaka who were recognized for their development of ionization methods for analysis of proteins using mass spectrometry.

NMR data for structure refinement in the early days can be attributed to this formidable refinement task, formally, an over-determined or under-determined global optimization problem.[13]

The pioneering efforts of deducing peptide and protein structures in solution by NMR techniques were reported between 1981 and 1986; they reflected year-long struggles in the laboratory. Only a decade later, with advances on the experimental, theoretical, and technological fronts, 3D structures of proteins in solution could be determined routinely for monomeric proteins with less than 200 amino acid residues. See [368, 1396] for texts by modern NMR pioneers, [487] for a historical perspective of biomolecular NMR, and [248, 249, 1184] for recent advances.

Today's clever methods have been designed to facilitate such refinements, from formulation of the target energy to conformational searching, the latter using tools from distance geometry, molecular dynamics, simulated annealing, and many hybrid search techniques [181,203,248,487]. The ensemble of structures obtained is not guaranteed to contain the "best" (global) one, but the solutions are generally satisfactory in terms of consistency with the data. The recent technique of residual dipolar coupling also has great potential for structure determination by NMR spectroscopy without the use of NOE data [1188, 1265].

1.4 Modern Era of Technological Advances

1.4.1 From Biochemistry to Biotechnology

The discovery of the elegant yet simple DNA double helix not only led to the birth of molecular biology; it led to the crucial link between biology and chemistry. Namely, the genetic code relating triplets of RNA (the template for protein synthesis) to the amino acid sequence was decoded ten years later, and biochemists began to isolate enzymes that control DNA metabolism.

One class of those enzymes, restriction endonucleases, became especially important for recombinant DNA technology. These molecules can be used to break huge DNA into small fragments for sequence analysis. Restriction enzymes can also cut and paste DNA (the latter with the aid of an enzyme, ligase) and thereby create spliced DNA of desired transferred properties, such as antibiotic-resistant bacteria that serve as informants for human insulin makers. The discovery of these enzymes was recognized by the 1978 Nobel Prize in Physiology or Medicine to Werner Arber, Daniel Nathans, and Hamilton O. Smith.

Very quickly, X-ray, NMR, recombinant DNA technology, and the synthesis of biological macromolecules improved. The 1970s and 1980s saw steady advances

[13] Solved NMR structures are usually presented as sets of structures since certain molecular segments can be over-determined while others under-determined. The better the agreement for particular atomic positions among the structures in the set, the more likely it is that a particular atom or component is well determined.

in our ability to produce, crystallize, image, and manipulate macromolecules. Site-directed mutagenesis developed in 1970s by Canadian biochemist Michael Smith (1993 Nobel laureate in Chemistry) has become a fundamental tool for protein synthesis and protein function analysis.

1.4.2 PCR and Beyond

The polymerase chain reaction (PCR) devised in 1985 by Kary Mullis (winner of the 1993 Nobel Prize in Chemistry, with Michael Smith) and coworkers [884] revolutionized biochemistry: small parent DNA sequences could be amplified exponentially in a very short time and used for many important investigations. DNA analysis has become a standard tool for a variety of practical applications. Noteworthy classic and current examples of PCR applications are collected in Box 1.2. See also [1044] for stories on how genetics teaches us about history, justice, diseases, and more.

Beyond amplification, PCR technology made possible isolation of gene fragments and their usage to clone whole genes; these genes could then be inserted into viruses or bacterial cells to direct the synthesis of biologically active products. With dazzling speed, the field of bioengineering was born. Automated sequencing efforts continued during the 1980s, leading to the start of the International Human Genome Project in 1990, which spearheaded many other sequencing projects (see next section).

Macromolecular X-ray crystallography and NMR techniques are also improving rapidly in this modern era of instrumentation, both in terms of obtained structure resolution and system sizes [874]. Stronger X-ray sources, higher-frequency NMR spectrometers, and refinement tools for both data models are leading to these steady advances. The combination of instrumentational advances in NMR spectroscopy and protein labeling schemes suggests that the size limit of protein NMR may soon reach 100 kDa [1404, 1423].

In addition to crystallography (see Fig. 1.1) and NMR, cryogenic electron microscopy (cryo-EM) contributes important macroscopic views at lower resolution for proteins that are not amenable to NMR or crystallography [418, 1252, 1286]. This technique involves imaging rapidly-frozen samples of randomly-oriented molecular complexes at low temperatures and reconstructing 3D views from the numerous planar EM projections of the structures. Adequate particle detection imposes a lower limit on the subject of several hundred kDa, but cryo-EM is especially good for large molecules with symmetry, as size and symmetry facilitate the puzzle gathering (3D image reconstruction) process. Though the resolution is low compared to crystallography and NMR, the resolution is becoming better and better with optimization of parameters and protocols used in the reconstruction process, as demonstrated for the 70S ribosome [710] shown in Figure 1.2.

Together with recombinant DNA technology, automated software for structure determination, and supercomputer and graphics resources, structure determination at a rate of one biomolecule per day (or more) is on the horizon.

Box 1.2: PCR Application Examples

- **Medical diagnoses of diseases and traits.** DNA analysis can be used to identify gene markers for many maladies, like cancer (e.g., BRCA1/2, *p53* mutations), schizophrenia, late Alzheimer's or Parkinson's disease. A classic story of cancer markers involves Vice President Hubert Humphrey, who was tested for bladder cancer in 1967 but died of the disease in 1978. In 1994, after the invention of PCR, his cancerous tissue from 1976 was compared to a urine sample from 1967, only to reveal the same mutations in the *p53* gene, a cancer suppressing gene, that escaped the earlier recognition. Sadly, if PCR technology had been available in 1967, Humphrey may have been saved.

- **Historical analysis.** DNA is now being used for genetic surveys in combination with archaeological data to identify markers in human populations.[14] Such analyses can discern ancestors of human origins, migration patterns, and other historical events [1070]. These analyses are not limited to humans; the evolutionary metamorphosis of whales has been unraveled by the study of fossil material combined with DNA analysis from living whales [1389].

