Undergraduate Colloquium Series

Biocomputational Immersive Learning

Drs. Ann Blakey, Jay Bagga and Munni Begum

Introduction: Biocomputational Methods

Biocomputational methods can broadly be viewed as information management systems at the interface of mathematics, statistics, computer science, and molecular biology/genetics. The main objective is to use the tools and methods from the disciplines of mathematics/statistics and computer science to turn vast, diverse, and complex life sciences data into useful knowledge. Biocomputational methods branch out to two major areas: bioinformatics and computational biology. These two terms are often used interchangeably but there is a fine line between them. Bioinformatics more closely addresses advancement of algorithmic and computational techniques based on computer science, mathematics and statistics to manage and analyze biological data efficiently. On the other hand, computational biology addresses hypothesis-driven investigation to a biological problem using computational, mathematical and statistical methods.

Biocomputational Immersive Learning (BIL)

We initiated the interdisciplinary Biocomputational Immersive Learning (BIL) program funded by a Ball State University Enhanced Provost Initiative (EPI) grant in Summer 2007. Biotechnology is a rapidly growing and economically important sector for Indiana. In May 2006, the state of Indiana was identified as *one of the nation's top four life science leaders* [3]. Our ultimate goal is to create an immersive experience in the biocomputational field through the BIL program that leverages Ball State's expertise in these areas. Inaugurating the program with a course on biocomputational methods is the first stepping stone towards fulfilling this goal.

Faculty Expertise

BIL is truly an interdisciplinary team approach to Biocomputational methodologies, represented by three actively engaged BSU faculty, where each brings a unique and integrative view of his/her field to the training of students. All three were selected as participants in the National Science Foundation funded "Workshop on Teaching Undergraduate Bioinformatics through Collaboration and Inquiry", which was held at the University of Akron in June 2006. The workshop focused on the development of undergraduate training in bioinformatics/biocomputational methodologies. Participation in the workshop afforded the investigators a valuable opportunity to experience first-hand the cutting-edge tools and technologies used for bioinformatics education and collaborative learning.

Each faculty member contributes to the program from his/her knowledge and expertise based on years of scholarship in overlapping and associated fields of research. Dr. Begum specializes in biostatistics and statistical computing methods. She has worked collaboratively as a statistical consultant in the biomedical research field both in academia and industry. Dr. Bagga has background in mathematics and computer science. His main areas of research are in graph theory, combinatorial and graph algorithms and their applications. He has also been involved in research in knowledge management, verification of software, and computational geometry. Dr. Blakey specializes in plant genomics, particularly comparative genomics. Her main areas of genetic research are in comparative plant genomics of the cereal grasses, including: a) floral genome analysis - gene expression products, microscopy of floral structures, and comparative DNA sequence analysis; and b) bioinformatics database structures and ontologies for use with computer system data interfaces.

Interfacing Molecular Biology/Genetics, Computer Science, and Mathematics/Statistics

The BIL initiative is interdisciplinary in the sense that it interfaces methods and principles primarily from three specific disciplines, molecular biology/genetics, computer science, and mathematics/statistics. The respective role of each of these three fields in biocomputation is presented briefly in the following subsections.

Molecular Biology/Genetics

The central dogma of molecular biology/genetics (Figure 1), when viewed from the perspective of information flow and input/output, is readily accessible to the computational sciences. Whether researchers are considering the number of DNA sequences associated with a particular function within a cell or those that result in a specific phenotype (appearance), computational analysis is required for the vast array of data and data types that could be correlated to the research question. Since its beginning in 1906, genetics has been a statistics-based science whether one considers: 1) the probability calculation for basic Mendelian inheritance of traits such as the gene for a specific disorder, 2) quantitative and epistasis using chi-square tests or regression analyses for obesity or behavioral traits, or 3) the computer-based high-throughput data production and analysis of full genome sequencing.

