Bioinformatics Research Paper
April 3, 2003

CS 692 Winter 2003
Dr. Jamal Alsabbagh

Jerry Evans
Ahmed Khan
Slava Pavlov
Phil Mayrose

ABSTRACT 4
INTRODUCTION 4
(I) BIOINFORMATICS OVERVIEW 5
(II) BRIEF HISTORY OF BIOINFORMATICS 5
(III) BIOLOGY 101 6
(IV) RESEARCH TECHNIQUES 7
Using public databases and data formats (not strictly speaking a method) 8
Sequence alignment and sequence searching 8
Multiple sequence alignment 10
Phylogenetic analysis 10
Extraction of patterns and profiles from sequence data 13
Protein sequence analysis 14
Protein structure prediction 14
Protein structure property analysis 15
Protein Structure Alignment and Comparison 16
Biochemical Simulation 17
Working with 3D Protein Structures 18
Whole Genome Analysis 18
Primer Design 20
DNA Microarray Analysis 20
Proteomics analysis 21
(V) BIOINFORMATICS-SPECIFIC LANGUAGES 21
The BioPerl Project 22
The BioJava Project 22
The BioPython Project 22
The BioCORBA Project 23
The BioDAS Project 23
Ruby and the BioRuby Project 23
The C Language and EMBOSS 23
(VI) SOME CURRENT PROJECTS 23
Pseudoknots in RNA Secondary Structure - Oxford University 24
Ensembl Genome Browser 24
Project VirGO 24
(VII) CURRICULUM 25
Bioinformatics Courses 26
Bioinformatics Courses 26
Biology/Biochemistry Courses 27
Mathematics/Statistics Courses 27
Computer Science 27
Law/Ethics 28

Seminar/Discussion/Workshop 28
(VIII) ANALOGIES 28
DNA, Protein and everything in-between, an analogy 28
Molecular shape, how is it determined? 30
The IT Problems and Solution: 33
(IX) CONCLUSION 35
(X) GLOSSARY 37
(XI) BIBLIOGRAPHY 38
(a) Publications 38
(b) Online Resources 39
(XII) APPENDICES 43
Appendix A Institutions 43
Appendix B Bioinformatics Courses 44
Appendix C – Amino Acids' 1- and 3-character codes 51
Appendix D – Biological Databases 51
Appendix E – Physics 52
Appendix F – Genetic Code Table 52
Appendix G – RNA Image 54

ABSTRACT

Bioinformatics is a comparatively new field involving a variety of disciplines including Biology, Computer Science, Genetics, Biochemistry, Physics, Statistics and other advanced mathematical topics. The complexity and novelty of the field make it a less common curriculum than other more established fields. Unlike those of more mature fields, Bioinformatics' Information Technology (IT) tools and techniques are still evolving – to the point where some institutions train their researchers to build tools themselves.

Bioinformatics shares many analysis techniques with Molecular and Cell Biology and has been called an evolution of Computational Biology. This paper will discuss some of the research techniques used in Bioinformatics, some of the languages used in analysis and tool building, some of the curricula for institutions that offer Bioinformatics degrees or training and some current projects underway in the field. We will then use analogies to help readers from Computer Science and Information Systems backgrounds learn about the field of Bioinformatics. Finally we will offer our suggestions for the future and conclusions.

INTRODUCTION

Why is Bioinformatics a field and not "Astro-informatics" or "Finance-informatics"? A primary reason is that unlike Finance and Astronomy, the field is relatively young and its tools immature. These fields can, within reason, take ordinary garden-variety Computer Science grads and teach them on the job any necessary industry-specific variations. Bioinformatics has such a life-science heavy problem domain that this is not possible. Our research suggests that the current trend in the field is to teach life-science specialists enough about the other necessary fields to function adequately within them within the bounds of their work. These cross-trained specialists are not as highly trained in IT as computer science or electrical engineering graduates, but the programs we have researched require enough IT course-work to make them more well-trained software developers than the average Biology or Chemistry major.

Bioinformatics researchers and practitioners have to be skilled in at least one major health science domain, e.g. Molecular Biology, and *also* conversant in several other domains as diverse as mathematics, computer science, statistics and physics. The primary purpose of this paper is to introduce the field of Bioinformatics to an information technologist and to provide information and references for further investigation.

The first three sections are introductory. They offer an overview and brief history of Bioinformatics and a basic Biology overview. Section IV describes some of the

computational techniques used in Bioinformatics research and development. Section V discusses some programming languages used in the field. Sections IV and V taken together provide a good example of why Bioinformatics is a field in its own right. Section VI reviews a few current Bioinformatics projects. Section VII talks about the curriculums currently offered by various academic institutions. Section VIII offers some analogies to help relate the world of Bioinformatics to the real world. Section IX wraps things up with our conclusions. The remaining sections include the Glossary, the Bibliography and the Appendices, which include many tables referenced in the paper.

(I) BIOINFORMATICS OVERVIEW

For the purpose of this paper Bioinformatics can be defined as a multi-disciplinary field involving aspects of Biology, Computer Science, Genetics, Biochemistry, Physics, Statistics and other advanced mathematical topics.

Bioinformatics is a new discipline that has its roots in Computational Biology. It is a field unto itself for several reasons. First, the data collected are a critical part of the research. This requires that the tools built for Bioinformatics research are suited specifically for it and are usually not suitable for other purposes. The data are heavily string-based and this influences the design of database systems and other tools used to analyze them. Second, Bioinformatics research is being performed at many schools and institutions in the public and private sectors. We believe that the trend is moving toward public databases and because of this, research will be optimized if a set of standardized tools is available. Third, Bioinformatics researchers will need to be educated in Biology and other fields including IT to carry out their research. It is possible that as the tool set evolves researchers will require less IT background but at the present time, most institutions we researched required IT coursework. Fourth, Bioinformatics research examines many problems that if solved, can represent very valuable medical advances including customized medication, cancer research, forensic medicine and cures or new treatments for chronic diseases.

(II) BRIEF HISTORY OF BIOINFORMATICS

The history of Bioinformatics begins with the history of biology, especially of cellular and molecular biology. The 18th century scientist Mendel is considered to be the father of genetics. He conducted cross-breeding experiments on peas in his garden and postulated a set of rules that govern inheritance. In 1869, Meischer isolated the first DNA. In the 1950s Alfred Sangler, the laureate of two Nobel Prizes, discovered the sequencing of insulin. He is considered to be the founder of genome sequencing. Also in the early 1950s, X-ray crystallography allowed researchers to determine the 3D structure of DNA (James Watson and Francis Crick) and proteins (Dr. Kendrew and Dr. Perutz). James

Watson formulated the Central Dogma of Molecular Biology, which determines the connection between DNA, RNA, and proteins.

Although the 1950s were marked by major achievements in molecular biology, the birth of Bioinformatics as a separate discipline would not be possible without the information technology revolution. The exponential growth of recorded biological data required development of specialized databases. In 1973 the Brookhaven Protein Data Bank was announced. In 1982 two major biological data banks, GenBank and EMBL database, were established. Development of the public biological databases and proliferation of the Internet made it possible to conduct scientific research not only in 'wet labs' but also through the online access to biological data. This change in the research environment led to the birth of Bioinformatics, which is concerned with acquisition, storage, retrieval, analysis, modeling, and distribution of the biological data.

(III) BIOLOGY 101

Bioinformatics deals with research into DNA, RNA, Amino Acids, Genes, Genomes, Proteins and Proteomes. This section gives the reader a basic understanding of these terms.

DNA (deoxyribonucleic acid) molecules, the building blocks of genes, contain two strands of four nucleotides (bases), and because of their double helix shape they resemble a spiral staircase. There are four DNA nucleotides named Adenine, Cytosine, Thymine and Guanine, commonly labeled A, C, T and G respectively.

RNA molecules consist of a single strand that is made of the four nucleotides (ribonucleotides) Adenine, Uracil, Cytosine, and Guanine. These form complementary pairs with the DNA nucleotides Thymine, Adenine, Guanine and Cytosine respectively.

Amino acids are made up of three nucleotides or bases. These do not include the amino acids used to make up DNA or RNA. See Appendix C for a list of the one-character and three-character codes used to encode the 20 types of amino acids.

A gene is complex structure made up of DNA. Many of the research techniques described later in the paper deal with analysis of genes and issues like how to identify the borders of a gene.

A Genome is a complete genetic map of an organism. One of the most famous recent Life Sciences projects is the Human Genome Project in which numerous researchers continue to map the human genome.

Proteins are made of amino acids. They typically contain between 100 and 500 amino acids of 20 different types. The composition and order of amino acids is the same for every type of protein e.g. insulin. The function of a protein is determined by not only its composition and sequencing of amino acids, but also its 3D shape, dictated by the protein's linear sequence.

A Proteome is "the entire protein profile of a cell" [Clark, 2001]. This is a biological description meaning all the proteins present in a given cell.

The major discovery of Molecular Biology was the interconnection between DNA, RNA and Proteins. The so-called "Central Dogma of Molecular Biology" which can be defined as follows: "DNA defines the synthesis of protein by way of an RNA intermediary."[Bergeron, 2003] When a cell replicates its DNA, the existing DNA strand splits. Each half then attracts a complimentary string of RNA. The result is a new DNA strand matching the original. Errors during the copy process are genetic mutations.

Both DNA and RNA are made up of data that can be treated similarly when collected, using two sets of four letters: {A, T, G, C}, {A, U, G, C}. "The process of converting the DNA to RNA is called transcription, which also refers to RNA synthesis."[Drlica, 1997] The RNA used in the synthesis is called messenger RNA.

Each amino acid is uniquely defined using three RNA nucleotides. The four RNA nucleotides, taken three at a time, can form 64 combinations, 3 of which are considered "stop bits". A triplet, or codon, is a set of three RNA nucleotides [Drlica, 1997].

