1

Statistical Contributions to Molecular Biology

Emmanuel N. Lazaridis and Gregory C. Bloom

1. Introduction

Developments in the field of statistics often parallel or follow technological developments in the sciences to which statistical methods may be fruitfully applied. Because practitioners of the statistical arts often address particular applied problems, methods development is consequently motivated by the search for an answer to an applied question of interest. The field of molecular biology is one area in which this relationship holds true. Even so, growth in application of statistical methods for addressing molecular biology problems has not kept pace with technological developments in the laboratory. Although the story of statistical contributions to the field of molecular biology is still unfolding, a consideration of its history can bring valuable insight into the hurdles—both technical and cultural—still to be overcome in interfacing the two fields. This is especially important given that recent technological advances have created a need for closer interaction among biologists and statisticians. Such considerations also motivated the selection of chapters for this text on statistical methods in molecular biology.

One important question to resolve at the start of such an exploration concerns the name of the field at the intersection of statistics and molecular biology. The term most widely employed by biologists, *bioinformatics*, seems quite inappropriate. The term *informatics* and its derivatives are commonly employed to describe studies of data acquisition and management practices. Evidence of this is the fact that until recently the bioinformatics literature was dominated primarily by computer science applications. Recently its literature has expanded to include areas of what has historically been called *computational biology*. In additional to computer applications, computational biology has

From: *Methods in Molecular Biology, vol. 184: Biostatistical Methods*
Edited by: S. W. Looney © Humana Press Inc., Totowa, NJ

historically focused on the interface of applied mathematics and molecular biology. As described in the following paragraphs, there are substantive reasons to differentiate statistical applications from these, as increased attention is paid to the stochastic nature of data. We considered the adjectival terms *statistical bioinformatics* and *statistical biology*, but discarded these as unsatisfactory owing to their ambiguity. *Statistical genetics* commonly refers to the application of statistical methods to the analysis of allelic data, by which genes directly related to an inherited condition are sought. We discarded this term as being too narrow. Although it includes a biological prefix, the term *biostatistics* has come to refer primarily to applications of statistics to medical research, and more specifically, to the conduct of medical studies. While many studies in molecular biology have eventual medical goals, only a minority have active clinical components. The term *biometry*, which for years has included statistical applications to the biological sciences in its definition, seems much superior. Because an argument can be made that biometry implies an emphasis on the measurement of biological phenomena, we slightly prefer the term *biometric modeling*, which explicitly recognizes the use of inferential models.

Our preference of the term biometry or a derivative such as biometric modeling is vindicated by the clear parallel between the use of images in molecular biology experiments and standard biometric applications such as aerial photographic surveys of wildlife populations. It is further supported by Stephen Stigler, the eminent statistical historian, who points out that by biometry Francis Galton and Karl Pearson meant "the application to biology of the modern methods of statistics" (*see* ***[1]*** for a more developed background concerning the field of biometry). Chapters 2 by Bloom et al., and 4 by Sieller-Moiseiwitsch et al., demonstrate especially well the substantive dependence of statistical models on processes of image analysis and quantitation as employed by molecular biologists in laboratory settings. Additional chapters relate biometric applications to biostatistical endeavors in the context of medical studies and clinical trials.

In this chapter we discuss biometric contributions to molecular biology research, and explore factors that have impeded the acceptance of these contributions by the field. We precede this discussion with a historical overview of the growth of molecular biology and of computational methods used therein.

2. Developments in Molecular Biology

Molecular biology encompasses the study of the structure and function of biological macromolecules and the relationship of their functioning to the structure of a cell and its internal components, as well as the biochemical study of the genetic basis for phenotypes at both cellular and systemic levels. Over the last half-century, such research has moved to the forefront of biology and medicine, so much so that molecular biology is sometimes called the "science of life."

Two major classes of questions have been posed by molecular biologists. These concern (1) evolutionary relationships within and across species of organisms and (2) issues of biological functioning in single cells and multicellular systems. Both sets of questions rely on describing and quantifying molecules that serve to characterize types of cells, cellular collections, or organisms, with regard to phenotypes or natural history. Molecules related to interesting phenotypes are called *biomarkers*. This book focuses on measurement and analysis of biomarker quantities because of their importance to our understanding of human disease. The process of identifying important and useful molecular markers is sometimes called *molecular fingerprinting*, *phenotyping*, or *profiling*, particularly when more than one marker is being considered simultaneously.

