1
Biomolecular Structure and Modeling: Historical Perspective

March 20, 2001

Chapter 1 Notation

<table>
<thead>
<tr>
<th>SYMBOL</th>
<th>DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Vectors</strong></td>
<td></td>
</tr>
<tr>
<td>h</td>
<td>unit cell identifier (crystallography)</td>
</tr>
<tr>
<td>r</td>
<td>position</td>
</tr>
<tr>
<td>$F_h$</td>
<td>structure factor (crystallography)</td>
</tr>
<tr>
<td>$\phi_h$</td>
<td>phase angle (crystallography)</td>
</tr>
<tr>
<td><strong>Scalars</strong></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>distance between parallel planes in the crystal</td>
</tr>
<tr>
<td>$I_h$</td>
<td>intensity, magnitude of structure factor (crystallography)</td>
</tr>
<tr>
<td>$V$</td>
<td>cell volume (crystallography)</td>
</tr>
<tr>
<td>$\theta$</td>
<td>reflection angle (crystallography)</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>wavelength of the X-ray beam (crystallography)</td>
</tr>
</tbody>
</table>
1.1 A Multidisciplinary Enterprise

1.1.1 Consilience

"... 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, organismic, and evolutionary biology. ... Such is the unifying and highly productive understanding of the world that has evolved in the natural sciences."


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 underscore its cross-disciplinary nature: computational biology, computational chemistry, in silico biology, structural biology, computational biophysics, theoretical biophysics, theoretical chemistry, and the list goes on.

As the pioneer of sociobiology Edward O. Wilson reflects above, some scholars believe in a unifying knowledge for understanding our universe and ourselves, or consilience that merges all disciplines in a biologically-grounded framework [1]. Though this link is most striking between genetics and human behavior — through the neurobiological underpinnings of states of mind and mental activity, and shaped by the environment — such a unification that Wilson advocates might only be achieved by a close interaction among the varied scientists at many stages of study. The post genomic era

\[1\] Consilience was coined in 1840 by the theologian and polymath William Whewell 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 recently took this notion further by advocating in his 1998 book Consilience [1] that the world is orderly and can be explained by a set of natural laws that are fundamentally rooted in biology.
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 21st-century society.

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 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. For example, experimental data have been important to interpretation of the funnel energy landscape for protein folding (a special shape for the energy governing the folding process) [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 naïveté 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 or 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 color stereographics and laser-made wooden sculptures. The computations encompass \textit{ab initio} and semi-empirical quantum mechanics, empirical (molecular) mechanics, molecular dynamics, Monte Carlo, free energy methods, quantitative structure-activity relationships (QSAR), chemical information and databases, and 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.

The questions being addressed with these tools today are as intriguing and as complex as the biological systems themselves. They range 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
1. Introduction to Biomolecular Structure

determination — from ions channels to single-molecule biochemistry\(^2\) — modeling approaches are needed to fill in many gaps and to further analyze the data.

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 dazzling speed. Advances are driven by improvements in instrumental 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.

Yet, as we witness the tantalizing progress, a cautionary usage of molecular modeling tools is warranted, as well as a critical perspective of the field’s strengths and limitations. 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. Indeed, the tools and programs developed by a handful of researchers thirty years 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, the elementary background for biomolecular modeling will be introduced: protein and nucleic-

---

\(^2\)Examples of recent triumphs in biomolecular structure determinations include elucidation of the nucleosome — essential building block of the DNA/protein spools that make up the chromosomal material \[3\]; ion channel proteins — regulators of membrane electrical potentials in cells, thereby generating nerve impulses and controlling muscle contraction, hormone production, and cardiac rhythm \[4, 5\]; and the ribosome — the cell’s protein-synthesis factory, the machine bundle of 54 proteins and three RNA strands that moves along messenger RNA and synthesizes polypeptides (complete system at low resolution \[6\], and the larger \[7, 8, 9\] and smaller subunits \[10, 11, 12, 13\] at moderate resolution (see perspective in \[14\]). Other important examples of experimental breakthroughs involve overstretched DNA — as seen in single-molecule force versus extension measurements \[15\], and competing folding and unfolding pathways for proteins — as obtained by kinetic studies using spectroscopic probes \[16, 17\].
acid structure tutorials (Chapters 3–6), overview of theoretical approaches (Chapter 7), details of force field construction and evaluation (Chapters 8 and 9), energy minimization techniques (Chapter 10), Monte Carlo simulations (Chapter 11), molecular dynamics methods (Chapters 12 and 13), and combinatorial chemistry/library design problems (Chapter 14).

