Genomic Engineering: Moving Beyond DNA Sequence to Function

J. PATRICK FITCH, SENIOR MEMBER, IEEE, AND BAHRAD SOKHANSANJ, STUDENT MEMBER, IEEE

Invited Paper

The DNA sequence of the human genome has been determined. This significant scientific milestone has been a multidisciplinary effort sponsored by both public and private investments. Engineering and computer science contributions have been essential to success, especially with the accelerated schedules of the Human Genome Project. A tutorial summary of the biological and health implications of sequencing the human genome is presented together with examples of how genome data from human and other organisms are already used. Engineering contributions to sequencing are identified as well as predictions of how engineering methods may contribute to the "postsequencing" era of biology.

Keywords*—DNA, expression profiling, functional genomics, genomics, protein, proteomics, stochastic simulation.*

I. INTRODUCTION

It has been 135 years since Gregor Mendel observed that several distinct traits of peas were inherited at statistical rates predicted by the traits of the parents [1]–[3]. More recently, we have learned that deoxyribonucleic acid (DNA) contains the biochemical codes for the inheritance that Mendel observed. The DNA that is associated with a specific trait or function is known as a gene. The entire set of information represented in the DNA is known as the genome. This combines the word "gene" with the suffix "ome" for mass. In humans, our DNA is packaged in 23 pairs of chromosomes. Each parent contributes one chromosome to each pair in the genome. The chromosomes are designated by X, Y, and the numbers 1–22. In normal cells, there are two copies of the numbered chromosomes. The X and Y chromosomes determine sex with the pairs X-to-X and X-to-Y resulting in female and male children, respectively. Therefore, the chromosomes determine gender as well as many other traits. Deviations of DNA from "normal" are known as mutations and may be inherited and/or derived from interactions with the environment. Some mutations impact health. As an example, in the 1960s it was determined that the presence of an extra chromosome 21 causes Down syndrome. Curiously, Dr. Langdon Down first described this syndrome in 1866—the same year as Mendel's famous observation [4].

Mendel observed many inherited traits of peas. It was not until 1944 [5] that inherited genes and DNA were linked. It is very important to note that the DNA is responsible for far more than passing static information from parents to children regarding inherited traits. The DNA has a significant role in the biochemical dynamics of every cell. The DNA contains the parts list and assembly instructions for cell activities including metabolism, growth, and reproduction. Every cell in a human has the same DNA sequence in its chromosomes. Even cells with very different structure and function, like brain cells and liver cells, have the same DNA sequence. Developmental processes differentiate the cells, changing which genes are on and which are off. For bacteria that cause human disease, a few different genes in the bacterial DNA can determine if the organisms cause sickness rather than death. The dry mass of a typical cell is less than 5% DNA. However, the DNA controls the production of proteins that make up 75% of the dry mass. The information contained in the DNA influences when and how cells respond to environmental conditions through the production of proteins. DNA stores the "parts list" for protein structure and function, and it dynamically interacts with proteins to regulate the timing and amount of their production. Therefore, DNA is a logical place to begin decoding how cellular function works at the molecular level.

The program to determine the information in the human genome was begun in 1986 [6]. As recently as 1998, completion was planned for 2003 [7]. Based on significant technological improvements and increased public and private investments, the sequence of the human genome has been de-

Manuscript received June 16, 2000; revised September 20, 2000. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

The authors are with the Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, University of California, Livermore, CA 94550 USA (e-mail: jpfitch@llnl.gov; sokhansanj@llnl.gov).

Publisher Item Identifier S 0018-9219(00)10764-9.

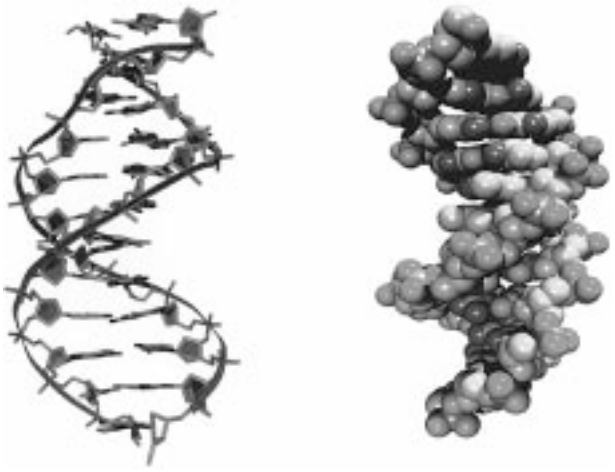
0018–9219/00$10.00 © 2000 IEEE

termined this year [8], [9]! These data provide the framework for revolutionary new biology, medicine, and human health and is available on the World Wide Web [10]. The accomplishment is a tribute to significant innovation in the life and physical sciences, engineering, and computing. In this paper, we summarize the significance of DNA sequence information, the need for additional data to increase our understanding of living systems, and the role of electrical and computer engineering in achieving these goals. We also examine the potential for engineering methods including simulation to integrate the current and emerging vast data sets so that our understanding of living systems can improve and we can apply the knowledge.

II. DNA Sequence Information

DNA is a macromolecule built from repeating subunits [11], [12]. The subunits are comprised of a nitrogenous base, a sugar, and a phosphate group generically denoted dNTP for deoxyribonucleotide triphosphate. The nitrogenous base is one of adenine (A), cytosine (C), guanine (G), or thymine (T) with the associated deoxynucleotides denoted dATP, dCTP, dGTP, and dTTP. The dNTPs can be joined along a sugar-phosphate backbone to form a single strand of DNA with the bases occurring in any order. The dNTPs and the strand have an orientation based on the orientation of the carbon atoms in the sugar. One end of the strand is designated five prime and the other three prime, 5′ and 3′, respectively. The list of bases in a strand of DNA is known as the DNA sequence and might appear as "5′-CGCGCTCCCTGAACC-3′." Single-stranded DNA is somewhat fragile and DNA usually occurs as a double strand with each nitrogenous base hydrogen bonding to a complementary base on the opposite strand. The base pairs in double-stranded DNA must occur as A-to-T or C-to-G. The strands are also antiparallel—e.g., "3′-GCGCGAGGGACTTGG-5′," for the earlier example. The two strands tend to twist into the familiar double helix shape associated with DNA shown in Fig. 1. In humans, each DNA molecule folds among various proteins into a compact package called a chromosome.

Genes produce messages like packets on an electronic network. The messages, called message (or messenger) RNA or mRNA, direct operations at network devices known as ribosomes. The ribosomes translate mRNA information into proteins (see Fig. 2). There is noncoding DNA called introns between genes that can be thought of as noise sources or unknown signals or messages that are no longer used. Some genes code for regulatory proteins that can inhibit the ability of other genes to put information on the "network." This may be accomplished by binding to the region of DNA that is needed to describe the protein and preventing messages from originating there. Regulatory proteins blocking the translation of specific gene messages at the ribosomes can also inhibit genes. And, proteins can be produced that bind to other proteins to inhibit or enhance function. These are only a few of the mechanisms to inhibit functional protein production that exist in cells. Regulatory proteins can also promote protein production by biochemically removing inhibitory proteins from DNA and making the gene available to the "network." Spatial organization within the cell can play a regulatory role as well, since mRNA and protein have to be transported to different locations in the cell to execute their functions. The inhibitors and promoters compete and complement each other in a complex feedback system that regulates protein production.



(a) (b)

Fig. 1. Two strands of DNA in the familiar double helix or twisted-ladder shape with a sugar-phosphate support backbone and nitrogenous base rungs. "Licorice and ribbons" (a) and space-filled (b) views of the Drew–Dickerson dodecamer, the first high-resolution measured crystal structure of B-DNA [13]. B-DNA is the dominant form of DNA under physiological conditions. Courtesy of D. Barsky, LLNL, using the VMD program (Humphrey) and with Rayshade 4.0 (Kolb and Bogart) and Raster3D (Merritt and Bacon).

Proteins are strings of amino acids assembled on a ribosome in an order specified by the DNA sequence from the associated gene. The DNA sequence information originates on chromosomes in the cell nucleus and is transmitted to the ribosome via single-stranded ribonucleic acid (mRNA). Three consecutive bases, known as a codon, specify which of the 20 amino acids in a cell is to be used next in the assembly of the amino acid chain. As shown in Table 1, there are 64 (4 cubed) possible codons that redundantly code for the 20 amino acids. Note that in RNA, thymine is replaced with the base uracil, i.e., T's in DNA becomes U's in RNA. RNA complements a template of single-stranded DNA with base pairs U-to-A, A-to-T, G-to-C, and C-to-G.

Deviations of genes from "normal" can be from inheritance and/or from exposure to environmental factors. For example, sickle-cell anemia is caused by a change in a single base of the DNA in an otherwise normal gene. Although it is just one base, the change causes a substitution of a single amino acid (valine for glutamine) in the protein that the gene encodes. The mutation results in abnormally shaped fragile red blood cells or "sickle cells." It is important to note that in many cases, changing only one base does not necessarily change the amino acid sequence of a protein, and changing one amino acid in a protein does not necessarily affect its structure or function.

In summary, the information stored in the DNA sequence of a gene is transcribed into a message of single-stranded

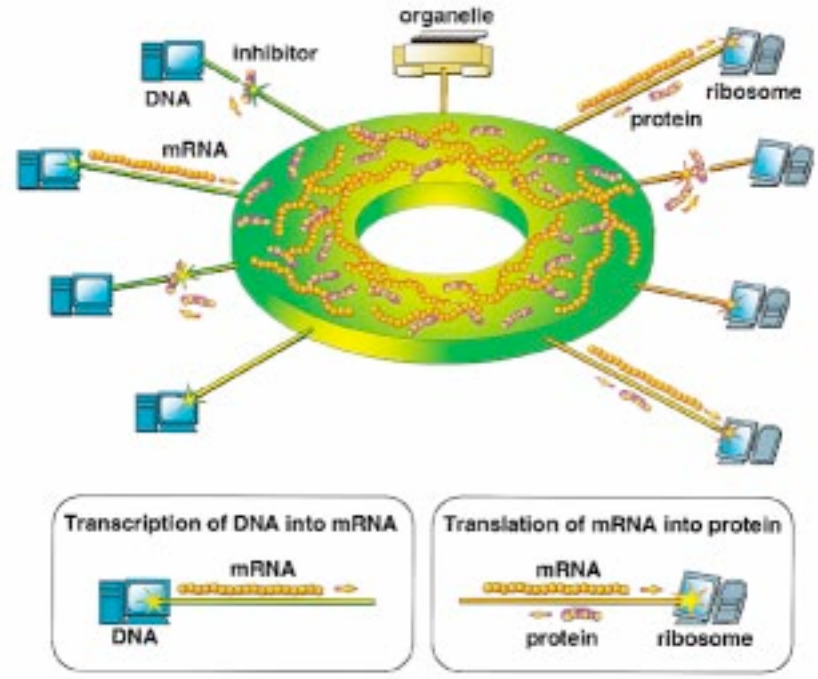


Fig. 2. Cells are organized and structured like an electronic network. Cells are not simply containers for biochemical reactions. Genes produce mRNA messages like packets on an electronic network. The messages direct operations at network devices known as ribosomes where information in the message is translated into proteins. Regulatory proteins known as inhibitors and promoters can modify message and protein production. The function of proteins and other molecules can change as interactions in the cell form complex macromolecules or degrade existing molecules.

RNA with complementary bases. The message RNA (mRNA) moves from the nucleus to the cytoplasm. On a ribosome in the cytoplasm, the mRNA is translated one codon (three bases) at a time into a code for one of 20 amino acids. The amino acids designated by the codons are chained together until a full-length protein is formed (see Fig. 3). Proteins are the workhorses of the cell. DNA and the environment control the quantity, timing, and selection of proteins expressed. As summarized in Fig. 4, producing proteins from the information stored in the chromosomes is mediated by a network system that includes feedback from biochemical inhibitors and promoters. This schematic is incomplete and we will add details later. The first step in providing a comprehensive understanding of this complex network is to obtain the entire human DNA sequence.

III. DNA Sequencing

DNA sequencing has become a high-throughput process for determining the ordered base pairs in a strand of DNA. Manufacturing techniques, including statistical process control, are now routine. An example throughput metric is the number of DNA bases sequenced per day per dollar. Some sequencing centers report daily and monthly production online [14]–[16]. In most DNA sequencing approaches, the DNA sequence is assembled from many shorter, overlapping subsequences. The sequence of a strand less than a couple thousand base pairs in length is measured using four-color electrophoresis. Our description of DNA sequencing begins with a simplified description of sequencing chemistry followed by discussions of electrophoresis, computerized base calling and assembly, and concludes with an overview of automation used to increase throughput and reduce cost. Submission of the sequence into the public DNA sequence database is the final step for most publicly funded sequencing projects.

Beginning with a purified template of single stranded DNA, the second complementary strand is generated using an enzyme known as DNA polymerase. Deoxynucleotides (dNTPs) for each of the bases are provided in solution so that the polymerase enzyme can assemble them along the template. First, a small chain of dNTPs, called a primer, is annealed to the template DNA. The primer is designed to be at a unique reference position on the template. Assembly of the complementary strand begins at the primer site and continues toward the five-prime ($5'$) end of the template. As each dNTP is added, a $3'$-hydroxyl group is left available for the next dNTP in the growing complementary second strand. The clever modification of a dNTP so that no $3'$-hydroxyl group is available for chain extension provides a means to terminate strand assembly. These synthesized molecules, known as dideoxynucleotides or ddNTPs, can also be labeled with a fluorescent dye specific to the base. By balancing the concentration of dNTPs and ddNTPs, an ensemble of DNA strands beginning at the same position on the template DNA but of different lengths can be generated. These strands are also terminated with a fluorescent label specific to the final base in the chain. The two strands are separated using temperature and/or biochemical techniques. Once the "chain termination" or "Sanger se-

Table 1
The 64 3-Base Codons (5′ to 3′ DNA) With the Corresponding Message RNA (5′ to 3′ mRNA) and the 20 Amino Acids. The Single Letter Abbreviations are Only Used in Long Lists. Note that the DNA Corresponds to the 5′ to 3′ Gene and so the mRNA Bases are Identical with the T to U Substitution. The mRNA is the Complement of the 3′ to 5′ DNA Strand that Participates in mRNA Transcription. Several Codons May Also Serve to Designate the Start (ATG) or Stop (TAA, TAG, TGA) of a Gene-Coding Region

DNA	mRNA	Amino Acid	DNA	mRNA	Amino Acid
AAA	AAA	K-Lys-Lysine	GAA	GAA	E-Glu-Glutamic Acid
AAC	AAC	N-Asn-Asparagine	GAC	GAC	D-Asp-Aspartic Acid
AAG	AAG	K-Lys-Lysine	GAG	GAG	E-Glu-Glutamic Acid
AAT	AAU	N-Asn-Asparagine	GAT	GAU	D-Asp-Aspartic Acid
ACA	ACA	T-Thr-Threonine	GCA	GCA	A-Ala-Alanine
ACC	ACC	T-Thr-Threonine	GCC	GCC	A-Ala-Alanine
ACG	ACG	T-Thr-Threonine	GCG	GCG	A-Ala-Alanine
ACT	ACU	T-Thr-Threonine	GCT	GCU	A-Ala-Alanine
AGA	AGA	R-Arg-Arginine	GGA	GGA	G-Gly-Glycine
AGC	AGC	S-Ser-Serine	GGC	GGC	G-Gly-Glycine
AGG	AGG	R-Arg-Arginine	GGG	GGG	G-Gly-Glycine
AGT	AGU	S-Ser-Serine	GGT	GGU	G-Gly-Glycine
ATA	AUA	I-Ile-Isoleucine	GTA	GUA	V-Val-Valine
ATC	AUC	I-Ile-Isoleucine	GTC	GUC	V-Val-Valine
ATG	AUG	M-Met-Methionine	GTG	GUG	V-Val-Valine
ATT	AUU	I-Ile-Isoleucine	GTT	GUU	V-Val-Valine
CAA	CAA	Q-Gln-Glutamine	TAA	UAA	Stop Codon
CAC	CAC	H-His-Histidine	TAC	UAC	Y-Tyr-Tyrosine
CAG	CAG	Q-Gln-Glutamine	TAG	UAG	Stop Codon
CAT	CAU	H-His-Histidine	TAT	UAU	Y-Tyr-Tyrosine
CCA	CCA	P-Pro-Proline	TCA	UCA	S-Ser-Serine
CCC	CCC	P-Pro-Proline	TCC	UCC	S-Ser-Serine
CCG	CCG	P-Pro-Proline	TCG	UCG	S-Ser-Serine
CCT	CCU	P-Pro-Proline	TCT	UCU	S-Ser-Serine
CGA	CGA	R-Arg-Arginine	TGA	UGA	Stop Codon
CGC	CGC	R-Arg-Arginine	TGC	UGC	C-Cys-Cysteine
CGG	CGG	R-Arg-Arginine	TGG	UGG	W-Trp-Tryptophan
CGT	CGU	R-Arg-Arginine	TGT	UGU	C-Cys-Cysteine
CTA	CUA	L-Leu-Leucine	TTA	UUA	L-Leu-Leucine
CTC	CUC	L-Leu-Leucine	TTC	UUC	F-Phe-Phenylalanine
CTG	CUG	L-Leu-Leucine	TTG	UUG	L-Leu-Leucine
CTT	CUU	L-Leu-Leucine	TTT	UUU	F-Phe-Phenylalanine

quencing" chemistry [17] has been completed, the DNA sequence can be obtained by ordering the new strands by size and fluorescent label (see Fig. 5).

