Bioinformatics

Travis Doom, Michael Raymer and Dan Krane

For thousands of years, humans have wondered about the mechanisms of inheritance, evolution, disease and the nature of life itself. Fifty years ago, James Watson and Francis Crick suggested that DNA is the genetic material responsible for inheritance. This observation led to a revitalization of the study of biology that has had stunning consequences. In recent years, breakthroughs in the experimental techniques of biology have allowed scientists to generate amounts of data unprecedented in previous generations.

The entire human genome—the complete set of genetic information within each human cell—has now been determined. Understanding these genetic instructions promises to allow scientists to better understand the nature of diseases and their cures, to identify the mechanisms underlying biological process such as growth and aging, to more clearly track our evolution and its relationship with other species, and so forth.

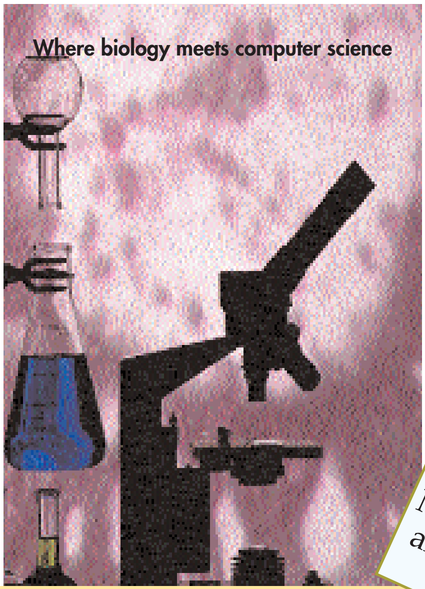
The key obstacle lying between investigators and the knowledge they seek is the sheer volume of data available. Biologists, like most natural scientists, are trained primarily to gather new information. Until recently, biology lacked the tools to analyze massive repositories of information such as the human genome data. Luckily, the discipline of computer science has been developing methods and approaches well suited to help biologists manage and analyze the incredible amounts of data that promise to profoundly improve the human condition.

An engineer's guide to molecular biology

Deoxyribonucleic acid (DNA) is the *genetic material*. In other words, the information stored in DNA allows the organization of inanimate molecules into functioning, living cells and organisms. These "groupings" can regulate their internal chemical composition, growth and reproduction. DNA also allows us to inherit our mother's curly hair, our father's blue eyes and winning smile and even our uncle's too-large nose. The various different units that govern those characteristics, be it chemical composition or nose size, are called *genes*.

Genes themselves contain their information as a specific *sequence of nucleotides* that are found in DNA molecules. Only four different nucleotides (or bases) are used in DNA molecules: adenine, guanine, cytosine, and thymine *(A, G, C* and *T)*. All the information within each gene comes simply from the order in which those nucleotides are found. Complicated genes can be many thousands of nucleotides long. Many genes code for proteins and best estimates are that it takes several tens of thousands of different proteins to make a human being.

©PHOTO DISC



The amount and type of proteins expressed in different individuals plays an important role in determining genetic differences between them, such as eye and skin color, nose size or health. Fundamentally, a protein is a long chain (a polymer) of building blocks called amino acids. Varying combinations of only 20 different *amino acids* are used to build all of the proteins in a human being.

Each of the 20 amino acids has chemical properties that make it distinct. Some amino acids contain atoms (such as oxygen) that have a negative charge; others contain atoms (such as nitrogen) that have a positive charge. Like magnets, amino acids avoid like-charges and are attracted to differently charged atoms. The attraction/repulsion of the amino acids from each other and the watery environment in which proteins exists causes the polymer of amino acids to "fold" up into a specific three-dimensional shape. This shape allows it to perform its mechanical function at a molecular level.

With only four different nucleotides in DNA molecules, encoding a sequence of 20 different types of amino acids is more complicated than a simple one-to-one correspondence. The way the encoding is accomplished is actually computationally elegant. There are 64 different three-character sequences of four characters (43 = 64). Cells use a *triplet code* (they read three nucleotides at a time) to translate the information stored in the DNA into the amino acid sequence of proteins. With only three exceptions, each group of three nucleotides (*a codon*) in the coding portion of a gene corresponds to a specific amino acid.

The three codons that do not correspond to specific amino acids are called *stop codons*; stop codons indicate the end of the coding sequence and cause the synthesis of the protein to complete. This same genetic code seems to have been in place since the earliest history of life on Earth and, with only a few exceptions, is universally used by all living things today.