Proteins are made up of collections of amino acids. "The information in (messenger) RNA (mRNA) is next changed into amino acid sequences in protein by a process called translation." [Drlica, 1997]

In summary, we have three sets of data: one letter from a set of four for each unique nucleotide's DNA, one letter from a set of four for each unique nucleotide's RNA, and 20 amino acids each abbreviated with three letters. Transcription is the process of converting DNA to RNA by replacing the T with a U. Translation is the process of using several mRNA to build a protein. The process of converting the information in messenger RNA into a protein is called protein synthesis. [Drlica, 1997]

(IV) RESEARCH TECHNIQUES

Our research identified a number of Bioinformatics research techniques. In this section, we give a brief overview of each and discuss the tools we think are suited to it. Some of these are common to Biological research techniques and may even predate

Bioinformatics as a separate discipline. Many of the methods described below have associated databases. The first section talks about some of the major databases used in Bioinformatics research.

Using public databases and data formats (not strictly speaking a method)

PubMed is a free public database of citations from biomedical periodicals. It is sponsored by the National Center for Biotechnology Information (NCBI) and the National Library of Medicine (NLM). In addition to the citation index, it has links to full-text articles available online. It is also connected to MEDLINE – the most comprehensive medical database that contains over 12 million references to articles from over 4,600 biomedical journals in over 30 languages and dating from 1966.

The largest biological databases are free and available online to the general public. Many of these databases specialize in specific types of biological data. For example, SWISS-PROT stores protein sequences and TIGR – genome sequences. Perhaps the best-known databases are GenBank and Protein Data Bank (PDB). GenBank is supported by NCBI, the DNA DataBank of Japan (DDBJ), and the European Molecular Biology Laboratory (EMBL) and is accessible through the same portal as PubMed. GenBank holds an annotated collection of DNA sequences. According to the GenBank sponsors, "GenBank grows at an exponential rate, with the number of nucleotide bases doubling approximately every 14 months. Currently, GenBank contains more than 28 billion bases from over 250,000 species" (GenBank online note dated February 12, 2003).

The Protein Data Bank is maintained by the Research Collaboratory for Structural Bioinformatics (RCSB) consortium, which consists of three members - Rutgers University, the San Diego Supercomputer Center at the University of California, and the National Institute of Standards and Technology. Among the biological web sites, the PDB is probably the most exciting place to visit because of its colorful collection of the 3D structures of proteins. The "Molecule of the Month" link takes you directly to the collection of selected molecules where you can see their structures and access further details. See Appendix D for a list of major biological databases used in Bioinformatics research.

Sequence alignment and sequence searching

Since DNAs, RNAs and proteins can be presented as sequences of characters, comparing these sequences is a natural technique for finding their similarities. Similar proteins or gene sequences are called homologues. To be homologous, two proteins must have at least 25% of their amino acids identical. Similarity in sequences can indicate their

common ancestry, as well as their functional and structural similarity, and therefore sequence alignment is commonly used to:

- determine common ancestry of organisms
- determine functions to unknown proteins
- identify important sites in proteins
- predict 3D structure of proteins

Based on the number of the compared sequences, alignments are classified as either pairwise sequence alignment when only two sequences are compared, or as multiple sequence alignment, when more than two sequences are compared. Depending on the scope of the alignment, it can be local or global. In the global alignment, the total score of the end-to-end comparison is found, while in the local alignment only the regions with highest local scores are determined. Sequence alignments are also classified as gapped or ungapped. Inserting gaps into sequences can result in stronger similarity of the compared sequences and can emphasize the differences as results of evolution.

To determine the level of similarity in sequences, various methods of scoring have been developed. Typically, all of these methods award scores for character matching, and apply small penalties for character mismatches and large penalties for the gaps. Most of the computational methods that are used for sequence alignment can be equally applied to pairwise, multiple, local and global alignment. Among these methods are:

- Bayesian methods
- dot-matrix analysis
- dynamic programming
- genetic algorithms
- hidden Markov models
- neural networks
- scoring matrices
- word-based techniques

Pairwise sequence alignment is the most common way to compare sequences. Computationally it is much cheaper than multiple sequence alignment. In the dot-matrix method, one sequence is aligned along the top of the matrix, and the other along its side. Dots are used to mark the intersections:

	A	T	C	C	G
A	*				
T		*			
C			*		

	A	T	C	C	G
A					
T					

Even though this method displays all possible matches between the compared sequences, it is efficient only for pairwise alignments and for sequences with a high level of similarity. Among the tools that use this method are DOTLET[1] and Dotter[2].

The most popular tools for sequence alignment, BLAST and FAST, use word-based algorithms. FASTA is the predecessor of BLAST (Basic Local Alignment Sequence Tool) and is not as commonly used nowadays as BLAST. BLAST is optimized for speed and can be successfully used to scan large databases. Online BLAST searches can be done at www.ncbi.nlm.nih.org/BLAST.

Multiple sequence alignment

Multiple sequence alignment is used with homologous sequences that are suspected to have evolutionary, structural, and functional similarities. It is most commonly applied to proteins. Multiple sequence alignments can be used to:

- identify a protein's family and to discover new family members
- reconstruct the history of the aligned proteins
- determine functional regions of proteins (binding sites)
- predict proteins' structure

Most of the methods that are used for pairwise alignment can be extended to use for multiple sequence alignments. Nonetheless, multiple sequence alignment has its own commonly used technique – progressive alignment of pairs of sequences. ClustalW, a very popular tool that is used for multiple sequence alignment, applies this technique. Besides ClustalW, other multiple alignment tools can be found at Baylor College of Medicine http://searchlauncher.bcm.tmc.edu/multi-align/multi-align.html
and at Centre INRA deToulouse http://protein.toulouse.inra.fr/multalin/multalin.html .

Phylogenetic analysis

Phylogenetic analysis deals with the analysis of the natural evolutionary relationship of protein or nucleic acid sequences linking individual organisms, populations or "taxa". The idea is to try to determine how many changes or mutations are necessary to convert one sequence to another. The use of this data will aid in mapping diseases to the location

[1] http://www.isrec.isb-sib.ch/java/dotlet/Dotlet.html
[2] http://www.cgr.ki.se/cgr/gropus/sonnhammer/Dotter.html

of the chromosome, which is associated with the transmission of the diseases. In particular, the study of recent human evolution has been providing more genetic evidence that can be used to trace our evolution. Furthermore, sophisticated analysis of genome variation will lead to understanding disease variations. Tracking the changes in man and the corresponding changes in the genetic evidence will provide a more diverse collection of data to analyze.

A view of the migration patterns of man can evidence the varied opportunity to collect these valuable data. This is a Bioinformatics application to the field of Anthropology. The image below is from Handprint.com [Handprint, 2002] and the subsequent quotation is from Oxford University's Mathematical Genetics and Bioinformatics Department web site [Oxford, 2002].

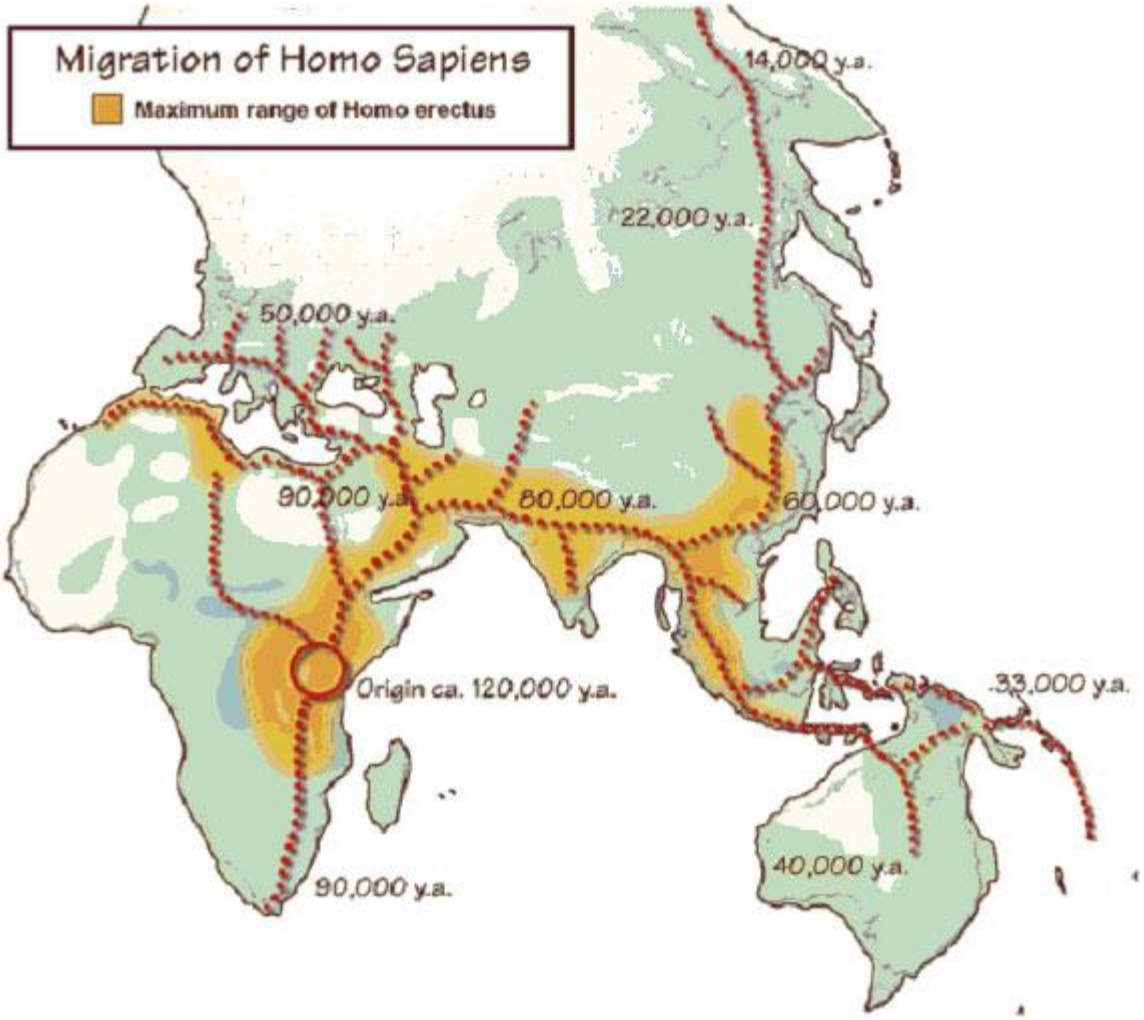

Figure 1 – From http://www.handprint.com/LS/ANC/disp.html

Around 120,000 years ago *Homo sapiens* emerged as a new species, most likely in central East Africa, and from there migrated into the Middle East, south Africa, Europe, central Asia, and finally into the New World. To reach the Bering Strait from Africa by 14,000 years ago, humans would have had to wander no more than one mile every eight years. -- The timing of Ice Age coolings, and the amount they lowered ocean levels, specifies the geologic periods in which it was possible to migrate to land masses otherwise separated by water. [Oxford, 2002]

The migration of homo sapiens from small groups into larger populations brings about changes in gene mapping. Furthermore, the mutations found in homo sapiens over time will contribute to gene mapping. The study of disease recombination and mutation may complicate the gene mapping process making identification of the associated disease vulnerable at each step of the process. Even so, this method of collecting data and finer-scale mapping analysis will "ultimately be more powerful than linkage analysis for isolating genes"[Oxford, 2002].