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 or molecular dynamics simulations) are complementary.

The remainder of this chapter and the next provides 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 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 ‘road map’ is shown in Table 1.1. The experimental section 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 the perspective by describing the computational challenges that naturally emerge from the dazzling 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 confluent experimental, theoretical, and technological advances.

Since the material presented in these introductory chapters is changing rapidly (e.g., the status of the genome projects, theoretical and instrumental progress), this chapter will be periodically updated and placed on the text web page.
### 1. Introduction to Biomolecular Structure

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1865</td>
<td>Genes discovered by Mendel</td>
</tr>
<tr>
<td>1910</td>
<td>Genes in chromosomes shown by Morgan's fruitfly mutations</td>
</tr>
<tr>
<td>1920s</td>
<td>Quantum mechanics theory develops</td>
</tr>
<tr>
<td>1926</td>
<td>Early reports of crystallized proteins</td>
</tr>
<tr>
<td>1930s</td>
<td>Reports of crystallized proteins continue and stimulate Pauling &amp; Corey to compile bond lengths and angles of amino acids</td>
</tr>
<tr>
<td>1944</td>
<td>Avery proves genetic transformation via DNA (not protein)</td>
</tr>
<tr>
<td>1946</td>
<td>Molecular mechanics calculations reported (Westheimer, others)</td>
</tr>
<tr>
<td>1949</td>
<td>Sickle cell anemia identified as 'molecular disease' (Pauling)</td>
</tr>
<tr>
<td>1950</td>
<td>Chargaff determines near-unity A:T and G:C ratios in many species</td>
</tr>
<tr>
<td>1951</td>
<td>Pauling &amp; Corey predict protein α-helices and β-sheets</td>
</tr>
<tr>
<td>1952</td>
<td>Hershey &amp; Chase reinforce genetic role of DNA (phage experiments)</td>
</tr>
<tr>
<td>1952</td>
<td>Wilkins &amp; Franklin deduce that DNA is a helix (X-ray fiber diffraction)</td>
</tr>
<tr>
<td>1953</td>
<td><strong>Watson &amp; Crick report the structure of the DNA double helix</strong></td>
</tr>
<tr>
<td>1959</td>
<td>Hemoglobin &amp; myoglobin deciphered by X-ray (Perutz &amp; Kendrew)</td>
</tr>
<tr>
<td>1960s</td>
<td>Systematic force-fields develop (Allinger, Lifson, Scheraga, others)</td>
</tr>
<tr>
<td>1963</td>
<td>Genetic code established (Crick, Brenner, and coworkers)</td>
</tr>
<tr>
<td>1969</td>
<td>Levinthal paradox on protein folding posed</td>
</tr>
<tr>
<td>1970s</td>
<td>Biomolecular dynamics simulations develop (Stillinger, Karplus and co-workers)</td>
</tr>
<tr>
<td>1970s</td>
<td>Site-directed mutagenesis techniques developed by M. Smith</td>
</tr>
<tr>
<td>1971</td>
<td>Protein Data Bank established</td>
</tr>
<tr>
<td>1974</td>
<td>t-RNA structure reported</td>
</tr>
<tr>
<td>1975</td>
<td>Fifty solved biomolecular structures available in the PDB</td>
</tr>
<tr>
<td>1977</td>
<td>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)</td>
</tr>
<tr>
<td>1980s</td>
<td>Dazzling progress realized in automated sequencing, protein X-ray crystallography, NMR, recombinant DNA, and macromolecular synthesis</td>
</tr>
<tr>
<td>1985</td>
<td>PCR devised by Mullis; numerous applications follow</td>
</tr>
<tr>
<td>1985</td>
<td>NSF establishes five national supercomputer centers</td>
</tr>
<tr>
<td>1990</td>
<td>International Human Genome Project starts; spurs other genome projects</td>
</tr>
<tr>
<td>1994</td>
<td>RNA hammerhead ribozyme structure reported; other RNAs follow</td>
</tr>
<tr>
<td>1995</td>
<td>First non-viral genome completed (bacterium <em>Haemophilus influenzae</em>), 1.8 Mb</td>
</tr>
<tr>
<td>1996</td>
<td>Yeast genome (<em>Saccharomyces cerevisiae</em>) completed, 13 Mb</td>
</tr>
<tr>
<td>1997</td>
<td>Chromatin core particle structure reported; confirms earlier structure</td>
</tr>
<tr>
<td>1998</td>
<td>Roundworm genome (<em>C. elegans</em>) completed, 100 Mb</td>
</tr>
<tr>
<td>1998</td>
<td>Crystal structure of ion channel protein reported</td>
</tr>
<tr>
<td>1998</td>
<td>Private Human Genome initiative competes with international effort</td>
</tr>
<tr>
<td>1999</td>
<td>Fruitfly genome (<em>Drosophila melanogaster</em>) completed (Celera), 137 Mb</td>
</tr>
<tr>
<td>1999</td>
<td>Human chromosome 22 sequenced (public consortium)</td>
</tr>
<tr>
<td>1999</td>
<td>IBM announced petaflop computer to fold proteins by 2005</td>
</tr>
<tr>
<td>2000</td>
<td><strong>First draft of human genome sequence announced</strong> (June 26), 3300 Mb</td>
</tr>
<tr>
<td>2000</td>
<td>Structures of ribosome subunit reported; complement low-resolution of 1999</td>
</tr>
<tr>
<td>2000</td>
<td>12,777 solved biomolecular structures available in the PDB (July 25)</td>
</tr>
</tbody>
</table>
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. More intricate cross terms and additional potentials are discussed in detail in Chapter 8.