The sum total of an organism's genetic material is referred to as its *genome*. The human genome contains approximately three billion (3,000,000,000) nucleotides. What's more, virtually every cell that makes up our bodies (of which there are also billions) contains a nearly perfect copy of that genome. If the 23 pieces of DNA that contain those nucleotides, the **chromosomes**, were unraveled and stretched out end-to-end they would be almost one yard long.

Occasionally, errors, or *mutations*, occur when making copies of the long stretches of A's, G's, C's and T's in DNA to be passed on to subsequent generations. Just as tinkering with the insides of an expensive watch, sometime these mutations result in improvements; but, more often, disastrous, significant changes result. These mutations give rise to all the variability we see between individuals. But they can also result in genetic diseases (such as cystic fibrosis, sickle cell anemia, and muscular dystrophy).

Mutations in coding regions often do not allow an organism to pass on its genes (or even survive) as effectively as individuals that do not have them. Thus, regions that are important to an organism's survival are usually the same from one individual to another and

0278-6648/04/$20.00 © 2004 IEEE

sometimes even between one kind of species and another. Such regions are said to be *conserved*.

For most of the genome, changes are accumulated slowly over many millions of years even when they occur in regions that are not conserved. It has been estimated that 98.5% of the DNA in chimpanzees and other closely related primates is identical to that found in humans. Unrelated humans are usually about 99.5% similar to each other at the DNA level. Still, due to the huge size of the human genome, the 0.5% difference between individuals amounts to literally millions of nucleotide differences (*polymorphisms*). Characterizing those differences makes it possible to uniquely identify virtually any diploid organism—including individual humans (barring identical twins).

Determining the complete sequence of an organism's genome is not a trivial task. Nonetheless, the highly publicized *Human Genome Project* reports an almost complete determination of 3.12 billion characters that represent the DNA blueprints needed to make a human being. The Human Genome Project was a bold and ambitious attempt to compile a kind of address list of where all the genes reside along human chromosomes. Having such a map is the foundation for ultimately being able to repair a defective gene or replace a missing one. As a result, we can cure or perhaps prevent numerous genetic-based diseases such as cancer.

But having an address/employee list for a giant office building doesn't tell you what job an employee does, how or if the employee works alone or with others to accomplish tasks, or if a worker's presence saves the company or sabotages it. Which of these three billion characters are responsible for the creation of which proteins? What is the role of each encoded protein in maintaining human health and function? Which genetic differences are responsible for changes in observable genetic traits? The data provided by the human genome project promises to answer many important questions, yet interpreting this data is remains a largely open challenge to biologists and computer scientists alike.

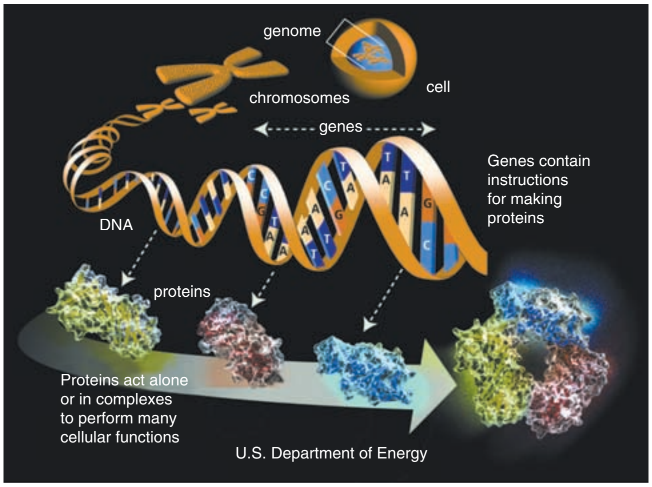

genome
cell
chromosomes
genes
DNA
Genes contain instructions for making proteins
proteins
Proteins act alone or in complexes to perform many cellular functions
U.S. Department of Energy

Fig. 1 DNA encodes a sequence of amino acids that fold to become a protein.

The birth of a new discipline

Clearly, biology has become an increasingly data-driven science. Modern experimental techniques, including automated DNA sequencing, gene expression microarrays (microscope slide-sized chips with different specific nucleotide sequences attached to thousands of spots), and X-ray crystallography (a means of determining the relative three-dimensional position of all the individual atoms in a given protein molecule) are producing molecular data at a rate that has made traditional data analysis methods impractical. Computational methods are becoming an increasingly important aspect of the evaluation and analysis of experimental data in molecular biology. Computational modeling and prediction methods, such as comparative modeling of protein structure, are now reaching a level of sophistication that allows some experimentation to take place entirely within a computational framework. The use of computational methods towards solving problems in biology is known as *bioinformatics*.