The computationally efficient inference of using computer aided numerical methods such as Markov Chain Monte Carlo (MCMC) to perform "coalescent methods to model the genealogy of a random sample from the population"[Oxford, 2002] utilizes the work of Griffiths and Tavare. The coalescent methods deal with how the human population unites into larger groups as they become one whole. "Measure-valued processes arise naturally in modeling the composition of evolving populations"[Oxford, 2002]. So, if we have an individual, that individual will belong to a small group. And several small groups together define the general population. Now each small group has its own set of DNA sequences and associated unique ancestral history, and we can have any number of groups making up the general population. In addition, the amount a small group contributes to the population can be counted. This Measured-value diffusion is how each individual belonging to a group is counted as it enters the general population.

For the engineer: It is sort of like using superposition in time invariant differential systems. The measured-value algorithms will take a homogeneous continuous system and form a discrete solution method something like using a discrete z-transform instead of using a continuous La Place transform.

The two simplest models offered were the "Fleming-Viot process" and the "Dawson_Watanabe Process". The Dawson-Watanabe method uses Brownian motions and therefore only obtains a steady state population of infinite or zero, (either exploding or dying out) which may not be what we want. The Fleming-Viot process results in a constant population and was shown to be useful by the Perkins Disintegration Theorem [Oxford, 2002] if the total population is used in the process. 'Peter Donnelly and Tom Kurtz have developed and exploited countable representations for measured-valued population models'[Oxford, 2002].

Extraction of patterns and profiles from sequence data

During the evolutionary process that results in related gene sequences, there are some fundamental structures that remain constant through time. They are conserved and are members of a group or sequence family. These "sequences of amino acids that define a substructure in a protein that can be connected to function or to structural

stability"[Gibas, 2001] are called motifs. They are the fundamental pattern used to identify the evolution of related gene sequences. Motifs remain when the selection pressure or energy shapes modify other less significant structures of the gene sequence through time. They are the beacon that guides the way through the sea of 'mutational noise'. Motifs, because they remain, are the fundamental definition of the protein as it evolves. "Sequence profiles are statistical descriptions of these motif signals…" [Gibas, 2001]

Multiple sequence alignment is one of the most successful tools used in discovering new related sequences.

Protein sequence analysis

Analysis of protein in 3D structures has been going on since the early 1970's and since that time many software packages such as the BMERC PSA Protein Structure Prediction Server at Boston University and HMMER[3] 2.2 from Washington University in St. Louis have been developed to help render proteins in 3D. The ability to view a 3D rendering of a protein allows the researcher to make educated guesses where to focus his study. Knowing the protein's detailed molecular shape allows the researcher to locate catalytic sites and interaction sites, where protein mutation is most likely to occur. The protein's functional chemistry from 3D molecular modeling enables the researcher to identify possible problems, three examples include:

- 'Molecular modeling of an allergy-causing protein'[Gibas, 2001]
- "Characterization of the mutagenic active site in DNA reverse transcriptase from the HIV virus; this site is thought to be responsible for the ability of the HIV virus to mutate rapidly" [Gibas, 2001]
- "Modeling of a DNA binding protein involved in Boom syndrome, and characterization of the mutations that cause the disease" [Gibas, 2001]

Protein structure prediction

Protein structure prediction is the idea that by reading the protein's sequence, you can determine its shape. There are two similar approaches to approximating a solution: homology modeling and ab-initio prediction.

Homology Modeling uses the idea of collecting the attributes of the sequence you want to model and finding another known sequence that has similar characteristic structure. Each protein has a "backbone", a structure on which chemical side chains are attached. Each of these side chains defines which amino acid is used in that particular building block.

[3] Hidden Markov Model

"These chemical properties [electro-magnetic forces] of the side chains help determine how the proteins fold; thus the arrangement of amino acids dictates the three-dimensional structure of the protein."[Drlica, 1997] The prediction process works by finding the best match of a total backbone between the known protein and the predicted protein. Then segments that are missing or incorrect are added or replaced by segments of backbone to make the total backbone the correct total length. Side chains are added to the backbone to complete the structure. Then piece-by- piece, the side chains are optimized to match the desired shape. The model is then optimized using energy minimization. "The key to a successful homology-modeling project isn't usually the software or server used to produce the 3D model. Your skill in designing a good alignment to a template structure is far more critical". [Gibas, 2001]

Ab initio structure prediction can be defined as structure prediction "without the input of experimental data."[Erk, 2000] According to Gibas, "Since ab-initio structure prediction from sequence has not been done with any great degree of success so far, we can't recommend software for doing this routinely." [Gibas, 2001]

Protein structure property analysis

Protein structure property analysis is used for the following tasks:

- Protein structure validation
- Calculating physicochemical properties
- Analysis of surface characteristics
- Analysis of molecular dynamics

There are many measurable properties in protein structure that can be explored further by crystallographers and structural biologists. There are tools available to validate protein structural models. Protein structures should conform to the rules that are based on existing structures or chemical models.

One of the tools available is PredictProtein server. This may be the best site where protein structure analysis can be done. The following US sites are accessible but their response time is little higher:

http://cubic.bioc.columbia.edu/predictprotein

http://www.sdsc.edu/predictprotein

They return analysis in the form of prediction of various structural properties and description of sequences.

Physicochemical properties and internal geometry of protein can be measured by other set of tools available. With the help of those tools, models of the protein's catalytic mechanism can also be developed. 'Some of the most interesting properties of protein structures are the locations of deeply concave surface clefts and internal cavities, both of which may point to the location of a cofactor binding site or active site.' [Gibas, 2001]

To carry out physicochemical property analysis, biologists, with the help of different tools like GRASS, GRASP etc., visualize molecular surfaces with mapped properties.

Electrostatic potential field and other electrostatically controlled parameters such as individual amino acids, pKas protein solvation energies, and binding constant around the protein are of significant importance. There are tools that can compute hydrogen-bonding patterns and analyze intramolecular controls.

One more method, Protein Structure Optimization is used, 'that is the process of bringing a structure into agreement with some "ideal" set of geometric parameters.' [Gibas, 2001]

Protein Structure Alignment and Comparison

Protein structure alignment and comparison are required useful when:

- Sequences appear to be non-homologous
- Structures may be similar (convergent evolution)

'Even when two gene sequences aren't apparently homologous, the structures of the proteins they encode can be similar.'[Gibas, 2001] By doing structural similarity tests using specific computing tools, it is possible to figure out distant homologies. Already discovered protein structures can be compared with structured homology models using these tools.

There are few considerations to do comparison of protein structures like:

RMSD: The difference between two protein structures in atomic positions can be expressed as RMSD or 'root mean squared deviation'. 'RMSD can be computed as a function of all the atoms in a protein or as a function of some subset of the atoms, such as the protein backbone or the alpha-carbon positions only.' [Gibas, 2001]

Superimposition : The first step in superimposition is to do sequence comparison. When one-to-one relationship between pair of atoms is established by using sequence comparison methods, RMSD can be measured. 'Atom-to-atom relationships, for the purpose of structure comparison, may actually occur between residues that aren't in the same relative position in the amino acid sequence. Sequence insertions and deletions can push two sequences out of register with each other, while the *core architecture* of the two structures remains similar.' [Gibas, 2001]

ProFit is an easy-to-use tool to superimpose two protein structures. ProFit measures RMSD and the coordinates for the superimposed proteins. 'ProFit allows the option of superimposing only selected regions of each protein so that domains can be examined independently.'[Gibas, 2001] ProFit compiles and runs on any Unix workstation. ProFit may be downloaded from Andrew's web site (http://www.bioinf.org.uk/).

One of the tools available to do Protein Structure Analysis is PROCHECK http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html)

Biochemical Simulation

The purpose of biochemical simulation is to simulate chemical reactions involved in metabolism e.g., metabolic pathways, transmembrane transport processes, entire cells, entire tissues etc. There are many tools for simulation that can extend from individual metabolic pathways to transmembrane transport processes and even properties of whole cells or tissues.

Earlier scientists who had mathematical background developed these tools. Before the tools became available, they used to describe the system mathematically – for example using complex differential equations to represent different chemical reactions.

Now there are new tools available that can build a simulation model interactively. So the biologists who have knowledge about dynamic system modeling can easily utilize these tools.