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 7) 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 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 Shneor Lifson at the Weizmann Institute of Science (Rehovot, Israel), 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 began to develop force field parameters for families of chemical compounds by testing calculation results against experimental observations regarding structure and energetics.

In the early 1970s, Rahman and Stillinger reported the first molecular dynamics work of a polar molecule, liquid water [20, 21]; results shed insights into the structural and dynamic properties of this life sustaining molecule. These researchers built upon the simulation technique described
1. Introduction to Biomolecular Structure

**TABLE 1.2. The evolution of molecular mechanics and dynamics**

<table>
<thead>
<tr>
<th>Period</th>
<th>System and Size*</th>
<th>Trajec. Length [ns]</th>
<th>CPU Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1973</td>
<td>Dimucleoside (GpC) in vacuum (8 flexible dihedral angles)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>1977</td>
<td>BPTI, vacuum (58 residues, 885 atoms)</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>1983</td>
<td>DNA, vacuum, 12 &amp; 24 bp (754/1530 atoms)</td>
<td>0.09</td>
<td>several weeks each Vax 780</td>
</tr>
<tr>
<td>1984</td>
<td>GnRH, vacuum (decapptide, 161 atoms)</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>1985</td>
<td>Myoglobin, vacuum (1423 atoms)</td>
<td>0.30</td>
<td>50 days VAX 11/780</td>
</tr>
<tr>
<td>1985</td>
<td>DNA, 5 bp (2800 atoms)</td>
<td>0.50</td>
<td>20 hrs Cray X-MP</td>
</tr>
<tr>
<td>1989</td>
<td>Phospholipid Micelle (≈ 7,000 atoms)</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>1992</td>
<td>HIV protease (25,000 atoms)</td>
<td>0.10</td>
<td>100 hrs Cray Y-MP</td>
</tr>
<tr>
<td>1997</td>
<td>Estrogen/DNA (36,000 atoms, fast multipoles)</td>
<td>0.10</td>
<td>22 days HP-735 (8)</td>
</tr>
<tr>
<td>1998</td>
<td>DNA, 24 bp (21,000 atoms, Ewald)</td>
<td>0.50</td>
<td>1 year, SGI Challenge</td>
</tr>
<tr>
<td>1998</td>
<td>β-heptapeptide in methanol (≈ 5000/9000 atoms)</td>
<td>200</td>
<td>8 months, SGI-Challenge (3)</td>
</tr>
<tr>
<td>1998</td>
<td>Villin headpiece (36 residues, 12,000 atoms, cutoffs)</td>
<td>1000</td>
<td>4 months, 256-proc. Cray T3D/E</td>
</tr>
<tr>
<td>1999</td>
<td>βc1 complex in phospholipid bilayer (91,061 atoms, cutoffs)</td>
<td>1</td>
<td>75 days, 64 450-MHz-proc. Cray T3E</td>
</tr>
</tbody>
</table>

