Chapter 2

The Impact Of Mathematics On Cellular and Molecular Biology

The application of mathematics to cellular and molecular biology is so pervasive that it often goes unnoticed. The determination of the dynamic properties of cells and enzymes, expressed in the form of enzyme kinetic measurements or receptor-ligand binding are based on mathematical concepts that form the core of quantitative biochemistry. Molecular biology itself can trace its origins to the infusion of physical scientists into biology with the inevitable infusion of mathematical tools. The utility of the core tools of molecular biology was validated through mathematical analysis. Examples include the quantitative estimates of viral titers, measurement of recombination and mutation rates, the statistical validation of radioactive decay measurements, and the quantitative measurement of genome size and informational content based on DNA (i.e., base sequence) complexity.

Several of the "classic experiments" in microbial genetics involved mathematical insights into experimental results. For example, the Luria and Delbruck fluctuation analysis, which clearly established that mutation was independent of selection, was a mathematical argument upon which a simple but elegant experimental design was based.

These examples are cited not to document the accomplishments of mathematical biologists but to bring focus to the fact that mathematical tools are intrinsic to biological fields. The discussion that follows focuses more clearly on the more sophisticated development of new mathematical concepts and statistical models to explain the complexity of biological systems. Biological complexity derives from the fact that biological systems are multifactored and dynamic.

Quantitative research in these fields is based upon a wide variety of laboratory techniques, with gel electrophoresis and enzyme-based assays among the most common. Measurements include activity, molecular weight, diameters, and sizes in bases, and with all these an understanding of the accuracy, precision, sources of variation, calibration, etc. In short, the quality of the measurement process is of central significance.

With the greatly increased amount of data being generated by laboratory techniques, and the pressure to move to more automated analysis, it is becoming even more important to understand the statistical aspects of these laboratory procedures. Such statistical work will involve the analysis of routinely collected data, the design and analysis of special studies, the development of new calibration and analysis techniques, and theoretical studies of the procedures in use, with an emphasis on robustness and the capacity to automate procedures. Furthermore, it will require a familiarity with the biology and the mathematical foundations of the analyses.

While the experimentalist strives to isolate single variables in order to make statistically significant measurements, many systems are not amenable to such single factor examination. Therefore, mathematically based computational models are essential to meaningful analyses. The goal of the present discussion is to provide a framework in which ongoing research in mathematical cell and molecular biology may be logically placed, and future opportunities can be described. This framework will provide for the analysis of the resource needs for future development and carries implications for current shortfalls. One factor is that undergraduate and graduate training in biology treats mathematics too superficially, especially in light of its role as an underpinning for quantitative research.

2.1 Accomplishments of the Past

2.1.1 DNA Structure

Differential geometry is the branch of mathematics that applies the methods of differential calculus to study the differential invariants of manifolds. Topology is the mathematical study of shape. It defines and quantizes properties of space that remain invariant under deformation. These two fields have been used extensively to characterize many of the basic physical and chemical properties of DNA. Specific examples of particular note follow.

The recent review of Dickerson (1989) summarizes how geometric concepts of tilt, roll, shear, propeller twist, etc. have been used to describe the secondary structure of DNA (i.e., the actual helical stacking of the bases that forms a linear segment of DNA). In addition, these concepts can be used to describe the interaction of DNA with ligands such as intercalating drugs (Wang et al. 1983).

From the time that closed circular DNA was discovered (see 3. below), it has been clear that such DNA exhibits both physical and chemical properties that differ in fundamental ways from those of related linear (or open circular) DNA. Using differential geometry and topology, both molecular biologists and mathematicians have been able to explain many of the properties of these molecules from two basic characteristics of the linking number: first, that it is invariant under deformations; and second, that it is the sum of the two geometric quantities, twist and writhe (White 1969). Among the major applications are:

a. the explanation for and extent of supercoiling in a variety of closed DNAs (Bauer 1978);
b. the analysis of the enzymes that change the topology of a DNA chain (Cozzarelli 1980, Wasserman and Cozzarelli 1986);
c. the estimation of the extent of winding in nucleosomes (Travers and Klug 1987);
d. the determination of the free energy associated with supercoiling (Depew and Wang 1975);
e. the quantitative analysis of the binding of proteins and of small ligands to DNA (Wang et al. 1983);
f. the determination of the helical repeat of DNA in solution and DNA wrapped on protein surfaces (White et al. 1988); and,
g. the determination of the average structure of supercoiled DNA in solution (Boles et al. 1990).