The following tools are available for biochemical dynamic modeling:

- Gepasi (http://www.gepasi.org/)
- XPP (http://www.math.pitt.edu/~bard/xpp/xpp.html)

Working with 3D Protein Structures

Protein 3D structure models can be viewed on a computer screen using the tools available on Protein Data Bank (PDB). 3D data sets can be retrieved from any PDB (Protein Data Bank) sites. These tools remove the complexity involved in combining solid structure models in labs and deriving information about sequences and 3D structure. A great amount of effort has been put in this area of Bioinformatics, now called structural Bioinformatics. Now most biologists use these experimental 3D structures available on the Internet.

There are tools available to predict 3D structures and these tools are divided into two sub-groups.

- *Homology Modeling*-

  Modeller (http://guitar.rockefeller.edu/modeller/modeller.html) is a program for homology modeling. [Gibas, 2001]

  ModBase – a database of automatically generated models.

  SWISS-MODEL is an automated homology modeling web server based at the Swiss Institute of Bioinformatics

- *Tools for Ab-Initio Prediction* – RAMP and ROSETTA

Whole Genome Analysis

In 1977, an efficient method was developed to sequence DNA. In 1995, the microbe *Hemophilus influenzae*'s sequence was entirely discovered. This remarkable achievement in 1990's made the Bioinformatics community look more into the technologies that could help them to determine the entire genome sequences. During this period of heavy genomics research scientists carried out tasks like genetic mapping, physical mapping and sequencing of entire genomes. As a result, the available length of the fragments of nucleotides now measure in the millions for microbes and the billions for animal and humans compared to the few thousands of fragments available before the invention of these computational tools.

These tools can store, query, analyze and display these millions of objects interactively. These tools also allow scientists to analyze sequences without gene discovery phase. Thus both steps, i.e., gene discovery and sequencing are now done in one shot. [Claverie, 2003]

With these advances, more and more genomes are sequenced. These genome data can be describe as linear sequences. Now scientists are more interested in analyzing raw genome data as long linear sequences and integrating derived information with presently available genetic and physically mapped data. This analysis of raw sequence data is called *basecalling.* There are some protocols that are used to increase the throughput of DNA sequencing e.g., Applied Biosystems sequencers and the Pharmacia ALF instrument.

In order to do entire genome sequencing, two methods are being frequently used

- The Shotgun Approach
- The Clone Contig Approach

There are few more methods for this purpose and one, called LIMS (Laboratory Information Management System), can track all mini-sequences.

All these genome sequence resources are accessible via tools available on the Internet. Some of them are:

- Genome Information
- Map Viewer
- ORF (open reading frame) Finder
- Locus Link
- Homolo Gene
- COG (Cluster of Orthologous Group)
- TIGR – one of the main resources of newly generated genome data
- EnsEMBL – a collaborative project of EMBL, EBI and Sanger Center

Many more exist, including an organism-specific genome resource, TAIR (The Arabidopsis Information Resource).

Institutes like the National Center for Biotechnology Information (NCBI) and other organizations are providing useful web-based applications to carry out this genome analysis so that other scientists can start from a high-level map and then determine the location of a specific gene sequence.

There are some issues that genome data present because of very large sequences:
difficulty to access genome data in a whole and meaningful way, and conflicts in overall functionalities that genome sequence data show. To overcome these issues, genome annotation and comparison are desirable, though they are difficult tasks.

Primer Design

Primer Design is a term given to the task of choosing from within a very large sequence the subsequence of DNA to be further researched. The goal is to choose a high-quality amino acid subsequence that does not include gaps, overlapped, folded genes or borders between genes. The subsequence is then fed into another process for further research. Polymerase Chain Reaction or PCR is a relatively new process that scientists can use to replicate a DNA sequence [Powledge, 2003]. PCR depends on having good quality "primer pairs" to then bio-chemically replicate. The researcher uses the PCR process to create thousands or millions of copies of the desired DNA sequence and then can perform experiments upon the resulting set. The combination of Primer Design as a methodology and PCR has drastically cut the time necessary for the researcher to build the data set to be studied. There are now even mechanical tools to help automate and further speed the process. PCR, coupled with the basis provided by Primer Design can "do in a week work that used to take a year." [Powledge, 2003]

Primer Design is itself a tool for PCR, which is used to create larger specimen sets more quickly than previously possible. Some common Primer Design applications include Primer from Whitehead, Primegen from Scientific Computing Services in Seattle, WA., eprimer3 by Gary Williams of Cambridge, OSP by LaDeana Hillier of the Washington University Medical School.

DNA Microarray Analysis

Microarray analysis is an expansion on existing Biology blotting techniques. The idea is that samples are evaluated in parallel and then experiments are performed on them. One risk of this is that to ensure quality of the initial data, the scientist must verify all the samples in the array prior to performing the experiment.

If the analysis of the base array can be done, then the experiment can be performed en masse so that many result sets can be evaluated. The sheer number of samples involved in microarray analysis requires that the researchers use statistical analysis techniques to evaluate the results of experiments. A benefit of this is that the results, provided the initial samples are certified to be of sufficient quality, can potentially be more reliable. Another benefit is that the statistical analysis process will exclude or give proper weighting to true outliers and unique data points.

The Medical Physics at the University of Naples describes how to create a DNA Microarray. That discussion cannot easily be paraphrased without reader knowledge of the terminology and techniques. They (MPUN) describe how to transfer information from

DNA onto the Microarray construct and use the wet lab technique of "Blotting" as an analog.

O'Hearn gives a nice English summary of Microarray analysis: [O'Hearn, 2002]

- Expression pattern DNA from entire genomes can be explored
- A microarray (gene chip) is a glass slide
- The slide contains about 20,000 precisely placed spots containing different probes--DNA (or cDNA) oligomers
- Any reaction changing the microarray signal can be assigned to a specific DNA sequence
- A given complementary DNA or RNA sequence is detected by each spot as a DNA sample is run over the chip.

Proteomics analysis

A Proteome is "the entire protein profile of a cell" [Clark, 2001]. Proteomics Analysis, by extension, is the study of proteins within cells. Since proteins are made up of amino acids, this is an extension of the DNA analysis techniques described elsewhere in this paper.

Like the DNA Microarray Analysis method described above, the automation of some of the work in Proteomics Analysis has resulted in much greater researcher analysis throughput. The technology available for Proteomics Analysis is a combination of software tools and machines.

(V) BIOINFORMATICS-SPECIFIC LANGUAGES

A great number of programming languages have been used in Bioinformatics. The different languages are at different levels of Bioinformatics library development. There is usually a group, in many cases an open source initiative, working within each language's constraints to equip it for effective Bioinformatics research.

Some languages, like Perl, are more popular among the Bioinformatics specialists than others. Redundancy in software development, especially in the development of programs aimed to perform very common tasks, such as representing sequences and structures, translating from DNA sequences into amino acids, has led to the creation of open source groups that share the goal of reducing the waste of development efforts through software reusability and standardization. One of such goups is the Open Bioinformatics Foundation . It grew out from the Bioperl, the BioJava, and the Biopython projects. At the present time, in addition to these projects, it includes BioRuby, BioPipe, BioPython,

BioSQL / OBDA, BioMoby, and DAS. Besides managing resources for the bio- projects, the OBF supports various developer-centric events and the BOSC conferences. By browsing the web sites of each of the bio-projects, you can notice that only BioPerl and BioJava have a sizable amount of work accumulated. The BioMoby and the BioSQL web sites show some initial achievements, while the BioPipe site doesn't contain any information about its project. The BioLisp project is not listed on the OBF site, possibly because it seems to be very new. The BioXML project is mentioned on several bio- web sites, but because the links to the BioXML site are broken, it is impossible to determine the project's state.

The BioPerl Project

Low learning curve, compactness of code and a good support for string manipulation (which is important for working with the DNA and protein strings) are among the qualities that made Perl a favorite programming language among the life science professionals. Even though Perl is not as fast as C/C++, it is the language of choice for small scientific projects.

The Bioperl Project provides reusable modules written in Perl that can be used to process results of various Bioinformatics programs including Blast, Clustalw, TCoffee, Genscan, ESTscan, and HMMER. It does not include complete programs that can be compared to commercial packages. Its Auxiliary Libraries include modules for creating graphical interfaces (bioperl-gui), persistent storage in RDMBS (bioperl-db), and CORBA bridges to the BioCORBA (bioperl-corba-client).

The BioJava Project

Since Java is a full-fledged programming language, the BioJava developers' goal goes beyond development of modules auxiliary to existing applications. Currently, the BioJava libraries include objects for manipulating sequences, file parsers, CORBA interoperability, dynamic programming, visualizing, external file formats and programs, and training models. With the growth and maturity of the BioJava libraries, it will be possible to use them for development of both free and commercial applications.

The BioPython Project

The BioPython Project contains several extensions to the Python language that have been developed at the Molecular Graphics Laboratory of the Scripps Research Institute. These extensions were used for generating two tools: the Python Molecule Viewer (PMV) and the AutoDockToolKit and (ADT).

The BioCORBA Project

The BioCORBA Project provides an object-oriented, language neutral, and platform independent environment that allows interoperability between different bio-frameworks like BioPerl, BioJava, BioPython, etc. Among the current projects of BioCORBA are Novella and BSANE. Novella stands for 'NO-Valuetype Enhanced Liaison to appLab Analyses', and the goal of this project is to eliminate valuetypes that are not implemented in some languages used in Bioinformatics. The BSANE project is a new specification requested by Life Sciences Research (LSR) group, which works under the OMG umbrella. It seems that both projects are on their initial stage; the BioCORBA web site and the sites that it has links to provide very cursory information about the projects.

The BioDAS Project

The BioDAS project is dedicated to the development of the distributed annotation system (DAS), which will allow a single client to integrate information from multiple servers. This system "allows a single machine to gather up genome annotation information from multiple distant web sites, collate the information, and display it to the user in a single view".

Ruby and the BioRuby Project

The BioRuby project aims to implement integrated environment for Bioinformatics with the help of Ruby - an interpreted scripting language with object-oriented features. Currently, BioRuby has classes for sequence manipulations, data I/O, database parsers, and some others.