  Historical analysis by French and American viticulturists also showed that the entire gene pool of 16 classic wines can be conserved by growing only two grape varieties: *Pinot noir* and *Gouais blanc*. Depending on your occupation, you may either be comforted or disturbed by this news . . . .

  PCR was also used to confirm that the fungus that caused the Irish famine (since potato crops were devastated) in 1845–1846 was caused by the fungus *P. infestans*, a water mold (infected leaves were collected during the famine) [1053]. Studies showed that the Irish famine was not caused by a single strain called US-1 which causes modern plant infections, as had been thought. Significantly, the studies taught researchers that further genetic analysis is needed to trace recent evolutionary history of the fungus spread.

- **Forensics and crime conviction.** DNA profiling — comparing distinctive DNA sequences, aberrations, or numbers of sequence repeats among individuals — is a powerful tool for proving with extremely high probability the presence of a person (or related object) at a crime, accident, or another type of scene. In fact, in the two decades since DNA evidence began to be used in court (1988), about 250 prisoners have been exonerated in the U.S., including from death row and one after 35 years behind bars, and many casualties from disasters (like airplane crashes and the 11 September 2001 New York World Trade Center terrorist attacks) were identified from DNA analysis of assembled body parts. In this connection, personal objects analyzed for DNA — like a black glove or blue dress — made headlines as crucial 'imaginary witnesses'[15] in the O.J. Simpson and Lewinsky/Clinton affairs. In fact,

[14] Time can be correlated with genetic markers through analysis of mitochondrial DNA or segments of the Y-chromosome. Both are genetic elements that escape the usual reshuffling of sexual reproduction; their changes thus reflect random mutations that can be correlated with time.

[15] George Johnson, "OJ's Blood and The Big Bang, Together at Last", *The New York Times*, Sunday, May 21, 1995.

a new breed of high-tech detectives is emerging with modern scientific tools, for example, by using bugs in crime research. See also the Anthrax attack case solved in 2008 described at the end of this chapter.

- **Family lineage / paternity identification.** DNA fingerprinting can also be used to match parents to offspring. In 1998, DNA from the grave confirmed that President Thomas Jefferson fathered at least one child by his slave mistress, Sally Hemmings, 200 years ago. The mystery concerning the remains of Tsar Nicholas II's family, executed in 1918, was solved after nine decades, by DNA analysis of bone shards found in a forest; the latest remnants analyzed in 2008 were determined to be those of his two children Alexksei and Maria, therefore completing the findings of the entire family. In April 2000, French historians with European scientists solved a 205-year-old mystery by analyzing the heart of Louis XVII, preserved in a crystal urn, confirming that the 10-year old boy died in prison after his parents Marie Antoinette and Louis XVI were executed, rather than spirited out of prison by supporters (Antoinette's hair sample is available). Similar post-mortem DNA analysis proved false a paternity claim against Yves Montand. In 2008, Egypt announced DNA analysis projects of 3500-year-old mummies, including fetuses of the mummies, to determine lineage connections to King Tutankhamun.

See also the book by Reilly [1044] for many other examples.

1.5 Genome Sequencing

1.5.1 Projects Overview: From Bugs to Baboons

Spurred by this dazzling technology, thousands of researchers worldwide have been, or are now, involved in hundreds of sequencing projects for species like the cellular slime mold, roundworm, zebrafish, cat, rat, pig, cow, and baboon. Limited resources focus efforts into the seven main categories of genomes besides *Homo sapiens*: viruses, bacteria, fungi, *Arabidopsis thaliana* ('the weed'), *Drosophila melanogaster* (fruitfly), *Caenorhabditis elegans* (roundworm), and *M. musculus* (mouse). For an overview of sequence data, see www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome and for genome landmarks, readers can search the online collection available on www.sciencemag.org/feature/plus/sfg/special/index.shtml. The Human Genome Project is described in the next section.

The first completed genome reported was of the bacterium *Haemophilus influenzae* in 1995 (see also Box 1.3). Soon after came the yeast genome (*Saccharomyces cerevisiae*) (1996, see www.yeastgenome.org), the bacterium *Bacillus subtilis* (1997), and the tuberculosis bacterium (*Mycobacterium tuberculosis*) in 1998. Reports of the worm, fruitfly, mustard plant, and rice genomes (described below) represent important landmarks, in addition to the human genome (next section).

Roundworm, *C. elegans* (1998)

The completion of the genome deciphering of the first multicellular animal, the one-millimeter-long soil worm *C. elegans*, made many headlines in 1998 (see the 11 December 1998 issue of *Science*, volume 282, and www.wormbase.org/). It reflects a triumphant collaboration of more than eight years between Cambridge and Washington University laboratories.

The nematode genome paves the way to obtaining many insights into genetic relationships among different genomes, their functional characterization, and associated evolutionary pathways. A comparison of the worm and yeast genomes, in particular, offers insights into the genetic changes required to support a multicellular organism.

A comparative analysis between the worm and human genome is also important. Since it was found that roughly one third of the worm's proteins (>6000) are similar to those of mammals, automated screening tests are already in progress to search for new drugs that affect worm proteins that are related to proteins involved in human diseases. For example, diabetes drugs can be tested on worms with a mutation in the gene for the insulin receptor.

Opportunities for studying and treating human obesity (by targeting relevant proteins) also exist: in early 2003 [66] biologists have identified the genes that regulate fat storage and metabolism in the roundworm (i.e., 305 that reduce and 112 that increase body fat) in a landmark experiment that inactivated nearly all of the animal's genes (i.e., expression was disrupted for 16,757 worm genes out of the the predicted total of 19,757 that code for proteins) in a single experiment using new RNA interference technology [626]. Such studies are now routine (e.g., to study regulation of immunity to pathogenic bacterial infections [273], or of programmed cell death (or *apoptosis*) [229]). (see Chapter 7 on RNA).