The idea that molecular fingerprints could be derived for use in characterizing cells and cellular collections has been around for quite some time. In 1958, 2 yr after the sequencing of insulin, Francis Crick recognized that "before long we shall have a subject which might be called 'protein taxonomy'—the study of the amino acid sequences of the proteins of an organism and the comparison of them between species. It can be argued that these sequences are the most delicate expression possible of the phenotype of an organism..." ***(2)***. Technological advances over the latter part of the 20th century continuing through the present day have substantially enlarged the availability of molecular data for phenotyping at a number of levels.

Even prior to Crick's remark, peptide fingerprinting techniques, whereby proteins were partially digested into amino acids and peptides and separated by chromatography or electrophoresis, had already been popularized as a means to seek evolutionary similarities across species. The sequencing of insulin, and soon thereafter, of ribonuclease and cytochrome *c*, opened the door to the large-scale characterization of organisms based on protein families. In 1967, the first computer algorithms were developed to seek phylogenetic relationships among a diverse assortment of organisms using a sizable database of protein sequence information. The programs generated binary trees based on a distance metric involving the mutational steps required to move from one protein to another ***(3)***. A "residue exchange matrix" was formed from distances between all pairs of measured organisms, and this was employed to represent the overall likelihood of mutations during evolution. In contrast to previous work on molecular data, this approach was very computationally intensive, requiring the use of a digital computer to seek the best binary tree to represent the calculated matrix. Thus was born the field of computational biology.

When Frederick Sanger published the basic chemistry for DNA sequencing in 1975, it became even more apparent that sophisticated database and analytic tools to work with sequence information would be needed. Sanger's approach

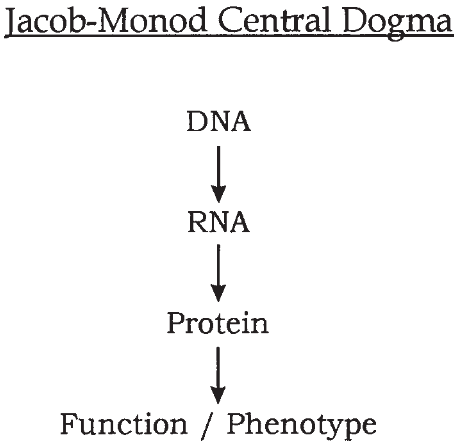

Fig. 1. Central dogma of molecular function and information flow.

is still the primary sequencing technology in use today. In this approach, nested sets of progressively longer DNA fragments are produced. These are then tagged with fluorescent dyes, separated by gel electrophoresis and scanned for identification of basepair sequences *(4)*. Applied Biosystems introduced the first automated sequencing system in 1986. With the addition of robotics to perform the preseparation reaction chemistry, the molecular biology laboratory would become increasingly automated, resulting in corresponding needs for database and analytic tools.

Development of the Southern blot in 1965 and of the Northern blot shortly thereafter marked the passage of another milestone in molecular biology. Biologists use the latter technique to measure the relative amount of RNA message being produced by any specific known gene for which a complementary target has been cloned. By the early 1970s, the basic techniques for studying cellular functioning at the molecular level had been established, so that this functioning could be investigated at each of the levels in the *central dogma* of molecular function and information flow, diagrammed in **Fig. 1**. This states that the direction of cellular information flow is primarily unidirectional, proceeding from DNA to RNA to protein. The genetic program, stored in the DNA, leads to the formation of biological macromolecules—RNAs and proteins—which through their biochemical function result in observable cellular behavior or phenotypes.

Of particular interest to molecular biologists is the manner in which groups of molecules interact to perform a specific function. A set of interacting molecules is said to form a *biological pathway*. **Figure 2** illustrates a particular pathway describing the signaling and functioning of signal transducers and