One popular alternative to dye terminator chemistry uses fluorescent labels at the primer site and is known as the dye primer method. The template DNA is separated into four aliquots and a fluorescent label is incorporated as part of the primer. Chain extension using dNTPs is performed as described above. However, chain termination differs in that ddNTPs for only a single base are used in each aliquot and the ddNTPs are not labeled. After chain extension and termination in the dye primer method each aliquot has fragments labeled at the primer site and terminated at the same base. The DNA sequence can be obtained by ordering the new strands by size in each aliquot separately and then computationally combining the results from the four aliquots. If a different color primer label is used in each of the four aliquots, the aliquots can be pooled before DNA sizing and the process continues as with dye terminator chemistry. One of the principal advantages of dye terminator chemistry is the ability to do the chain extension in a single aliquot.

The method of choice for determining the size of the DNA strands is four-color electrophoresis [18]–[22]. Electrophoresis to separate biomolecules began with the Nobel Prize winning work by Tiselius on proteins in 1937. In DNA sequencing, electrophoresis uses the force from an applied electric field to move the negatively charged single-stranded DNA molecules through a separation medium. DNA has roughly a constant charge to mass ratio. The sieving medium and the electric field are engineered to produce differential drift velocities proportional to the length of the DNA usually for DNA less than one thousand bases long. Limitations arising from diffusion and convection led to the use of polyacrylamide or agarose sieving media in many instruments [23], [24]. High-throughput instruments have utilized several approaches including gels spread thinly across slabs of glass and gels injected into glass capillary arrays, glass microchannels and plastic microarrays [25]–[36]. Each of these types of instruments is in use today with instruments using arrays of glass capillaries currently dominating sequencing in the large centers. We describe the electrophoresis process with a bias toward the capillary

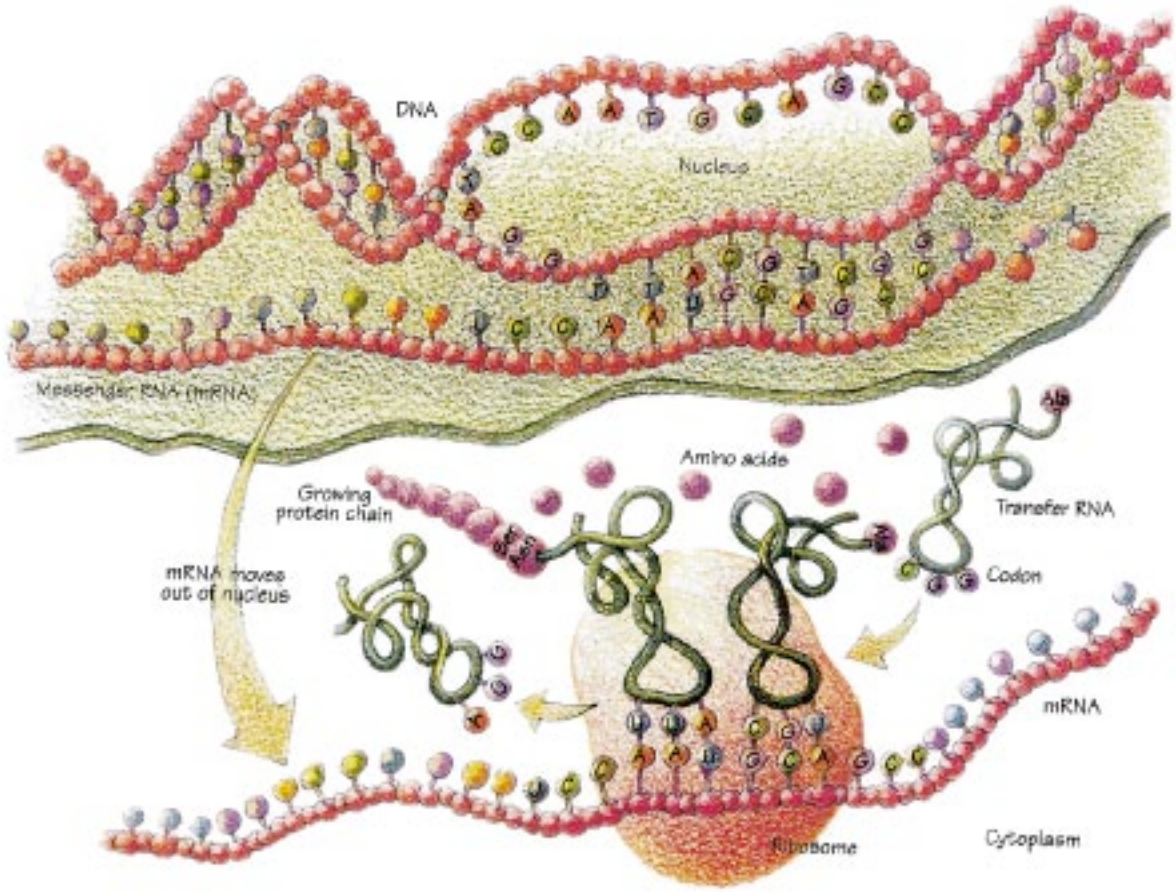


Fig. 3. Proteins are chains of amino acids assembled in an order specified by the sequence of DNA bases located in chromosomes in the cell nucleus. Single stranded RNA molecules are the messages that move from the nucleus to the ribosomes. The ribosomes assemble proteins by matching three-base sets (codons) in message RNA with complementary codons on transfer RNA (tRNA) attached to individual amino acids [6].

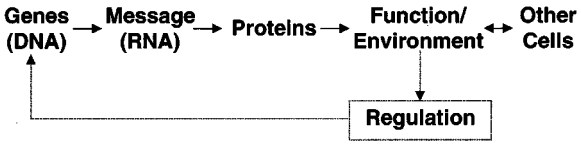


Fig. 4. Schematic showing the transcription of information from genes (DNA) to message (RNA) and the translation of mRNA to proteins. The environment and the DNA change the scheduling of protein production through a complex feedback network.

systems. Fig. 6 provides a generic capillary electrophoresis DNA sequencing instrument schematic.

Before DNA can be loaded into an electrophoresis instrument, gel is pumped from the data collection end of the system into the capillaries using a syringe-type pump or compressed gas at over 1000 psi. Usually several capillary-volumes are pumped through the system. The excess gel is aspirated from the sample-end of the capillaries. The loading well around each capillary entrance is then filled with a buffer solution. The sample (usually a few microliters) is introduced into the loading buffer. The goal in DNA loading is to create a thin stack of DNA in the gel [37]. If the stack spreads out before electrophoresis, the resolution of the system degrades and fewer DNA bases can be deduced from the run. In electrokinetic injection, an electric field (anode at the detection end of the system) is applied and the negatively charged DNA in the sample is moved into the gel in the capillary. The loading buffer is then aspirated out of the system and a running buffer is introduced to promote migration of the DNA down the capillary. Commercially available sequencing instruments now require very little operator intervention. In one of the commercial systems, the Applied Biosystems 3700 DNA Analyzer [38], a robotic arm performs sample loading and some of the aspiration operations [39]. The input sources for all high-throughput commercial systems are standard laboratory 96 or 384 well plastic microtiter plates.

An innovative alternative to injecting the sample into the gel electrically is to create a narrow cross channel (see Fig. 7) that moves DNA across the gel in the sequencing channel [40]. After loading, the stack of DNA in the channel is roughly the width of the cross channel. After isolating the cross channel electrically, electrophoresis begins in the sequencing channel. Although electric fields have been used

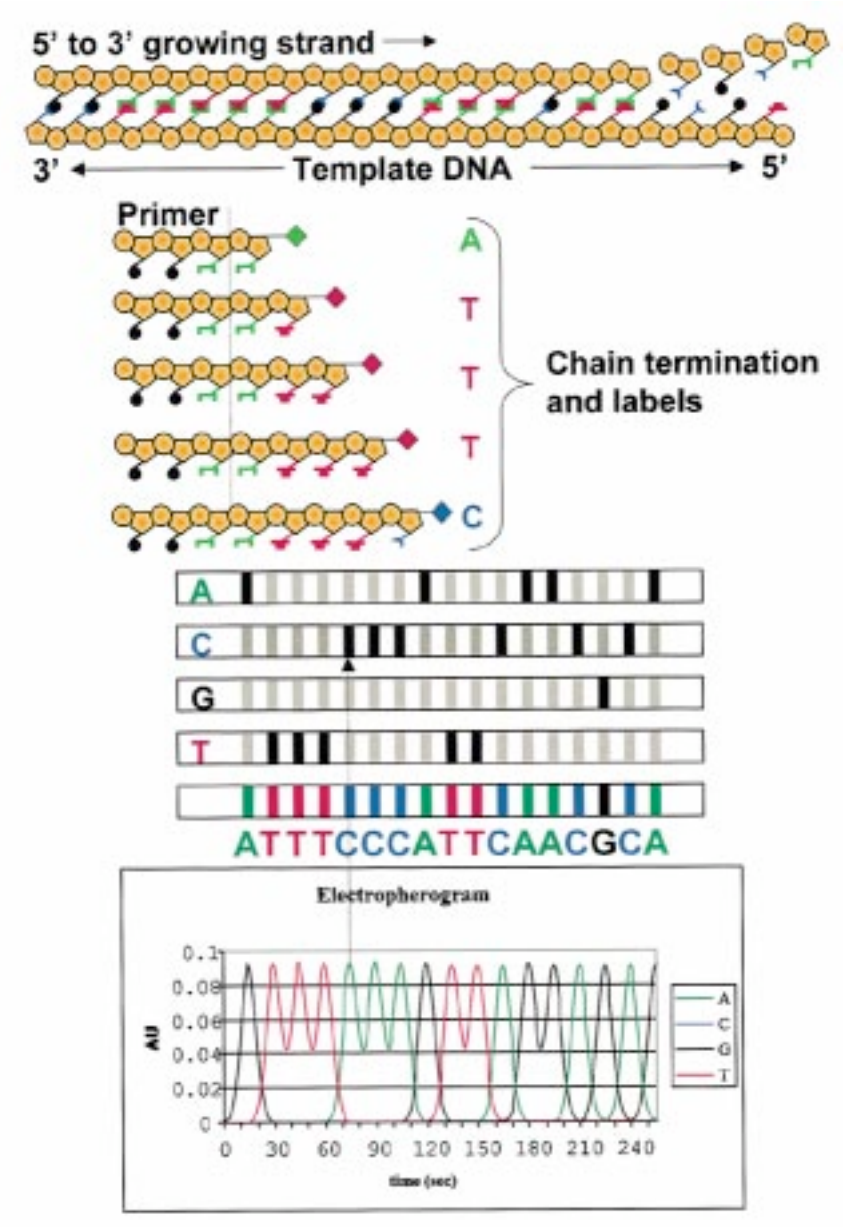


Fig. 5. DNA sequencing using four-color electrophoresis and the Sanger chain termination chemistry. A template of DNA is copied into many random length pieces of DNA that start at the same primer and terminate with an optical label specific to the last base in the chain. The strands are separated by length using electrophoresis allowing the DNA base sequence to be deduced.

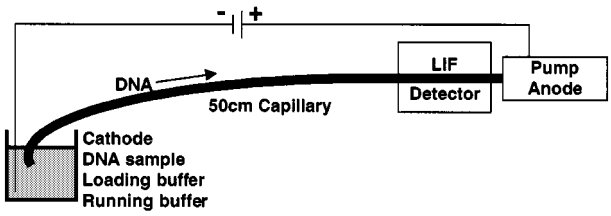


Fig. 6. Capillary-based electrophoretic DNA sequencing instrument with Laser Induced Fluorescence (LIF) detection system.

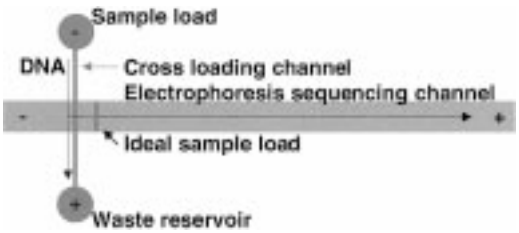


Fig. 7. Cross-channel loading for microchannel electrophoresis.

to move the DNA for this loading scheme, the geometry has mechanically isolated the DNA for electrophoresis. The cross-channel loading has been implemented in several systems using micro electro mechanical systems (MEMS) techniques including lithographic patterning of the channel and cross-channel. For a recent review of MEMS applied to genetic diagnostics including cross-channel loading, see [34]. Compared to electrokinetic sample loading, cross-channel loading requires additional electrical circuitry.

The voltage and current must be controlled in both the main and the crossing channels to minimize diffusion of the sample into the gel. Although it has had several promising demonstrations, it is currently not available in any commercial system.

Once the DNA sample is loaded into the gel in the capillary, an electric field of 100–200 V/cm is applied to move the DNA through the gel toward the detector. The shorter fragments and surplus primer from the enzymatic reaction arrive first at the detector. Surplus template DNA without attached fluorescent labels can contaminate a run by arriving at the detector at the same time as shorter DNA fragments that have an attached label that slows migration. It is also possible for the single-stranded DNA to fold on itself and cause poor electrophoretic separations. As with aliasing in an analog to digital sampling system, it is not possible to determine from the electrophoresis data alone if a peak is due to a short fragment or a longer fragment that has folded on itself and migrates faster than it would without the fold. To reduce some of these noise sources, the samples are often purified before loading, the gel and buffer chemistries are engineered to keep single-stranded DNA from hybridizing to complementary DNA strands, and the temperature and running conditions are optimized for electrophoretic resolution. As an example, the Applied Biosystems 3700 DNA Analyzer often loads 2 μl from a 25-μl source in microtiter format with a 30-s electrokinetic load at 1 kV. This represents approximately 20 ng of DNA loaded onto the column. The run voltage is often 6.5 kV. The 50-cm long and 50-μm inner diameter capillaries are filled with a polydimethylacrylamide (PDMA) sieving gel. The run duration is about 2 h for 500 bases with single base resolution [38], [39].

A standard measure of resolution of DNA sequence is the ratio of the peak width to the peak spacing. The peak width is usually taken as the full width at half the maximum value. The number of bases where this resolution is unity is known as the crossover point for the system (loosely similar to the Rayleigh diffraction limit of an optical system). Signal-to-noise ratio and other parameters influence performance, but the crossover point is a good measure of the inherent capability of an instrument to resolve DNA fragments differing in length by one base. When the DNA fragments are resolvable, the fluorescent labels on the terminating ddNTP will designate the last base on the fragment of interest and there will be many copies of that label (one for each DNA fragment) allowing for optical detection. For most of the commercially available ddNTP labels, an Argon ion laser (488- and 514-nm wavelengths) induced fluorescence (LIF) system is used for detection. The optics of the two most common detection systems are a scanning confocal microscope with photomultiplier tubes [41] and a fixed CCD imaging system that collects multiple wavelengths simultaneously through a prism [42]. The detection system can operate through the glass capillary or through a liquid that creates a "sheath flow" around the end of the capillary [43]. Although mechanically complex, the sheath flow eliminates refraction through the capillaries and allows the laser to illuminate many channels simultaneously without scanning. Numerous other detection methods have been proposed including electrochemical [44], [45], but they have not been adopted in commercially available sequencers.

The fluorescent labels used in most DNA sequencing instruments have emission spectra that overlap. Example spectra are presented in Fig. 8. Color correction is needed before beginning data analysis to detect DNA bases. Other characteristics of the electrophoretic separation that must be corrected include length-dependent changes in velocity of the DNA fragments and velocity differences due to the four different fluorescent labels. For a fixed detector system, the DNA that arrives later appears to have a broader distribution. This is due to the slower velocity and not necessarily a more spatially dispersed ensemble of DNA fragments. Corrections for the velocity dependent artifacts are referred to collectively as mobility correction [18]. In general, there are two approaches for color and mobility correction. The first approach is system calibration with known samples using the same loading and running parameters as will be used with the unknown samples [46]. The second approach is system compensation by estimating the parameters of a correction model dynamically from the data [47]. Because of the high-throughput nature of DNA sequencing, the system parameters remain fixed for many runs making the first approach (the use of calibrated test runs) preferred. Ideally the traces from each of the DNA fragments would have the same shape, samples would be evenly distributed by size in the electropherogram, and the fluorescent labels would not overlap spectrally. Unfortunately, the electropherogram has many distortions, and the signal environment is similar to a digital communication system with fading channels and crosstalk. When the biochemistry or temperature is suboptimal, the folding of the single strand of DNA can also change the electropherogram as if the signal had multipath artifacts.