Topology, and in particular knot and link theory of closed space curves, has been used extensively to elucidate additional intertwining of closed DNA caused by catenation of two closed duplexes or knotting of a single duplex. In particular, the recent developments in polynomial invariants for links and knots have been used to describe the structure of DNA and to characterize the action of recombinases (Wasserman and Cozzarelli 1986, White et al. 1987).

2.1.2 Macromolecular Sequences

DNA sequences are collected in the GenBank database, and protein sequences are collected in the Protein Identification Resource (PIR). When a new DNA sequence is determined, GenBank is searched for approximate similarities with the new sequence. Translations of the DNA sequence into the corresponding amino acid sequence are used to search the protein database. Sensitive search methods require time and space proportional to the product of the sequences being compared. Searching GenBank (now more than 40ÊxÊ106 bases) with a 5000 bp sequence requires time proportional to 2 x 1011 with traditional search techniques. Lipman and Pearson (1985) have developed techniques that greatly reduce the time needed. Using their techniques, one can screen the databases routinely with new sequences on IBM PCs, for example. These methods rapidly locate diagonals where possible similarities might lie and then perform more sensitive alignments. This family of programs, FASTA, FASTN, etc., are the most widely used sequence analysis programs and have accounted for many important discoveries. An example of the impact of such analysis is the unexpected homology between an oncogene and a growth factor. This discovery became the basis of the molecular theory of carcinogenesis.

More sensitive sequence analysis can be obtained by dynamic programming methods. In part they are used after the diagonals are located in the FASTN and FASTA programs. Here similar sequence elements are aligned with positive scores and dissimilar elements are aligned with negative scores. Complicating the analysis are the insertions and deletions that also receive negative scores. The challenge of the problem is to arrange two sequences into the maximum scoring alignments. Additional difficulty arises from the fact that slightly similar regions of DNA or protein sequences might lie in otherwise unrelated sequences. In spite of the complex nature of the problem, an efficient algorithm (Smith and Waterman 1981) has been devised and is in wide usage.

The problem of sequence comparison creates a related statistical problem of estimating p-values (attained significance levels) for the alignment scores. The set of possible alignment scores from two sequences are dependent random variables since they result from overlapping sequence segments. Motivated by the problems of sequence comparison, investigators have refined and extended the Chen-Stein method (Arratia et al. 1989). This method is a powerful tool for approximating the distribution of sums of dependent indicator random variables by the Poisson distribution. In addition to sequence analysis, this method is being used in regression analysis and random graphs.

2.1.3 Genetic Mapping

Genetic mapping deals with the inheritance of certain "genetic markers" within the pedigree of families. These markers might be genes, sequences associated with genetic disease, or arbitrary probes determined to be of significance (e.g., Restriction Fragment Length Polymorphism [RFLP] probes). The sequence of such markers and probabilistic distance (measured in centiMorgans) along the genome can often be determined by hybridizing each family member's genome against the predetermined probes. In essence, the genetic map most likely to produce the observed data is constructed. Only a few years ago our knowledge of the mathematics involved and the computational complexity of algorithms based on that mathematics allowed us to analyze no more than five or six markers. As our knowledge of approximations to the formulas and likelihood estimation has improved, we have been able to produce software capable of producing maps for 60 markers or more (Lander and Botstein 1986). Progress in this area has been based on mathematical areas such as combinatorics, graph theory and statistics.

2.1.4 Cell Motility

Cells can move, monitor changes in their environment, and respond by migrating towards more favorable regions. It is a remarkable fact that a bacterial flagellum is driven at its base by a reversible rotary motor powered by a transmembrane proton flux, and analysis of models for this device has been prolific. The study of bacterial chemotaxis (the migration of bacteria in chemical gradients) has been particularly rewarding, in part because organisms such as Escherichia coli are readily amenable to genetic and biochemical manipulation, and in part because their behavior is closely tied to the constraints imposed by motion at low Reynolds number and by diffusion (of both the cell and the chemoattractant). Mathematics has helped us learn how a cell moves (Brokaw 1990, Dembo 1989), how it counts molecules in its environment (Berg and Purcell 1977), and how it uses this information (Berg 1988). It also has made it possible to relate the macroscopic behavior of cell populations to the microscopic behavior of individual cells (Rivero et al. 1989).

Studies of eukaryotic cell motility (and of the motion of intracellular organelles) has been revolutionized by in vitro assays in which motor molecules (myosin, dynein, kinesin) and the polymers along which they move (actin and microtubules) are linked to glass or plastic surfaces. Following the addition of ATP, one can observe, for example, the motion of individual actin filaments over a glass slide bearing only the heads of the myosin molecules. Statistical analysis is playing an important role in determining how such assays can be extended to the study of single motor molecules (Howard et al. 1989).