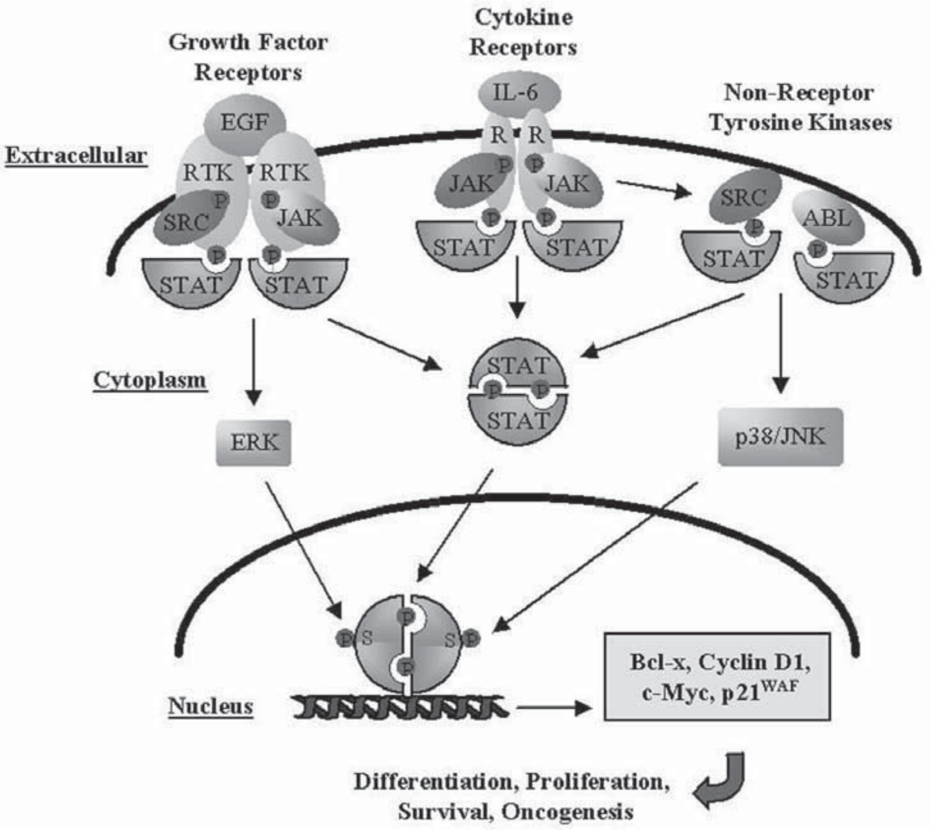

Fig. 2. Biological pathway describing the signaling and functioning of signal transducers and activators of transcription (STATs, *see* **ref. 5**).

activators of transcription (STATs). In brief, STATs are latent cytoplasmic transcription factors that are activated by cytokines, growth factors, and oncogenic tyrosine kinases. They are critical signaling molecules and have been linked to various cellular processes including proliferation, differentiation, and apoptosis. STAT proteins become activated by phosphorylation and dimerization, which allows them to translocate to the nucleus where they bind to specific DNA sequences and in conjunction with other factors control gene transcription. As a result of STAT activation, specific genes are known to be induced that contribute to regulation of cell cycle progression and survival of tumor cells. Experiments and methods that can identify and describe biological pathways such as the one shown in **Fig. 2** are very important to biologists.

The science of *genomics*—the study of the complete set of an organism's genes—was declared when the entire sequence of the bacterium *Haemophilus* *influenzae* was deduced in 1995. The sequence of baker's yeast, *Saccharomyces cerevisiae*, was completed in 1997. In 1998, a soil-dwelling nematode worm, *Caenorhabditis elegans*, was the first complex animal to have its genome sequenced. The Human Genome Project, with the ambitious goal of mapping the entire human genome in a matter of years, is ongoing.

In spite of this, studies employing one-gene-at-a-time technologies such as the Northern blot form the vast majority of the published investigations up through the present day. In part this is due to the expense and complexity associated with technologies for conducting multigene studies. A more fundamental barrier to the use of comprehensive profiling methods lies in molecular biology training, which emphasizes the full characterization of small networks of interrelated molecules. Recent initiatives by the NIH, such as the NCI Director's Challenge to discover molecular profiles important to cancer biology, seek to change this reticence.