The remarkable roundworm also made headlines in the Fall of 2002 when the Nobel Prize in Physiology or Medicine was awarded to Sydney Brenner, Robert Horvitz, and John Sulston for their collective work over 30 years on *C. elegans* on programmed cell death, *apoptosis*. This process by which healthy cells are instructed to kill themselves is vital for proper organ and tissue development and also leads to diseases like cancer and neurodegerative diseases. Better knowledge of what leads to cell death and how it can be blocked helps to identify agents of many human disorders and eventually to develop treatments.

Fruitfly, *Drosophila* (1999)

The deciphering of most of the fruitfly genome in 2000 by Celera Genomics, in collaboration with academic teams in the Berkeley and European *Drosophila* Genome projects, made headlines in March 2000 (see the 24 March 2000 issue of *Science*, volume 287, and www.fruitfly.org), in large part due to the groundbreaking "annotation jamboree" employed to assign functional guesses to the identified genes.

Interestingly, the million-celled fruitfly genome has fewer genes than the tiny, 1000-celled worm *C. elegans* (though initial reports of the number of worm's genes may have been over-estimated) and only twice the number of genes as the unicellular yeast. This is surprising given the complexity of the fruitfly — with wings, blood, kidney, and a powerful brain that can compute elaborate behavior patterns. Like some other eukaryotes, this insect has developed a nested set of genes with alternate splicing patterns that can produce more than one meaning from a given DNA sequence (i.e., different mRNAs from the same gene). Indeed, the number of core proteins in both fruitflies and worms is roughly similar (8000 vs. 9500, respectively).

Fly genes with human counterparts may help to develop drugs that inhibit encoded proteins. Already, one such fly gene is *p53*, a tumor-suppressor gene that, when mutated, allows cells to become cancerous. The humble baker's yeast proteins are also being exploited to assess activity of cancer drugs.

Mustard Plant, *Arabidopsis* (2000)

Arabidopsis thaliana is a small plant in the mustard family, with the smallest genome and the highest gene density so far identified in a flowering plant (125 million base pairs and roughly 25,000 genes). Two out of the five chromosomes of *Arabidopsis* were completed by the end of 1999, and the full genome (representing 92% of the sequence) published one year later, a major milestone for genetics. See the 14 December 1999 issue of *Nature*, volume 408, and www.arabidopsis.org/, for example.

This achievement is important because gene-dense plants (25,000 genes versus 19,000 and 13,000 for brain and nervous-system containing roundworm and fruitfly, respectively) have developed an enormous and complex repertoire of genes for the needed chemical reactions involving sunlight, air, and water. Understanding these gene functions and comparing them to human genes will provide insights into other flowering plants, like corn and rice, and will aid in our understanding of human life. Plant sequencing analysis should lead to improved crop production (in terms of nutrition and disease resistance) by genetic engineering and to new plant-based ingredients in our medicine cabinets. For example, engineered crops that are larger, more resistant to cold, and that grow faster have already been produced.

Arabidopsis's genome is also directly relevant to human biological function, as many fundamental processes of life are shared by all higher organisms. Some common genes are related to cancer and premature aging. The much more facile manipulation and study of those disease-related genes in plants, compared to human or animal models, is a boon for medical researchers.

Interestingly, scientists found that nearly two-thirds of the *Arabidopsis* genes are duplicates, but it is possible that different roles for these apparently-duplicate genes within the organism might be eventually found. Others suggest that duplication may serve to protect the plants against DNA damage from solar

radiation; a 'spare' could become crucial if a gene becomes mutated. Intriguingly, the plant also has little "junk"[16] (i.e., not gene coding) DNA, unlike humans.

The next big challenge for *Arabidopsis* aficionados is to determine the function of every gene by precise experimental manipulations that deactivate or overactivate one gene at a time. For this purpose, the community launched a 10-year gene-determination project (a "2010 Project") in December 2000. Though guesses based on homology sequences with genes from other organisms have been made (for roughly one half of the genes) by the time the complete genome sequence was reported, much work lies ahead to nail down each function precisely. This large number of "mystery genes" promises a vast world of plant biochemistry awaiting exploration.

Mouse (2001, 2002)

The international Mouse Genome Sequencing Consortium (MGSC) formed in late fall of 2000 followed in the footsteps of the human genome project [288]. The mouse represents one of five central model organisms that were planned at that time to be sequenced. Though coming in the backdrop of the human genome, draft versions of the mouse genome were announced by both the private and public consortia in 2001 and 2002, respectively.

In the end of 2002, the MGSC published a high-quality draft sequence and analysis of the mouse genome (see the 5 December 2002 issue of *Nature*, volume 420). The 2.5 billion size of the mouse genome is slightly smaller than the human genome (3 billion in length), and the number of estimated mouse genes, around 30,000, is roughly similar to the number believed for humans. Intriguingly, the various analyses reported in December 2002 revealed that only a small percentage (1%) of the mouse's genes has no obvious human counterpart. This similarity makes the mouse genome an excellent organism for studying human diseases and proposed treatments. But the obvious dissimilarity between mice and men and women also begs for further comparative investigations; why are we not more like mice? Part of this question may be explained through an understanding of how mouse and human genes might be regulated differently.

Related to this control of gene activation and function are newly-discovered mechanisms for transcription regulation. Specifically, the mouse genome analyses suggested that a novel class of genes called *RNA genes* — RNA transcripts that do not code for proteins — has other essential regulatory functions that may play significant roles in each organism's survival (see discussion on RNAs in Chapter 7 on non-coding RNAs). As details of these mechanisms, as well as comparisons among human and other closely-related organisms, will emerge, explanations may arise. In the mean time, genetic researchers have a huge boost of resources and

[16]The term "junk DNA", coined early in the genomics game, turned out to be misleading. Non-coding DNAs are now known to have important functions, far from useless DNA and more like a reservoir for rearranging genes. These repetitive elements are thus essential components of eukaryotic organisms [817].

are already exploiting similarities to generate expression patterns for genes of entire chromosomes as a way to research specific human diseases like Down's syndrome, which occurs when a person inherits three instead of two copies of chromosome 21.