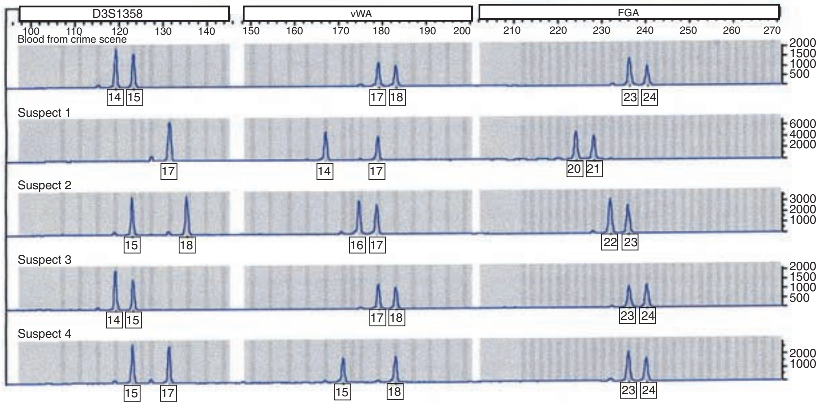


Fig. 2 The results of the STR analysis of five samples: blood from a crime scene and reference samples of four suspects. This analysis includes three STR regions, labeled "D3S1358," "vWA," and "FGA." Each person has two alleles (peaks) at each locus, one from each parent. The position of the "peaks" on each graph (known as an electropherogram) indicates the length (and hence the number of core sequence repeats) of each STR. As can be seen, the profile of suspect 3 corresponds to that of the crime scene sample, indicating he is a possible source. Suspects 1, 2 and 4 are eliminated as possible sources.

Bioinformatics provides a path to understanding genomes as well as the complete set of proteins they encode. As a discipline, bioinformatics is currently at the point that computer science was at in the 1960s. The problems are known, but the solutions to many of even the most basic problems are still under development and constant refinement. The Djikstras, Turings, Knuths and Bill Gates' of bioinformatics have yet to be universally recognized. In short, bioinformatics is an ideal discipline for talented visionaries who want to make their mark in a burgeoning new field.

Bioinformatic problems generally fall into one of just a handful of classes. *Proteomics* deals with understanding the process of how proteins fold into particular structures and towards understanding the functions and interactions of these molecular machines. *Genomics* emphasizes the chemical and physical properties of the flow of genetic information from DNA to proteins. This class includes a broad array of problems, ranging from evolutionary issues rooted in gene replication and recombination to health issues such as gene repair and the diagnosis and treatment of genetic diseases.

Although bioinformatics problems are fundamentally biological in nature, the principle means of attacking these problems is fundamentally computational. Computational techniques such as database theory, data mining, pattern recognition, neural networks, and computer modeling are the basic tools used to solve contemporary bioinformatic problems.

Rational drug design

An excellent example of how bioinformatics is changing things is its effect on the process of drug design. Traditionally, drugs to treat various ailments have been discovered primarily by identifying some physiological process involved in the ailment. Then a large library of proprietary "lead compounds" are individually tested to determine if each aids or hinders the process in question.

Problems to be solved in bioinformatics

- What defines a gene in the genome?
- How do you get from a gene (a one-dimensional representation of nucleotides in DNA) to the correct corresponding folded protein (a three-dimensional structure)?
- What is the relation between structure (3D shape) and biological function?
- What are the relationships between biological molecules (networks and pathways)?
- How do all these biological machines work together to create life?

The amazing thing is how few "basic drug compounds" 2,000-plus years of Western medicine has been able to identify. If you have a headache, take an aspirin. If your blood clots too quickly, no problem, take an aspirin. Want to reduce your risk of colon cancer or a stroke? We've got a drug for that (aspirin). Rheumatoid arthritis? You're in luck: take an aspirin. In fact, there are only around 400 different basic chemical structures in the entire *pharmicopia* of western medicine. Even though, we can be reasonably confident that there is just one compound that would be most effective in remedying defects in each of our tens of thousands of proteins while not interfering with the normal function of others. (Some people with headaches would rather not have their blood clotting times simultaneously reduced.)