Comprehensive methods for profiling at the levels of RNA and protein matured throughout the 1990s. Chapter 3 presents a history of the development of microarray technology, an extension of Northern blotting whereby an investigator can simultaneously measure the expression of thousands of genes whose complementary sequence has been arrayed on a glass slide or chip. Briefly, the spotted microarray was pioneered at Stanford University in the early 1990s ***(6)***. In this implementation, full-length cDNA corresponding to a known gene or an expressed sequence tag (EST) is layered onto a solid surface, usually a treated glass slide, using a robotic arrayer. Total RNA or mRNA is isolated from the sample of interest and labeled cDNA is constructed. The labeled cDNA is hybridized to the cDNA on the surface of the slide and visualized via the incorporated fluorescence tag. Currently, up to 30,000 genes and ESTs can be arrayed on a small glass slide using this technique. A second technology was pioneered by Steven Foder and colleagues in 1991 and has been further developed by Affymetrix ***(7)***. The Affymetrix approach uses photolithography and light-activated chemistry to array probes corresponding to different regions of a known mRNA transcript on a solid surface. By combining the signal intensity of the probe sets that query specific transcripts, values for gene expression are obtained. The study of molecular profiles at the level of RNA is sometimes called *transcriptomics*, to parallel the corresponding term at the level of the genome. Chapter 3 by Gieser et al. discusses methods for the design and analysis of microarray studies.

Similarly, *proteomics* is the science that deals with gene end-products, namely, proteins, concerning itself with the set of proteins (the *proteome*) produced by a particular cell, a cellular collection, or an organism. Important information can be derived from experiments seeking to establish whether specific proteins are made in higher or lower concentrations in response to disease, drug treatment, or exposure to toxicants. The most commonly used approach for studying a proteome is two-dimensional (2-D) gel electrophoresis, which combines a first dimension separation of proteins by isoelectric focusing with a second dimension separation by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE). The first separation is according to charge (different proteins are focused at their respective isoelectric points) and the second by size (molecular weight). The orthogonal combination of two separations results in a distribution of proteins from a biological sample across the 2-D gel. One or more images of the 2-D gel are collected and analyzed. Chapter 4 by Seillier-Moiseiwitsch et al. covers techniques for finding proteins in 2-D gel images.

To illustrate the potential power of comprehensive measures of gene functioning, consider measurements of mRNA in cancerous and noncancerous tumors or cell lines. Such measurements can elucidate which genes are relatively more or less active in one expression profile as compared to the other, giving an investigator the means to suggest which genes or groups of genes may be important in carcinogenesis. This approach can also be used to understand the effect of drug treatment on a particular tumor, whereby a researcher can investigate which genes change in expression and to what degree as a result of the drug (*see* Chapters 7 by Jones et al. and 9 by Dignam et al.).

3. Developments in Computational Biology

The rapid growth in availability of sequence data at the level of DNA particularly motivated the growth of computational biology. In this context, two basic sets of computational problems were addressed in the late 20th century, both of which were needed to derive possible biological functioning of molecular componentry using computers and mathematics. Each set of problems had implications at every level of molecular phenotyping: genomic, transcriptomic, and proteomic.

The first set of problems depended on the hypothesis that one could predict the functioning of a biological macromolecule if one could only predict its molecular structure. Work in the 1950s and early 1960s had demonstrated that the requisite three-dimensional (3-D) modeling of macromolecules was not feasible using physical, brass-wire models. Cyrus Levinthal demonstrated in 1965 that virtual, computer models of 3-D structures were substantially more amenable to exploration. For two reasons, the elegant theory that the information for the 3-D folding and structure of a protein is uniquely contained in its sequence of amino acids has proven unwieldy in practice. The first impediment has been that various computational complexities arise from its mathematics, whereby solutions require a long series of high-dimensional minimizations often surpassing available computer resources. Although the impact of this impediment could be reduced by providing sufficient correlative information derived from protein crystallographic studies, the cost and difficulty of these studies has been a second major hurdle. To illustrate the early state of this science, consider that the Biological Macromolecule Crystallization Database of the Protein Data Bank (a resource established in 1971 to collect, standardize, and distribute atomic coordinates and other data from crystallographic studies) contained only 12,000 verified protein entries at the time of this writing.

In contrast, perhaps 10 million basepairs are sequenced every day. The wide availability of primary sequence information suggested that tools for rapidly and confidently identifying homologous sequences among those contained in the increasingly large sequence databases could lead to substantive information concerning functioning, without relying on molecular structure. Various search and scoring procedures were developed and applied. For example, in the protein arena, the residue exchange matrix approach referred to previously was modified for use as a scoring matrix in sequence alignment procedures. In this case the matrix is used to determine the likelihood that two residues occur at equivalenced positions in a sequence alignment. Another example relates to the BLAST programs for fast database sequence searching. The name stands for Basic Local Alignment Search Tool. BLAST-style programs use a heuristic search algorithm, seeking to quickly search databases while making a small sacrifice in sensitivity for distantly related sequences ***(8,9)***. Databases are compressed into a special format, and the program compares a query sequence to each sequence in the database in the following manner. Sequences are first abstracted by listing exact and similar words within them. BLAST uses these abstracted words to find regions of similarity between the query and each database sequence, after which the words are extended to obtain high-scoring sequence pairs (HSPs). This approach was extended to include gaps in the alignments (GAPPED BLAST), and combined with the scoring matrix approach to increase the sensitivity of hits (PSI-BLAST).