Rice (2002)

The second largest genome sequencing project — for the rice plant (see a description of the human genome project below) — has been underway since the late 1990s in many groups around the world. The relatively small size of the rice genome makes it an ideal model system for investigating the genomic sequences of other grass crops like corn, barley, wheat, rye, and sugarcane. Knowledge of the genome will help create higher quality and more nutritious rice and other cereal crops. Significant impact on agriculture and economics is thus expected.

By May 2000, a rough draft (around 85%) of the rice genome (430 million bases) was announced, another exemplary cooperation between the St. Louis-based agrobiotechnology company Monsanto (now part of Pharmacia) and a University of Washington genomics team.

In April 2002, two groups (a Chinese team from the Bejing Genomics Institute led by Yang Huanming and the Swiss agrobiotech giant Syngenta led by Stephen Goff) published two versions of the rice genome (see the 5 April 2002 issue of *Science*, volume 296), for the rice subspecies *indica* and *japonica*, respectively. Both sequences contain many gaps and errors, as they were solved by the whole-genome 'shotgun' approach (see Box 1.3), but represent the first detailed glimpse into a world food staple. The complete sequences of chromosomes 1 and 4 of rice were reported in late 2002 (see the 21 November 2002 issue of *Nature*, volume 420).

Pufferfish, *Fugu* (2002)

The highly poisonous delicacy from the tiger pufferfish prepared by trained Japanese chefs has long been a genomic model organism for Sydney Brenner, a founder of molecular biology and recipient of the 2002 Nobel Prize in Physiology or Medicine for his work on programmed cell death (see above, under Roundworm). The compact *Fugu rubripes* genome is only one-ninth the size of the human genome but contains approximately the same number of genes: shorter introns and less repetitive DNA account for this difference. The whole-genome shotgun approach (see Box 1.3) was used to sequence *Fugu* (see the 23 August 2002 issue of *Science*, volume 297). Through comparative genomics, analyses of this ancient vertebrate genome and of many others help understand the extent of protein evolution (through common and divergent genes) and help interpret many human genes.

Homo Sapiens (2003)

See the separate section.

Other Organisms

Complete sequences and drafts of many genomes are now known. (Check websites such as www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome for status reports). Included are bacterial genomes of a microbe that can survive environments lethal for most organisms and might turn useful as a metabolizer of toxic waste (*D. radiodurans* R1); a nasty little bacterium that causes diseases in oranges, grapes, and other plants (*Xylella fastidiosa*, decoded by a Brazilian team); and the bugs for human foes like the common cold, anthrax, cholera, syphilis, tuberculosis, malaria, the plague, typhus, and SARS (severe acute respiratory syndrome). Proteins unique to these pathogens are being studied, and disease treatments will likely follow (e.g., cholera vaccine).

The third mammalian genome, that of the rat (Brown Norway rat) was completed in early 2004 (see the 1 April 2004 issue of *Nature*, volume 428), allowing us to explore characteristics that are specific to rodents but also common to all mammals.

By early 2010, the genomes of several mammals have been characterized, including the human, chimp, Rhesus macaque, dog, cow, horse, opossum, platypus, giant panda, and Tibetan antelope. The platypus genome, for example, reveals both reptilian and mammalian features and helps define the ancestral line of animal evolution. The cow genome is useful for studying human diabetes, since bovine insulin is a model for studying many human endocrine diseases. The giant panda genome helps explain the animal's bamboo-chomping habit (malfunction of a gene related to digestion). The genome of the Tibetan antelope, also an endangered species, may guide researchers in understanding the animal's ability to adapt to harsh environments (extreme cold and low oxygen) and the pathogenesis of chronic plateau sickness.

1.5.2 The Human Genome

The International Human Genome Project was launched in 1990 to sequence all three billion bases in human DNA [288]. The public consortium has contributions from many parts of the world (such as the United States, United Kingdom, Japan, France, Germany, China, and more) and is coordinated by academic centers funded by NIH and the Wellcome Trust of London, headed by Francis Collins and Michael Morgan (with groups at the Sanger Institute near Cambridge, UK, and four centers in the United States).

In 1998, a competing private enterprise led by Craig Venter's biotechnology firm Celera Genomics and colleagues at The Institute for Genomic Research (TIGR), both at Rockville, Maryland (owned by the PE Corporation; see www.celera.com), entered the race. Eventually, this race to decode the human genome turned into a collaboration in part, not only due to international pressure but also because the different approaches for sequencing taken by the public and private consortia are complementary (see Box 1.3 and related articles [2,268,476, 888,1351,1352] comparing the approaches for the human genome assembly).

Milestones

A first milestone was reached in December 1999 when 97% of the second smallest chromosome, number 22, was sequenced by the public consortium (the missing 3% of the sequence is due to 11 gaps in contiguity); see the 2 December 1999 issue of *Nature*, volume 402. Though small (43 million bases, $<2\%$ of genomic DNA), chromosome 22 is gene rich and accounts for many genetic diseases (e.g., schizophrenia).

Chromosome 21, the smallest, was mapped soon after (11 May 2000 issue of *Nature*, volume 405) and found to contain far fewer genes than the 545 in chromosome 22. This opened the possibility that the total number of genes in human DNA is less than the 100,000 previously estimated. Chromosome 21 is best known for its association with Down's syndrome; affected children are born with three rather than two copies of the chromosome. Learning about the genes associated with chromosome 21 may help identify genes involved in the disease and, eventually, develop treatments.

Completion of the first draft of the human genome sequence project broke worldwide headlines on 26 June 2000 (see, for example, the July 2000 issue of *Scientific American*, volume 283). This draft reflects 97% of the genome cloned and 85% of it sequenced accurately, that is, with 5 to 7-fold redundancy.

Actually, the declaration of the 'draft' status was arbitrary[17] and even fell short of the 90% figure set as target. Still, there is no doubt that the human genome represents a landmark contribution to humankind, joined to the ranks of other 'Big Science' projects like the Manhattan project and the Apollo space program. The June 2000 announcement also represented a 'truce' between the principal players of the public and private human genome efforts and a commitment to continue to work together for the general cause.