*The examples for each period are representative. The first five systems are modeled in vacuum and the others in solution. Except for the dimucleoside, simulations refer to molecular dynamics (MD). The computational time is given where possible; estimates for the vacuum DNA and heptapeptide simulations [18, 19] were kindly provided by M. Levitt and W. van Gunsteren, respectively. The two system sizes for [19] reflect two (temperature-dependent) simulations. See text for definitions of abbreviations and further entry information.
much earlier (1959) by Alder and Wainwright but applied to hard spheres [22].

It took a few more years, however, for the field to gain some ‘legitimacy’. 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 they were quite curious about the predictions that molecular mechanics could make. In time, it became evident that force constants are indeed not generally transferable; still, the molecular mechanics approach was sound since nonbonded interactions are included, terms that spectroscopists omitted.³

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 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 and their coworkers.

1.2.2 Biomolecular Simulation Perspective

Table 1.2 and Figures 1.1 and 1.2 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) landscape of Figure 1.1 shows ‘buildings’ with heights proportional to system size; Figure 1.2 offers molecular views of the simulation subjects.

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 [23]; 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!

The small protein BPTI (Bovine Pancreatic Trypsin Inhibitor) was the subject of a 1977 pioneering dynamic simulation applied to a protein [24]. It showed substantial atomic fluctuations on the picosecond timescale.

³Personal 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 his sharing these experiences with me.
The first nucleic-acid simulation in 1983 [18] was 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 [23]. The stability of DNA simulations was problematic in the early days — the strands untwisted and separated — and required the advanced treatments of solvation and electrostatics introduced a decade later, as discussed, in [26], for example.

The linear decapeptide GnRH (gonadotropin-releasing hormone) was studied in 1984 for its pharmaceutical potential, as it triggers LH and FSH hormones [27].

The 300 ps dynamics simulation of the protein myoglobin in 1985 [28] 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 [29] was an ambitious undertaking; it incorporated a hydrated micelle (i.e., a spherical aggregate of phospholipid molecules) containing 85 LPE (lysophosphatidyl ethanolamine) and 1591 water molecules.

The HIV protease system simulated in solution in 1992 [30] captured an interesting flap motion at the active site. See also Figure ?? and a discussion of this motion in the context of protease inhibitor design.

The 1997 estrogen/DNA simulation [31] 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 [32].

The 1998 DNA simulation [33] used the alternative, Ewald treatment for consideration of long-range electrostatics and uncovered interesting properties of A-tract sequences.

The 1998 peptide simulation in methanol used periodic boundary conditions and captured reversible, temperature-dependent folding [19]; the 200 ns time reflects four 50 ns simulations at various temperatures.

The 1 µs villin-headpiece simulation of the same year (using periodic boundary conditions) [34] is considered longer by three orders of magnitude than simulations to date. A folded structure close to the native state was approached; see also [35].

The solvated protein bc, embedded in a phospholipid bilayer [36] 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.

Trends

Note from the table and figure the transition from simulations in vacuum (first five entries) to simulations in solvent (rest). Note also the steady increase in simulated system size as well as computational time, with a
FIGURE 1.1. The evolution of molecular dynamics simulations with respect to system sizes and simulation lengths (see also Table 1.2).

A leap increase in simulation lengths made only recently. Large system sizes or long simulation times can only be achieved by sacrificing other simulation aspects. For example, truncating long-range electrostatic interactions makes possible the study of large systems over short times [36], or small systems over long times [34]. In fact, with the increased awareness of the sampling problem in dynamic simulations, we now see the latter trend more often, namely studying smaller solvated molecular systems for longer times; one long simulation is often replaced by several shorter trajectories, leading to overall better sampling statistics. See a related recent perspective on long-timescale simulations of peptides and proteins in solution, discussing the progress to date and the future challenges that remain ahead [37].
FIGURE 1.2. The evolution of molecular dynamics simulations with respect to simulation lengths (see also Table 1.2 and Figure 1.1).
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: John Sumner, urease; 1934: J. D. Bernal and Dorothy Crowfoot-Hodgkin, pepsin; 1933: Crowfoot-Hodgkin, insulin). However, only in 1959 did Max Perutz and John Kendrew succeed in deciphering the X-ray diffraction pattern from the crystal structure of the protein (hemoglobin and myoglobin, respectively). Earlier (1953), Perutz found that a structure could be solved by comparing the X-ray diffraction pattern of crystals of native hemoglobin with that of hemoglobin combined with the heavy atom mercury. When this was demonstrated a few years later, modern structural biology was founded.