The two approaches to predicting biological function of a molecule—based on its molecular structure or on its primary sequence—could also be combined. For example, it was noted that evolutionary mutations are not equally likely to occur at different positions in a protein, and that a single scoring matrix for all positions in the sequences to be aligned may be inadequate. Overington et al. ***(10)*** extended Dayhof's idea by using multiple matrices to reflect different mutation probabilities in different regions of a sequence. Similarly, Bowie et al. ***(11)*** created an exchange matrix for each position in the sequence. Inhomogeneous scoring approaches such as these require reference to a three-dimensional protein structure for at least one of the family members, in order to estimate the parameters of the exchange matrix. The discovery process could also be reversed. Sander and Schneider ***(12)*** described a procedure to estimate protein structure based on sequence profiles derived from multiple sequence alignments.

Toward the end of this chapter we mention several statistical approaches to sequence alignment and search problems as well as to problems of structural modeling; however, these approaches have had relatively little impact on molecular biology practices to date. In the next section we explore one barrier to the integration of statistical thinking into molecular biology practice.

4. Statistical Content of Academic Programs in Computational Biology

The growth of research in the field of computational biology was accompanied by the initiation of corresponding academic programs. As of October, 2000, the International Society for Computational Biology listed 44 university programs in bioinformatics and computational biology, 29 of these at 23 universities in the United States and Canada. Of these, 13 offer degree programs at the undergraduate or graduate level with required curricula in bioinformatics or computational biology.

To understand the curricular content of these training programs we created an *ad hoc* scoring system, which we used to rate the focus of the programs on a standardized, multidimensional scale. Curricula were obtained from eight graduate programs in the United States. Explicitly required upper-level prerequisite courses were included in this analysis. Each course from each curriculum was categorized according to whether its primary focus was most closely aligned with statistics, nonstochastic mathematics, physics/imaging sciences, biology, medicine, computer science, medical informatics, or bioinformatics. Bioinformatics courses were those that evidenced a multifaceted syllabus including biological databases, sequence searching and alignment technologies, and general techniques for genomics, transcriptomics, or proteomics. In the "other" category we included general seminars and courses on such topics as law and ethics. We based our assignments primarily on course syllabus, course title, the listing department or its code, and lastly on the training and interests of the primary faculty instructor when available over the web. Elective courses were valued at half the weight assigned to required courses in a curriculum, to reflect their relatively lower impact on training. Required upper-level prerequisite courses were given the same weight as required courses.

Although our scoring system is admittedly arbitrary, the results of our analysis are interesting in that they demonstrate that the statistical training offered to students in these programs is secondary to the trinity of required and recommended biology, mathematics, and computer science coursework. On average, statistics coursework accounted for about 10% of the curriculum of these training programs (range: 5–25%). The bulk of required statistics courses were either introductory or focused exclusively on probability theory. **Table 1** summarizes our analysis of the average curricular focus for these graduate programs.

Table 1
Average Percentage of Computational Biology Curricula Devoted to Various Subject Areas

Primary course focus	Percentage of curriculum
Biology	37.4%
Mathematics	14.0%
Computer Science	13.6%
Biostatistics/statistics	10.5%
Bioinformatics	9.8%
Physics/imaging sciences	3.4%
Medical	2.5%
Medical informatics	2.1%
Other	6.7%