Fig. 9 shows an electropherogram before and after color correction, background subtraction, and filtering for mobility and shape correction [48]. The ultimate metric to compare against is not the homogeneity of the electropherogram, but rather the accuracy of the assignment of DNA bases. This process is known as base calling and the state of the art has been defined by the early work in industry for supporting the ABI 373 sequencing instrument and more recently by Phil Green's group with codes named Phred and Phrap [49], [50]. The Phred and Phrap codes are arguably the gold standards for base calling, assessing the quality of a base call and assembling sequence data from many DNA fragments into an estimate of a longer contiguous DNA sequence. The code performs all of the necessary filtering mentioned above, does peak tracking to identify potential locations of bases in the electropherogram, and then performs a model fit to call a base and to look up a probability of error in a calibrated table for the particular instrument. "Phred 20 bases" has become the industry standard for identifying the number of bases that have roughly a 1 in 10 000 probability of error.

Electrophoresis allows determination of the sequence of the template DNA of lengths of around 1000 bases. So how are entire genomes with over a billion bases sequenced? The sequence of longer segments of DNA is assembled from

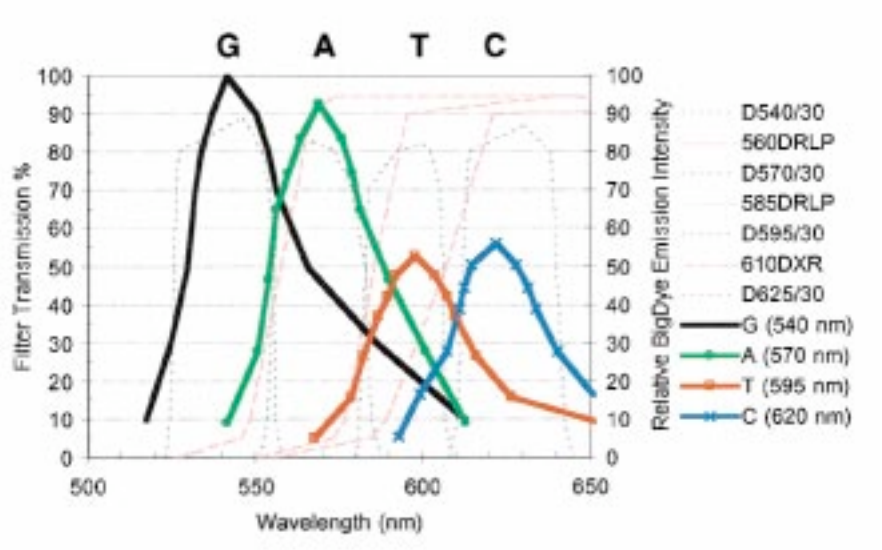


Fig. 8. Example spectra for four-color fluorescent DNA labels and the optical transmission characteristics of the LLNL microchannel DNA sequencer. The dichroic mirrors are designated xxxD, where xxx is the cutoff wavelength in nm. The optical bandpass filters are designated Dxxx/xx, where xxx is the center wavelength and xx is the bandwidth in nm. The fluorescent terminator labels were sold under the trademark PE Big Dye.

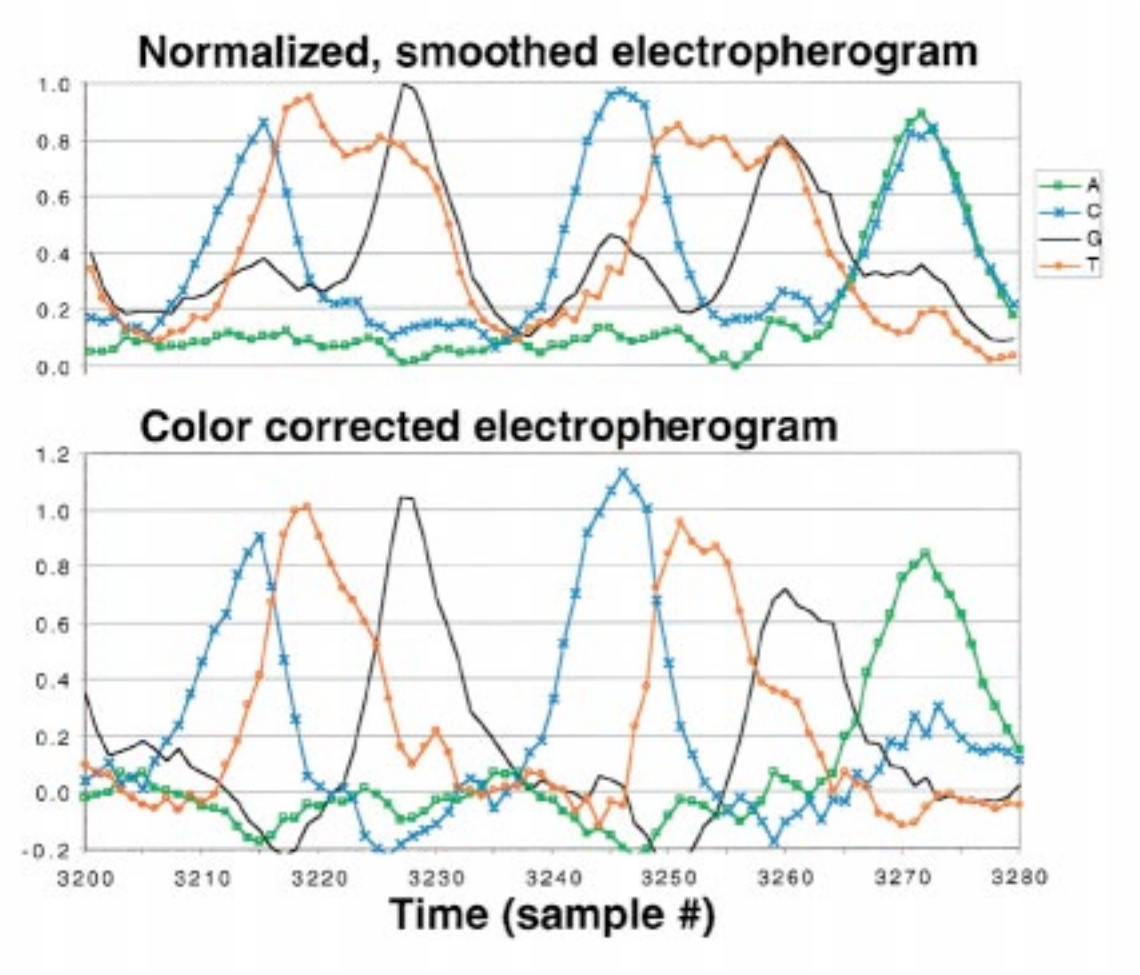


Fig. 9. Electropherograms before and after the color correction step. The sample spacing is roughly 1 s. Note that the width of the peaks and peak-to-peak spacings are not uniform.

many overlapping shorter sections of DNA. The average number of times each base appears in a different DNA fragment is called coverage. Sequencing projects range from "draft" quality with 3–5× coverage to "finished"

quality with order 10× coverage and on average less than one incorrect base call per 10 000. The overlapping, but not identical, DNA fragments are the initial input for the sequencing chemistry. These fragments are usually generated mechanically or biochemically. In the mechanical approach, a purified source of DNA is sheared or fragmented into many random-length subsequences, often by forcing the DNA through a pore. In the biochemical approach, multiple restriction enzymes that digest DNA at fixed sites are used to generate different fragments depending on the order the enzymes are applied. In the public sequencing project, both techniques have been used.

There are two different strategies on how to get 1000-base template DNA from multimegabase chromosomes. In the mapping approach, mechanical shearing is used on the chromosomes with fragments selected to be around 150 000 base pairs. The fragments are coarsely assembled into a rough map that identifies the location of the particular DNA fragment on the chromosome. Sometimes it is necessary to "end-sequence" these large fragments to facilitate assembling the map. The size of the fragment and the limited sequence information are used to identify a tiling pattern that covers the section of interest on the chromosome. Once the 150 kbase clones are mapped, the selected clones in the tiling pattern are fragmented randomly into DNA short enough (less than two kbases and known as subclones) to use as templates for electrophoretic sequencing. In sequencing smaller genomes, like microbes [51], and in the PE/Celera "shotgun" approach to human genome sequencing [52], the mapping step is skipped and the whole genome is fragmented. Eliminating mapping saves time and effort but assembly of the significantly larger number of DNA fragments is more difficult.

Automation has been applied to many steps of the DNA sequencing process. Parallel aspiration out of and dispensing into microtiter format plates with 96, 384, or 1536 wells allows a large number of samples to be processed simultaneously. The plastic plates have been standardized at 12.8 by 8.6 cm. The plates are available in different depths to help accommodate volume and surface area effects on the biochemistry. As the number of wells increases in each plate, the reduced cross section of each well requires more and more accuracy from the automation systems. A 384-well plate, for example, has 4.5-mm center-to-center spacing of the wells. Flexible tips are often used so that slight misalignments do not damage the plates or the robots. DNA samples remain in the plates for many operations including centrifugation, thermal cycling, template purification, and sequencing reaction setup allowing many DNA samples to be processed in parallel. Only a few nanograms of DNA would be required for DNA sequencing if it were possible to reliably make and handle volumes that small.

With many molecular biology instruments now accommodating microtiter format plates, tracking and moving plates around the laboratory has become important to efficiency. Plate shuttling is often accomplished with a conveyor belt having stations along the track. Shuttling can also be done with a robotic arm that picks and places plates on subsystems around the arm. The first step in many DNA sequencing centers is the growing of many copies of bacteria with an inserted piece of DNA on a flat culture plate. As the bacteria replicates, many copies of the inserted DNA are made. "Picking" robots can harvest bacterial plaques and colonies into microtiter plates. This requires sophisticated imaging systems to identify the location of the bacteria to be harvested and high-speed positioning systems that can direct the tip that picks up the bacteria. In summary, automation has allowed significant time and cost savings, reduced sample volumes, improved protocol consistency, and allowed for more accurate sample tracking [53], [54].

For most publicly funded projects, the final step for sequence data is submission to the public database GenBank. GenBank is managed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM) of the National Institutes of Health (NIH) [10]. There are a variety of ways to submit data. In the end, a unique accession number is assigned to each submission so that the data may be appropriately referenced. High-throughput sequencing centers can also submit data at different levels of completion. Finished sequence is known as phase 3 status and draft data is either phase 0, 1, or 2 depending on the base call quality and the number of gaps in the data [55].

It is instructive to work through a specific example of how to access DNA sequence data. We selected as an example, the inherited disorder known as myotonic dystrophy, which is the most common form of muscular dystrophy that affects adults [56]. Symptoms range in severity from male-pattern baldness to lethal. The cause of myotonic dystrophy is a set of CTG repeats that occur in the 3′ untranslated region of the dystrophia myotonica (DM) protein kinase gene on the long arm of chromosome 19 (19q13.2–19q13.3) [57]–[59]. The CTG pattern repeats 5–20 times in the normal population. Affected individuals have 50 to thousands of CTG repeats and symptoms appear stronger with each affected generation.

We could find information about myotonic dystrophy in the scientific literature and we could find it in the DNA sequence database. To use the sequence data itself, go to NCBI online at http://www.ncbi.nlm.nih.gov/ and enter "dystrophia myotonica" in the GenBank search window. A list of hyperlinks will be returned. Select the link for accession number NM_004409 "Homo sapiens dystrophia myotonica protein kinase (DMPK), mRNA." If for some reason the search is not working, the accession number should be able to link to the same data. The information available at NM_004409 includes published information related to the gene, the source (Homo sapiens, 19q13.3), a list of the 629 amino acids in the expressed protein (MSAEVRL. . .PGAARAP), a list of the 3407 bases from the DNA sequence, the position in the sequence that codes the protein (bases 777 to 2666), and the location of the 3′ untranslated CTG repeat (begins at base location 2890). The last seven amino acids in the protein and the associated DNA sequence are shown in Table 2. Using a GenBank search is just one of many ways to find and compare DNA and protein sequence information. As an example of another approach, the Online Mendelian Inheri-

Table 2
The Last Seven of 629 Amino Acids in the Dystrophia Myotonica Protein Kinase (DMPK) and the Associated DNA Bases in the Gene. Both the Amino Acids and the DNA Sequence can be Verified Online at NCBI

Position	623	624	625	626	627	628	629
Letter	P	G	A	A	R	A	P
Abbrev.	Pro	Gly	Ala	Ala	Arg	Ala	Pro
Acid	Proline	Glycine	Alanine	Alanine	Arginine	Alanine	Proline
mRNA	CCA	GCA	GCC	GCC	CGC	GCU	CCC
DNA	CCA	GGA	GCC	GCC	CGC	GCT	CCC

tance in Man or OMIM is also accessible through the NCBI web site. This database links together significant scientific information with the gene databases. A search for "myotonic dystrophy" will provide significant detailed information on many muscular disorders and is an excellent launching point for reviewing the scientific literature.

The gene associated with myotonic dystrophy was discovered as part of research being conducted on the disease. In contrast, the location and function of genes is largely unknown for the DNA sequence being submitted by the large centers. Regions of the DNA that are amenable to transcription are called open reading frames (ORFs). ORF finding software has been developed [60] and can be customized for a particular organism [61]. For instance, bacterial and human/mouse ORF finders usually use different algorithms. Generic markers like start and stop codons may often designate the beginning and end, respectively, of an ORF. However, just looking for these codons is generally not sufficient, and algorithm complexity has grown to include techniques like hidden Markov models [62]. As the DNA sequence for more organisms becomes available, the accuracy of ORF finders can be evaluated and improved [63]. A continuing opportunity exists to leverage algorithms developed for other applications like speech recognition and customize those techniques to ORF finding. Once the ORF finding problem is addressed and genes are provisionally located and identified, the next major step is to determine the function of the genes.

IV. From Genes to Function

The amount of DNA sequence data available is growing at a significant rate. The DNA sequence of the human genome and the genomes of many other organisms are completed (see Table 3). With the raw sequencing power of the large centers, it is now possible to draft sequence an entire bacterial genome in a single day. The availability of these data is changing the approach to many biological research questions. Just as it was possible to take a whole-organism approach to DNA sequencing, it is now possible to consider a whole-organism approach to getting at the mechanisms that control the biochemistry in a cell and therefore the basis of how to prevent or treat disease. High-throughput whole-organism approaches include genomics, proteomics, functional genomics, and structural genomics for the study of genes, proteins, gene function, and three-dimensional (3-D) protein structures, respectively. These approaches contrast with and complement the hypothesis-driven, single isolated phenomenon, research tradition in biology. The next generation of hypothesis will address an entire complex activity like metabolism, which requires information about multiple protein-DNA interactions of the cell's regulatory mechanisms. The rest of this paper presents approaches to collecting gene, protein, and regulatory information. We conclude with a description of how computer modeling and simulation might facilitate data interpretation and understanding of complex biochemical pathways and mechanisms. The potential impact of appropriate modeling tools parallels the historic role of circuit simulation in electrical engineering. Genomic engineers need simulations of complex biochemical networks that can reduce the amount of experimentation needed to understand changes in the network or to introduce deliberate changes that influence function. This level of control will profoundly influence our ability to safely engineer new crops, medicines, and genetic treatments.

Monozygotic identical twins arise from the same fertilized egg and share exactly the same genetic code. Despite sharing many physical characteristics, they are not truly identical. For example, twins have similar but noticeably different fingerprints. More generally, events will shape twins differently. One twin may get a viral infection that causes an immune system disorder like multiple sclerosis. The other twin may eat carcinogens in grilled meat and develop cancer. Furthermore, every cell in a human shares exactly the same DNA, but nerve cells and white blood cells have radically different shapes and functions. Therefore, knowing the genetic code tells us what *might* happen, but it does not tell us what *will* happen.

In principle, knowing every gene in an organism provides the sequence of every protein that organism can produce. A nerve cell and a white blood cell in a human are distinguished because a different subset of genes is expressed, i.e., producing RNA messages for protein synthesis. Expression patterns also change with time and environment. When a nerve cell receives a signal from another cell across the synapse, there is a change in which genes are expressed. The changed expression results in protein products that signal the next cell in the brain's neural network.