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 [38, 39]. 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 Cavendish Laboratory in Cambridge in 1947, created in 1962 the Laboratory of Molecular Biology there. Perutz and Kendrew received the Nobel Prize for Chemistry for their accomplishments in 1962.4

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.

Intrigued by 1928 experiments of the British medical officer Fred Griffith, in which mice became fatally ill upon infection from a live but harmless

---

4See 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 (www.nobel.se). This virtual museum was recently constructed to mark the 100th anniversary in 2001 of Alfred B. Nobel’s legacy.
(coatless) strain of pneumonia-causing bacteria mixed with the DNA from heat-kill pathogenic bacteria (which surprisingly transformed live harmless into live pathogenic bacteria, Oswald Avery and coworkers Colin MacLeod and Maclyn McCarty continued to study pneumonia infections; their mixing of DNA from virulent strains of pneumococci with harmless strains, and use of enzymes that digest DNA but not proteins, led to the cautious announcement that the 'transforming agent' of traits is made exclusively of DNA. This finding was held with skepticism until the breakthrough, Nobel prize-winning phage experiments of Alfred Hershey and Marsha 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. 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 information in the DNA is stored by pairs, with the relative amount of each kind of pair (adenine/thymine and guanine/cytosine) depending on the DNA source. Combining these crucial data with the X-ray fiber diffraction photographs of hydrated DNA taken by Maurice Wilkins and Rosalind Franklin that suggested a helical arrangement, Watson and Crick ingeniously elucidated the structure of DNA in 1953.

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. Though many other DNA forms besides the classic Crick and Watson (B-DNA) form are

---


6 A wonderful introduction to the rather reclusive Hershey, who died at the age in 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 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 *amine* group (=N-H₂) can tautomerize to an *imino* form (=N-H),
now recognized, including triplexes and quadruplexes, the B-form is still the most prevalent under physiological conditions.

RNA crystallography is at a more early stage, but has recently made quantum leaps with the solution of several RNA molecules (see Chapter 6) [40, 41]. These developments followed the exciting discoveries in the 1980s that established that RNA, like protein, can act as a catalyst in living cells. Thomas Cech, recipient of the 1995 U.S. National Medal of Science for his ground-breaking research with RNA enzymes or ribozymes, was co-winner of the 1989 Nobel Prize for chemistry for this discovery, with Sidney Altman.

The next two subsections elaborate upon the key techniques for solving biomolecular structures: X-ray crystallography and NMR. This section on experimental progress then concludes 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! 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 upon slow precipitation of supersaturated solutions. See [42, 43] 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.

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. Together, the amplitudes and phases of the diffraction data are used to calculate the electron density map; the

---

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.

Bragg’s law specifies the conditions for diffraction and the relation among three key quantities: \( d \) (distance between parallel planes in the crystal), \( \lambda \) (the wavelength of the X-ray beam), and \( \theta \) (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 \( \lambda \). The associated relationship is \( \lambda = 2d \sin \theta \).
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 [44] and the necessary mathemat-
ical analysis of the diffraction data limit the amount of accurate bio-
" molecular data available. Well-ordered crystals of biological macromolecules
are difficult to grow, because large globular molecules have irregular sur-
faces with cavities filled with disordered solvent molecules. 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 mathematical phase problem in crystallography [45, 46] involves re-
solving the phase angles $\phi_h$ associated with the structure factors $F_h$ when
only the intensities (squares of the amplitudes) of the scattered X-ray pat-
tern, $I_h = |F_h|$, are known. The structure factors $F_h$, defined as

$$ F_h = |F_h| \exp(i\phi_h), $$

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

$$ \rho(\mathbf{r}) = V^{-1} \sum_h F_h \exp(-2\pi i \mathbf{h} \cdot \mathbf{r}). $$

Here $\mathbf{r}$ denotes position, $\mathbf{h}$ identifies the defining planes of the unit cell, $V$
is the cell volume, and $\cdot$ denotes a vector product.