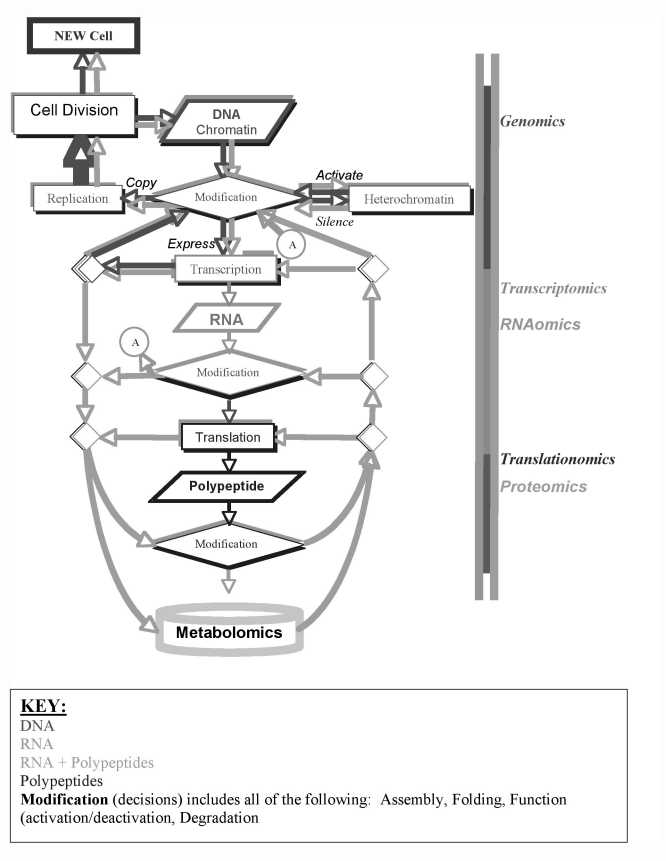

Figure 1: Central dogma as information flow in molecular biology/genetics

Molecular biology and genetics are faced with a deluge of data as entire genomes of more organisms are sequenced and released to databases. How does the geneticist analyze, interpret, or even store these data? Mathematics and statistics are of pivotal importance to the analysis of genetic data. The use of computer science designs built on the information flow analogy of biological systems has led to innovations in computational data analysis when coupled with the mathematical sciences and applied to understanding biological systems. Technology has now given us the capabilities to generate high-definition images of growth and development at sub-cellular levels, graphical algorithms provide a means for taking those images and studying the progression of each stage of growth. High-throughput genomics produces millions of sequences per year across the USA. Biological databases are overflowing with information. Mathematical/statistical tools and computer algorithms allow us to access, analyze, and interpret the information in meaningful ways. Thus we are able to cross the traditional science disciplinary barriers to demonstrate the importance of collaborative team learning and research. This also leads to insights into new applications of statistics and computer science.

Computer Science

An *algorithm* can be thought of as a sequence of steps (or instructions) that solve a well-formulated problem. Computer science and computer engineering are primarily concerned with creating computing machines and efficient design and analysis of algorithms that can be carried out on those machines. In its most basic form, a computer can be considered as a machine which takes as input a string of characters from some predefined alphabet. The string encodes the sequence of steps of an algorithm, as well as an instance of the problem the algorithm is trying to solve. The computer processes the problem instance according to the encoded instructions, and then outputs a result in the form of another string. This notion of computer was formalized by Alan Turing in 1936 who first defined a formal model which is now known as a Turing Machine (TM) [2]. Turing's model proposed the input string to be stored on a tape which is divided into cells such that each cell contains a single character. A read-write head can move across the tape, scanning the cells, and based on the instructions, can change or write certain cells. Figure 2 [8] shows a graphical depiction of a TM.

As described above, the process of DNA replication can be considered as a biological algorithm, analogous to those considered in computer science. Replication employs various "machines" to perform the tasks of establishing "start" sites of the programmed processes (DNA helicase), complementation checking and assembly (DNA polymerase), and linking of large components of processed information (DNA ligase). While this may not have been quite the way Watson and Crick [6, 7] visualized the basic mechanism in 1953, the analogy to modern computer hardware and software design is uncanny. In [6], Watson and Crick said: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material...".