The C Language and EMBOSS

EMBOSS (European Molecular Biology Open Software Suite) is an extensive open source collection of the C programs and libraries for the molecular biology. Currently, the EMBOSS suite contains about 100 programs for a variety of tasks such as sequence alignment, database searching with sequence patterns, protein motif identification, nucleotide sequence pattern analysis, codon usage analysis for small genomes, and much more.

(VI) SOME CURRENT PROJECTS

This section describes some non-language projects currently underway. These are research initiatives and are just a few of the field's many ongoing projects.

Pseudoknots in RNA Secondary Structure - Oxford University

A Pseudoknot is defined as two non-disjoint base pairs that span an RNA sequence where neither base pair contains the other. If the pairs were disjoint, they could be studied with existing tools. If one contained the other, they could be studied independently, again with existing tools.

Oxford's web site [Oxford, 2002] discusses the problem of pseudoknot analysis and compares the efficiencies of the algorithms used to analyze pseudoknots. Their claim is that these algorithms are less predictive than the Bioinformatics models and algorithms used to analyze other biological units. The Oxford project intends to try to define an analysis methodology to improve the study of Pseudoknots.

They note that Pseudoknots are found in only 5% of a given set of RNA base pairs in an RNA structure. This means that a researcher can study the other 95%. The site goes on to state "almost all RNA structures contain one or more pseudoknots."[Oxford, 2002] This means a complete analysis of an RNA structure requires that this problem be solved.

The major challenge they define is that the existing algorithms are $O(n^6)$ for execution time and $O(n^4)$ for storage space. What this means is that as the size (n) of the entities studied increases, the execution time and space required increase to the point where the algorithm is infeasible.

Ensembl Genome Browser

Ensembl[Ensembl, 2003] is a web-based system offering researchers quick and easy access to sequence information for a number of genomes including the common fruit fly, homo sapiens and the Zebra Fish. It is a joint venture of the Sanger Institute and the European Bioinformatics Institute and its database is publicly available.

Project VirGO

Project VirGO[VirGO, 2003] is based at the University of Victoria, British Columbia, Canada. Its research focuses on the genomes of viruses and related entities. The project began in 1998. The VirGO web site describes a tool (VGO) that takes a GenBank data file and translates it and executes a series of operations against it that combine several common tools including FASTA and BLAST described elsewhere in the paper.

(VII) CURRICULUM

This section describes some of the Bioinformatics curricula for some institutions with formal programs. More detailed lists of courses and institutions are available in the Appendices. This section can be a starting point for a student interested in Bioinformatics programs or for an institution interested in starting a new program.

Numerous universities in the U.S. and abroad now offer Bioinformatics programs at a variety of levels, ranging from diplomas to Ph.D. The table in Appendix A lists some of the institutions that offer Bioinformatics programs. Some institutions offer Masters and Ph.D. programs while others offer only Bachelors' programs, certificates or research opportunities.

Of those that offer research only, some, like North Carolina State University, have Bioinformatics research done in their Biology departments. Others, like the National University of Australia, are in the early stages of building an academic department. These may eventually evolve into full programs.

The Baskin School of Engineering at the University of California, Santa Cruz (UCSC) offers an undergraduate degree in Bioinformatics. UCSC's program is limited to honors students because of the breadth of coursework that must be completed for the degree. The program description cautions potential applicants that they are expected to do a lot of research and learning outside the course.

The UCSC B.S program offers one course with a specific definition. The other courses are seminars, research courses or tutorials with vague definitions. The specific course is their "100" course. According to the definition, it emphasizes "DNA and protein sequence alignment and analysis."[UCSC, 2003]

Georgia Tech offers some undergraduate Bioinformatics coursework as part of its undergraduate Molecular Cell Biology program. A review of the course descriptions for its required courses lists no Bioinformatics-specific courses.

The University of Minnesota offers a summer internship in Bioinformatics. This program offers workshops and tutorials in various topics related to Bioinformatics including Sequence Analysis and DNA Microarrays. The stated goal of this program is to give the students a "beginning understanding of the area."[UMinn, 2003]

UCLA offers an Undergraduate Major in Bioinformatics/Cybernetics. The program description asserts that this program prepares students equally well for graduate studies in Bioinformatics and for continued Bioinformatics research and development. UCLA's

recommendation for graduate students suggests the school prefers graduate applicants with a multidisciplinary background including Biology and either Computer Science, Mathematics or Statistics.

The University of Alberta, Canada offers a Bioinformatics undergraduate program as a joint venture between its Biology and Computer Science Departments. Students in this program are required to take two Biochemistry courses, two Biology courses and one Genetics course in their second year. The third- and fourth-year options include two specific Bioinformatics courses and two of four Genetics courses. The first Bioinformatics course deals with computational tools and databases as well as sequence analysis from a molecular biology perspective. The second course is described as including "advanced topics" of Bioinformatics including team assignments to create new tools for Bioinformatics research.

A few institutions, like Boston University, offer everything from Seminars for current researchers to Masters and Ph.D. programs. The purpose of the Seminars is to educate specialists in related fields (e.g. Biology or Genetics) in the tools and techniques used in Bioinformatics research. Since a Biologist may have to be taught much about the commonalities between Biology and Bioinformatics, the Boston seminars can get such a researcher up to speed in the Bioinformatics-specific aspects of the field.

There are also traditional institutions that offer Masters and Ph.D. programs and even one that offers a diploma. The Masters and Ph.D. institutions include Iowa State University and the University of Michigan.

Bioinformatics Courses

The table in Appendix B lists courses offered in various Bioinformatics graduate programs. Any quotations in the comments are from the appropriate institution's curriculum description on the web site listed in Appendix B. This list was curtailed because of the sheer number of courses. The discussion below discusses the types of courses and their relevance to the field.

The courses above fit into one of six basic categories: Bioinformatics, Biology/Chemistry, Mathematics/Statistics, Computer Science, Seminar/Discussion/Workshop, Law/Ethics.

Bioinformatics Courses

These courses generally give the graduate student hands-on experience with the tools and techniques best suited to Bioinformatics research. The courses can range from something

very close to tool training e.g. “DNA Microarray Bioinformatics” to more theoretical courses like “Research in Bioinformatics”. The institutions above have a range of courses available to give the Ph.D. candidate significant freedom to choose a direction. As is found in many graduate disciplines, the introductory graduate courses appear to differ only slightly from their undergraduate counterparts.

Biology/Biochemistry Courses

Whatever the technology applied to the problem, the problem domain surrounds biology and biochemistry. The Ph.D. candidate must be familiar with the problem domain and be equipped to perform the necessary biological research or the Bioinformatics tools and techniques are pointless. Many institutions, for example the University of Michigan and Virginia Tech, have requirements for a Ph.D. candidate’s background or undergraduate degree. If the candidate lacks the appropriate background, usually in Biology and Biochemistry, that candidate is required to take additional courses, often not for graduate credit, to attain sufficient field knowledge.

Mathematics/Statistics Courses

The mathematics and statistics courses are required and offered for two primary reasons. First, much of the analysis portion of Bioinformatics research involves statistical analysis tools and techniques. This means that the Ph.D. candidate must have a solid grounding in this field. Second, because the tools available for Bioinformatics research are still relatively new, the researcher is better served with an ability to perform the analysis using conventional tools. This can give the researcher the means to verify the statistical or numerical analysis portion of an experiment independent of a potentially suspect tool.

Computer Science

These courses range from pure programming, e.g. ‘Java I’ or ‘Perl and Unix for Bioinformatics,’ to algorithm design and analysis to user interface design. The institutions noted above do not require many pure CS courses for Bioinformatics Ph.D. candidates, but the courses are available. We believe that these are made available for similar reasons to those of the Mathematics/Statistics courses: the tools used in Bioinformatics research are not yet mature and it serves the field well if some of those doing the field’s practical research have the ability to construct additional tools or to recommend enhancements to or analyze existing tools’ efficacy.

Law/Ethics

The required Law and Ethics courses are intended to give the researchers and students an understanding of the legal and ethical issues having to do with Bioinformatics. A common application of Bioinformatics research is genetic engineering. Therefore, there is a clear need to ensure that researchers are well informed as to the current laws. The ethics courses seemed to be similar in content to those of the medical schools. Not all institutions required an Ethics or Law course in their program. Of those that did, usually one or at most two courses were required.

Seminar/Discussion/Workshop

Each of the institutions with courses listed above had at least one seminar-type course in the curriculum. We think this is because the field is still expanding and new discoveries are relatively more frequent than in a more mature field. In addition, related research techniques and tools are evolving. The seminar format allows a Bioinformatics department to have a course on the curriculum that can quickly be adapted to include a brand-new technique or to consider a new discovery.

(VIII) ANALOGIES

The Analogies section will provide a simplified view or mental picture of some of the molecular biological processes of life. Also included are some suggestions about data storage that maybe useful for tool efficiency enhancement.

DNA, Protein and everything in-between, an analogy

The purpose of this section is to provide a basic understanding of how the blueprint of life, DNA, is transformed into proteins, the machinery of life. The construction process of building proteins can be compared with creating a brass casting.

We have a model of what we want cast, DNA. We want a brass figurine, a protein. The casting process requires us to make a mold of the model, the biologist call this mold RNA. A cast of the model is made and the brass is poured into this mold and allowed to cool into our brass figurine. But it is not just that simple is it? What material do we need to add to the casting so it will come apart cleanly? What are the details of the process?

The DNA

The molecular biologist is concerned with more than just pouring brass into a physical shape, they are concerned with the electrical forces and energy between atoms, molecules

and other building blocks that become the final outcome of there observation or craft. The DNA is the blueprint, but what kind of blueprint? DNA is a blueprint of electrical forces and energy boundaries. It is a database of all the blueprints needed for all the proteins needed for life.