The fact that, to date, the field of statistics has had relatively little impact on the practice of molecular biology is not unrelated to its relative underemphasis in computational biology coursework. Not only do regular biology programs provide even less quantitative training, but the most quantitative, modeling-oriented subset of biology is infused with a constructionist philosophy that contrasts sharply with the reductionist training that is the hallmark of statistics programs. *Constructionists* prefer the creation of reasonable models from consideration of the underlying science to immediate consideration of data, while *reductionists* prefer to generate a model from observed data, perhaps illumined by a small subset of scientific considerations, than to deal with the full-scale intricacies of the underlying science. The constructionist approach largely informs both mathematics and computer science training. Although it is true that computer scientists also perform "data mining" tasks, typical computer science approaches to data-driven analysis tend toward "black box" methods. Neural networks are prime examples of black box methods, as they typically result in little or no acquisition of generalizable, structural knowledge. It has been noted that there are benefits to be derived from the integration of both philosophies into applied work ***(13)***. Without the stochastic component, however, quantitation of uncertainty in inferential results is not possible. We believe it is essential that more statisticians be encouraged to support molecular biology studies, another key reason for providing this book as an introductory reference. As we describe in the next section, such statisticians will be rewarded by rediscovering the roots of their field.

5. A Brief History of Statistical Genetics

Having suggested that computer science and constructionist modeling approaches inform computational biology in practice, it is worthwhile to note that the history of statistics in evolutionary biology is a long one. The term *statistical genetics* has come to refer to the application of statistical methods to the analysis of allelic data, by which genes directly related to an inherited condition are sought.

Starting in the early 1870s, Francis Galton became widely known for his championing of eugenics, the science of increasing human happiness through the improvement of inherited characteristics. The creation of a science of eugenics required of Galton that he attempt to solve a number of complex problems, including that of how hereditary traits were transmitted by reproductive processes. Galton's energies gradually focused on statistical reasoning about hereditary processes. This work informed the so-called "Biometrical" school of thought, which believed that continuously varying traits exhibited *bleeding* inheritance. With the rediscovery of Mendel's work on the genetics of dichotomous traits in the early 20th century, fierce debates between Mendelians and biometricians focused on whether discrete and continuous traits shared the same hereditary and evolutionary properties. It is well known that this clash was influenced more by personalities than by facts, but it did serve to motivate further development in the statistical thinking needed to address genomics questions.

By the 1920s, the basic ideas of statistical genetics were developed by Fisher and Wright. These ideas formed a synthesis of the statistics, Mendelian principles and evolutionary biology needed to extend genomic modeling to continuously varying characteristics, which may or may not also exhibit polygenic (multilocus) etiology. These ideas were almost immediately embraced by plant and animal breeders. Extension into human models was developed theoretically, but its rewards would await later developments in computer science and molecular biology. The study of statistical genetics laid the foundation for many advances in theoretical and applied statistics, such as regression and correlation analyses, analysis of variance (ANOVA), and likelihood inference.

An excellent survey article by Elston and Thompson ***(14)*** breaks the study of statistical genetics into four major areas—population genetic models, familial correlations, segregation analysis, and gene mapping. Chapters 6–9 discuss some of these approaches. We note the availability of a software package called SAGE (Statistical Analysis for Genetic Epidemiology), a collection of more than 20 programs for use in genetic analysis of family and pedigree data. SAGE is available through the Human Genetic Analysis Resource of the National Center for Research Resources.

6. Biometric Modeling: Interfacing Molecular Biology and Statistics

There are three general areas in which molecular biology and statistics are interfacing at the present time. The first is in the context of clinical trials using molecular biomarkers. Most standard methodologies for the design and execution of clinical trials apply in this context. Chapter 9 by Dignam et al. in particular illustrates the use of a biomarker in a clinical trials setting. This setting is somewhat complicated by the fact that many such clinical trials involve multiple biomarker studies, resulting in little conclusive power for individual tests. An additional complication arises from the fact that many molecular biomarkers are quantitative summaries derived from images. Inter- and intraobserver variability associated with biomarkers is frequently studied prior to the initiation of a clinical trial, with methods such as those described in Chapter 5 by Looney; however, variability and bias resulting from image analysis is frequently ignored once a biomarker has entered the clinical trials situation. Chapter 2 by Bloom, et al. suggests a means whereby such information could be incorporated into these studies, increasing their generalizability.