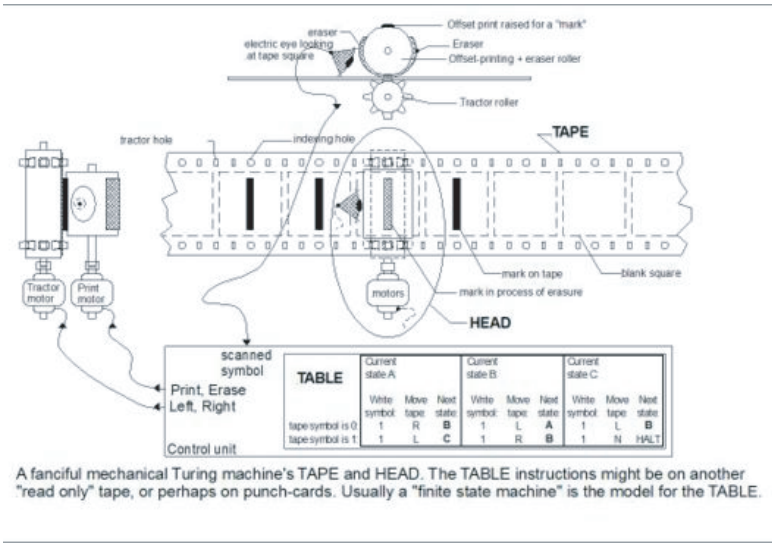

Figure 2: A Turing Machine (TM)

A copying algorithm can be easily programmed in a TM. Thus given a string `ACGGTATCCGAATGC`... with `A`: adenine, `T`: thymine, `G`: guanine and `C`: cytosine, our algorithm copies the given string to `TGCCATAGGCTTACG`..., with the transformations `A` $\leftrightarrow$ `T` and `C` $\leftrightarrow$ `G` (base-pairings) to get the complementary strings.

We thus see that information flow as seen in the central dogma of molecular biology has parallels in computer science. Computer science approaches to biocomputation include studies and applications of algorithms. Numerous algorithm design techniques in computer science find applications in biocomputation and bioinformatics. Such algorithms include exhaustive search, greedy algorithms, dynamic programming, divide-and-conquer, graph algorithms, combinatorial pattern matching, and clustering.

For an illustration, let us consider the DNA sequencing problem. The goal is to assemble individual short fragments into a single genomic sequence (superstring). Therefore, one way to approach the problem would be to treat it as a Shortest Superstring Problem (SSP): given a set of strings, find a shortest string that contains all of them. Another example is to consider biological networks. This could be approached similar to the classic Traveling Salesperson Problem: given a graph with edge costs, find a salesman tour of least cost. A salesman tour is a cycle in the graph that contains all the nodes of the graph. Thus, as we learn more about the fine molecular interactions of biological systems we find analogous computational structures and innovative approaches to developing new ideas in both fields.

Mathematics/Statistics

Genetic information can broadly be taken as a very long string of four letters `A`, `T`, `G`, and `C`, representing the four nitrogenous bases (Adenine, Thymine, Guanine and Cytosine, respectively) of a deoxyribonucleic acid (DNA) molecule as discussed above. This information occurs in pairs of complementary bases commonly known as base pairs. As an example of the vastness of this information, the human genome is approximately 3×10^9 base-pairs long. In practice, many overlapping small pieces, a contiguous sequences of genetic code termed *contig*, are sequenced and these fragments are assembled into a longer contig. Thus DNA sequencing can be compared to the problem of finding a shortest super string mentioned earlier. Addressing even the simplest problem of determining the actual sequence of DNA requires consideration of quantitative findings such as: proportion of genome covered by contigs, mean number of contigs, and mean contig size. One can efficiently handle these types of problems while analyzing a single DNA sequence with certain assumptions on the data generating mechanism. It is when the need arises for the comparison of multiple DNA or protein sequences that many of the problems of the biocomputational field result. All of these problems call for a deeper understanding of the underlying mathematical and statistical theories upon which the computing algorithms are based.