So look at it this way, DNA has a backbone[4] made up of molecules whose primary purpose is to create a docking location for information molecules. Something like a 'blank movie film'[5] waiting for the exposure of the story. 'Each frame' holds a segment of the action. The subunits that are this information are called nucleotides. Each nucleotide can be seen as an electrical energy signature that shapes its frame. Nucleotides also twist their backbone. There are four nucleotides labeled {A, T, G, C}[6].

The gene

So how do we define a gene? It is a part of the DNA (made up of nucleotides) that is the model to create a protein. It is the film, the scene, which describes how to make a protein. We need to see the movie move. We need several frames together to create the action. First, each action needs a beginning and an end. Second, the action needs several frames in order to move. "This is because proteins, the linear, chainlike molecules made from *information* in DNA, have subunits (amino acids) of 20 *types* rather than the four subunits (nucleotides) used to store information in DNA."[Drlica, 1997] Further it has been found that each 'action', each amino acid, that makes up the protein is derived form three nucleotides. The DNA model has three nucleotides grouped together called codons that are the model for each amino acid. The grouping of the three nucleotides define an electrical shape or "socket" that forms the "surface" of the DNA that is to be copied. The machinery of the cell copies the electrical force, the energy signature of the codons, to the mold called RNA (ribonucleic acid). This process of converting information from the DNA to the RNA is called transcription, or RNA synthesis.

The enzyme

The energy of formation, the energy needed to temporally reshape a "socket" so other energy shapes can be inserted is facilitated by a catalyst, or as the biologist call it, an enzyme. A catalyst is like a special tool that allows chemical parts to be assembled or dissembled. It enters into the transaction, makes it possible, and is not consumed in the outcome. It is like using a pair of pliers to put together plastic building blocks that are so tight that large forces are needed to put the blocks together. On the other hand when they

[4] most of the data about the life process were taken from Drlica [3]
[5] The analogy of the movie film is an extension presented by Drlica [3]
[6] The set of four DNA nucleotides are Adenine (A), Cytosine (C), Guanine (G), and Thymine (T)

are together it is so hard to take them apart; we would want to use a pair of pliers to take them apart. An enzyme is this tool.

Pouring amino acids into the RNA mold creates a protein figurine in much the same way as the DNA was used to create RNA. Enzymes are the lubricant that allows the protein to be formed and be able to be separated from the mold.

Molecular shape, how is it determined?

The DNA, the RNA, the protein are all twisted. How can we view this behavior?

Electrical energy, the glue that holds it all together

But what is all this electrical energy stuff about? This glue is like a bunch of rubber bands holding some tinker toys together. Each rubber band influences how it all fits together, even though any one of them, taken by itself is insufficient to give you a dim reflection of how the result will look. However, each rubber band is responsible for changes made in the molecular structures (twisted shape), the change in its electrical signature. Let us look at a simple energy model that may help us extrapolate a workable solution to answer the question.

Room x, a model of superposition[7] of light

Visualize a rectangular room with seven lights equally spaced across the ceiling. Directly below the center light there is centered a rectangular table. Centered on the table is one piece of 8.5 by 11 paper and centered on the paper is a single letter x. All the internal surfaces found in the room have been painted with magic paint that reflects no light. We have a light meter that measures the light intensity at the x location. The x is illuminated by each of the seven bulbs. In the experiment we in turn light only one bulb at a time and take the illumination reading and look to see if we can read the letter x. At the end we light all the bulbs. We find the sum of all the light from each bulb equals the illumination with all the lights together being turned on. We notice that the illumination from each of the bulbs is proportional to the inverse square of the distance from the bulb to the x. We will call this room x. Let us now visualize the room expanded infinitely long with an infinite number of lights. We will replace the x with a y and call this room y. In room x let us place a pencil pointing straight up at the x and move it away from all the lights toward the bottom of the paper. We turn on all the lights and we see seven shadows pointing down away from the pencil. We observe that the further the lights are from the

[7] Room x and room y are an example of superposition, for the shadows can be analyzed separately and then assembled together for the total outcome just as the electrical forces are done for molecules.

pencil, the dimmer the shadow becomes. That is to say the further the light source is from the pencil the less influence the light has on producing a shadow. If we did the same thing in room y we would see the same thing except there would be more shadows. Now for the point, or should we say a few questions. At what point in the y room do the shadows make no difference? Next, in the x room and y room let us start by turning on the center light, and then the next two outer lights and then the next two outer lights and so on. At what point does it make any difference if you turn on the next set of lights? What do you conclude?

Electrical Forces have the greatest influence

We need to look at what is most significant in the changes, or electrical forces, to be able to solve the question of how any given structure will appear. What causes the changes? Sometimes it is not the physical item but the energy related to the object we need to model. Of the molecular behavior of DNA, RNA, proteins, etc. it is the electrical forces and their related energies that needs to be modeled. After all, the equations of physics demonstrate the electrical forces are greatly more influential than gravitational in the total force equations at the molecular level.[8] Just as the light room model demonstrated that the closest lights to the pencil had the greatest influence on a shadow, we can see that the greatest influence on molecular structure is the "closest" electrical force to the molecules. It is the closest rubber bands that have the greatest influence. This is the answer to the question above. Let us use this analogy to help us model the data structures we need.

There are 20 Amino acids, the building blocks of proteins

First, there are three objects of interest: DNA, RNA, and proteins. There are four DNA nucleotides {A, T, G, C}. There are four RNA nucleotides {A, U, G, C}. From mathematics we can calculate the total number of DNA codons. There are four DNA nucleotides {A, T, G, C} that are taken three at a time (called Codons) to yield 64 combinations. There are 64 ways we can form a codon. These 64 codons in some way map to 20 Amino acids that are the building blocks of proteins. See Appendix F for details on the Codons and their amino acids.

Electrical Forces influence the protein's 3D shape

Please focus on the reality of space in which we are discussing. The physical location of an electric force container, the atomic structure, seems in some way to space the electric forces and their shape. It is if the backbones (of the DNA, of the RNA, and of the proteins) space each nucleotide or each amino acid apart from each other. They hang off

[8] The gravitational and electrical force equations can be found in appendix

the side of the backbone like a separate wing of a large mansion. Please look at the DNA and protein image in Appendix G to see this is true. Further there is a twisting of the structures from the electrical forces, not seen easily in the pictures, which make some components more influential in the atomic electrical force interactions. The building blocks are not just in two dimensions, but also in the third depth dimension. This is just a way of looking at the table and description of the codons in Appendix F. From this perspective it may make more since why some of the nucleotides in the codons in the above table have more significant roles to play then others. It is a two-way path, each energy signature affects the molecular structures of the others around it, and the others affect its structure. Maybe now it makes since why fifty percent of the codons do not need the third nucleotide for their definition. Why only three codon combinations need all three nucleotide energy signatures for their total definition. Further these groups of three nucleotides are isolated in there own wing of the structure, further separating the shape. The remaining codons need the last nucleotide to make the call in a close race. Why such detail?

Sequence data only sometimes can determine 3D rendering of proteins

The purpose of the last few paragraphs is to paint a picture of understanding for the next few statements. Each codon, each amino acid, has its own shape. However when a protein is formed only sometimes the protein's linear sequence of amino acids can be used to model the 3D rendering. "In multiple alignment generated from sequence data alone, regions that are similar in sequence are usually found to be super imposable in structure as well" [Gibas, 2001] "In reality, amino acids which are far apart in the linear chain may be physically close to each other when a protein folds. Chemical and electrical interactions between them cannot be predicted with a linear model." [Karchin, 1999] Superposition[9] or tinker toy construction for 3D rendering of molecular structures can be accomplished some of the time by using their sequence.

Other complications make 3D rendering of proteins complex to model

There are other details that will make it complicated. Some proteins may bind to a gene and block its expression. For example, my large right toe does not need to know what color my eyes are and so that gene may be blocked in my large right toe. There are other complexities, for example 'transfer RNA molecules that serve as adapters that convert the four letter alphabet of DNA and RNA into the 20-letter alphabet of proteins.'[Drlica, 1997] The complex processes that messenger RNA uses to build proteins is another example of the biological complexity that will dynamically change the shape of the cell's machinery. The dynamic changes taking place in the cell result from many proteins doing

[9] Room x and room y are an example of superposition, for the shadows can be analyzed separately and then assembled together for the total outcome just as the electrical forces are done for molecules.

their task in real time together. Each protein modifying their neighborhood through one or more locations where their close proximity to other protein structures influences that locations protein shapes. Further there is the complexity shown in the Research Techniques section that demonstrates the current art of research for Protein sequence analysis, Protein structure prediction, Protein structure property analysis and the list goes on. The list of complexity is overwhelming and may be hard to model. It seems the researcher is resolving the complexity problem by using data to define their research.

The Information technologist helps the researcher

Each of the above processes will temporally or permanently modify or convert molecular structures as the process of life goes on. The energy signature, the physical shapes seen by the researcher, help them predict future transactions. The researchers seem to be using sequence analysis and 3D rendering of the objects they are studying. Understanding little of the complexity of what the researcher is doing maybe we can still help by examining what little we do know and propose some improved data structures. Looking at some of the simpler underlying manipulations that the data undergoes may also help.

First, the complexity of 3D rendering is beyond the scope of the paper and so nothing more than acknowledging its complexity is saying too much.

Second we find that most of the other research is based on comparing part of a DNA or RNA or protein sequence with another sequence of a like kind. The process integration leads into big oh's of expediential order. "Dynamic programming alignment of two sequences takes seconds. Alignment of four relatively short sequences takes a few hours. Beyond that, it becomes impractical to align sequences…" [Gibas, 2001] The researchers have resorted to other approaches to do their work. They search smaller sequences. They search fewer sequences at a time. They use the Greedy algorithms and accept the lower quality of its results. We have a few suggestions.

The IT Problems and Solution:

There are large volumes of sequenced data. "DNA molecules can be very long, sometimes containing more than a hundred million nucleotides." [Drlica, 1997] The sheer volume of disk space needed to save such data is only part of the problem. Each large string must also be read from input, translated and stored with the associated overhead.