The phase problem was solved by direct methods for small molecules
(roughly $\leq 100$ atoms) by Jerome Karle and Herbert Hauptman in the late
1940s and early 1950s; they were recognized for this feat in the 1985
Nobel Prize in Chemistry. For larger molecules, biomolecular crystallogra-
phers 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 metals. The combination of diffraction patterns for the bio-
molecule, heavy metals, and biomolecule/heavy-metal complex offers more
information for estimating the desired phases. The differences in diffracted
intensities between the native and derivative crystals are used to pinpoint
the heavy atoms, whose waves serve as references in the phase determina-
tion 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
[42, last chapter]. Techniques to facilitate the phase determination process
by adding heavy metals and using multi-wavelength anomalous diffraction
(MAD) [47, 48] 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 [49], and improved methods for model refinement and phase determination are continuously being reported [50]. Such advances are leading to highly refined biomolecular structures\(^9\) (resolution \(\leq 2\ \text{Å}\)) at much greater numbers [51], even for nucleic acids [52].

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. 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). This requirement for nonoverlapping signals to produce a clear picture limits the protein sizes that can be studied by NMR to proteins with masses less than \(\sim 35\ \text{kDa}\) at this time. The biomolecular NMR future is bright, however, with novel strategies for isotopic labeling of proteins [53] and solid-state NMR techniques, the latter of which 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 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

\(^9\)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 detail that can be discerned.
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. 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 NMR data for structure refinement in the early days can be attributed to this formidable refinement task — formally, an overdetermined global optimization problem.

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 [54] for a historical perspective of biomolecular NMR and [55] 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 [56, 54, 57]. 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.

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, as well as to cut and paste DNA (the latter with the aid of an enzyme, ligase) and thereby create spliced DNA of desired transferred properties (e.g., 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 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 Chemistry Nobel Prize, with Michael Smith) and coworkers [58] 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 1.1.

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 [59]. Stronger X-ray sources, higher-frequency NMR spectrometers, and refinement tools for both data models are leading to these steady advances. The combination of instrumental advances in NMR spectroscopy and protein labeling schemes is suggesting that the size limit of protein NMR may soon reach 100 kDa [60, 53].

In addition to crystallography and NMR, cryogenic electron microscopy (cryo-EM) contributes important macroscopic views at lower resolution for proteins that are not amenable to NMR or crystallography (see Box 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 1.1: PCR Application Examples

- **Medical diagnoses of diseases and traits.** DNA analysis can be used to identify gene markers for many maladies, such as cancer (e.g., BRCA1/2, p53 mutations), 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 treated early and cured.

- **Historical analysis.** DNA is now being used for genetic surveys in combination with archaeological data to identify markers in human populations. Such analyses can discern ancestors of human origins, migrations patterns, and other historical events [61]. Historical analysis by French and American viticulturists also recently 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...

- **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 or another type of scene. In fact, since PCR was invented, dozens of prisoners have been exonerated, some from death row, and many casualties from disasters were identified from DNA analysis of assembled body parts. In this connection, personal objects analyzed for DNA — like a white glove or blue dress — made recent headlines as crucial ‘imaginary witnesses’ in the O.J. Simpson and Lewinsky/Clinton affairs. In fact, a new breed of high-tech detectives is emerging with modern scientific tools; see www.uio.no/~mostarke/forens_ent/forensic_entomology.html, for example, for the use of bugs in crime research, and the 11 August issue of *Science*, volume 289 for related news articles.

- **Family lineage / paternity identification.** DNA fingerprinting can also be used to match parents to offsprings. In 1998, DNA from the grave was used to confirm that President Thomas Jefferson fathered at least one child by his slave mistress, Sally Hemmings, 200 years ago. The remains of Tsar Nicholas’ family, executed in 1918, were recently identified by DNA, and French historians with European scientists solved in April 2000 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 was also used to prove false a paternity claim against Yves Montand.