Genes are categorized by the function of the protein produced. Structural genes code for enzymes that catalyze a

Table 3
DNA Sequencing Milestones and the Status of Many Publicly Funded Projects

1995	The 1.8 million base DNA sequence comprising the genome of *Haemophilus influenzae* was determined by The Institute for Genomic Research (TIGR) [51]. This gram-negative bacterium was the first living organism to be completely sequenced. *H. influenzae* cause over 90% of systemic infections in children including meningitis, pneumonia and middle ear infections. *H. influenzae* is named due to the mistaken assumption that it causes flu (now known to be a viral infection).
1996	The 12.1 Mbase genome of the first eukaryotic organism *Saccharomyces cerevisi* also known as Brewers' or Bakers' yeast was sequenced by an international team [143].
1998	The first animal sequenced was *Caenorhabditis elegans*, a nematode (worm) with a genome of 97.1 Mbase [144].
1999	Chromosome 22 with 33.4 Mbases was the first human chromosome completed at finished quality and was announced in December [145].
2000	The genome of the fly *Drosophila melanogaster* was finished in March with about 120 Mbases [146]
2000	In April, PE Celera, Inc. announced that it has completed the data collection phase of its whole genome shotgun sequencing strategy for human DNA. Also in April, The Department of Energy Joint Genome Institute announced completion of human Chromosomes 5, 16 and 19 to draft quality [8].
2000	The 33.5 Mbase finished sequence of Chromosome 21 was announced in May [147]. An extra copy of Chromosome 21 causes Down syndrome, which affects about 1 in 700 live births.
2000	On June 26, 2000, President Clinton announced the milestone draft sequence of the human genome as "the most wondrous map ever produced by humankind." [9].

Continuing Updates Available Online

"The Comprehensive Microbial Resource Home Page," TIGR description of about microbial genomes and sequencing projects [Online] http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl. This site currently has about 30 genomes.

Entrez database part of NCBI provides access to 6 Archaea and 23 Bacteria [103].

S. Beck and P. Sterk, "Genome Monitoring Table," provides daily updates on all public sequencing projects [Online] http://www.ebi.ac.uk/~sterk/genome-MOT/.

reaction performing some function for the cell, like energy production, environmental sensing, and cell defense. Regulatory genes code for proteins that bind to DNA in ways that stimulate or suppress the expression of structural genes. We define a pathway as a series of reactions that perform some function for the cell. For example, the breaking down of starch to produce energy is a pathway. An enzyme catalyzes each step of the pathway. The amount of the enzyme is controlled by the expression of its structural gene, which is in turn controlled by the regulatory genes associated with the pathway. A pathway does not have to be sequential. Reactions can occur simultaneously and there can be branching. Pathways are generally self-regulated by feedback loops. A pathway to digest a particular molecule typically turns itself off when that molecule is no longer present. Some pathways turn on if they sense a different external temperature or chemical concentration. Consider *Yersinia pestis*, the bacterium that causes plague in humans. The virulence mechanism of *Y. pestis* that causes plague in humans is not activated when *Y. pestis* bacteria are in fleas at 25°C. When the same bacteria enter human hosts, the temperature increases to 37°C and the calcium concentration falls, the bacteria begin to produce virulence proteins [64]. Fig. 4 shows the interaction of genes (DNA), message (RNA), proteins, function outcome, and environment in a pathway. Pathways are not independent; they often share enzymes and can stimulate or suppress each other, and they are not necessarily confined to a single cell.

Traditionally, a pathway is studied by a series of knock out experiments. In each experiment, a single structural or regulatory gene is mutated or removed from the genome. In a simple example, if a pathway is responsible for digesting fructose, its failure means the cell can no longer use fructose as a source of energy and cannot grow if fed only fructose. So, if a gene is mutated and the cell continues to survive, the mutation did not affect its fructose metabolism pathway. What biologists have found repeatedly is that different combinations of genes may lead to different results. For obvious evolutionary reasons, pathways often have redundant branches. The loss of one gene may reduce efficiency or

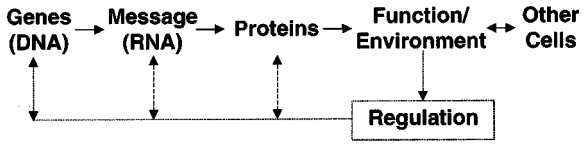


Fig. 10. Genes, messages, and proteins in a complex system with multiple feedback loops.

have no effect at all. So while losing either gene A or gene B might produce no observable changes, losing both genes A and B would result in cell death. In the case of regulatory genes, the situation is more complex. Experiments have shown that there are regulatory genes where losing either gene A or B may kill a cell, but losing both will not! So, to fully understand a pathway requires testing every possible case of gene expression under all environmental conditions. Even a very small pathway contains about ten genes or 2^{10} possible gene deletion experiments if every combination of genes is deleted. Some human pathways result from the interaction of hundreds of different genes, and the collective cell contains hundreds of interconnected pathways.

Even given the whole genetic code, it is obvious that traditional molecular biology would take centuries to tackle even the 470 genes of the smallest known genome of any free-living organism (the bacterium *Mycoplasma genitalium*, [65]). *Functional genomics* is a group of massively parallel, high-throughput experimental and computational techniques to study the function of every gene in an organism. This includes measuring the mRNA concentration for every gene, determining the function and structure of every protein, and finally being able to model the interconnected regulatory network of the whole cell. In the following sections, we will discuss each component of the system in order. We will focus on a few key technologies of particular interest to engineers and computer scientists: DNA microarrays used for expression profiling, computational protein structure prediction, and mathematical models of pathway control.

Given that the smallest free-living organisms have about 500 genes and a human may have about 120 000 genes, what we have outlined is already a great challenge. Unfortunately, the picture of Fig. 4 is incomplete. It shows the transition from genes to proteins by simple arrows. However, each arrow actually represents a complex series of processes. Human genes are usually not continuous segments of DNA [11], [12]. RNA is spliced across several separate regions of DNA to form message RNA. The mRNA is then transported out of the nucleus to the ribosomes where protein is synthesized. After the protein is synthesized, it is often modified before it assumes its functional role. Regulation and feedback can occur at each step of the process going from DNA to functional protein. As a result of the modifications from regulation and feedback, some complex human genes can produce hundreds of different proteins [66]. The genomic approach necessarily depends on the DNA code that is now accessible through high-throughput technologies. However, researchers are just beginning to develop the tools required to do similarly high-throughput study of the nongenome interactions shown in Fig. 10.

V. Gene Expression: DNA Microarrays

The variation of cell behavior with changing conditions is a function of differential gene expression. Under a given internal and external state, each gene is copied to mRNA at a particular rate. Thus, in principle, measuring mRNA concentrations under a set of conditions provides a "snapshot" of genetic activity. After the cell is subjected to an external perturbation, the genetic activity changes as some pathways are turned on, some are turned off, others are "tuned" up or down, and many might not change at all. These changes are dynamic, so snapshots after the initial perturbation show continued changes as the first pathways produce intermediate products that stimulate the next wave of pathways. Finally, as shown in Fig. 10, the cell's response can interact with the environment and other cells. For example, during intense exercise a human muscle cell runs out of oxygen. The cell responds by activating the much less efficient pathways responsible for anaerobic respiration (energy production without oxygen). Anaerobic respiration produces lactic acid that cannot be broken down fast enough. The accumulation of lactic acid stimulates a signaling pathway that sends a chemical message to a nearby nerve cell, which then sends a "pain" signal to the human brain. When exercise stops because of muscle pain, oxygen becomes available, aerobic respiration resumes, the lactic acid is removed, and the signaling pathway turns off [67].

Obviously, measuring the expression of just a few genes is not enough to characterize these complex changes. Several technologies have been developed for the simultaneous measurement of the concentration of thousands or more different mRNA sequences. DNA chips and microarrays separate a mixture of mRNA molecules based on knowing their sequences. If the sequence is not known, a method called Serial Analysis of Gene Expression (SAGE) can identify mRNA transcript that did not come from a known gene [68], [69]. Given the rapid acquisition of sequence data discussed earlier, sequence knowledge is typically not a problem. However, as described above, it is difficult to identify what parts of the sequence actually code for genes. In a typical DNA chip or microarray experiment [70]–[73], the mRNA is isolated from a sample of cells in the state of interest. The mRNA is then processed by a reverse transcription reaction ($5'$ to $3'$ on a DNA strand), which produces a complementary single strand of DNA (cDNA). A fluorescent marker is attached to the cDNA. Now, the cDNA target can bind with a single strand of DNA with the complementary sequence. The binding of a single strand of DNA with its complement is called hybridization. DNA chips and arrays have surfaces covered by thousands of spots, where each spot can contain billions of cDNA probes corresponding to a particular known gene. The targets are poured onto the probe array, the targets hybridize with the complementary probes (if present in the array), and the array is washed removing targets that did not hybridize. The intensity of fluorescence at a spot then indicates how much mRNA with the corresponding sequence was present in the

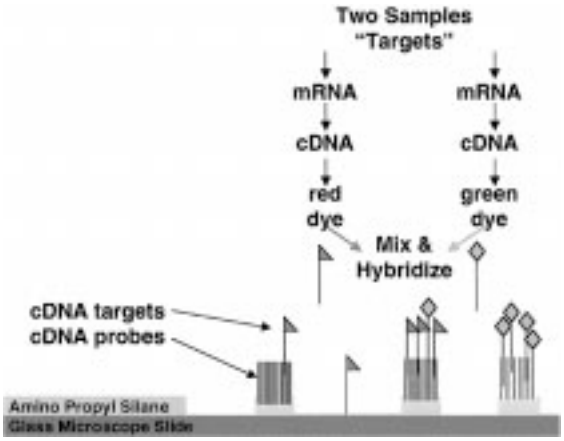


Fig. 11. DNA microarray experiment to measure changes in gene expression. Two samples are separately labeled and compete for hybridization with complementary DNA on a glass slide.

original sample of cells. It is currently not possible to quantitatively determine the original mRNA concentration from this fluorescence signal. Therefore, DNA array experiments usually measure the competitive hybridization of mRNA extracted from two samples. A different fluorescent label is attached to the mRNA from each sample (e.g., red Cy5 and green FITC). The ratio of fluorescence corresponding to each sample then indicates the relative mRNA concentration difference between the two samples. Fig. 11 outlines a typical DNA microarray experiment.

DNA chips, developed by Affymetrix, use oligonucleotide probes: 20 or 25 base subsequences unique to each gene. These probes are synthesized on a 1.3 cm $\times$ 1.3 cm surface using photolithographic techniques originated in semiconductor manufacturing [74], [75]. The probes are synthesized by a series of masks and chemical reactions that sequentially extend the oligonucleotide probes exposed through the mask by a specific base. In contrast to DNA chips where the probe DNA is synthesized, microarrays use cDNA probes copied from actual DNA and amplified using polymerase chain reaction, also known as PCR [76]. Each probe can be a section of or the entire gene (about 1000 bases is typical). Nylon microarrays with radioactive probes have been used for analyzing the simultaneous expression of almost all *E. coli* genes [71]. Glass microarrays with optical fluorescence detection, pioneered by Patrick Brown's lab at Stanford [70], are now more frequently used because of their greater sensitivity. In general, microarrays are relatively easy to customize and public protocols are available on the web [77]. A series of reviews on DNA microarrays and chips are presented in [78]. One significant problem with all DNA array experiments is that the hybridization is not perfect. Errors in hybridization become particularly acute at 90% sequence similarity, which is the case for many important regulatory genes. In these cases, redundant probes specific to unique subsequences are necessary to separately identify mRNA targets.

Reliable image analysis of microarrays is challenging. Fig. 12 shows the raw pixel data from the red channel of a microarray made in our lab. Data acquisition considerations are similar to other optical imaging systems including

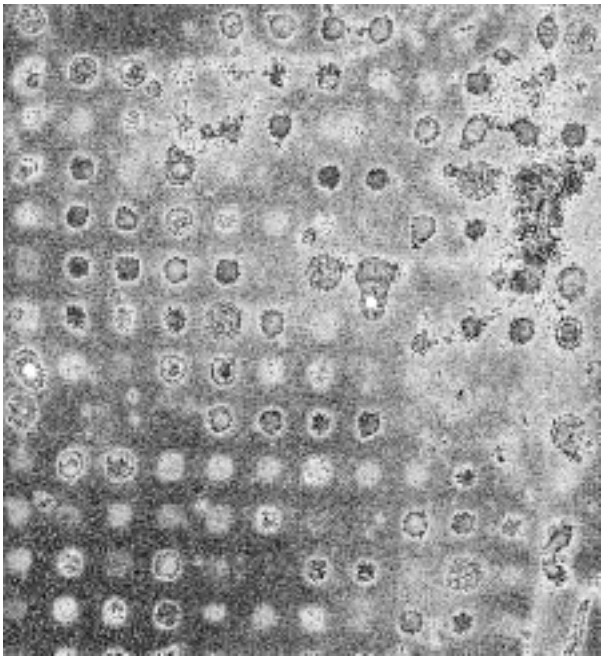


Fig. 12. Raw microarray image from one channel (unpublished results courtesy A. Wyrobek and E. Garcia, LLNL).

integration time, optical cross talk of fluorescent labels, and nonuniform illumination. Microarray images often have a very low signal-to-noise ratio. There are generally many more probes than targets, so the spots are only partially fluorescent. Since the technology to print microarrays cannot form consistent spots, exact *a priori* assignment of spot regions is impossible. Thus, the spots have to be recognized after the experiment from a weak and irregular signal. Also, some targets will bind to the wrong spot, and some will bind to the substrate and fail to be washed off. Stray targets and other sources of fluorescence including the substrate and coatings contribute to a significant nonlinear background that must be removed in order to retrieve the signal. Image analysis issues and solutions for typical microarray experiments are discussed in [79].

A recent innovation that improves spot finding and background estimation and compensation is the inclusion of a third dye, blue DAPI, in the microarray experiment. DAPI is a DNA counterstain that binds to cDNA that failed to hybridize with the target DNA. Thus, the blue channel reveals the shape of each spot, making spot recognition simpler and more accurate. The DAPI stain also helps estimate the background noise from DNA-surface binding. After the signals for each sample are normalized for background and the relative intensity of the fluorescent dyes, the final outcome of a microarray experiment for each probe spot is a ratio of the mRNA concentration in one sample relative to the other. Perhaps the most important obstacle for obtaining useful microarray data is quantifying and reducing the large error inherent in defining this ratio.

Fig. 13 shows the final processed image from Fig. 12. Along with controls, it contains 85 genes from *Yersinia pestis*. The virulence mechanisms of *Y. pestis* become active at 37°C. In this picture, mRNA from cells at 25°C are

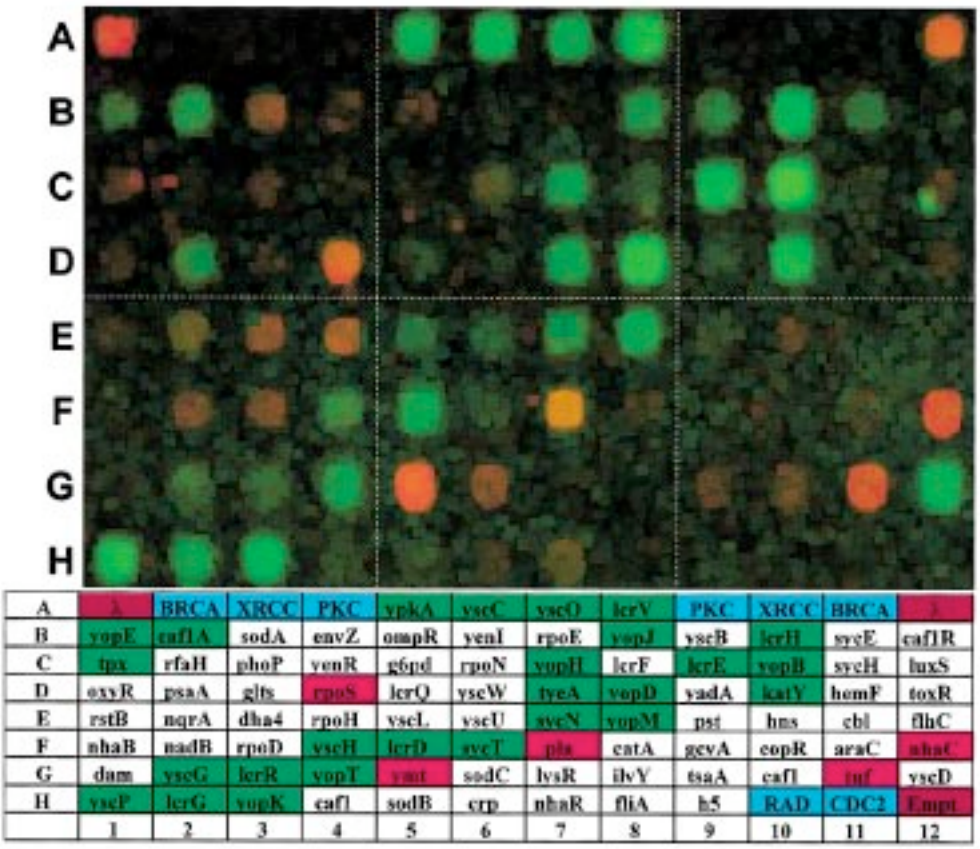


Fig. 13. Microarray data for an 85-gene *Y. pestis* microarray (unpublished results courtesy E. Garcia and A. Wyrobek, LLNL).

labeled red and mRNA from cells at 37°C are labeled green. The more intense the color, the more mRNA from that sample hybridized to the spot relative to the other sample. Below the microarray data is a table of the gene names corresponding to each spot. Cells colored green are *Y. pestis* genes expressed more at 37°C than 25°C, red are genes expressed more at 25°C. Cells colored blue are control spots from mouse and human genes that are absent in *Y. pestis*.