The multiple sequence alignment process is by nature "big oh" of exponential order.[Gibas, 2001] Gene manipulation made up of hundreds or thousands of characters can be very expensive.

Both DNA and RNA have logically the same format. Therefore, we suggest a simple modification of the data structures. During the transcription process, DNA is bio-chemically "parsed" into RNA. Transcription requires a read, a copy a translation T to U throughout the selected string. We suggest using a pointer to the start of the gene in the DNA that can be used to represent RNA we want. We can then use the pointer to quickly get to the start of the DNA string we want. Instead of the read-copy-translate steps shown above, we simply read it into memory and use an alternate character set to display it according to its domain.

We suggest creating a character set optimized for the characters and attributes necessary to describe the data of Bioinformatics. Such a custom character set would leverage the work already done to interpret a common character set (e.g. Western European) to translate the characters from the user interface(s) to storage and back. Devices and software can then be built to understand this character set thereby speeding graphical rendering of Bioinformatical data.

The table below shows a DNA fragment spanning from pointer P1 to pointer P2. The only difference between them is that each T in the DNA has been translated into a U in the RNA. Since computing devices must translate the binary code from storage into human readable form in a user interface, we can switch the character set used to do the translation and use no more time or cycles than if we had done a straight read operation.

Table 1 - DNA Transcription Example

			P1											P2		
DNA[10]	…	A	C	C	T	C	G	G	T	A	C	C	G	C	T	…
RNA	…	A	C	C	U	C	G	G	U	A	C	C	G	C	U	…

This can be extended to more complex structures like proteins. Of course, such a character set must allow for special formatted datasets:

- Headers to tell what domain is coming – DNA, RNA or something else
- 2 bit (0,1,2,3) RNA nucleotides or DNA nucleotides
- 2+2+2 bit nucleotides for amino acid codes (with stop and start codes) we just list the nucleotides

If we were to implement such a solution, it will not always increase performance. Storage and caching will be improved. Storage can be optimized by as much as a factor of four.

[10] One side only

If an alternate character set adds too much complexity to the use of the binary codes then object-oriented polymorphic solutions applied to each owner class may be appropriate. The java toString method is a simple implementation of this process.

(IX) CONCLUSION

As future directions, we see the field shifting in emphasis from genes (genomics) to gene products (proteomics; structural genomics). Eventually, the available database and toolsets could allow researchers to create things like designer drugs – personalized medicine based on the analysis of a person's genome. If the database technology and standardization sufficiently evolves, we could even see something like real-time data mining and transaction monitoring. This could be used to detect drug-to-drug interactions and side effects.

Bioinformatics is a relatively young field compared to other life science fields. We see several challenges that must be met as it matures. The first challenge is a lack of a standard data storage format. Many of the tools and databases in use today began as proprietary applications or are still proprietary and therefore use different presentation formats including Gene Expression Markup Language (GEML), based on XML, and Microarray Markup Language (MAML). As in other fields, it will take time for a de facto standard to emerge and still longer, if ever, for something like an IEEE standard. Until the data formats are standardized, the available tools and languages cannot be easily compared for performance and efficiency will be lost converting from one format to another. As long as the data storage is not standardized, nor can the retrieval. This means that reports or other presentation techniques may have to be rewritten for each particular data format.

The second major challenge is related to storage. Because of the three-dimensional nature of many items being studied, there are efficiency issues in how to store information describing them in a relational database. Additionally, the entities in many cases behave based on external adjacent entities or conditions. The human proteome contains about 30,000 proteins (3D structures). Static visualization is therefore difficult. Dynamic visualization is even more challenging because conditions like temperature and proximity can make proteins change their shape. This places further importance on storage techniques. Object-Oriented (OO) databases are best for genomic data because genes can then be stored as single units. There is however a big challenge in moving data from existing relational to OO databases. Another issue is the relative immaturity of OODBMSs compared to RDBMSs. For three-dimensional items, we recommend a compromise: use a deductive model as an extension of relational databases. For two-dimensional items, we recommend the custom structure proposed in Section VIII. This description, from the IT Problems and Solution section, offers a 4:1 storage advantage.

Without a standard data format, data mining across multiple databases is much less efficient. Each format would need to have a translation set that would convert it to a common format with the associated performance hit.

With standard data storage, we believe that developers building tools can focus their energy on new and better tools rather than on building several equivalent systems. We believe that this may reduce the need to cross-train Molecular or Cell Biologists into IT people and may eventually allow Bioinformaticians to be IT people cross-trained in the fundamental concepts of Biology and Mathematics. These people would be analogous to IT people with expertise in other non-IT problem domains.

(X) GLOSSARY

Glossary[11]

Coalescent – to grow together or to arise from the combination of distinct elements

DNA – Deoxyribonucleic Acid

Evolutionary tree or Phylogenetic tree[12] – A diagram that depicts the evolutionary relationship of protein or nucleic acid sequences.

Exons – The protein-coding DNA sequences of a gene.

ho·mol·o·gy *n. pl.* ho·mol·o·gies[13] – The relation of the elements of a periodic family or group. The relation of the organic compounds forming a homologous series.

PCR – Polymerase Chain Reaction – A method for amplifying a DNA base sequence using a heat-stable polymerase and two 20-base primers, one complementary to the (+)-strand at one end of the sequence to be amplified and the other complementary to the (-)-strand at the other end. Because the newly synthesized DNA strands can subsequently serve as additional templates for the same primer sequences, successive rounds of primer annealing, strand elongation, and dissociation produce rapid and highly specific amplification of the desired sequence. PCR also can be used to detect the existence of the defined sequence in a DNA sample.

Pseudoknot – two non-disjoint base pairs that span an RNA sequence where neither base pair contains the other.

RNA – Ribonucleic Acid

Splicing Site – The joining of separate strands of DNA or RNA.

Stochastic – involving chance or probability

Phylogenetic[14] – based on natural evolutionary relationships

---

[11] BioTech Life Science Dictionary – University of Texas – http://biotech.icmb.utexas.edu/search/dict-search.mhtml Copyright BioTech Resources and Indiana University
[12] page167 McGraw – HILL DICTIONARY of BIOSCIENCE ISBN 0-07-052430-0
[13] http://dictionary.reference.com/
[14] http://www.m-w.com/cgi-bin/dictionary

(XI) BIBLIOGRAPHY

(a) Publications

[Bergeron, 2003]
Bergeron, B. *Bioinformatics Computing*. Upper Saddle River, NJ: Prentice Hall, 2003.

[Claverie, 2003]
Claverie, J.M., and C. Notredame. *Bioinformatics for Dummies*. New York, NY: Wiley Publishing, 2003.

[Doom, 2002]
Doom, T., M. Raymer, D. Krane, and O. Garcia. "A Proposed Undergraduate Bioinformatics Curriculum for Computer Scientists." ACM SIGCSE 2002 February 27-March 3, 2002, Covington, KY 2002 ACM.

[Drlica, 1997]
Drlica, K. *Understanding DNA and Gene Cloning: A Guide for the Curious.* 3d ed. John Wiley & Sons, 1997.

[Erk, 2000]
Erk, P., Nancy 19th European Crystallographic Meeting – August 25-31, 2000. (http://www.lcm3b.u-nancy.fr/ecm19/pdf/09102%20-%20oral.pdf)

[Gibas, 2001]
Gibas, C., and P. Jambeck. *Developing Bioinformatics Computer Skills*. Sebastopol, CA: O'Reilly & Associates, 2001.
(http://www.sglab.org/bioinformatics/shortcourse/dbcs/index.html)
(http://www.sglab.org/bioinformatics/shortcourse/dbcs/contents.html)

[Hughley, 2001]
Hughley, R., and K. Karplus. "Bioinformatics: A New Field in Engineering Education." IEEE 31st ASEE/IEEE Frontiers in Education Conference, October 10-13, 2001 Reno, NV, 2001 IEEE.

[Parker, 1997]
Parker, S.P, Ed. *Dictionary of Bioscience*. McGraw-Hill, 1997.

[Sears, 1967]
Sears, F.W., and M.W. Zemansky. *University Physics*. Addison-Wesley, 1967.

[Tisdall, 2001]
Tisdall, James. *Beginning Perl for Bioinformatics*. Sebastopol, CA: O'Reilly, 2001.

(b) Online Resources

[Abate, 2003]
Abate, L., E. Bertolucci, M. Conti, C. Montesi, and P. Russo. Medical Physics at University of Naples, Italy (MPUN), 2003 - http://www.na.infn.it/mfa/med/autoradiography.html

[Bergeron, 2002]
Bergeron, Bryan. "Applied Bioinformatics Computing: An Introduction." Nov 29, 2002 - http://www.informit.com/.../st~%7B822202F4-E865-472D-B32F-BE77D3B7ACE1%7D/ content/index.asp

[Biocorba, 2003]
Biocorba.org - http://biocorba.org/

[Biodas, 2003]
Biodas.org - http://biodas.org/

[Bioinformatics, 2003]
Bioinformatics.org - http://bioinformatics.org/

[Biojava, 2003]
Biojava.org - http://www.biojava.org/

[Biolisp, 2003]
Biolisp.org - http://www.biolisp.org/

[Biomoby, 2003]
Biomoby.org - http://biomoby.org/

[Bioperl, 2003]
Bioperl.org - http://www.bioperl.org/

[Biopython, 2003]
Biopython.org - http://biopython.org/

[Brown]
Brown, Stuart M. "Bioinformatics Tools." Department of Cell Biology, NYU School of Medicine - http://www.med.nyu.edu/people/S.Brown.html

[Chang, 2001]
Chang, J., B. Chapman, and I. Friedberg. "Biopython Tutorial", 2003 - http://www.biopython.org/Tutorial.pdf

[Clark, 2001]
Clark, M. W., E. Henderson, W. Henderson, A. Kristmundsdottir, M. Lynch, C. Mosher, and S. Nettikadan. "Nanotechnology Tools for Functional Proteomics Analysis." International Scientific Publications, Inc., 2001 - http://www.iscpubs.com/articles/abl/b0103.cla.pdf