Bioinformatics provides an opportunity for scientists to identify and even to customize molecules that might aid or hinder the activities of specific proteins involved in an aliment. Computer models can estimate the effectiveness of each potential drug lead in a fraction of the time (1-2 years instead of the traditional 5-7 years) and at a fraction of the cost of traditional methods (not only saving billions of dollars in development and clinical testing, but also increasing revenues by effectively increasing the number of years that the drug is protected by patents). More importantly, computer models of the dysfunctional system allow for the possibility of creating completely synthetic molecules previously unknown to medicine. The first of these *rationally designed drugs* are just now appearing, such as Relenza (developed to treat influenza), Ritonivir (developed to treat HIV by interfering with HIV protease), TNX-901 (developed to treat peanut allergies by suppressing immunoglobulin-E), and Viagra (originally developed to treat angia, but having noticeably commercializable side effects!).

Forensic DNA testing

Another exciting application comes from the analysis of forensic DNA testing results in criminal investigations. The late 1990s saw the advent of *short tandem repeat* (STR) DNA testing. STR tests are phenomenally sensitive. (Results can actually be obtained from as little material as a single cell and fingerprints typically contain hundreds of shed skin cells.) At the same time, they can provide profoundly compelling evidence. Chances of coincidental matches between unrelated individuals are usually described in terms of one in quadrillions.

STRs are particularly useful for human identification because they tend to be associated with *length polymorphisms*—regions that differ from one person to another due to how many nucleotides are strung together in that part of the genome. At each STR region that is examined, people have two alleles (one from each parent) that vary in length depending on the number of repetitions of a four-nucleotide core sequence. A person with *genotype* 14, 15 for an STR has one allele with 14 repeating units, and another with 15 repeating units.

In 1997, the Federal Bureau of Investigation identified 13 STR regions it deemed appropriate for forensic testing purposes. Commercial firms quickly developed test kits and automated equipment for typing these STR regions. Computer-controlled electronic sensors detect the DNA fragments that are produced by these kits during the testing process. The resulting electronic data is interpreted with software that identifies alleles and displays the results (as shown in Fig. 2). When no suspects have been identified, the DNA profile information from an evidence sample can be used to query databases that contain DNA profile information for millions of convicted felons within the United States to determine if one of them left material at the scene of a crime. Test results often allow a clear-cut determination of whether a particular individual could be the source of an evidentiary sample. However, experts have differed over interpretation of results in some cases, particularly those involving mixed samples (DNA from more than one person) and low quantities of DNA.

Emerging issues in forensic DNA testing appear to be questions regarding the appropriate statistics to apply in increasingly common cases where suspects are initially identified by DNA testing results ("cold hits" or "database trawls"). Also, there is concern over the lack of independence of most DNA testing labs from law enforcement/prosecutorial agencies. Bioinformatics can help with both kinds of problems through complex simulations and the objectivity that comes from automated reviews of testing results.

Career opportunities

There is a high demand for professionals with a background in bioinformatics. Currently, there are more jobs in bioinformatics than qualified people to fill them. The sequencing and analysis of the human genome is currently being studied on a worldwide scale. Pharmaceutical companies are coming to rely more and more upon computational systems to integrate, store and analyze data from a wide array of experimental sources. Computer scientists are needed to analyze, index, represent, model, display, process, mine and search large biological databases. Computational scientists with a strong knowledge of biology, or biologists with the proper computational background, are currently a rarity rather than the norm. Industry analysts forecast that the market for bioinformatics professionals will make bioinformatics of the ten hottest careers of the 21st century.

As a discipline, bioinformatics is still in the process of defining itself. Biotechnology is one of the fastest growing sectors of the economy. Bioinformatics may very well be the "computer science" of the next generation. Undergraduate students who have interests in both computer science and biology should consider a course of study that will prepare them for opportunities in this upstart field.

Educational opportunities

Largely due to the inherently interdisciplinary nature of bioinformatics research, academia has been slow to respond to demands. Currently, most bioinformatics training is done in graduate-school programs.

Undergraduate students at universities that do not yet have a baccalaureate degree program in bioinformatics should consider developing a personalized program of study. Such programs of study must incorporate a specific biology sequence with a more focused computer science foundation. The entrance requirement for job openings and graduate school admittance in bioinformatics look for graduating students having significant interdisciplinary familiarity with both biology and computer science.

Fundamentally, bioinformatics professionals must be capable of communicating effectively in the languages of both computer science and biology. Both disciplines are rich in technical terminology. The defining trait of a successful bioinformatician is not necessarily complete mastery of both fields, but rather a traditional mastery of one field and a comfortable familiarity with the other.

With this in mind, any undergraduate student can prepare themselves for advanced studies or a career in bioinformatics by majoring in biology or in computer science while earning a minor/cognate in the other discipline. In addition, highly motivated students should take advantage of summer bioinformatics training opportunities such as those provided by internships or Research Experiences for Undergraduates (REUs) sponsored by the National Science Foundation (http://www.nsf.gov).