The challenge of visualizing microarray data is hinted at in Fig. 13. In particular, a microarray containing every gene in *Y. pestis* would contain over 4300 different spots, making an image like Fig. 13 impossible to interpret. In general, DNA array experiments generate large, complex data sets. For any one experiment, each gene has three measurements associated with it: the intensities of the two competing samples, and the ratio of those intensities. Typically a series of assays are taken over time. Thus, an array of 10 000 genes can be thought of as a set of 10 000 3-D vectors that are changing in time. There is a need for ways to store the data in conveniently accessible, public data warehouses, visualize experimental results, and interpret the relative expression of the genes to identify pathways and common regulatory mechanisms.

To date, most microarray experiments have been published in scientific journals with the expression data either included or referenced at the URL of the authors' website where the data are posted [80]. There is no current standard for microarray data warehousing, so they are stored in a variety of database formats or even large spreadsheet or text files. Currently, there are different database frameworks under development, one of many examples is ArrayDB developed at the NHGRI [81], but there are many other public and private efforts. In the very near future, as microarray experiments become increasingly frequent, there will be a need for a central public facility for submitting and accessing experimental data sets, similar to the current NIH GenBank for storing DNA sequences. Coupled to data warehousing, there need to be creative ways for biologists to visualize experimental results. Microarrays with thousands of grid cells can be viewed as similar to multidimensional geographical information and work is being done to extend tools from that area to be useful in biology [82].

Even with better visualization and data storage, manually processing tens of thousands of data points is very difficult. Pattern recognition methods developed for imaging can be extended to automatically classify gene expression data. The goal is to divide genes into categories based on expression levels. Classification by expression level can indicate which genes are involved in the same pathway [83] and potentially identify common regulatory mechanisms [84], [85]. The problem is that a 50% increase in expression could be as important to the biology of one gene as a 300% increase in expression in another gene. Simple threshold-based classification approaches erroneously classify these genes

in separate categories. Furthermore, experimental error is sufficiently high that differences in relative expression are not very well quantified. Initial approaches to the problem included unsupervised learning methods, hierarchical clustering algorithms [86], [87], and self-organizing maps [88]. Often the function of at least some genes is known or suspected with a high degree of confidence. A support vector machine (SVM) method has been presented that takes advantage of this prior knowledge [89]. The challenge facing algorithm developers is that even biologists cannot currently distinguish biologically significant features and clusters from artifacts. Progress depends on developing an appropriate biological framework and translating it to a mathematical model.

VI. Structure to Function: Structural Genomics

Differential expression data can provide an initial clue of an unknown gene's role in an organism. The specific function of a gene is executed by the protein for which it codes. We now turn our attention to the next component of Fig. 10: protein function. The typical first step upon obtaining the DNA sequence of a suspected gene is to search for its sequence in a database of all publicly available genetic sequences like GenBank [10]. Approximate search algorithms like BLAST [90]–[92] are used to find similar genes in previously sequenced organisms. If a known gene is found to be "similar," usually 25% or more identical, its corresponding protein or function is assigned to the unknown gene. This is called sequence annotation. Generally, over 40% of all suspected genes in newly sequenced bacteria are not found in the database [93]. Also, the annotation is not guaranteed until independent supporting evidence is found. In fact, many of the genes already in the database have unknown functions themselves and are merely annotated as "hypothetical proteins." However, BLAST searches provide a useful first indication of protein function and large numbers of genes can be searched rapidly using parallel algorithms [94]. More definite knowledge about protein structure and functions is required for reliable and comprehensive genome annotation.

Structural genomics is an effort to do high-throughput identification of the 3-D protein structures corresponding to every gene in the genome [93]. The effort combines high-throughput experimental structure determination along with computational structure prediction. Known protein structures are stored in Protein Databank (PDB), a large public database on the web similar to GenBank [95]. Structural biology is based on the paradigm that the 3-D structure of a protein will define its function. Knowing the surface chemical structure of an enzyme will suggest how the substrates bind to the enzyme, and how the enzyme interacts with other proteins. Fig. 14 shows an electrostatic potential map of an antibody and the protein it binds, suggesting the detailed structure of the chemical reaction that takes place when the molecules bind [96].

There are two experimental methods for determining 3-D protein structure: nuclear magnetic resonance (NMR) and X-ray diffraction [97], [98]. NMR measures the coupling

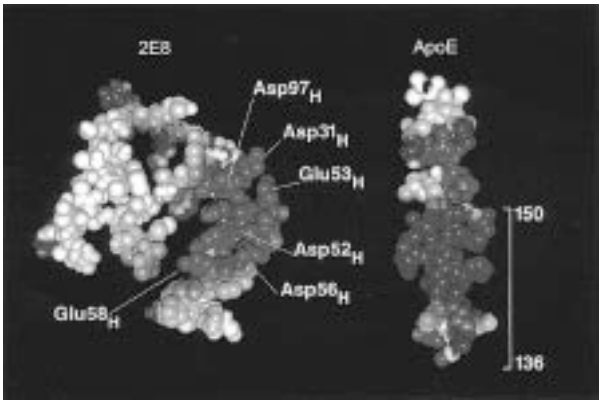


Fig. 14. Electrostatic potential map of the solvent-accessible surface of the 2E8 Fab antibody and the corresponding receptor-binding region of the apoE protein: red is negative potential, blue positive potential, and white neutral. The labels indicate important amino acids (courtesy of B. Rupp, LLNL, and K. Weisgraber, Gladstone Institutes).

of atoms across chemical bonds and short distances through space under the influence of a magnetic field. A typical NMR experiment with a 600 MHz field takes four–five weeks for data collection and is limited to a protein with 120 amino acids. Data analysis that once took months can now occur in a single day using new algorithms [99]. New NMR instruments with fields as high as 1 GHz and more sensitive techniques allow faster analysis of larger proteins [100], but there is still a size limit of a few hundred amino acids. X-ray crystallography has no theoretical size limit. A protein crystal diffracts X-rays, generating the Fourier transform of the atomic structure (electron density) of the protein. Since the phase is unknown, techniques like replacing selected atoms by heavy metals and using different wavelengths are required. Data analysis used to be as complicated as with NMR, but recent algorithms have made it much faster [101]. X-ray crystallography is, however, sharply limited because it is often very difficult and sometimes impossible to obtain a high-quality protein crystal. Because of the limitations of both measurement techniques, little structural data are known for the important class of membrane proteins, about 30% of the total protein in a cell. Improvements in NMR magnets and methods and protein crystallization protocols are needed.

Even with the implementation of high-throughput techniques, protein structure experiments are not keeping up with the amount of sequence data. Currently, there are 10 675 proteins with known 3-D structures from X-ray crystallography and NMR experiments in the Protein Databank (PDB) [102]. This is a tiny fraction of the millions of gene sequences in GenBank from 820 different species [103]. However, while there is a great diversity of sequences, there are certain structural features that arise repeatedly in many different proteins. The 10 000 protein structures in PDB share about 1800 folds, or common structural features. It is estimated that fewer than 5000 folds occur naturally. Thus, with careful selection of experimental targets, perhaps only a few thousand more protein structures are required so that every possible structural feature is in the database [92]. The current prediction is that by 2003, there will be 35 000 protein sequences in PDB [104]. In

principle, if every fold that can be put together to build a protein is known, it should be possible to compare the sequence of an unknown protein to the sequences corresponding to these folds and predict its final structure. Comparative modeling is a computational approach to doing this. In general, structure is assigned by identifying sequence similarities (homologies) between the unknown protein and proteins with known structures. There are several different approaches; a good review is [105]. Generally, 30% amino acid sequence similarity is thought to be sufficient for accurate structure prediction.

A significant limitation of this approach is that alignment is an unsolved problem. Algorithms like BLAST can align sequences to compare them as a whole, but protein structure requires aligning different parts of the protein sequence. Fig. 15 is a simple illustration of how similar subsequences can be located in different regions of the overall sequence, and the subsequences themselves may not be completely identical. In some cases, equally suitable alignments can be found in which every amino acid is at a different position in the predicted structure [106]. Furthermore, most experimental structure data ignores the amino acids at the end of the protein chain, limiting what is available in databases.

Moreover, many amino acids share the same basic chemical properties, so exchanging them does not significantly affect structure or function. Therefore two proteins with very different sequences may in fact have the same structure. There are hundreds of known hemoglobin protein sequences in mammals, all of which share similar structure and the same function. One approach is to consider the evolutionary history of the organism in question and when its sequence diverged from those in the databank. For example a human gene will have a more similar sequence to a chimpanzee gene than to a mouse gene, and will be much less similar to an *E. coli* gene [107]. Another approach is to combine the results of sequence homology and evolutionary history along with the results of microarray experiments [108]. To ensure that algorithms can be applied generally to unknown sequences, algorithm performance is measured in biannual Critical Assessment of Protein Structure Prediction (CASP) experiments. Before each CASP meeting, the experimental community provides a list of structures that are about to be determined. The sequences are distributed to the computational community, which analyzes them without knowing the structure beforehand. Results from the last CASP 3 meeting are contained in [109]. Overall, while the accuracy of predictions is improving, computational structure prediction is still limited to subsequences of an unknown gene that have high sequence similarity.

The shortcomings of comparative modeling would be avoided if it were possible to predict protein structure *ab initio* based on amino acid chemistry alone. Almost all protein structure is the result of the interaction of amino acids with water in cells [110]. Thus, *ab initio* simulation of the folding of a 1000 amino acid protein requires a 10 000-body calculation of the interactions of the amino acids and 9000 surrounding water molecules. Fundamental quantum chemistry calculations are limited to studying

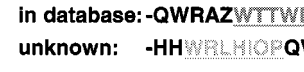


Fig. 15. A simple example of the structure alignment problem in protein homology assessments.

the area around a single amino acid or DNA base [111]. Classical molecular dynamics treats amino acids and water molecules as "billiard balls" and can model a subregion of the protein, up to about 40–60 amino acids in length [112]. IBM has launched a new effort to produce a computer that is capable of one petaflop, about a hundred times faster than the most powerful supercomputers currently under development [113]. However, even if technological obstacles can be overcome, the computer algorithms currently used for *ab initio* structure prediction do not scale well and will need to be redesigned.

VII. Proteomics

The orderly sequence of Fig. 4 implies that knowing when a gene is expressed means we know when its corresponding protein is active in the cell. Fig. 10 suggests that because of complex regulation, this is not the case. Experimental studies in yeast [114] and the human liver [115] indicate that protein concentration is not a linear function of mRNA transcript concentration. In many cases, even if gene A has more transcript than gene B, protein A has a lower concentration than protein B. This paradox is inevitable, since mRNA and protein are very different molecules. The message is intended to be short lived and the product durable, so RNA decays far more rapidly than protein. Thus, a sharp burst of mRNA transcription results in the long-term presence of many proteins, continuing after the mRNA is gone. Moreover, proteins can be modified after they are translated from the mRNA template at the ribosome. Like genomics is for genes, proteomics is the automated, high-throughput simultaneous study of every protein in a cell [116]–[118].

Unlike genomics, there is no universal gene chip or microarray for measuring the concentration of many proteins simultaneously. There is no general tool for protein production like PCR produces copies of DNA. Producing multiple copies of a protein usually requires finding its coding gene, inserting it in the bacterium *E. coli*, growing the cells, breaking the cells apart and harvesting the protein. Harvesting natural protein from a cell is more difficult than DNA and RNA, since as much of 10% of the cell protein can not be extracted. The biggest problem lies with detecting specific proteins. There is no general protein binding like the complementary sequence binding of DNA and RNA. If protein is known to bind to a specific antibody or to a DNA molecule, it can be detected with an array of those specific targets [119]. The only general way to separate different proteins is with mass, since different amino acid sequences correspond to different masses.

The most common tool in proteomics is two-dimensional (2-D) gel electrophoresis [118]. The separations follow the same principles as 1-D gel electrophoresis described earlier in the paper under sequencing. The electric field is applied

along one axis, separating the proteins based on how much was present in the sample, and then the other axis, identifying the protein species by mass. Finally, each 2-D spot on the gel represents the concentration of a particular protein in the sample. A public database of images and data from 2-D gel electrophoresis experiments is available at [120]. Recently, tandem mass spectrometry has also been extended to high-throughput proteomics. Some enzymes digest a protein and break it up into subunits in highly predictable, well-known ways. These fragments can then be analyzed with a mass spectrometer, generating a fingerprint for the protein. This fingerprint can then be searched against a database of known proteins, and the protein can then be identified and its concentration in the sample measured. A review of the current state of automation in proteomics can be found in [118].

VIII. Genetic Regulation: Biological Basics

We have now outlined approaches to measuring gene and protein expression and identifying protein structure and function. As shown in Fig. 10, it is impossible to study genes and proteins in isolation from each other. Therefore, it is necessary to determine and model the regulation of a pathway to fully understand it. As we discussed earlier, a pathway of just ten genes has 2^{10} states of on/off genetic activity, more if one includes rates of genetic activity. The effect of changing the activity of a gene can be studied in simulation by using a pathway model. If successful, this simulation would obviate the need for an impractical number of biochemical experiments. Ultimately, it will be possible to use pathway simulation to design novel control systems in organisms to produce useful proteins in a regulated fashion. The hardware tools for implementing an artificial genetic control system are emerging. For example, a genetic "switch" was recently demonstrated in *E. coli* bacteria [121]. One of the biggest obstacles to gene therapy is that it is almost impossible to regulate a dose [122]. The patient has to be given a large quantity of genes through virus vectors, with a high probability of negative immune response and possible death. Thus, the ability to implement stable regulatory pathways for gene expression is critical for the success of one of the Human Genome Project's most important promises. In this section, we will describe a simple example of an actual biological pathway to show which reactions have to be modeled in a pathway simulation.

Regulation in biological pathways is often described using the *operon*, a model proposed by Jacob and Monod in 1960 [123]. The example they first described is the lactose pathway of *E. coli* (the *lac* operon). Fig. 16(a) shows how the operon is arranged on the *E. coli* chromosome [11]. Lactose metabolism requires enzymes coded by the genes *lacZ* and *lacY*. For transcription of mRNA to occur, RNA polymerase must bind to a short *promoter* sequence. In Fig. 16 the promoters are labeled P1 and P2. If the RNA polymerase binds to P2, the DNA sequence of the lacZ, lacY, and lacA *structural* genes will be transcribed into mRNA molecules that are then translated on the ribosome to form proteins.

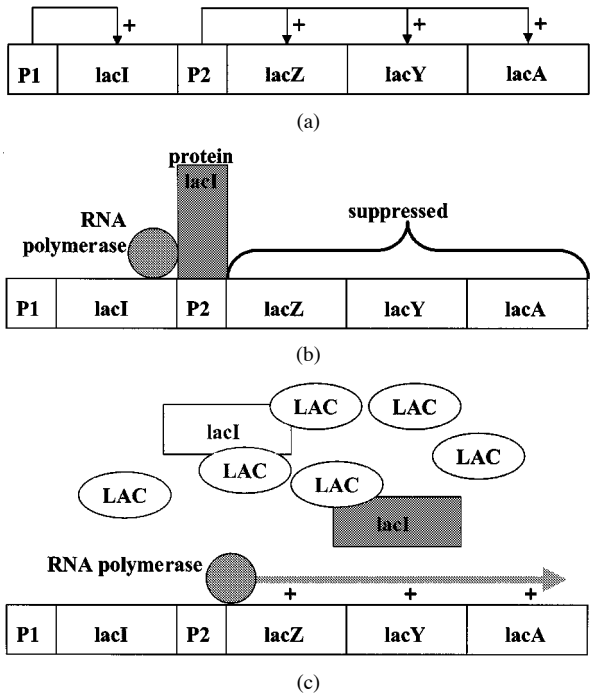


Fig. 16. Organization of the *lac* operon in *E. coli*. (a) Promoters P1 and P2 and genes *lacZ* and *lacY* code for lactose metabolism enzymes. (b) The repressor protein coded by *lacI*, binds to P2 preventing *lacZ*, *lacY* and *lacA* transcription. (c) Lactose binds with *lacI*, allowing RNA polymerase to bind to P2 and transcribe the structural genes.