2.1.5 Structural Biology

Mathematics has made perhaps its most important contribution to cellular and molecular biology in the area of structural biology. This area is at the interface of three disciplines Ñ biology, mathematics and physics Ñ because its success has involved the use of sophisticated physical methods to determine the structures of biologically important macromolecules, their assembly into specialized particles and organelles, and even at higher levels of organization more recently. A wide array of methods has been employed, but we focus on the two most powerful of these, x-ray crystallography and Nuclear Magnetic Resonance spectroscopy (NMR), but with a mention of other methods.

Mathematics plays three roles. First, computational methods lie at the heart of these techniques because a large amount of information about local areas or short distances are encrypted in the raw data, and it is a major computational task to deduce a structure. Second, new mathematical methods of analysis are continually being developed to improve ways of determining the structure. Third, increasingly sophisticated computer graphics have been developed in response to the need to display and interpret such structure.

In crystallography the actual process of data collection has been enhanced by modern methods of detection (e.g., area detectors) and the use of intense synchrotron sources so that data collection per se is rarely rate limiting. Also, the use of modern techniques of recombinant DNA have greatly facilitated the isolation of material for crystallization. The rate-limiting step is often the preparation of isomorphous derivatives. As computational methods improve, fewer and sometimes no derivatives need to be analyzed.

Until the development of 2D-NMR in 1978 by Richard Ernst, the use of nuclear magnetic resonance for studying the structure of biological macromolecules was limited by the need to represent too much information in a limited space. With the pioneering development of the ability to represent NMR spectra in two frequency domains, it became possible to resolve the spectra of small proteins and oligonucleotides. A key benefit was that cross peaks, resulting from magnetic interactions of nuclei close to one another, could be measured. Since these cross peaks contained spatial information, there was an immediate movement to determine the structure of these molecules at atomic resolution. The technique has been remarkably effective. The structures of a number of proteins and oligonucleotides have been determined. The use of NMR to determine structures has proven to be an important complement to x-ray crystallography because the structures of many biologically important molecules (e.g., zinc fingers by Klevit 1991, Summers 1991, and Lee et al. 1991) have resisted attempts at crystallization; these structures must be studied in solution. The success of this technique has been critically dependent on mathematics beginning with the theoretical underpinnings by Ernst. The determination of structures is dependent on the mathematical technique of distance geometry that calculates all structures consistent with the distance constraints obtained from the NMR experiment. Other methods have included molecular dynamics and more recently the use by Altman and Jardetzky (1989) and Altman et al. (1991) of a Kalman filter to sample conformational space. There are significant limitations to 2D-NMR for structure determinations. First of all, the resolution obtained from NMR is less than that obtained from the best x-ray structures and is insufficient to see in detail active sites of biologically important molecules. A major mathematical challenge is to obtain such detailed structural information from structures that are basically underdetermined. One important approach is to use the structure to back-calculate the NMR data and by iteration improve resolution. A second limitation is that the determination of structures is limited to molecules with a weight of less than about 15,000. Better computational techniques could extend the limit.

One cannot over-estimate the importance of solving structures at atomic resolution. It has led directly to an understanding of the replication of DNA and its supercoiling in chromatin; the basis of protein and nucleic acid secondary, tertiary, and quaternary structures, how proteins act as enzymes and antibodies; and how electron transfer is achieved.

2.2 Grand Challenges

The grand challenges at the interface between mathematics and computation and cellular and molecular biology relate to two main themes: genomics, which is critical for example to support efforts at sequencing and mapping the human and other genomes, and structural biology, including structural analysis, molecular dynamic simulation, and drug design. These two areas have developed rapidly in the recent past because of the contributions of mathematics and computation, and they will continue to derive particular benefit from an enhanced interaction.

2.2.1 Structural Analysis of Macromolecules

The area of molecular geometry and its interface with visualization has been under-represented in research to date. This research, which would benefit from the involvement of geometers and would likely contribute to new mathematics, is a major limiting area in structural biology, especially in drug design and protein folding. As noted above, new methods will enhance the use of NMR for the determination of structures. Significant advances for solving mathematically the phase problem are being pursued. Important advances are being made in the field of computer-aided drug design.