A New York Times editorial by David Baltimore on the Sunday before the Monday announcement was expected underscored this achievement, but also emphasized the work that lies ahead:

> The very celebration of the completion of the human genome is a rare day in the history of science: an event of historic significance is recognized not in retrospect, but as it is happening . . . . While it is a moment worthy of the attention of every human, we should not mistake progress for a solution. There is yet much work to be done. It will take many decades to fully comprehend the magnificence of the DNA edifice built over four billion years of evolution and held in the nucleus of each cell of the body of each organism on earth.
>
> David Baltimore, *New York Times*, 25 June 2000.

[17] It has been said [1238] that this day happened to be free in the diaries of U.S. President Bill Clinton and Britain's Prime Minister Tony Blair, who proclaimed victory over the genome along with leading scientists.

Baltimore further explains that the number of proteins, not genes, determines the complexity of an organism. The gene number should ultimately explain the complexity of humans. In June 2000, the estimated number of total human genes was 50,000, compared to 14,000 in a fly or 18,000 in a worm. Several months after the June announcement, this estimate was reduced to 30,000–40,000 (see the 15 February 2001 issue of *Nature*, volume 409, and the 16 February 2001 issue of *Science*, volume 291). This implies an 'equivalence' of sorts between each human and roughly two flies . . . . However, the estimated number has declined since then to around 20,000–25,000.

This astonishingly low number suggests that hidden levels of complexity exist in the human genome from networks of genes rather than individual genes. Clearly, the final word on the number of human genes and the conserved genes that humans share with flies, mice or other organisms awaits further studies.

Three years after the 'working draft' of the human genome sequence was announced with much fanfare, the Human Genome Project as originally devised was declared complete, to an accuracy of 99.9%; the international consortium of genome sequencing centers put all the fragments of the 3.1-billion DNA units of the human genome in order, and closed nearly all of the gaps. The month of April 2003 for this declaration was timed to coincide with the 50th anniversary of Watson and Crick's report of the structure of the DNA double helix.

Some chromosome segments of the human genome, like in chromosome Y, are more difficult to characterize as they are highly repetitive; fortunately, these segments may be relatively insignificant for the genome's overall function.

Of course, we still have little clue on what to make of the DNA book of life presented before us in terms of greater predispositions for specific diseases and individualized response to therapy. These tasks will undoubtedly occupy us in the coming decades.

With the determination of the human genome sequence for several individuals, including Craig Venter (in 2007) [753], James Watson (in 2008) [1368], a 4000-year-old man from Greenland and five southern Africans including Archbishop Desmond Tutu (in 2010), variations (polymorphisms) in the DNA sequence that contribute to disease in different populations are being investigated and analyzed. The ancient DNA analysis also sheds light on migration trends and ancestry by revealing unsuspected movements from Siberia to Greenland about 5500 years ago. The African genomes also help understand human genetic history because the number of variations is relatively high. The establishment of the Personal Genome and 1000 Genomes Projects promises to deliver more such insights.

In addition to identifying these variations, the next task is to define the proteins produced by each gene and understand the cellular interactions of those proteins. This is opening new avenues for disease diagnostics and development of designer drugs. Undoubtedly, the determination of sequences for 1000 major species in the next decade will shed further insights into the human genome, but clearly we are only beginning to understand what it all means. Until then, knowing and interacting with an individual may be more informative than sorting through her/his DNA, as M. Olson suggests in his commentary on Watson's genome sequencing [936].

A Triumph of Technology

It should be emphasized that none of these studies would be possible without remarkable advances in sequencing technology in terms of speed and efficiency [1323]. The HGP took 13 years to complete in 2003 at a cost of $2.7 billion. Four years later, Venter's DNA was sequenced after only four years of work at a cost of $100 million [753]. Just one year later, in 2008, Watson's DNA was determined after about 4 months of work at a fraction of the cost (<$1.5 million) [936, 1368].

In 2009, BioNanomatrix designed a nanofluidic chip approach to sequencing that could lower DNA sequencing costs down to <$100 within the next 5 years (see Figure 1.5). Another innovative 'DNA Sudoku' approach to sequencing was soon after reported — using combinatorial pooling strategies to sequence multiple genomes for the purpose of analyzing rapidly specific regions of sequence variants, such as associated with disease [366]. This strategy of mixing many geonome samples and using logic and combinatorial rules as used in number puzzles to search for specific patterns in the pool of samples is best suited for genotype analysis of short segments to diagnose genetic diseases such as Tay-Sachs or cystic fibrosis that tend to occur in certain ethnic groups.

Many other companies (e.g., Illumina, Life Technologies, Oxford Nanopore Technologies, Pacific Biosciences, and IBM) have entered the personal sequencing service with innovative approaches; efficiency will certainly increase and cost dramatically decrease in the very near future. Next-generation machines could also make possible rapid sequencing of disease-causing microbes to allow infection diagnosis or tracking, as well as lead to important engineering applications (e.g., bacteria that produce more hydrogen). The increasing challenge of handling and processing large amounts of sequence data may also be alleviated by the use of cloud computing — computational resources distributed over the Internet.

It is no surprise that the era of large-scale sequencing projects has been supplanted by many private, for-profit companies like Navigenics, 23andMe, GeneWize, Knome, and others who are offering individuals personal atlases of their DNA. Will this type of direct-to-consumer genetics be the norm in the near future?

Box 1.3: Different Sequencing Approaches

Two synergistic approaches have been used for sequencing. The public consortium's approach relies on a 'clone-by-clone' approach: breaking DNA into large fragments, cloning each fragment by inserting it into the genome of a bacterial artificial chromosome (BAC), sequencing the BACs once the entire genome is spanned, and then creating a physical map from the individual BAC clones. The last part — rearranging the fragments in the order they occur on the chromosome — is the most difficult. It involves resolving the overlapped fragments sharing short sequences of DNA ('sequence-tagged sites').