The enzymes will break down lactose and generate energy for the cell.

If RNA polymerase binds to P1, the DNA sequence of the lacI regulatory gene will be transcribed to an mRNA molecule, which is then translated to form its protein. The gene lacI produces a repressor protein. The protein binds to P2, as shown in Fig. 16(b), resulting in the RNA polymerase being blocked from binding P2 and starting transcription of the structural genes, including lacZ and lacY. The binding of RNA polymerase to P1 has blocked the transcription of the lactose digestive enzymes. Fig. 16(b) shows the equilibrium case. In fact, the repressor protein is rapidly binding and unbinding the DNA. The binding of RNA polymerase with DNA has a much lower rate constant, so it cannot compete with the repressor and cannot bind to the promoter. But, if the repressor-DNA rate constant decreases, the binding of RNA polymerase to DNA will become favored. The lacI repressor protein also binds to lactose with an even higher rate constant than with DNA. In the presence of lactose, the repressor preferentially binds with it and leaves the DNA free. This leaves P2 open for binding with RNA polymerase, as shown in Fig. 16(c), and allows the production of lactose digestive enzymes that break down lactose for energy. When all of the lactose is exhausted, the lacI repressor protein will be free to bind to P2 and suppress further enzyme production, and the system will return to the state of Fig. 16(b).

The *lac* operon represents a simple negative feedback system familiar to electrical engineers. Positive feedback loops also exist for some operons, whereby a protein binding to DNA promotes the production of downstream genes. An

example of this is the *ara* operon for arabidose regulation [11]. In these cases, the protein might change the 3-D structure of the DNA, making it easier for RNA polymerase to attach to the promoter. In general, promotion and repression occurs through a variety of DNA-protein interactions.

Soon after the operon idea was introduced, theoretical biologists tried to model it as a Boolean system: a gene was turned on (1) or off (0) by the promoter or repressor protein [112]. This led to important discoveries about the nature of network dynamics, autocatalytic sets, and how order emerges from the interconnected reactions of a cell [114], but ultimately failed to model the actual behavior of the biological system of interest [116]. However, a full treatment of a complicated biological system is still difficult, so a Boolean representation of the reactions is still used when the pathway size is large [117]. Modern simulations incorporate as much detail as possible: the transcription of DNA to mRNA, the translation of mRNA to protein at the ribosome, binding of the RNA polymerase to the DNA, promoter and repressor kinetics, and the decay of mRNA and protein molecules.

IX. Stochastic Pathway Simulation

Until recently, most modeling work consisted of representing the reactions in a pathway by coupled chemical kinetic ordinary differential equations (ODEs), with rate constants derived from experimental data. While this method is used extensively with quantitative success in industry for bulk enzyme production, it has only qualitatively reproduced pathway behavior in biological systems [128], [129]. Success appears to depend on the accuracy of the experimental parameters. These parameters come from traditional single-molecule biochemical experiments, since most of the high-throughput methods we described in previous sections are not reliably quantitative. However, microarray data are now being used to validate models, and to suggest the genes that should be included in the simulation.

These ODE models can successfully predict the average behavior of large numbers of cells. They fail, however, to simulate basic biological features that arise from stochastic effects. Recent studies have shown that biological systems are very noisy, and that much of cell behavior occurs because of that noise [130]–[132]. In the case of cancer, random effects with a low individual probability accumulate, causing dramatic changes. Moreover, a continuous representation of protein concentration fails for a single cell. An *E. coli* cell is so small that one protein molecule is equivalent to a 1-nM concentration. In some cases, fewer than ten copies of a particular regulatory protein are produced. And most importantly, there is just one molecule of DNA in a cell that can bind with proteins and produce mRNA transcripts. Therefore, any complete pathway model must necessarily be stochastic and discrete. McAdams and Arkin [133] presented a review of simulation for bacterial cells, including an extensive discussion of stochastic techniques.

A typical pathway can include thousands of proteins, so the state space is too large for an exact solution of the stochastic process of coupled reactions. Gillespie [134], [135] proposed a Monte Carlo method to exactly simulate the stochastic time evolution of a reaction system. The probability of each reaction occurring is a function of its rate constant from experimental kinetics data and the number of available reactants. At each point in time, there exists a joint probability distribution function for both the reaction and the time at which it would occur. The simulation generates an exactly correct random trajectory through the state space. Multiple simulations can be used to estimate the expected variance of the system as a function of the number of molecules participating in the pathway.

Arkin *et al.* [136] recently applied the Gillespie method to a fully stochastic model of *E. coli* infected by the λ phage virus. The operon model can be applied to virus genes that determine whether the cell is in the lysogeny or lysis pathway. Lysogeny is the state where the virus is reproducing by using the cell's machinery and lysis is the explosion of the cell and release of virus. The simulation incorporated the transcription and translation of five genes, as well as protein-protein and DNA-protein reactions of four regulatory proteins and two proteases (enzymes that destroy proteins). It represented a total of 32 chemical reactions. Some of these 32 reactions were compound transcription and translation reactions, consisting of hundreds of reaction events for the processing of each base. Implementing the simulation required parallel array supercomputers. Algorithmic improvements can reduce the number of required random numbers and reduce the time for state updates in the case where reactions are only coupled to a few others [137]. Part of the strategy is to group sequential, independent reactions into a single step. For example, the hundreds of transcription and translation reactions for a gene to produce one protein would be combined into a single random step by assuming that the reaction of each base is independent.

There are some important deficiencies with including only coupled volume chemical reactions. For example, the effects of diffusion and transport of the proteins through the cell are neglected. Also, many important reactions occur on membrane surfaces or other cellular structures. Fortunately, Gillespie's method has already been applied to surface chemistry [138] and may be adapted for membrane simulations. A more significant drawback of stochastic simulation is that the number of time steps that must be calculated grows at least $O(N)$ and on average $O(N^{\sqrt{2}})$ with the total number of molecules.

X. Reality Check: Pathway Simulation in the Real World

Operons are the most common kind of regulation in prokaryotic organisms (bacteria). However operons are the least complex kind of biological pathway regulation. Operons are found infrequently in the well studied and sequenced eukaryotic organisms *S. cerevisiae* (the single-celled baker's yeast) and *C. elegans* (the tiny nematode) [11]. As shown in Fig. 10, regulation occurs at every step of gene expression: the production of mRNA, the transport of mRNA, the translation of mRNA to protein, and

posttranslational protein modification. This adds considerably to the number of reactions that must be included if we want to model human systems using the simulations we just described.

The basic reason for the difference is that bacteria are *prokaryotes*: they have no nuclei isolating the DNA from the other components of the cell. In fact, the chromosome is usually attached to the cell wall. Furthermore, the prokaryotic chromosome is compact and genes are almost always continuous sequences. In prokaryotic gene transcription, as soon as the mRNA is produced it is translated on the ribosomes into proteins. The cells are compact, so molecules do not have far to go. Plants, fungi, and animals are *eukaryotes*: in their cells, DNA is contained entirely within the nucleus and reactions are compartmentalized. Also, genes are scattered throughout multiple chromosomes, and individual genes are typically coded in segments separated by noncoded regions. Promoter sequences also have complex structure, and there are additional "enhancer" sequences that control the rate of transcription. Promotion and repression requires a complex of different proteins called transcription factors working together. This adds enormous complexity. For example, *transcription factors* are commonly activated by *phosphorylation* (adding a phosphorus ion) at various sites in the protein. There are usually multiple phosphorylation sites on a single protein, and it is only necessary for a specific subset of these sites to be phosphorylated for the transcription factor to be active and the gene to be expressed. Thus, for N sites, the state space includes 2^N combinations: the activation of a site in one combination does not imply that it must be active in all combinations [139]. This presents a daunting task for modeling. For example, an important transcription factor in cancer research has 16 phosphorylation sites or 65 636 possible states.

Furthermore, after transcription, the pieces of mRNA have to be spliced together, the mRNA has to be stabilized for its journey through the nucleus, and it has to be transported to a ribosome. Regulation occurs by controlling mRNA stability, its ability to leave the nucleus, speed of binding to the ribosome, and speed of release from the ribosome [140]. After a protein is synthesized, it usually undergoes modification and has to be transported to a particular site in the cell. An analysis of a sea urchin developmental gene provides an example for the challenge of modeling regulation of just a single eukaryotic gene [141].

No matter what the complexity of the system, simulation requires experimental data. Most genomic data has been static: sequence and structure. Expression profiling with DNA chips and microarrays is a shift to dynamic information, but as we have discussed, these experiments are still qualitative. Expression profiling can help validate models, but cannot build them from scratch. Much of the data that simulation requires will come from enzymology experiments which can accurately measure reaction kinetics. New advances in molecular imaging allow observations of single molecule interactions *in vivo* that will accelerate the collection of reaction kinetic information [142]. Further progress in obtaining quantifiable data is needed, since without accurate pathway simulation, millions of mutation experiments will have to be conducted to analyze the data gathered in the Human Genome Project.

XI. Summary

Electrical and computer engineering has had an important role in completing the DNA sequence for the Human Genome Project. Laboratory automation and computer tools and systems have contributed to the Project's success. As biologists move forward to challenges in discerning genetic variation and function, the technologies and tools from electrical and computer engineering are increasingly important. Tools for larger databases, sharing and mining diverse data sets, digital image processing and pattern recognition, complex system simulation, and new measurement technologies are all needed for the challenges of functional and structural genomics. In the end, the data will profoundly change how we view the living world and ourselves.

Acknowledgment

The authors wish to thank the biologists of the LLNL Genomics Division, both at the Livermore Lab and the Walnut Creek production DNA sequencing facility, for many educational conversations and the opportunity to make a significant contribution to the Human Genome Project. The leadership and guidance of A. Carrano (former Director of Biology) and P. McCready (former Sequencing Director) are greatly appreciated. The authors also thank L. Ashworth, K. Fitch, and the reviewers for providing valuable comments on the manuscript.

References

[1] G. Mendel, "Versuche uber Pflanzen–Hybriden," *Verhandlungen des Naturforschenden Vereines, Abhandlungen, Brunn*, vol. 4, pp. 3–47, 1866.

[2] C. T. Druery and W. Bateson. Experiments in plant hybridization. [Online] an English translation of [1]. Available: http://www.stg.brown.edu/webs/MendelWeb/Mendel.html

[3] G. Mendel, *Experiments in Plant Hybridization*, J. H. Bennett, Ed. London, U.K.: Oliver and Boyd, 1965.

[4] S. E. Antonarakis, "Down syndrome," in *Princ. Molecular Medicine*, J. L. Jameson and E. W. Jabs, Eds. Totowa, NJ: Humana, 1998, ch. 119, pp. 1069–1078.

[5] O. T. Avery, C. M. MacLeod, and M. McCarty, "Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III," *J. Exp. Med.*, vol. 79, no. 2, pp. 137–158, Feb. 1 1944. [Online] http://www.profiles.nlm.nih.gov/CC/A/A/B/Y/_/ccaaby.pdf.

[6] A. Patrinos, *To Know Ourselves*, D. Vaughan, Ed: U.S. Dept. Energy Rep., PUB-773, July 1996, pp. 2–3. [Online] http://www.lbl.gov/Publications/TKO.

[7] F. S. Collins, A. Patrinos, E. Jordan, A. Chakravarti, R. Gesteland, and L. Walters, "New goals for the U.S. Human Genome Project: 1998–2003," *Science*, vol. 282, pp. 682–689, Oct. 1998.

[8] E. Pennisi, "DOE team sequences three chromosomes," *Science*, vol. 288, pp. 417–419, Apr. 2000.

[9] E. Marshall, "Rival genome sequencers celebrate a milestone together," *Science*, vol. 288, pp. 2294–2295, June 2000.

[10] National Center for Biotechnology Information home page. [Online]. Available: http://www.ncbi.nlm.nih.gov/

[11] M. Watson, J. D. Hopkins, N. H. Roberts, J. W. Steitz, J. A. , and A. M. Weiner, *Molecular Biology of the Gene*, 4th ed. Menlo Park, CA: Benjamin/Cummings, 1987.