Conclusion

The field of bioinformatics is constantly redefining itself as methods for collecting biological data are developed and refined. While the future directions of the field are impossible to predict, one conclusion seems to be evident: computational techniques have changed the way in which biologists collect and analyze experimental data. Computation will continue to be a prominent component of biochemistry and molecular biology research for the foreseeable future.

While early studies developed the techniques necessary to sequence entire genomes, scientists are now investigating the interacting mechanisms that control the expression of genes. Ambitious new efforts are underway to identify the complex biological pathways of interaction between genes, the proteins for which they code, and the various metabolic intermediates acted upon by these proteins. Advances in understanding these sorts of large scale biological problems bear enormous promise for improving the human condition.

Read more about it

- D. Krane and M. Raymer, *Fundamental concepts of bioinformatics*. San Francisco: Benjamin Cummings, '02.
- A. Cambell and L. Heyer, *Discovering genomics, proteomics, and bioinformatics*. Addison Wesley, 2003.
- J. Setubal and J. Medinus, *Introduction to computational molecular biology*. Brooks Cole, 1997.
- T. Doom, M. Raymer, D. Krane, and O. Garcia, "Crossing the interdisciplinary barrier: A baccalaureate computer science option in bioinformatics." *IEEE Transactions on Education*, Volume 46, No. 3, pp 387-393, August 2003T.
- T. Doom, M. Raymer, D. Krane, and O. Garcia, "A proposed undergraduate bioinformatics curriculum for computer scientists." *Proceedings of the 2002 ACM Special Interest Group on Computer Science Education (SIGCSE 2002)*, Covington (KY), February 2002.
- B. Schachter, "Bioinformatics moves to the head of the class," *Bio-IT World*, pp. 62 – 67, Jun. 2002.
- C. Henry, "The hottest job in town," *Chemical and Engineering News*, vol. 79, no. 1, pp. 47 – 55, January 2001.
- S. Moore, "Understanding the human genome," *IEEE Spectrum*, vol. 37, no. 11, pp. 33 – 35, November 2000.
- National Center for Biotechnology Information (NCBI). URL: http://www.ncbi.nlm.nih.gov.
- Forensic Bioinformatics, Inc. URL: http://www.bioforensics.com

About the authors

Travis Doom (Senior Member IEEE, '03) joined Wright State University in 1998 as an assistant professor in the Department of Computer Science and Engineering and is a member of the graduate faculty in Biomedical Sciences. Dr. Doom earned his Ph.D. (1998) and M.S. (1994) in computer science and engineering from Michigan State University and holds B.S. degrees in computational mathematics (1992) and computer science (1992) from Bowling Green State University.

Michael Raymer (Member IEEE, '03) received the B.S degree in Computer Science from Colorado State University in 1991, and the M.S. and Ph.D. degrees from Michigan State University, East Lansing, in 1995 and 2000, respectively. He is currently an Assistant Professor in the Department of Computer Science and Engineering at Wright State University (WSU) in Dayton, Ohio and an associate member of WSU's Biomedical Sciences Ph.D. program.

Dr. Dan Krane received the B.S degree with a dual major in Biology and Chemistry from John Carroll University in 1985 and the Ph.D. degree in Molecular Biology from Penn State University in 1990. He pursed post-doctoral research at Washington University and Harvard before accepting a faculty appointment at Wright State University in 1993. Since 1991 he has also testified as an expert witness in approximately 50 criminal trials in which DNA evidence has been presented.

Drs. Krane, Doom, and Raymer are also co-founders and partners of Forensic Bioinformatics, Inc. (http://www.-bioforensics.-com). Drs. Krane and Raymer have authored the first biologist/computer scientist co-authored textbook that is specifically designed to make bioinformatics accessible to undergraduates and prepare them for more advanced work. Drs. Doom, Raymer, and Krane have addressed issues on the incorporation of genomics education into the computer science curriculum at national conferences. Drs. Raymer and Doom are the co-directors of the bioinformatics research group in the Computer Science Department at Wright State University.

Undergraduate courses that help prep for a career in bioinformatics

- Computer science (6-9 courses, including introductory programming and data structures; artificial intelligence, databases, formal language theory, and operating systems are recommended)
- Biology (4-6 courses, including a introductory biology sequence, molecular biology, and genetics)
- Chemistry (3-6 courses, including inorganic and organic chemistry)
- Mathematics (3-4 courses, including discrete mathematics and statistics; introductory calculus and graph theory are recommended)