---

*Time can be correlated with genetic markers through analysis of mitochondrial DNA or segments of the Y-chromosome. Both of these are genetic elements that escape the usual reshuffling of sexual reproduction, and hence their changes reflect random mutations that can be correlated with time.*

Box 1.2: Cryogenic Electron Microscopy (Cryo-EM)

Proteins that are difficult to crystallize or study by NMR may be studied by
cryogenic electron microscopy (cryo-EM) [62]. This technique involves imaging
rapidly-frozen samples of randomly-oriented molecular complexes at low tempera-
tures 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, new biological
insights may be gained, as demonstrated for the ribosome [63, 14] and the recent
cryo-EM solution of the 520-kDa tetramer of α-latrotoxin at 14 Å resolution [64],
as shown in Figure 1.3. (This solution represents an experimental triumph, as the
system is relatively small for cryo imaging).

This toxic protein in the venom of black widow spiders (so called because the cruel
females eat their mates!) forms a tetramer only in the presence of divalent cations.
This organization enables the toxin to adhere to the lipid bilayer membrane and
form channels through which neurotransmitters are discharged. This intriguing
system has long been used to study mechanisms of neurotransmitter discharge
(synaptic vesicle exocytosis), by which the release of particles too large to diffuse
through membranes triggers responses that lead to catastrophic neuro and
cardiovascular events.

With faster computers and improvements in 3D reconstruction algorithms, cryo-EM
should emerge as a greater contributor to biomolecular structure and function
in the near future.

1.5 Resulting 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 dozens 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 genome landmarks,
readers are invited to search the online collection available on
FIGURE 1.3. Top and side cryo-EM views of the 520-kDa tetramer of α-latrotoxin solved at 14 Å resolution [64]; images kindly provided by Yuri Ushkaryov.
The first completed genome reported was in 1995 (bacterium *Haemophilus influenzae*). Soon after came the yeast genome (*Saccharomyces cerevisiae*) in 1996 (see http://genome-www.stanford.edu/Saccharomyces/), the bacterium *Bacillus subtilis* in 1997, and the tuberculosis bacterium (*Mycobacterium tuberculosis*) in 1998.

Roundworm (1998)

The completion of the genome deciphering of the first multicellular animal, the one-millimeter-long roundworm *C. elegans*, made many headlines in December 1998 (see the 11 December issue of *Science*, volume 282, and http://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 comparison of 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 (e.g., a diabetes drug can be applied to worms with a mutation in the gene for the insulin receptor).

Fruitfly (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 issue of *Science*, volume 287, and http://www.fruitfly.org), in large part due to the ground-breaking “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 overestimated) 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.

**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%) published one year later, a major milestone for genetics. see the 14 December issue of *Nature*, volume 408, and http://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 more resistant to cold and grow faster and large 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 few “junk” (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.

Rice

The second largest genome sequencing project, for the rice plant, has been underway since April 1999 and is expected to run three years in its first phase. This project is a coordinated international effort orchestrated by the U.S. National Science Foundation and the United States Departments of Agriculture and Energy. 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, wheat, rye, and sugarcane. By May 2000, a rough draft (around 85%) of the rice genome (400 million bases) was announced, another model of cooperation between the St. Louis-based company Monsanto (now part of Pharmacia) and a University of Washington genomics team headed by Leroy Hood. Completion of the rice sequencing is expected in 2003.

Other Organisms

Dozens of genomes are now known, and the list grows quickly (see websites below 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 cholera, syphilis, tuberculosis, malaria, and typhus. Proteins unique to these pathogens are being studied, and disease treatments will likely follow (e.g., cholera vaccine).

Implications

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 [65] but also lead to medical breakthroughs.

Already, the practical benefits of genomic deciphering have emerged [66]. 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 [67].

Completion of the mammalian mouse genome sequence (rough draft produced by Celera in early 2001) is considered crucial for annotating the
human genome and also likely to be invaluable for testing new drugs for humans. The Mouse Sequencing Consortium (MSC) formed in late fall of 2000 follows in the footsteps of the human genome project and also reflects a collaboration between the private and public sectors.

1.5.2 The Human Genome

The International Human Genome Project was launched in 1990 to sequence all three billion bases in human DNA. 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 Saenger Center near Cambridge, UK, and four centers in the United States); see www.nhgri.nih.gov/HGP/. 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 competition to decode the human genome turned into a collaboration, 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).