[12] B. R. Glick and J. J. Pasternak, *Molecular Biotechnology: Principles and Applications of Recombinant DNA*. Washington, DC: Am. Soc. Microbiology, 1998.
[13] H. R. Drew, R. M. Wing, T. Takano, C. Broka, S. Tanaka, K. Itakura, and R. E. Dickerson, "Structure of a B-DNA dodecamer: Conformation and dynamics," in *Proc. Natl. Acad. Sci. USA*, vol. 78, Apr. 1981, pp. 2179–2183.
[14] DOE Joint Genome Institute. [Online]. Available: http://www.jgi.doe.gov
[15] Sanger centre sequencing statistics. [Online]. Available: http://www.sanger.ac.uk/HGP/stats.shtml
[16] Finished sequence totals. Genome Sequencing Center, Washington Univ., St. Louis, MO. [Online]. Available: http://genome.wustl.edu/gsc/breakdown.html
[17] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," in *Proc. Natl. Acad. Sci. USA*, vol. 74, Dec. 1977, pp. 5463–5467.
[18] P. D. Grossman and J. C. Colburn, *Capillary Electrophoresis: Theory & Practice*. San Diego, CA: Academic, 1992.
[19] T. Hunkapiller, R. J. Kaiser, B. F. Koop, and L. Hood, "Large-scale and automated DNA sequence determination," *Science*, vol. 254, pp. 59–67, Oct. 1991.
[20] L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. H. Heiner, S. H. B. Kent, and L. E. Hood, "Fluorescence detection in automated DNA sequence analysis," *Nature*, vol. 321, pp. 674–679, June 1986.
[21] L. M. Smith, S. Fung, M. W. Hunkapiller, and L. E. Hood, "The synthesis of oligonucleotides containing and aliphatic amino group at the 5′ terminus: Synthesis of fluorescent DNA primers for use in DNA sequence analysis," *Nucleic Acids Res.*, vol. 13, pp. 2399–2405, Apr. 1985.
[22] L. M. Smith, J. Z. Sanders, and L. E. Hood, "The synthesis and use of fluorescent oligonucleotides in DNA sequence analysis," *Meth. Enzymoogyl.*, vol. 155, pp. 260–301, 1987.
[23] R. S. Madabhushi, "Separation of 4-color DNA sequencing extension products in noncovalently coated capillaries using low viscosity polymer solutions," *Electrophoresis*, vol. 2, pp. 224–230, Feb. 1998.
[24] R. S. Madabhushi, M. Vainer, V. Dolnik, S. Enad, D. L. Barker, D. W. Harris, and E. S. Mansfield, "Versatile low-viscosity sieving matrices for nondenaturing DNA separations using capillary array electrophoresis," *Electrophoresis*, vol. 1, pp. 104–111, Jan. 1997.
[25] A. S. Cohen and B. L. Karger, "High-performance sodium dodecyl sulfate polyacrylamide gel capillary electrophoresis of peptides and proteins," *J. Chromatography*, vol. 397, pp. 409–417, June 1987.
[26] A. S. Cohen, A. Paulus, and B. L. Karger, "High-performance capillary electrophoresis using open tubes and gels," *Chromatographia*, vol. 24, pp. 15–24, 1987.
[27] A. Manz, D. J. Harrison, E. M. J. Verpoorte, J. C. Fettinger, A. Paulus, H. Ludi, and H. M. Widmer, "Planar chips technology for miniaturization and integration of separation techniques into monitoring systems: Capillary electrophoresis on a chip," *J. Chromatography*, vol. 593, pp. 253–258, Feb. 1992.
[28] D. J. Harrison, A. Manz, Z. Fan, H. Ludi, and H. M. Widmer, "Capillary electrophoresis and sample injection systems integrated on a planar glass chip," *Anal. Chem.*, vol. 64, pp. 1926–1932, Sept. 1992.
[29] D. J. Harrison, K. Fluri, K. Seiler, Z. Fan, C. S. Effenhauser, and A. Manz, "Micromachining a miniaturized capillary electrophoresis-based chemical analysis system on a chip," *Science*, vol. 261, pp. 895–897, Aug. 1993.
[30] I. Kheterpal, J. R. Scherer, S. M. Clark, A. Radhakrishnan, J. Ju, C. L. Ginther, G. F. Sensabaugh, and R. A. Mathies, "DNA sequencing using four-color fluorescence capillary array scanner," *Electrophoresis*, vol. 17, pp. 1852–1859, Dec. 1996.
[31] P. Lindberg, M. Stjernstrom, and J. Roeraade, "Gel electrophoresis of DNA fragments in narrow-bore capillaries," *Electrophoresis*, vol. 18, pp. 1973–1979, Oct. 1997.
[32] D. Schmalzing, L. Koutny, A. Adourian, P. Belgrader, P. Matsudaira, and D. Ehrlich, "DNA typing in thirty seconds with a microfabricated device," in *Proc. Natl. Acad. Sci. USA*, vol. 94, Sept. 1997, pp. 10 273–10 278.
[33] P. C. Simpson, A. T. Woolley, and R. A. Mathies, "Microfabrication technology for the production of capillary array electrophoresis chips," *Biomed. Microdevices*, vol. 1, Apr. 1998.
[34] C. H. Mastrangelo, M. A. Burns, and D. T. Burke, "Microfabricated devices for genetic diagnostics," *Proc. IEEE*, vol. 86, pp. 1769–1787, Aug. 1998.
[35] D. Schmalzing, A. Adourian, L. Koutny, L. Ziaugra, P. Matsudaira, and D. Ehrlich, "DNA sequencing on microfabricated electrophoretic devices," *Anal. Chem.*, vol. 70, pp. 2303–2310, June 1998.
[36] S. Swierkowski, J. W. Balch, L. R. Brewer, A. C. Copeland, J. C. Davidson, J. P. Fitch, J. R. Kimbrough, R. S. Madabhushi, R. L. Pastrone, P. M. Richardson, L. A. Tarte, and M. Vainer, "Large microchannel array fabrication and results for DNA sequencing," *Proc. SPIE*, vol. 3606, Jan. 1999.
[37] Y. Xiong, S.-R. Park, and H. Swerdlow, "Base stacking: pH-mediated on-column sample concentration for capillary DNA sequencing," *Anal. Chem.*, vol. 70, no. 17, pp. 3605–3611, Sept. 1998.
[38] Applied Biosystems specifications for the ABI Prism 3700 DNA analyzer. Applied Biosystems, Foster City, CA. [Online] (formerly PE Biosystems). Available: http://www.appliedbiosystems.com/ga/3700/specs.html
[39] D. Swanson, "The art of the state of nucleic acid sequencing: Highly refined automated sequencing systems keep up with demand for large-scale genomic projects," *Scientist*, vol. 14, no. 3, p. 23, Feb. 2000.
[40] S. C. Jacobson, R. Hergenroder, L. B. Koutny, R. J. Warmack, and J. M. Ramsey, "Effects of injection schemes and column geometry on the performance of microchip electrophoresis devices," *Anal. Chem.*, vol. 66, pp. 1107–1113, Apr. 1994.
[41] I. Kheterpal, J. Ju, G. S. Brandt, C. L. Ginther, S. M. Clark, J. R. Scherer, G. F. Sensabaugh, and R. A. Mathies, *Ultrasensitive Biochemical Diagnostics*, G. E. Cohn, S. A. Soper, and C. H. W. Chen, Eds. Bellingham, WA: SPIE, 1996, pp. 204–213.
[42] A. J. Kostichka, M. Marchbanks, R. L. Brumley, H. Drossman, and L. M. Smith, "High speed automated DNA sequencing in ultrathin slab gels," *Bio/Technol.*, vol. 10, no. 1, pp. 78–81, 1992.
[43] H. Swerdlow, S. L. Wu, H. Harke, and N. J. Dovichi, "Capillary gel electrophoresis for DNA sequencing, laser-induced fluorescence detection with the sheath flow cuvette," *J. Chromatography*, vol. 516, pp. 61–67, 1990.
[44] R. A. Wallingford and A. G. Ewing, "Capillary zone electrophoresis with electrochemical detection," *Anal. Chem.*, vol. 59, pp. 1762–1766, July 1987.
[45] A. T. Woolley, K. Lao, A. N. Glazer, and R. A. Mathies, "Capillary electrophoresis chips with integrated electrochemical detection," *Anal. Chem.*, vol. 70, pp. 684–688, Feb. 1998.
[46] W. Huang, Z. Yin, D. R. Fuhrmann, D. J. States, and L. J. Thomas Jr., "A method to determine the filter matrix in four-dye fluorescence-based DNA sequencing," *Electrophoresis*, vol. 18, pp. 23–25, Jan. 1997.
[47] Z. Yin, J. Severin, M. C. Giddings, W. Huang, M. S. Westphall, and L. M. Smith, "Automatic matrix determination in four dye fluorescence-based DNA sequencing," *Electrophoresis*, vol. 17, pp. 1143–1150, June 1996.
[48] J. P. Fitch, J. W. Balch, M. S. Bass, L. R. Brewer, A. C. Copeland, J. C. Davidson, L. M. Kegelmeyer, J. R. Kimbrough, R. S. Madabhushi, P. M. McCready, D. O. Nelson, R. L. Pastrone, P. M. Richardson, S. P. Swierkowski, L. A. Tarte, M. Vainer, and T. E. Warth, "A microchannel electrophoresis DNA sequencing system," in *Proc. 33rd CISS*, Baltimore, MD, Mar. 17–19, 1999.
[49] B. Ewing, L. Hillier, M. C. Wendl, and P. Green, "Base-calling of automated sequencer traces using *Phred*: I. Accuracy assessment," *Genome Res.*, vol. 8, pp. 175–185, Mar. 1998.
[50] B. Ewing and P. Green, "Base-calling of automated sequencer traces using *Phred*: II. Error probabilities," *Genome Res.*, vol. 8, pp. 186–194, Mar. 1998.
[51] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J. D. Gocayne, J. Scott, R. Shirley, L. Liu, A. Glodek, J. M. Kelley, J. G. Weidman, C. A. Phillips, T. Spriggs, E. Hedblom, M. D. Cotton, T. R. Utterback, M. C. Hanna, D. T. Nguyen, D. M. Saudek, R. C. Brandon, L. D. Fine, J. L. Fritchman, J. L. Fuhrmann, N. S. M. Geoghagen, C. L. Gnehm, L. A. McDonald, K. V. Small, C. M. Fraser, H. O. Smith, and J. C. Venter, "Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*," *Science*, vol. 269, pp. 496–512, July 1995.
[52] J. C. Venter, M. D. Adams, G. G. Sutton, A. R. Kerlavage, H. O. Smith, and M. Hunkapiller, "Shotgun sequencing of the human genome," *Science*, vol. 280, pp. 1540–1542, June 1998.

[53] D. Meldrum, "Automation for genomics, Part One: Preparation for sequencing," *Genome Res.*, vol. 10, pp. 1081–1092, Aug. 2000.
[54] ——, "Automation for genomics, Part Two: Sequencers, microarrays, and future trends," *Genome Res.*, vol. 10, pp. 1288–1303, Sept. 2000.
[55] B. F. Ouellette and M. S. Boguski, "Database divisions and homology search files: A guide for the perplexed," *Genome Res.*, vol. 7, Oct. 1997.
[56] H. Y. Zoghbi, "Spinocerebellar ataxia and other disorders of trinucleotide repeats," in *Princ. Molecular Medicine*, J. L. Jameson and J. B. Martin, Eds. Totowa, NJ: Humana, 1998, ch. 100, pp. 917–918.
[57] H. G. Harley, K. V. Walsh, S. Rundle, J. D. Brook, M. Sarfarazi, M. C. Koch, J. L. Floyd, P. S. Harper, and D. J. Shaw, "Localisation of the myotonic dystrophy locus to 19q13.2–19q13.3 and its relationship to twelve polymorphic loci on 19q," *Human Genetics*, vol. 87, pp. 73–80, May 1991.
[58] D. J. Shaw, M. McCurrach, S. A. Rundle, H. G. Harley, S. R. Crow, R. Sohn, J. P. Thirion, M. G. Hamshere, A. J. Buckler, and P. S. Harper, "Genomic organization and transcriptional units at the myotonic dystrophy locus," *Genomics*, vol. 18, pp. 673–679, Dec. 1993.
[59] M. S. Mahadevan, C. T. Amemiya, G. Jansen, L. Sabourin, S. Baird, C. E. Neville, N. Wormskamp, B. Segers, J. Lamerdin, P. de Jong, B. Wieringa, and R. G. Korneluk, "Structure and genomic sequence of the myotonic dystrophy (DM Kinase) gene," *Human Molecular Genetics*, vol. 2, pp. 299–304, Mar. 1993.
[60] J. H. Badger and G. J. Olsen, "CRITICA: Coding region identification tool invoking comparative analysis," *Molecular Biol. Evol.*, vol. 16, pp. 512–524, Apr. 1999.
[61] NCBI ORF Finder (Open reading frame finder) . [Online]. Available: http://www.ncbi.nlm.nih.gov/gorf/orfig.cgi
[62] A. V. Lukashin and M. Borodovsky, "GeneMark.hmm: New solutions for gene finding," *Nucleic Acids Res.*, vol. 26, pp. 1107–1115, Feb. 1998.
[63] M. Reese, G. Hartzell, N. L. Harris, U. Ohler, J. F. Abril, and S. E. Lewis, "Genome annotation assessment in Drosophila melanogaster," *Genome Res.*, vol. 10, pp. 483–501, Apr. 2000.
[64] R. D. Perry and J. D. Fetherston, "Yersinia pestis: Etiologic agent of plague," *Clinical Microbiol. Rev.*, vol. 10, pp. 35–66, Jan. 1997.
[65] C. M. Fraser *et al.*, "The minimal gene complement of *Mycoplasma genitalium*," *Science*, vol. 270, pp. 397–403, Oct. 1995.
[66] M. Missler and T. C. Sudhof, "Neurexins: Three genes and 1001 products," *Trends Genetics*, vol. 14, pp. 20–26, Jan. 1998.
[67] A. C. Guyton and J. E. Hall, *Textbook of Medical Physiology*. Philadelphia, PA: Saunders, 1996, ch. 7, pp. 81–84.
[68] V. E. Velculescu, L. Zhang, W. Zhou, J. Vogelstein, M. A. Basrai, D. E. J. Bassett, P. Hieter, B. Vogelstein, and K. W. Kinzler, "Characterization of the yeast transcriptome," *Cell*, vol. 88, pp. 243–251, Jan. 1997.
[69] Serial analysis of gene expression. [Online]. Available: http://www.sagenet.org
[70] J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, pp. 680–686, Oct. 1997.
[71] C. S. Richmond, J. D. Glasner, R. Mau, H. Jin, and F. R. Blattner, "Genome-wide expression profiling in *Escherichia coli* K-12," *Nucleic Acids Res.*, vol. 27, pp. 3821–3835, Oct. 1999.
[72] D. D. L. Blowtell, "Options available—from start to finish—for obtaining expression data by microarray," *Nature Genetics*, vol. 21, pp. 25–32, Jan. 1999.
[73] S. Granjeaud, F. Bertucci, and B. R. Jordan, "Expression profiling: DNA arrays in many guises," *BioEssays*, vol. 21, pp. 781–790, Sept. 1999.
[74] S. P. A. Fodor, "DNA sequencing: Massively parallel genomics," *Science*, vol. 277, p. 393, July 1997.
[75] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobyashi, H. Horton, and E. L. Brown, "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nature Biotechnol.*, vol. 14, pp. 1675–1680, Dec. 1996.
[76] K. B. Mullis, F. A. Faloona, S. J. Scharf, R. K. Saiki, G. T. Horn, and H. A. Erlich, "Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction," in *Cold Spring Harbor Symp. Quant. Biol.*, vol. 51, 1986, pp. 263–273.
[77] The brown lab's complete guide to microarraying for the molecular biologist. [Online]. Available: http://cmgm.stanford.edu/pbrown/mguide/index.html
[78] *Nature Genetics*, Jan. 1999, vol. 21.
[79] Y. Chen, E. R. Dougherty, and M. L. Bittner, "Ratio-based decisions and the quantitative analysis of cDNA microarray images," *J. Biomed. Opt.*, vol. 2, pp. 364–374, Oct. 1997.
[80] Exploring the metabolic and genetic control of gene expression on a genomic scale. [Online]. Available: http://cmgm.stanford.edu/pbrown/explore/
[81] O. Ermolaeva, M. Rastogi, K. D. Pruitt, G. D. Schuler, M. L. Bittner, Y. Chen, R. Simon, P. Meltzer, J. M. Trent, and M. S. Boguski, "Data management and analysis for gene expression arrays," *Nature Genetics*, vol. 20, pp. 19–23, Sept. 1998.
[82] Sandia National Laboratories VxInsight software home page. [Online]. Available: http://www.cs.sandia.gov/projects/VxInsight.html
[83] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," in *Proc. Natl. Acad. Sci. USA*, vol. 95, Dec. 1998, pp. 14863–14868.
[84] P. Bucher, "Regulatory elements and expression profiles," *Current Opinion in Structural Biol.*, vol. 9, pp. 400–407, June 1999.
[85] W. Waserman and J. Fickett, "Identification of regulatory regions which confer muscle-specific gene expression," *J. Molecular Biol.*, vol. 278, pp. 167–181, Apr. 1998.
[86] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering gene expression patterns," *J. Computat. Biol.*, vol. 6, pp. 281–297, 1999.
[87] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation," in *Proc. Natl. Acad. Sci. USA*, vol. 96, Mar. 1999, pp. 2907–2912.
[88] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines," in *Proc. Natl. Acad. Sci. USA*, vol. 97, Jan. 2000, pp. 262–267.
[89] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nature Genetics*, vol. 22, pp. 281–285, July 1999.
[90] M. S. Waterman, *Introduction to Computational Biology*. Cambridge, U.K.: Chapman & Hall, 1995.
[91] D. Gusfield, *Algorithms on Strings, Trees, and Sequences*. Cambridge, U.K.: Cambridge Univ. Press, 1997.
[92] NCBI BLAST tutorial. [Online]. Available: http://www.ncbi.nlm.nih.gov/Education/blasttutorial.html
[93] S. K. Burley, S. C. Almo, J. B. Bonanno, M. Capel, M. R. Chance, T. Gaasterland, D. Lin, A. Sali, F. W. Studier, and S. Swaminathan, "Structural genomics: Beyond the human genome project," *Nature Genetics*, vol. 23, pp. 151–157, Oct. 1999.
[94] T. Kuczmarski *et al.*, unpublished.
[95] The Research Collaboratory for Structural Bioinformatics Protein Data Bank home page. [Online]. Available: http://www.rcsb.org
[96] B. S. Trakhanov, S. Parkin, R. Raffai, R. Milne, Y. M. Newhouse, K. H. Weisgraber, and B. Rupp, "Structure of a monoclonal 2E8 Fab antibody fragment specific for the low-density lipoprotein-receptor binding region of apolipoprotein E refined at 1.9 Å," *Acta Crystallographica D*, vol. 55, pp. 122–128, Jan. 1999.
[97] C. R. Cantor and P. R. Schimmel, *The Behavior of Biological Molecules. Part II: Techniques for the Study of Biological Structure and Function*. New York: Freeman, 1990.
[98] Nature of 3D structure data. [Online]. Available: http://www.rcsb.org/pdb/experimental_methods.html
[99] H. N. Moseley and G. T. Montelione, "Automated analysis of NMR assignments and structures for proteins," *Current Opinion in Structural Biol.*, vol. 9, pp. 635–642, Oct. 1999.
[100] A. Sali, "100 000 protein structures for the biologist," *Nature Structural Biol.*, vol. 5, pp. 1029–1031, Dec. 1998.
[101] T. C. Terwilliger and J. Berendzen, "Automated MAD and MIR structure solution," *Acta Crystallographica D*, vol. 55, pp. 849–861, Nov. 1999.
[102] PDB current holdings, May 9, 2000. [Online]. Available: http://www.rcsb.org/pdb/holdings.html
[103] NCBI Entrez-Genome statistics. [Online]. Available: http://www.ncbi.nlm.nih.gov/Entrez/Genome/main_genomes.html
[104] H. Weissig and P. E. Bourne, "An analysis of the protein data bank in search of temporal and global trends," *Bioinformatics*, vol. 15, pp. 715–716, Oct. 1999.
[105] R. Sanchez and A. Sali, "Comparative protein structure modeling in genomics," *J. Computat. Phys.*, vol. 151, pp. 388–401, May 1999.
[106] A. Godzik, "The structural alignment between two proteins: Is there a unique answer?," *Protein Sci.*, vol. 5, pp. 1325–1338, July 1996.

[107] D. L. Gerloff, M. Joachimiak, F. E. Cohen, G. M. Cannarozzi, S. G. Chamberlin, and S. A. Benner, "Structure prediction in a post-genomic environment: A secondary and tertiary structural model for the initiation factor 5A family," *Biochem. Biophys. Res. Commun.*, vol. 251, pp. 173–181, Oct. 1998.

[108] E. M. Marcotte, M. Pellegrini, M. J. Thompson, T. O. Yeates, and D. Eisenberg, "A combined algorithm for genome-wide prediction of protein function," *Nature*, vol. 402, pp. 83–86, Nov. 1999.

[109] *Proteins: Structure, Function, and Genetics*, 1999, vol. 37.