Related to the structure of crystalline and hydrated proteins is the question of how proteins fold. For many proteins the folded structure and organelle formation (e.g., ribosomes) are dictated by the sequence. Reduction of the folding code has resisted intense efforts, but very recently important new approaches have been developed that have revealed significant new information. For example, two laboratories have shown that relatively short polypeptides can have significant secondary structure. This finding is important because it validates a piecemeal approach to protein folding, where secondary structure can be considered apart from tertiary structure. The second is the minimalist approach of DeGrado et al. (1989), in which model structures with predicted motifs are synthesized by chemical means. Experimental advances such as these, together with the explosive expansion of the available data and the development of more powerful decoding methods, means that members of families of protein folding codes will soon be readily identifiable. Once again this area requires mathematical innovation.

Finally, we note that microscopy is undergoing a technical revolution after a long period of plateau. Two new microscopes, scanning tunneling and atomic force microscopes, can yield a picture of macromolecules at atomic resolution. Actually, for these computer age microscopes, the picture is represented via a computer graphics display of digital data stored on optical media. Additionally, computational methods are the heart and soul of electron microscopic tomography. For example, using this technique, one can obtain four-dimensional information on chromatin structure (e.g., Belmont et al. 1989).

It is worth repeating that mathematical biologists in structural biology are in great demand. The theoretical work also is highly important and frequently has immediate payoff. The medical and commercial importance of structural biology is obvious.

2.2.2 Molecular Dynamics Simulation

Three-dimensional structures as determined by x-ray crystallography and NMR are static since these techniques derive a single average structure. In nature, molecules are in continual motion; it is this motion that allows them to function (a static molecule is as functional as a static automobile). Mathematical and computational methods have been able to complement experimental structural biology by adding the motion to molecular structure. These techniques have been able to bring molecules to life in a most realistic manner, reproducing experimental data of a wide range of structural, energetic and kinetic properties. Systems studied have extended from pure liquid water, through small solutes in water, to entire proteins and segments of DNA in solution.

The methods used for these calculations provide a glimpse of how simulation can be used generally in biology. Starting with a three-dimensional structure, a mathematical formulation for the forces between atoms gives the total force on each atom. These net forces then are used in Newton's second law of motion to give the accelerations, which are then integrated to give a numerical trajectory. The trajectory provides a complete description of the system, giving the position and velocity of every atom as a function of time. It is remarkable that simple forces and classical mechanics seem to give such a faithful picture of molecular motion.

At present, some of the most extensive molecular dynamics simulations have been used to study proteins and segments of DNA in solution. Such calculations involve tens of thousands of atoms and generate trajectories containing hundreds of thousands of structures changing with time; they require hundreds of hours of computer time, yet simulate periods lasting less than a nanosecond. As computer power continues to increase, it should be feasible to run simulations lasting microseconds (a billion time steps) and deal with the largest biological structures (a million atoms). In the limit of these longer time-scales, there is a natural connection with analytical and stochastic theories. Indeed, such theories provide essential checks on the numerical methods used to generate trajectories. An area ripe for this combined approach involves ionic channels, where molecular dynamic simulations can provide the frictional constants used in analytical treatments. This provides a direct link to the extensively studied phenomenological equations of nerve conduction (Hodgkin-Huxley equations). The molecular dynamics method gives a fully detailed description of the system simulated; this in turn provides a unique opportunity to visualize these molecular systems at work. Such visualization often is accomplished by making a motion picture of the system as it changes with time. Numerical analysis of the trajectories also is necessary to calculate properties that relate to experimental data. Better techniques for this analysis are sorely needed.

2.2.3 Drug Design

Molecules interact strongly when they fit together well. This occurs when their three-dimensional shapes are complementary and when there are stabilizing interactions (hydrogen bonds, charged pairs, etc.). One of the most interesting and potentially useful molecular interactions concerns drugs that bind with very high affinities to protein and nucleic acid macromolecules and either block the normal function of the macromolecule or mimic other ligands for such structures as receptors and induce a normal physiological response. Inhibition can be advantageous if the protein is made in excess, or if normal cellular control of the protein's activity has been lost. Because drug binding involves spatial complementarity, and because the aim is to design a molecule that binds with the highest affinity possible, it should be possible to use the three-dimensional structure to aid design. Current work in this area has followed several directions. The most direct approach is to crystallize the protein together with the drug. Study of the structure of the complex can suggest modifications to the drug expected to enhance its affinity for the receptor or enzyme active site. For this method to work, one needs an initial drug known to bind to the protein.

Other methods aim to circumvent this requirement by deducing the structure of the drug directly from the structure of the protein. While these methods are able to suggest completely new drug molecules, they involve a search for structures that fit a binding site. The theoretical underpinnings of such searches require further theoretical development. More specifically, they would benefit from application of better methods in global optimization and graph theory.