The alternative approach pioneered by Venter's Celera involves reconstructing the entire genome from small pieces of DNA without a prior map of their chromosomal positions.

The reconstruction is accomplished through sophisticated data-processing equipment. Essentially, this gargantuan jigsaw puzzle is assembled by matching sequence pieces as the larger picture evolves.

The first successful demonstration of this piecemeal approach was reported by Celera for decoding the genome of the bacterium *Haemophilus influenzae* in 1995. This bacterium has a mere 1.8 million base pairs with estimated 1700 genes, versus three billion base pairs for human DNA with at least 30,000 genes. The sequence of *Drosophila* followed in 1998 (140 million base pairs, 13,000 estimated genes) and was released to the public in early 2000 (see the 24 March 2000 issue of *Science*, volume 287).

This 'shotgun' approach has been applied to the human genome, more challenging than the above organisms for two reasons. The human genome is larger — requiring the puzzle to be formed from ∼70 million pieces — and has many more repeat sequences, complicating accurate genome assembly. For this reason, the public data were incorporated into the whole genome assembly [476]. The whole-genome shotgun approach has also been applied to obtain a draft of the mouse (2001), rice (2002) and pufferfish (2002) genomes, for example.

The two approaches are complementary, since the rapid deciphering of small pieces by the latter approach relies upon the larger picture generated by the clone-by-clone approach for overall reconstruction. See a series of articles in 2002 [476, 888, 1351] that scrutinized those approaches, focusing on the extent of public-database information utilized in the

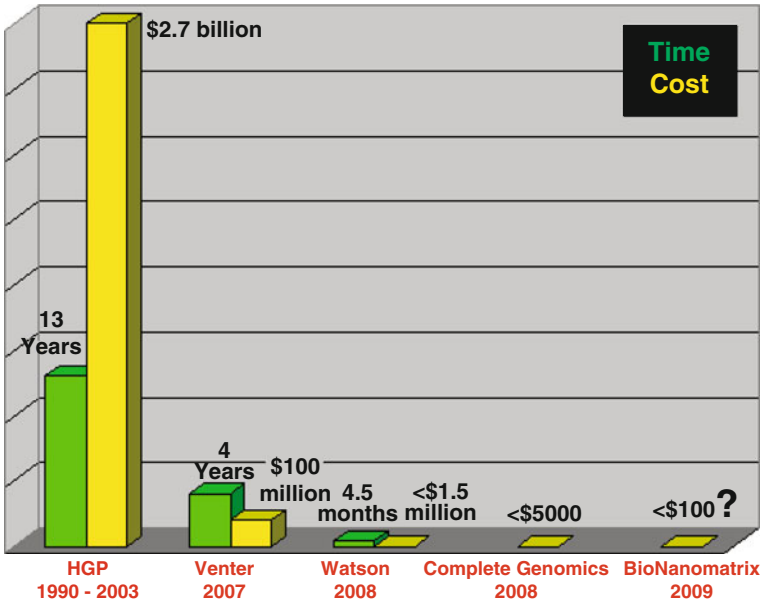

Figure 1.5. The progress of DNA sequencing technology. The sequencing time and cost associated with the human genome project, determination of Venter's and Watson's DNA, and biotechnology company prospects are given.

whole-genome shotgun approach to the human genome assembly, and the second round of debate in 2003 [2, 268, 1352]. The recent success of individual DNA sequencing involves rapid-sequencing methodologies that cut DNA into tiny segments of only 250 base pairs long. This makes genome assembly more technically challenging but the the entire sequencing process much faster and cheaper.

A Gold Mine of Biodata

The most up-to-date information on sequencing projects can be obtained from the U.S. National Center for Biotechnology Information (NCBI) at the U.S. National Library of Medicine, which is developing a sophisticated analysis network for the human genome data.

For information, see the Human Genome Resources Guide www.ncbi.nlm.nih.gov/genome/guide/human (click on Map Viewer), the U.S. National Human Genome Research Institute's site www.nhgri.nih.gov/, that of Department of Energy (DOE) at genomics.energy.gov, the site of the University of California at Santa Cruz at genome.ucsc.edu/, and others.[18]

Since 1992, NCBI has maintained the GenBank database of publicly available nucleotide sequences (www.ncbi.nlm.nih.gov). A typical GenBank entry includes information on the gene locus and its definition, organism information, literature citations, and biological features like coding regions and their protein translations. Many search and analysis tools are also available to serve researchers.

Implications – Some Application Examples

The genomic revolution and the comparative genomics enterprises now underway will not only provide fundamental knowledge about the organization and evolution of biological systems in the decades to come [672] but will also lead to medical breakthroughs.

Already, some practical benefits of genomic deciphering have emerged (e.g., [529, 892]). A dramatic demonstration in 2000 was the design of the first vaccine to prevent a deadly form of bacterial meningitis using a two-year gene-hunting process at Chiron Corporation. Researchers searched through the computer database of all the bacterium's genes and found several key proteins that in laboratory experiments stimulated powerful immune responses against all known strains of the *Neisseria meningitidis* Serogroup B Strain MC58 bug [1003].

[18] Some useful web sites for genomic data include www.arabidopsis.org, www.ncbi.nlm.nih.gov/Sitemap/, plant genomes for the specialist, the Agricultural Genome Information System, *Caenorhabditis elegans* Genetics and Genomics, Crop Genome Databases at Cornell University, FlyBase, The Genome Database, Genome Sequencing Center (Washington University), GenomeNet, U.S. National Agricultural Library, Online Mendelian Inheritance in Man, *Pseudomonas aeruginosa* Community Annotation Project, *Saccharomyces* Genome Database, The Sanger Institute, Taxonomy Browser, and UniGene.