[110] K. A. Dill, "Theory for the folding and stability of globular proteins," *Biochem.*, vol. 24, pp. 1501–1509, Mar. 1985.

[111] J. P. Erzberger, D. Barsky, O. D. Schärer , M. E. Colvin, and D. M. Wilson, "Elements in abasic site recognition by the major human and *Escherichia coli* apurinic/apyrimidinic endonucleases," *Nucleic Acids Res.*, vol. 26, pp. 2771–2778, June 1998.

[112] T. Schlick, E. Barth, and M. Mandziuk, "Biomolecular dynamics at long timesteps: bridging the timescale gap between simulation and experimentation," *Annu. Rev. Biophys. Biomolecular Structure*, vol. 26, pp. 181–222, 1997.

[113] IBM announces $100 million research initiative to build world's fastest supercomputer. IBM Press Release. [Online]. Available: http://www.research.ibm.com/bluegene

[114] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold, "Correlation between protein and mRNA abundance in yeast," *Molecular Cellular Biol.*, vol. 19, pp. 1720–1730, Mar. 1999.

[115] L. Anderson and J. Seilhamer, "A comparison of selected mRNA and protein abundances in human liver," *Electrophoresis*, vol. 18, pp. 533–537, Mar.–Apr. 1997.

[116] V. Hatzimanikatis, L. H. Choe, and K. H. Lee, "Proteomics: Theoretical and experimental considerations," *Biotechnol. Progress*, vol. 15, pp. 312–316, May–June 1999.

[117] I. P. James, "Of genomes and proteomes," *Biochem. Biophys. Res. Commun.*, vol. 231, pp. 1–6, Feb. 1997.

[118] J. M. Quadroni and P. James, "Proteomics and automation," *Electrophoresis*, vol. 20, pp. 664–677, Apr.–May 1999.

[119] M. L. Bulyk, E. Gentalen, D. J. Lockhart, and G. M. Church, "Quantifying DNA-protein interactions by double-stranded DNA arrays," *Nature Biotechnol.*, vol. 17, pp. 573–577, June 1999.

[120] Expert Protein Analysis System (ExPASy) molecular biology server.. [Online] HYPERLINK . Available: http://www.expasy.ch http://www.expasy.ch

[121] T. S. Gardner, C. R. Cantor, and J. J. Collins, "Construction of a genetic toggle switch in Escherichia coli," *Nature*, vol. 403, pp. 339–342, Jan. 2000.

[122] E. Marshall, "Improving gene therapy's tool kit," *Science*, vol. 288, p. 953, May 2000.

[123] F. Jacob and J. Monod, "Genetic regulatory mechanisms in the synthesis of proteins," *J. Molecular Biol.*, vol. 3, pp. 318–356, June 1961.

[124] M. Sugita, "Functional analysis of chemical systems *in vivo* using a logical circuit equivalent," *J. Theoretical Biol.*, vol. 1, pp. 415–430, Oct. 1961.

[125] S. A. Kauffman, *The Origins of Order*. New York: Oxford Univ. Press, 1993.

[126] L. Glass and S. A. Kauffman, "The logical analysis of continuous, nonlinear biochemical control networks," *J. Theoretical Biol.*, vol. 39, pp. 103–129, Apr. 1973.

[127] C. H. Schilling and B. O. Palsson, "The underlying pathway structure of biochemical reaction networks," in *Proc. Natl. Acad. Sci. USA*, vol. 95, Apr. 1998, pp. 4193–4198.

[128] B. J. Hammond, "Quantitative study of the control of HIV-1 gene expression," *J. Theoretical Biol.*, vol. 163, pp. 199–221, July 1993.

[129] D. Endy, D. Kong, and J. Yin, "Intracellular kinetics of a growing virus: A genetically structured simulation for bacteriophage T7," *Biotechnol. Bioeng.*, vol. 55, pp. 375–389, July 1997.

[130] H. H. McAdams and A. Arkin, "It's a noisy business! Genetic regulation at the nanomolar scale," *Trends in Genetics*, vol. 15, pp. 65–69, Feb. 1999.

[131] ——, "Stochastic mechanisms in gene expression," in *Proc. Natl. Acad. Sci. USA*, vol. 94, Feb. 1997, pp. 814–819.

[132] M. B. Elowitz and S. Leibler, "A synthetic oscillatory network of transcriptional regulators," *Nature*, vol. 403, pp. 335–338, Jan. 2000.

[133] H. H. McAdams and A. Arkin, "Simulation of prokaryotic genetic circuits," *Annu. Rev. Biophys. Biomolecular Structure*, vol. 27, pp. 199–224, 1998.

[134] D. T. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *J. Computat. Phys.*, vol. 22, pp. 403–434, Dec. 1976.

[135] ——, "Exact stochastic simulation of coupled chemical reactions," *J. Phys. Chem.*, vol. 81, pp. 2340–2361, Dec. 1977.

[136] A. Arkin, J. Ross, and H. H. McAdams, "Stochastic kinetic analysis of developmental pathway bifurcation in phage _infected Escherichia coli cells," *Genetics*, vol. 149, pp. 1633–1648, Aug. 1998.

[137] M. A. Gibson and J. Bruck, "Efficient exact stochastic simulation of chemical systems with many species and many channels," *J. Phys. Chem. A*, vol. 104, pp. 1876–1889, Mar. 2000.

[138] J. J. Lukkien, J. P. L. Segers, P. A. J. Hilbers, R. J. Gellen, and A. P. J. Jansen, "Efficient Monte Carlo methods for the simulation of catalytic surface reactions," *Phys. Rev. E*, vol. 58, pp. 2598–2610, Aug. 1998.

[139] A. Komeili and E. K. O'Shea, "Roles of phosphorylation in regulating activity of the transcription factor Pho4," *Science*, vol. 284, pp. 977–980, May 1999.

[140] J. E. G. McCarthy, "Post transcriptional control of gene expression in yeast," *Microbiol. Molecular Biol. Rev.*, vol. 62, pp. 1492–1553, Dec. 1998.

[141] C. H. Yuh, H. Bolouri, and E. H. Davidson, "Genomic Cis-regulatory logic: Experimental and computational analysis of a sea urchin gene," *Science*, vol. 279, pp. 1896–1902, Mar. 1998.

[142] R. S. Nie and R. N. Zare, "Optical detection of single molecules," *Annu. Rev. Biophys. Biomolecular Structure*, vol. 26, pp. 567–596, 1998.

[143] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver, "Life with 6000 genes," *Science*, vol. 274, pp. 546–567, Oct. 1996.

[144] "Genome Sequence of the Nematode *Caenorhabditis Elegans*. A platform for investigating biology," *Science*, vol. 282, pp. 2012–2018, Dec. 1998.

[145] I. Dunham, A. R. Hunt, J. E. Collins, R. Bruskiewich, D. M. Beare, M. Clamp, L. J. Smink, R. Ainscough, J. P. Almeida, A. Babbage, C. Bagguley, J. Bailey, K. Barlow, K. N. Bates, O. Beasley, C. P. Bird, S. Blakey, A. M. Bridgeman, D. Buck, J. Burgess, W. D. Burrill, J. Burton, C. Carder, N. P. Carter, Y. Chen, G. Clark, S. M. Clegg, V. Cobley, C. G. Cole, R. E. Collier, R. E. Connor, D. Conroy, N. Corby, G. J. Coville, A. V. Cox, J. Davis, E. Dawson, P. D. Dhami, C. Dockree, S. J. Dodsworth, R. M. Durbin, A. Ellington, K. L. Evans, J. M. Fey, K. Fleming, L. French, A. A. Garner, J. G. R. Gilbert, M. E. Goward, D. Grafham, M. N. Griffiths, C. Hall, R. Hall, G. Hall-Tamlyn, R. W. Heathcott, S. Ho, S. Holmes, S. E. Hunt, M. C. Jones, J. Kershaw, A. Kimberley, A. King, G. K. Laird, C. F. Langford, M. A. Leversha, C. Lloyd, D. M. Lloyd, I. D. Martyn, M. Mashreghi-Mohammadi, L. Matthews, O. T. Mccann, J. Mcclay, S. Mclaren, A. A. Mcmurray, S. A. Milne, B. J. Mortimore, C. N. Odell, R. Pavitt, A. V. Pearce, D. Pearson, B. J. Phillimore, S. H. Phillips, R. W. Plumb, H. Ramsay, Y. Ramsey, L. Rogers, M. T. Ross, C. E. Scott, H. K. Sehra, C. D. Skuce, S. Smalley, M. L. Smith, C. Soderlund, L. Spragon, C. A. Steward, J. E. Sulston, R. M. Swann, M. Vaudin, M. Wall, J. M. Wallis, M. N. Whiteley, D. Willey, L. Williams, S. Williams, H. Williamson, T. E. Wilmer, L. Wilming, C. L. Wright, T. Hubbard, D. R. Bentley, S. Beck, J. Rogers, N. Shimizu, S. Minoshima, K. Kawasaki, T. Sasaki, S. Asakawa, J. Kudoh, A. Shintani, K. Shibuya, Y. Yoshizaki, N. Aoki, S. Mitsuyama, B. A. Roe, F. Chen, L. Chu, J. Crabtree, S. Deschamps, A. Do, T. Do, A. Dorman, F. Fang, Y. Fu, P. Hu, A. Hua, S. Kenton, H. Lai, H. I. Lao, J. Lewis, S. Lewis, S.-P. Lin, P. Loh, E. Malaj, T. Nguyen, H. Pan, S. Phan, S. Qi, Y. Qian, L. Ray, Q. Ren, S. Shaull, D. Sloan, L. Song, Q. Wang, Y. Wang, Z. Wang, J. White, D. Willingham, H. Wu, Z. Yao, M. Zhan, G. Zhang, S. Chissoe, J. Murray, N. Miller, P. Minx, R. Fulton, D. Johnson, G. Bemis, D. Bentley, H. Bradshaw, S. Bourne, M. Cordes, Z. Du, L. Fulton, D. Goela, T. Graves, J. Hawkins, K. Hinds, K. Kemp, P. Latreille, D. Layman, P. Ozersky, T. Rohlfing, P. Scheet, C. Walker, A. Wamsley, P. Wohldmann, K. Pepin, J. Nelson, I. Korf, J. A. Bedell, L. Hillier, E. Mardis, R. Waterston, R. Wilson, B. S. Emanuel, T. Shaikh, H. Kurahashi, S. Saitta, M. L. Budarf, H. E. Mcdermid, A. Johnson, A. C. C. Wong, B. E. Morrow, L. Edelmann, U. J. Kim, H. Shizuya, M. I. Simon, J. P. Dumanski, M. Peyrard, D. Kedra, E. Seroussi, I. Fransson, I. Tapia, C. E. Bruder, and K. P. O'Brien, "The DNA sequence of human chromosome 22," *Nature*, vol. 402, no. 6761, pp. 489–495, Dec. 1999.

[146] M. D. Adams, S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, R. A. George, S. E. Lewis, S. Richards, M. Ashburner, S. N. Henderson, G. G. Sutton, J. R. Wortman, M. D. Yandell, Q. Zhang, L. X. Chen, R. C. Brandon, Y. H. C. Rogers, R. G. Blazej, M. Champe, B. D. Pfeiffer, K. H. Wan, C. Doyle, E. G. Baxter, G. Helt, C. R. Nelson, G. L. G. Miklos, J. F. Abril, A. Agbayani, H. J. An, C. Andrews-Pfannkoch, D. Baldwin, R. M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E. M. Beasley, K. Y. Beeson, P. V. Benos, B. P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M. R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K. C. Burtis, D. A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J. M. Cherry, S. Cawley, C. Dahlke, L. B. Davenport, P. Davies, B. de Pablos, A. Delcher, Z. Deng, A. D. Mays, I. Dew, S. M. Dietz, K. Dodson, L. E. Doup, M. Downes, S. Dugan-Rocha, B. C. Dunkov, P. Dunn, K. J. Durbin, C. C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A. E. Gabrielian, N. S. Garg, W. M. Gelbart, K. Glasser, A. Glodek, F. Gong, J. H. Gorrell, Z. Gu, P. Guan, M. Harris, N. L. Harris, D. Harvey, T. J. Heiman, J. R. Hernandez, J. Houck, D. Hostin, K. A. Houston, T. J. Howland, M. H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G.H. Karpen, Z. Ke, J. A. Kennison, K. A. Ketchum, B. E. Kimmel, C. D. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, A. A. Levitsky, J. Li, Z. Li, Y. Liang, X. Lin, X. Liu, B. Mattei, T. C. McIntosh, M. P. McLeod, D. McPherson, G. Merkulov, N. V. Milshina, C. Mobarry, J. Morris, A. Moshrefi, S. M. Mount, M. Moy, B. Murphy, L. Murphy, D. M. Muzny, D. L. Nelson, D. R. Nelson, K. A. Nelson, K. Nixon, D. R. Nusskern, J. M. Pacleb, M. Palazzolo, G. S. Pittman, S. Pan, J. Pollard, V. Puri, M. G. Reese, K. Reinert, K. Remington, R. D. C. Saunders, F. Scheeler, H. Shen, B. C. Shue, I. Sidén-Kiamos , M. Simpson, M. P. Skupski, T. Smith, E. Spier, A. C. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, A. H. Wang, X. Wang, Z. Y. Wang, D. A. Wassarman, G. M. Weinstock, J. Weissenbach, S. M. Williams, T. Woodage, K. C. Worley, D. Wu, S. Yang, Q. A. Yao, J. Ye, R. F. Yeh, J. S. Zaveri, M. Zhan, G. Zhang, Q. Zhao, L. Zheng, X. H. Zheng, F. N. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, H. O. Smith, R. A. Gibbs, E. W. Myers, G. M. Rubin, and J. C. Venter, "The genome sequence of Drosophila melanogaster," *Science*, vol. 287, pp. 2185–2195, Mar. 2000.

[147] M. Hattori, A. Fujiyama, T. D. Taylor, H. Watanabe, T. Yada, H.-S. Park, A. Toyoda, K. Ishii, Y. Totoki, D.-K. Choi, E. Soeda, M. Ohki, T. Takagi, Y. Sakaki, S. Taudien, K. Blechschmidt, A. Polley, U. Menzel, J. Delabar, K. Kumpf, R. Lehmann, D. Patterson, K. Reichwald, A. Rump, M. Schillhabel, A. Schudy, W. Zimmermann, A. Rosenthal, J. Kudoh, K. Shibuya, K. Kawasaki, S. Asakawa, A. Shintani, T. Sasaki, K. Nagamine, S. Mitsuyama, S. E. Antonarakis, S. Minoshima, N. Shimizu, G. Nordsiek, K. Hornischer, P. Brandt, M. Scharfe, O. Schon, A. Desario, J. Reichelt, G. Kauer, H. Blocker, J. Ramser, A. Beck, S. Klages, S. Hennig, L. Riesselmann, E. Dagand, T. Haaf, S. Wehrmeyer, K. Borzym, K. Gardiner, D. Nizetic, F. Francis, H. Lehrach, R. Reinhardt, and M.-L. Yaspo, "The DNA sequence of human chromosome 21," *Nature*, vol. 405, no. 6784, pp. 311–319, May 2000.

J. Patrick Fitch (Senior Member, IEEE) received the B.S. degrees in physics and engineering science from Loyola College, Baltimore, MD, in 1981, and the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, in 1984.

He is a Division Leader at the University of California, Lawrence Livermore National Laboratory. In the past decade, his division responsibilities have included genomics, bioengineering, and engineering research with over 200 scientific and technical staff members. His research interests include bioinformatics, bioinstrumentation (automation, MEMS, and photonic), and medical devices. Prior to life science applications, he was the principal investigator for a variety of imaging and computing projects applied to astronomy, nondestructive evaluation and national security. He authored *Synthetic Aperture Radar*, (New York: Springer-Verlag, 1988) and is working on *Introduction to Biotechnology for Nonbiologists* with the SPIE Press. He also successfully developed and marketed a medical device business strategy to venture investors.

Dr. Fitch is a Fellow of the American Society for Laser Medicine and Surgery, a Member of the SPIE, and an Editorial Board Member of Biomolecular Engineering, Elsevier. He received an IEEE best paper award in 1988 and national FLC awards for medical devices in both 1998 and 1999.

Bahrad Sokhansanj (Student Member, IEEE) received the B.E. degree in engineering physics from the University of Saskatchewan, Saskatoon, Canada, in 1998 and the M.S. degree in applied science from the University of California, Davis-Livermore, in 2000. He is currently pursuing the Ph.D. degree in applied science.

His thesis area is the mathematical modeling and large-scale computer simulation of biological networks. He holds a Student Employee fellowship at Lawrence Livermore National Laboratory (LLNL). He is performing his research at LLNL as a member of the Biology and Biotechnology Research Program and the Institute for Scientific Computing Research.