Technically known as alignments, these tools are based on statistical tests of hypotheses developed specially for comparing sequences of DNA nucleotides or amino acids. Exactly-Matching and Well-Matching subsequences [1] are two commonly used sequence alignment techniques. The Well-Matching subsequences technique is implemented in the widely used alignment tool known as BLAST for database searching. Alignment algorithms for two or more DNA sequences are based on a simple scoring scheme, but a more sophisticated statistical method requiring stochastic processes are used for aligning multiple protein sequences.

Novel statistical methods are required to analyze high throughput gene expression data generated by the advanced microarray technology. With the advent of this efficient technology one can study the effects of treatment, diseases or various environmental conditions on thousands of genes simultaneously. This aspect of microarrays has revolutionized the way scientists investigate gene expression. A graphical representation [4] of a typical microarray experiment is depicted in Figure 3.

Microarray technology is based on the use of fluorescently labeled molecules known as probes or hybridizing probes, to identify complementary base pair sequences. Experimental probes are prepared by extracting mRNAs, products of gene expression, from two samples of interest (e.g., diseased and normal cell) where the samples are labeled with "fluorescent tags" (usually red and green chemical compounds). These are mixed and incubated on a microarray which is a fixed-surface (slide) that contains thousands of immobilized genes or sequences of interest. If the genes of interest are expressed in a given cell, the labeled probes will bind to the arrayed sequences and the levels of expression based on fluorescence can be measured and analyzed.

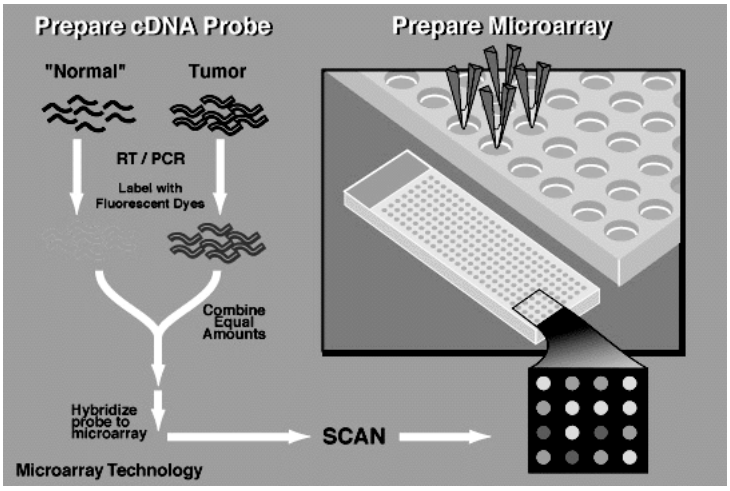

Figure 3: Microarray technology for measuring gene expression level

These data are to be adjusted for background noises that come from the use of different technology, and may not be due to the inherent biological variation. Data preprocessing methods such as "normalization" can be used for this purpose. In addition to the quality of microarray data, interpreting and analyzing these vast data meaningfully are daunting tasks for the scientists alone. Statisticians/computer scientists trained in the biocomputational field will be able to bridge the gap. Since microarrays measure expression levels for thousands of genes simultaneously, traditional statistical methods alone are not amenable to handle such data. Computation-based statistical methods, such as classification, clustering, bayesian modeling, networks and integrated models based on graph or other computing algorithms would be more appropriate in analyzing and interpreting such high throughput data.

Finally, mathematical modeling of the evolutionary process is of importance, particularly when addressing questions and hypotheses regarding genetic variations at the species level as well as at the individual level. Appropriate statistical methods are required for the testing of predictions based on evolutionary models and conducting tests of hypotheses.