The second area in which molecular biology and statistics are interfacing is in the context of the standard questions of computational biology, described previously, related to the functioning of biological macromolecules. Various statistical models have been proposed to perform sequence alignment ***(15–20)***. The work of Charles Lawrence and his colleagues is particularly interesting, in that they have employed Bayesian methods to address these problems. For example, the Bayesian solution to a product multinomial model has been proposed to perform multiple alignment, detecting subtle sequence motifs shared in common by a given set of amino acid or nucleotide sequences. One such model employs a Bernoulli motif sampler which assumes that each sequence could contain zero or more motif elements of each of a set of motif types. Starting with an alignment of motifs, the site sampler proceeds to follow two Gibbs sampling steps. First is a predictive update step that chooses one of the *N* sequences in order from first to last. The motif element for each motif type in the chosen sequence is added to the background and counts of discovered motifs are updated. Second is a sampling step, in which the probability associated with each possible motif starting position is estimated according to a model. Weighted sampling of a single motif element is then conducted for each motif type. This two-step process is repeated until a local maximum alignment has been obtained. Bayesian models have also been employed in the context of protein folding ***(21,22)*** and RNA structure ***(23)***. Although formal statistical modeling in problem areas in the traditional domain of computational biology has had little impact on molecular biology practice thus far, the fact that statistical model-based search methods are found to provide substantial improvement over current non-model-based methods ***(20)*** bodes well for the future.

A third area of interaction between statistics and molecular biology is slowly emerging because of recent advances in comprehensive, high-throughput laboratory methods for studies of gene expression at the levels of RNAs and proteins. Chapter 3 discusses methodological issues and approaches related to studies employing microarray technology, and Chapter 4 discusses approaches for 2-D protein gel analysis.

7. Conclusions

The infusion of statistical methods into the field of molecular biology promises to substantially enhance current scientific practices. Improved tools resulting in superior inference may be required to ensure important scientific breakthroughs. In this chapter we summarized statistical work in relation to the fields of molecular and computational biology, and explored some of the barriers still to be overcome. This book seeks to assist in the fusion of statistics and molecular biology practice by focusing on methods related to biomarker studies and molecular fingerprinting. We hope that it will prove useful to statisticians and biologists alike.

References

1. Stigler, S. (2000) The problematic unity of biometrics. *Biometrics* **56,** 653–658.
2. Crick, F. H. C. (1958) On protein synthesis. *Symp. Soc. Exp. Biol.* **12,** 138–163.
3. Dayhoff, M. O. (1969) Computer analysis of protein evolution. *Sci. Am.* **221,** 87–95.
4. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74,** 5463–5467.
5. Bowman, T., Garcia, R., Turkson, J., and Jove, R. (2000) STAs in Oncogenesis. *Oncogene* **19**, 2474–2488.
6. Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–470.
7. Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14,** 1675–1680.
8. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.
9. Karlin, S. and Altschul, S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* **87,** 2264–2268.
10. Overington, J., Donnelly, D., Johnson, M. S., Sali, A., and Blundell, T. L. (1992) Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.* **1,** 216–226.
11. Bowie, J. U., Luthy, R., and Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253,** 164–170.
12. Sander, C. and Schneider, R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9,** 56–68.

13. Lazaridis, E. N. (1999) Constructionism and reductionism: two approaches to problem-solving and their implications for reform of statistics and mathematics curricula. *J. Statist. Educ.* [Online] **7,** *http://www.amstat.org/publications/jse/secure/v7n2/lazaridis.cfm.*

14. Elston, R. C. and Thompson, E. A. (2000) A century of biometrical genetics. *Biometrics* **56,** 659–666.

15. Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1986) Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* **84,** 4355–4358.

16. Lawrence, C. E., Altschul, S. F., Bogouski, M. S., Liu, J. S., Neuwald, A. F., and Wooten, J. C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262,** 208–214.

17. Krogh, A., Mian, I. S., and Haussler, D. (1994) A hidden Markov model that finds genes in *E. coli* DNA. *Nucleic Acids Res.* **22,** 4768–4778.

18. Liu, J. S., Neuwald, A. F., and Lawrence, C. E. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Statist. Assoc.* **90,** 1156–1170.

19. Neuwald, A., Liu, J., Lipman, D., and Lawrence, C. (1997) Extracting protein alignment models from the sequence data database. *Nucleic Acids Res.* **25,** 1665–1677.

20. Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., and Chothia, C. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284,** 1201–1210.

21. Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992) A new approach to protein fold recognition. *Nature* **358,** 86–89.

22. Bryant, S. H. and Lawrence, C. E. (1993) An empirical energy function for threading protein sequence through the folding motif. *Proteins* **16,** 92–112.

23. Ding, Y. and Lawrence, C. E. (1999) A Bayesian statistical algorithm for RNA secondary structure prediction. *Comput. Chem.* **23,** 387–400.