In April 2003, just two months after the first inklings of a deadly disease called SARS emerged from Asia, a global effort coordinated by the World Health Organization announced that it had mapped the coronavirus genome that causes this highly infectious disease (see resulting papers in the 30 May 2003 issue of *Science*, volume 300). This was made possible by the new high-tech science era of internet links and sequencing methods. Soon after, Affymetrix released a SARS resequencing array encompassing the entire 30-kb genome of the virus, allowing analysis of variation among various virus versions and thus an understanding of how rapidly the virus changes and spreads. Having the viral genome sequence, work continues on understanding the roles of viral proteins in SARS pathogenesis, crucial for developing suitable drugs and vaccines [42]. Finding a treatment for SARS certainly presents a challenge, since the devastating economic losses due to the SARS outbreak underscored our vulnerabilities to infectious diseases. Yet the global virus hunt is a model *par excellence* for the potential of genomics initiatives. Indeed, three years later, a promising inhibitor of the SARS virus was identified by computer-aided molecular design [329].

Full genome sequencing and genomics association studies were once again celebrated when years of secret forensic investigations led in August 2008 to the source of the 2001 anthrax attacks in the U.S. It was then that the U.S. Federal Bureau of Investigation (FBI) finally produced conclusive evidence against Army Institute scientist Bruce Ivins (in Fort Detrick, MD) as the propagator of the anthrax murders in 2001. Not only does this case reveal the dangers of genomic information of killer organisms like *Baccilus anthracis*; it underscores the importance of full-genome sequencing and related technological breakthroughs as forensic tools.

The story begins in 2001, when in the wake of the horrid Al Qaeda attacks on New York and Washington, D.C., mysterious white powders mailed to several individuals caused several illnesses and five eventual deaths from various forms of anthrax. For years, the public was frustrated at the FBI's lack of resolution of the culprit, though Army scientist Steven Hatfill was implicated (and ultimately vindicated and compensated by a large sum of money). From the start, scientists specializing in biowarfare were suspected since using anthrax as a killer requires expert knowledge.

In July 2008, Ivins died from a drug overdose, apparently a suicide, raising public suspicions and re-igniting the public's desire for a conclusion to this affair. Little did we know that the FBI has been busy for years with meticulous scientific investigations, most of which was commissioned to The Institute for Genomic Research (TIGR) in MD. When the story was finally unveiled, it was concluded that the anthrax species at Ivins' possession, flask RMR-1027, matched conclusively the victims' anthrax. The technical complications also became clear: the nonuniform anthrax genome contained variants or *spores* that required special cultivation followed by sequencing to identify, isolate, and finally match perfectly to one source: only the Ivins source and seven directly-related isolates contained all the four mutations that were identified from the victims' anthrax genome (see [363] and N. Wade in *The New York Times*, Aug. 21, 2008). The dangerous nature

of the genome also required the FBI to make piecemeal requests for the scientific work involved. Though Ivins' motive for these deadly acts remains unclear, the possibility that Ivins propagated the scare in an effort to ensure continued government funding for anthrax vaccine development and other biological warfare developments remains viable.

The anthrax story emphasizes the long-recognized dangers of having genome information in the wrong hands. However, in theory, genomic information on deadly agents, from anthrax to the influenza virus, can bring about new treatments. In practice, this has turned to be a greater challenge than anticipated. For example, in our ongoing search for new potent antibiotics to fight pernicious infections that are resistant to many known antibiotics (like MRSA, or Methicillin-resistant *Staphylococcus aureus*), the availability of hundreds of sequenced bacterial genomes has hardly led to new antibiotic targets. Genomics may simply be an insufficient basis to serve as a foundation for developing better infection treatments since even a minute amount of resistant bodies can trigger resurgence of disease and/or drug resistance.

Indeed, our susceptibility to viruses, for example, was made clear in the 2009 pandemic of the swine flu (H1N1). This virus resulted in millions of infections throughout the world in a very short time and hundreds of deaths, despite genomics advances and modern screening and treatment tools.

Of course, one of the primary hopes for genomics applications has been the expectation for better diagnosis and treatment of human disease. However, genome association studies have generally shown limited value in predicting human diseases because genetic variations only explain a small part of the genetic/disease link. In other words, the genetic link to disease remains largely unclear and it is possible that rare variants account for most of the genetic basis for disease. Indeed, dissenting voices have suggested that our efforts to pinpoint the genetics of common diseases may not work. Still, successful examples of using a patient's DNA to help tailor medication has shown some promise, as in the case of Tamoxifen for breast cancer, Erbitux for colon cancer, or Iressa for lung cancer — all of which work only for patients with a particular genetic mutation. Nonetheless, these findings are complicated by the fact that genetic tests are not sufficiently accurate at present, as indicated by the breast cancer drug Herceptin, which is only effective on certain subtypes of the disease. See Chapter 15 for an expanded discussion of pharmacogenomics, including Herceptin.

Many still remain generally hopeful because exploiting genomic information for new medical breakthroughs is a new field that will take time to mature and succeed. There is no doubt that the many ideas and tools not available to us several decades ago but now at our disposal will ultimately lead to significant medical advances. And given the deadly infectious diseases, like HIV/AIDS, Ebola, Marbourg, and antibiotics-resistant 'super-bugs', that are responsible for more than 25% of the world's deaths, it is hoped that the new tools, together with new ideas and improved techniques, will lead to these needed biomedical developments.

Ongoing Challenges and Ramifications

As gene products are being identified, the biological revolution is beginning to affect many aspects of our lives [259], perhaps not too far away from Wilson's vision of consilience. A 'gold mine' of biological data is now amassing, likened to "orchards . . . just waiting to be picked".[19] This rich resource for medicine and technology also provides new foundations, as never before, for computational applications.

Consequently, in fifty years' time, we anticipate breakthroughs in protein folding, medicine, cellular mechanisms (regulation, gene interactions), development and differentiation, history (population genetics, origin of life), and perhaps new life forms, through analysis of conserved and vital genes as well as new gene products. See the 5 October 2001 issue of *Science* (volume 294) for a discussion of new ideas, projects, and scientific advances that followed since the sequencing of the human genome.