The BIL Immersive Program Experience

We began the BIL program initiative in response to the demand for highly trained biocomputational personnel by industries and government agencies. The overall goals of this initiative are:

1. To provide a creative and active learning environment in which our students can achieve the competencies necessary for careers in biocomputational field.

2. To strengthen and expand student learning opportunities through unique research and problem-solving training.

This comprehensive BIL program includes a biocomputational methods course, opportunities for guided independent/collaborative study and internships.

Biocomputational Methods Course

We have developed an immersive credit-bearing course as an outcome of the BIL initiative. This course is cross-listed in the departments of the three participating faculty and it will be offered for the first time in Fall 2008. The BIL course is available to both advanced undergraduate and graduate students. Course topics include, but are not limited to, current, state-of-the-art computational methods and strategies for analyzing biological data (DNA, RNA, and protein data). Biocomputational and bioinformatics tools and approaches will be covered in detail, such as sequence alignment, database searches, phylogenetic prediction, gene prediction and classification, gene expression analysis and microarray technology. The course is targeted at students interested in mathematical sciences, statistics, computer science, and biology, as well as biochemistry, biophysics, physiology, physical anthropology, psychology, and students from the Human Performance Laboratory.

Independent/Collaborative Study

The field of biocomputation is growing at a very rapid speed and it encompasses almost all life science areas. As such novel scientific as well as computational problems are arising at a similar pace. Our advanced students will be devoted to studying current topics independently and as collaborative teams under the supervision of BIL faculty advisors. The biocomputational problem-solving activities and projects will involve interdisciplinary teams of students working on a solution. A hands-on approach will guide the learning process. Students will be required to study and implement various algorithms. They will also be involved in field trips to companies that are involved in biocomputation/bioinformatics. Each student will complete an interdisciplinary team project. The teamwork approach will be strongly encouraged as an essential aspect of the program.

Internships

The BIL initiative will strive to include industry and government partnerships in the development of student-driven projects. Involvement in the active learning environment of the BIL initiative will aid students in determining a career path which best suits their talents and provides connections through internships and fellowships with industry partnerships and government agencies.

There are a number of bioinformatics-related companies and institutions in our region. These include Eli Lilly, DowAgro Sciences, Roche, Strand Laboratories, and the IU Medical Center. Partnerships with these companies will

provide collaborative research projects involving teams of faculty, students, and researchers/experts from these institutions as well as internship and employment opportunities for our students.

Current Research and Future Directions

There is a broad range of current research in the biocomputational field, from the study of Genomics, Transcriptomics, RNAomics, to Proteomics and Metabolomics. Genomics basically studies the networks of genes, transcriptomics studies networks of transcriptional factors, RNAomics deals with networks of RNAs and Proteomics and Metabolomics study polypeptides/proteins and their associated processes. The BIL team is actively engaged in collaborative research activities by bringing their own areas of respective expertise into the biocomputational field. The future directions of the BIL program rely on maintaining an active research team in the fields of computational biology and bioinformatics in order to train students in this highly competitive and technologically demanding field.

References

[1] W. Ewens and G. Grant, Statistical methods in bioinformatics: An introduction, Springer (2001).

[2] R. Herken (Editor), The universal Turing machine: A half-century survey, Oxford University Press (1992).

[3] Life Science Report, Indiana Economic Development Corporation (2007).

[4] Talking glossary of genetics: Microarray technology, National Human Genome Research Institute ⟨`http://www.genome.gov/Pages/Hyperion/DIR/VIP/Glossary/`⟩ (5 July 2008).

[5] Neil C. Jones and Pavel A. Pevzner. *An Introduction to Bioinformatics Algorithms.* The MIT Press, Cambridge, Massachusetts, USA, 2004.

[6] J. Watson and F. Crick, *A structure for deoxyribose nucleic acid*, Nature **171** (1953) 737–738.

[7] J. Watson and F. Crick, *Genetical implications of the structure of deoxyribonucleic acid*, Nature **171** (1953) 964–967.

[8] ⟨`http://en.wikipedia.org`⟩