2.2.4 Nucleic Acid Sequence and Structural Analyses of Nucleic Acids

When a DNA sequence is determined, it is examined for a variety of sequence features known to be important: tRNA's, rNA's, protein coding regionsÑintrons and regulatory regions, promoters and enhancers. Since these sequence features are not identical in all organisms, it is often quite difficult to identify them. Even the widely studied bacterium Escherichia coli promoter sequences cannot be identified with certainty. As more and more DNA is sequenced, it becomes increasingly important to have accurate methods to identify these regions without many false positives. Statistics and mathematics should make significant contributions in this area.

As described above, pairwise alignment of sequences using dynamic programming is a well developed area. However, alignment of more than two sequences remains a serious problem, with high computation time. Some recent advances reduce the computation time so that 10 sequences might be practical, but many problems are not approachable. Heuristic methods that align by building up pairwise alignments have been proposed, but they often fail to give good multiple alignments. Closely coupled with multiple alignment is the construction of evolutionary trees. Closely related sequences should be neighbors with few changes between them.

In the area of DNA structure, several subareas are particularly amenable to mathematical analysis: (1) A complete analysis of the packaging of DNA in chromatin. Only the first order coiling into core nucleosomes is understood. By far the largest compaction of DNA comes from higher order folding. (2) Presentation of the topological invariants that describe the structure of DNA and its enzymatic transformations. The goal is to be able to predict the structure of interstate or products from enzymatic mechanisms and in turn to predict mechanisms from structure. (3) An analysis of the reciprocal interaction between secondary and higher order structures. This includes the phenomena of bending, looping, and phasing.

This work has implications for both biology and mathematics. Mathematics will be impacted in both topology and geometry. Renewed interest in the study of imbedding invariants for graphs has occurred because of the enumeration and classification topoisomers; the study of random knots has been used to study macromolecules in dilute solution, and tangle calculus and Dehn surgery theory have been used in the study of DNA enzyme mechanisms.

In the study of kinetoplast DNA, topology and the theory of interacting particles have been brought together in a unique way. Finally, in the study of DNA-protein interactions, theorems from differential geometry and differential topology have been recast in different frameworks to solve helical periodicity problems. The determination of the configuration of closed circular DNA brings together the fields of geometry and topology and non-linear partial differential equations, or topology and Monte Carlo techniques. These will involve extensive use of computational techniques including the creation of new codes to use non-linear partial differential equations to solve elasticity problems for closed circular rods.

2.2.5 Structural Analysis of Cells

Mathematical models have played, and will continue to play, an important role in cell biology. A major goal of cell biology is to understand the cascade of events that controls the response of cells to external ligands (hormones, transport proteins, antigens, etc.). The problem begins with understanding the interaction of the ligand with the cell's surface receptors. For some types of receptors, binding of the ligand to the receptor will lead to the generation of a transmembrane signal. For others, aggregation among the receptors must occur before a cell response can be triggered. The receptors themselves are under dynamic control, up or down regulating in response to external ligands, changing their rate of capture by coated pits, altering their recycling pattern, changing their rate for new receptor synthesis, changing their rate of delivery of old receptors to lysosomes for degradation, etc. The signaling pathways that are now being elucidated are equally, or more, complex. The role of mathematical models in studying these processes is to help rigorously test ideas about mechanisms and pathways, aid in analyzing experiments, determine parameter values, and help in the design of new experiments. Mechanistic models for some of the stages of the receptor pathway already have been developed, e.g., aggregation of receptors on cell surfaces (Dembo and Goldstein 1978, Perelson and DeLisi 1980), capture of receptors by coated pits (Goldstein et al. 1988), receptor-ligand sorting in endosomes (Linderman and Lauffenburger 1988), and have been useful in understanding receptor dynamics. Kinetic models have been used to analyze studies of ligand binding and internalization for a variety of receptor systems. With models it should be possible to dissect the relationship between structure and function. Thus, for example, a large number of mutants of the epidermal growth factor receptor have been generated. Determining whether the induced change in structure then affects ligand binding, tyrosine kinase activity, receptor aggregation, capture of the receptor by coated pits, etc. can best be done via collaborative experimental modeling efforts. A major challenge that lies ahead is to build mathematical models of specific cell types that incorporate all the known biochemistry, and that can be used to answer questions about the normal and disease states of the cell. Such an attempt is underway for the red blood cell (Yoshida and Dembo 1990), but here the effect of the biochemistry on the biomechanics of the cell also is important since the shape of the red blood cell is so critical for normal function. Predicting cell shape and the dynamic changes that occur in the cell's cytoskeleton due to interactions at the cell surface, which may lead to calcium influxes, receptor phosphorylation events, etc., are challenges for future models.

.