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% is due to 11 gaps in contiguity); see the 2 December 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 to identify genes involved in the disease and, eventually, treatments. See a full account of the chromosome 21 story in http://www.sciam.com/explorations/2000/051500chrom21/.

Completion of the first draft of the human genome sequence project broke worldwide headlines on June 26, 2000 (see, for example, the July 2000 issue of Sci. Amer., volume 283). This draft reflects 97% of the genome cloned and 85% of it sequenced accurately, that is, with 5 to 7-fold redundancy.
The declaration of the ‘draft’ status was actually arbitrary and even fell short of the 90% figure set as target, but 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.”


Baltimore further explains that the number of proteins, not the genes, determines the complexity of an organism. This number should ultimately explain the complexity of humans, for whom the estimated number of total genes (50,000 in June 2000), not too far away from the number in a fly (14,000) or a worm (18,000). Interestingly, several months after the June announcement, the estimated number of human genes was reduced from 50,000 to 30,000–40,000 (see the 15 February issue of *Nature*, volume 409, and the 16 February issue of *Science*, volume 291), ‘equivalence’ of sorts between each human and roughly two flies...; the final word, however, on the number of human genes and the conserved genes that humans share with flies, mice or other organisms awaits further studies.

The public consortium projects completed version of the human genome in 2002 or before [68]. Some chromosome segments of the human genome will likely be impossible to characterize (at least with current technology) as they are too repetitive; fortunately, these segments may be relatively insignificant for the genome’s overall function.

Celeraw expects to generate a consensus human sequence from five individuals, from which variations (polymorphisms) will be determined (important for research on designer drugs). Their wider plan consists of sequencing the genomes of 1000 major species in the next decade, beginning with the mouse (see http://www.informatics.jax.org/), *Arabidopsis* mustard plant, and rice
in the near future; the chimp and cow genomes are especially interesting future candidates.

---

**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 gene, versus three billion base pairs for human DNA with at least 50,000 genes. The sequence of *Drosophila* followed in 1998 (140 million base pairs, 13,000 estimated genes) and released to the public in March 2000 (see March 24 issue of *Science*, volume 287).

This 'shotgun' approach has also worked for 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.

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.

---

**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 [http://www.ncbi.nlm.nih.gov/genome/guide/](http://www.ncbi.nlm.nih.gov/genome/guide/) (click on Map Viewer), the U.S. National Human Genome Research Institute’s site [www.nhgri.nih.gov/](http://www.nhgri.nih.gov/), that of Department of Energy (DOE) at
www.ornl.gov/hgmis/, the site of the University of California at Santa Cruz at http://genome.ucsc.edu/, and others. Since 1992, NCBI has maintained the GenBank database of nucleotide sequences, and many search and analysis tools have been developed to serve researchers.

As the sequencing of each new human chromosome is being completed, the biological revolution is beginning to affect many aspects of our lives [68], perhaps not too far away from Wilson’s vision of consilience. A ‘gold mine’ of biological data is now amassing, a rich resource not only for medicine and technology but also 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. Among the promising medical leaps are personalized and molecular medicine, perhaps in large part due to the revolutionary DNA microarray technology (see Box 1.4), and gene therapy. Of course, since information is not knowledge, but rather a road that can lead to perception, these achievements will require concerted efforts to extract information from all the sequence data concerning gene products.

Many societal, ethical, economic, legal, and political issues will have to be addressed with these developments. Still, like the relatively minor Y2K (Year 2000) anxiety, these issues could be resolved in stride through multidisciplinary networks of expertise. Visit www.nhgri.nih.gov/ELSI for more on the ethical, legal, and social implications of human genome research.

In a way, sequencing projects make the giant leap from sequence to function, but the 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. 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 [69].

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, for example [70].

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. Technical challenges remain concerning verification of the DNA sequences and ensuring their purity, amplifying the DNA samples, and quantitating the results accurately. 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 accurate 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) may 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 recently for the bacterium *M. tuberculosis*, based on experimental profiles obtained before and after exposure to the tuberculosis drug isoniazid [71].

For further information on microarray technology and available databases, see www.gene-chips.com, www.ncbi.nlm.nih.gov/geo, and industry.ebi.ac.uk/~alan/MicroArray, for example.