Among the promising medical leaps are personalized and molecular medicine, perhaps in large part due to the revolutionary DNA microarray technology (see [400] and Box 1.4) and gene therapy. Of course, *information is not knowledge*, but rather a road that can lead to perception. Therefore, these aforementioned achievements will require concerted efforts to extract information from all the sequence data concerning gene products.

Many initiatives are underway to process genetic data in the goal of understanding, and eventually treating, human diseases. For example, in 2003 Britain launched a 'genetic census' *Biobank* project — assembly of a database of medical information based on 500,000 Britons representing Britain's demographics aimed at quantifying the combined genetic and environmental (e.g., pollution, smoking, exercise, diet) influence on common human ailments.

Other national genetic database projects (with corresponding numbers of participants) are underway in Iceland (275,000), Sweden (80,000), Estonia (1 million), and Latvia (50,000). Private genomic database projects are also being assembled by the American Cancer Society (110,000), Mayo Clinic (200,000), and CARTaGENE (50,000). Companies like the pioneering Icelandic DeCode Genetics went on a hunt to search for disease genes in these genealogies.[20] Many other international consortia and large-scale projects have been formed, including ENCODE, 1000 genomes project, Cancer Atlas, HapMap, Personal Genome Project, and much more, to interpret the human genome, and many private companies have started to exploit many biomedical and technological issues of genomic sciences.

[19] B. Sinclair, in *The Scientist*, 19 March 2001.

[20] In November 2009, DeCode Genetics, which was founded in 1996, filed for bankruptcy. Though quickly becoming the world leader in the race to identify genetic connections to common diseases like cancer, diabetes and schizophrenia, experts believe that — regardless of business strategies used by the company — the genetic nature of human disease has turned out to be much more complex that originally envisioned. In January 2010, the company announced that it emerged from bankruptcy and will continue its genetics research as a private company, though it will abandon its drug development efforts.

When leading scientists were queried in 2002 by the publisher of the website Edge.org devoted to science to advise on action to take concerning the most pressing scientific issues in the world, physicist Freeman Dyson boldly suggested "a planetary genome sequencing project to identify all the segments of the genomes of all the millions of species that live together in the planet". Dyson's vision for completing the sequencing of the biosphere within less than half a century aims to profoundly increase our understanding of the ecology of the planet, which could lead to environmental improvements and cures for human diseases. There is no doubt that creative and well engineered projects combining technological innovations with biological data can have enormous ramifications on our lives. See, for example, a vision for the future of genomics research by Collins and coworkers [258].

Some of the ongoing challenges that face us now include establishing gene number, location, and function; understanding the interaction of protein networks; understanding non-coding DNA (amount, distribution, function); determining protein structure and function evolution; correlating single nucleotide polymorphisms to health and disease predisposition; establishing evolutionary trends among organisms; and exploiting genome information for environmental restoration via designer organisms.

Many societal, ethical, economic, legal, and political issues will also have to be addressed with these developments. Still, like the relatively minor Y2K (Year 2000) anxiety, these problems could be resolved in stride through multidisciplinary networks of expertise.

In a way, sequencing projects make the giant leap directly from *sequence to function* (possible only when a homologous sequence is available whose function is known). However, the crucial middle aspect — *structure* — must be relied upon to make systematic functional links. This systematic interpolation and extrapolation between sequence and structure relies and depends upon advances in biomolecular modeling, in addition to high-throughput structure technology ('the human proteomics project').

The next chapter introduces some current challenges in modeling macromolecules and mentions important applications in medicine and technology.

Box 1.4: Genomics & Microarrays

DNA microarrays — also known as gene chips, DNA chips, and biochips — are becoming marvelous tools for linking gene sequence to gene products. They can provide, in a single experiment, an expression profile of many genes [400]. As a result, they have important applications to basic and clinical biomedicine. Particularly exciting is the application of such genomic data to *personalized medicine* or *pharmacogenomics* — prescribing medication based on genotyping results of both patient and any associated bacterial or viral pathogen [370]. Prescribing specific diets to affect health based on genetic responses to diet (*nutritional genomics* or *nutrigenomics*) is another application gaining momentum.

Essentially, each microarray is a grid of DNA oligonucleotides (called *probes*) prepared with sequences that represent various genes. These probes are directed to a specific gene or mRNA samples (called *targets*) from tissues of interest (e.g., cancer cells). Binding between probe and target occurs if the RNA is complementary to the target nucleic acid. Thus, probes can be designed to bind a target mRNA if the probe contains certain mutations. Single nucleotide polymorphisms or SNPs, which account for 0.1% of the genetic difference among individuals, can be detected this way [848].

The hybridization event — amount of RNA that binds to each cell grid — reflects the extent of gene expression (gene activity in a particular cell). Such measurements can be detected by fluorescence tagging of oligonucleotides. The color and intensity of the resulting base-pair matches reveal gene expression patterns.

Different types of microarray technologies are now used (e.g., using different types of DNA probes), each with strengths and weaknesses. The technique of principal component analysis (PCA, see Chapter 15) has shown to be useful in analyzing microarray data (e.g., [1009]). Technical challenges remain concerning verification of the DNA sequences and ensuring their purity, amplifying the DNA samples, and assessing the results. For example, false positives or false negatives can result from irregular target/probe binding (e.g., mismatches) or from self-folding of the targets, respectively. The problem of accuracy of the oligonucleotides has stimulated various companies to develop appropriate design techniques. Affymetrix Corporation, for example, has developed technology for designing silicon chips with oligonucleotide probes synthesized directly onto them, with thousands of human genes on one chip. All types of DNA microarrays rely on substantial computational analysis of the experimental data to determine absolute or relative patterns of gene expression.

Such patterns of gene expression (induction and repression) can prove valuable in drug design. An understanding of the affected enzymatic pathway by proven drugs, for example, may help screen and design novel compounds with similar effects. This potential was demonstrated for the bacterium *M. tuberculosis*, based on experimental profiles obtained before and after exposure to the tuberculosis drug Isoniazid [1378].