[Dugan, 2001]
Dugan, Jonathan, "Open Source Initiatives in Bioinformatics" (report). August 2001. http://www-smi.stanford.edu/pubs/SMI_Reports /SMI-2001-0902.pdf

[Emboss, 2003]
Emboss.org - http://www.emboss.org/

[Emsembl, 2003]
Emsembl.org – http://www.ensembl.org

[Felsenstein, 2002]
Felsenstein, J. Department of Genome Sciences, University of Washington, 2002 http://evolution.genetics.washington.edu/phylip/software.html http://evolution.genetics.washington.edu/phylip.html

[GenBank, 2003]
GenBank - http://www3.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed

[Handprint, 2003]
Handprint.com – http://www.handprint.com/LS/ANC/disp.html

[Hanekamp, 2002]
Hanekamp, T. "Protein Sequence Alignments" (lecture notes). Department of Molecular Biology, University of Wyoming, 2002. www.uwyo.edu/molecbio/LectureNotes/ MOLB5650/Hanekamp1.ppt

[Hessling, 2002]
The Hessling Editor – Hessling M., 2002 http://hessling-editor.sourceforge.net/

[Just, 2002]
Just, W. "Complexity Issues in Bioinformatics." Department of Mathematics and

Quantitative Biology Institute, Ohio University, Apr 2002.
www.math.ohiou.edu/~just/TALKS/Complexitytalk/Complexitytalk.PPT

[Karchin, 1999]
Karchin, R. "Hidden Markov Models and Protein Sequence Analysis." Baskin School of Engineering, University of California, Santa Cruz, 1999 -
http://www.cse.ucsc.edu/research/compbio/ismb99.handouts/KK185FP.html

[Mallery]
Mallery, Charles H. "A Brief and Incomplete History of Cell and Molecular Biology." Department of Biology, University of Miami
http://fig.cox.miami.edu/~cmallery/150/gene/hist_CMB.htm

[O'Hearn, 2002]
O'Hearn, S., National Research Council Canada, OCT Tech (OCCTECH)
http://www.octtech.com/research/OHEARN/c475ERASED/page13.html, 2002

[Open, 2003]
Open Bioinformatics Foundation - http://open-bio.org/

[Oxford, 2002]
Oxford University Department of Mathematical Genetics and Bioinformatics, 2002 –
http://www.stats.ox.ac.uk/mathgen/evolve.html
http://www.stats.ox.ac.uk/mathgen/research.html
http://www.stats.ox.ac.uk/mathgen/bioinformatics/projects/pseudoknots

[Powledge , 2003]
Powledge, T. M. "The Polymerase Chain Reaction." 2003 -
http://www.faseb.org/opar/bloodsupply/pcr.html

[PDB, 2003]
Protein Data Bank - http://www.rcsb.org/pdb/
National Center for Biotechnology Information - http://www.ncbi.nlm.nih.gov/

[Richon]
Richon, Allen B. "A Short History of Bioinformatics", Network Science -
http://www.netsci.org/Science/Bioinform/feature06.html

[Rubi, 2003]
Rubi-lang.org - http://www.ruby-lang.org/en/

[Sanner]
Sanner, Michel, and Sophie Coon. “Python for Structural Bioinformatics.” The Molecular Graphics Laboratory, The Scripps Research Institute, La Jolla, CA - http://www.scripps.edu/pub/olson-web/people/sanner/html/talks/PSB2001talk.html

[UBC, 2003]
University of British Columbia (UBC) - Training Program in Bioinformatics for Health Research, 2003 - http://www.genetics.ubc.ca

[UCSC, 2003]
Baskin School of Engineering, University of California, Santa Cruz, Bioinformatics Undergraduate program, 2003-
http://www.cse.ucsc.edu/programs/bioinformatics/undergraduate
http://www.cse.ucsc.edu/programs/bioinformatics/undergraduate/courses.html

[UMinn, 2003]
University of Minnesota – Summer Bioinformatics Institute, 2003 - http://www.bsi.umn.edu/education.html

[VirGO, 2003]
VirGO Project, 2003 - http://athena.bioc.uvic.ca/genomes

[Wolf, 2002]
Wolf, M. Wolf Demo – Multiprocessor Bio-array, 2002 - http://conferences.oreillynet.com/pub/w/21/track_enduser.html

(XII) APPENDICES

Appendix A Institutions

The list below shows some institutions that offer Bioinformatics programs at various levels.

Institution	Program(s)	Web Site
Boston University	Seminar Masters Ph.D.	http://bioinfo.bu.edu/index.shtml
Georgia Institute of Technology	Masters of Science in Bioinformatics Ph. D.	http://www.biology.gatech.edu/bioinformatics/index.html Ph.D. submitted for approval to Board of Regents. Approval is anticipated in the 2002-2003 academic year.
International Graduate School in Bioinformatics and Genome Research	Ph.D.	University of Bielefeld, North-Rhine-Westfalia, Germany http://www.cebitec.uni-bielefeld.de/IoB/index.html
Iowa State University	Ph.D., M.S.	http://www.bcb.iastate.edu
Johns Hopkins University and Kennedy Krieger Institute	Research	http://www.kennedykriger.org/kki/kki_2nd_inside.jsp?pid=9 http://pevsnerlab.kennedykrieger.org/bioinformatics/bioinf.htm
Medical College of Wisconsin	M.S. in Bioinformatics	http://brc.mcw.edu
National University of Australia	Research	http://biology.anu.edu.au
North Carolina State University	Research	http://genomics.ncsu.edu http://statgen.ncsu.edu/bioinformatics/index.html
Technical University of Denmark	Ph.D.	http://www.cbs.dtu.dk/index.html
University of Alberta, Canada	B.S.	http://www.biology.ualberta.ca/programs/undergraduate/index.php?Page=730
University of British Columbia Simon Fraser University British Columbia Cancer Agency	Diploma M.Sc. Ph.D.	http://www.genetics.ubc.ca
University of California at Los Angeles	B.S., M.S., Ph.D.	http://www.bioinformatics.ucla.edu
University of California, Santa Cruz	B.S.	http://www.cse.ucsc.edu/programs/bioinformatics/undergraduate
University of Michigan	Ph.D.	http://www.bioinformatics.med.umich.edu
University of Minnesota	Internship	http://www.bsi.umn.edu/program.html
University of Texas	Research Training	http://www.bioinformatics/uthscsa.edu/www.welcome.html
University of Waterloo (Ontario, CA)	BSCS with Bioinformatics Option	http://www.cs.uwaterloo.ca
Virginia Bioinformatics Institute	Ph.D.	http://www.vbi.vt.edu
Washington University in St. Louis	Research	http://www.genetics.wustl.edu

Appendix B Bioinformatics Courses

The table below lists some of the courses available at various institutions that offer Bioinformatics degrees. The Comment section includes a brief description of the course or some comment as to its content.

Course Name	Comment	Institution	Degree Program
Advanced Algorithms in Computational Biology	This is essentially an algorithm design course.	Iowa State	Ph.D. / M.S.
Advanced Biochemistry	This is one of the standard Biology courses.	Boston University	Ph.D. / M.S.
Advanced Bioinformatics	This course is more general than those at many of the U.S. institutions. The Technical University of Denmark had few courses compared to Iowa State or University of Michigan.	Technical University of Denmark	Ph.D. / M.S.
Advanced Cell Biology	This is a standard Biology course.	Boston University	Ph.D. / M.S.
Advanced Database Systems	This is a computer science course and demonstrates that U.M. thinks that advanced knowledge of database systems is important to Bioinformatics researchers.	University of Michigan	Ph.D.
Advanced Discrete Mathematics	This course covers selected topics in discrete mathematics. This course is important for algorithm analysis and for coding efficiency.	Boston University	Ph.D. / M.S.
Advanced Physical Chemistry of Biological Macromolecules	This is a standard Biology course.	Boston University	Ph.D. / M.S.
Advanced Topics in Genetic Modeling	This course is a hybrid of IT and Biology-related issues. The modeling tools include some Bioinformatics tools.	University of Michigan	Ph.D.
Advanced Topics in Physical Chemistry: Experimental Probes of Biomaterial and Biosensor Processes	This is a standard Biology course.	Boston University	Ph.D. / M.S.
An Introduction to Complex Systems	This course is from the EECS15 department. It shows that UM thinks the IT topics in this course are relevant to Ph.D.s.	University of Michigan	Ph.D.
Applied Biostatistics	This course involves Biological applications of Statistics. This course requires some computational techniques and would be required or available to Biology Ph.D.s as well as to the Bioinformaticists.	University of Michigan	Ph.D.
Applied Multivariate Analysis	This is an advanced statistics course and suggests that UM believes that Bioinformatics Ph.D.'s need this type of training.	University of Michigan	Ph.D.
Applied Statistics I & II	These statistics courses are from the Mathematics-Statistics department and reinforce the importance UM puts on statistical training and knowledge in Bioinformatics Ph.D.s.	University of Michigan	Ph.D.
Bioinformatics and Gene Expression	This course is specific to Bioinformatics although it would be available and important to Biologists and Biochemists as	University of Michigan	Ph.D.

[15] Electrical Engineering Computer Science

Course Name	Comment	Institution	Degree Program
	well.		
Bioinformatics Graduate Seminar	This is a discussion group that reviews current issues in Bioinformatics.	Boston University	Ph.D. / M.S.
Biological Database Systems	This course teaches the student about relational database models and techniques and then extends the discussion and research into some of the available online Bioinformatics databases.	Boston University	Ph.D. / M.S.
Biological Sequence Analysis	This is a technical course covering Sequence Analysis	Technical University of Denmark	Ph.D. / M.S.
Biomolecular Architecture	This is a standard Biology course.	Boston University	Ph.D.
Biophysical Chemistry I & II	These courses cover thermodynamics as it relates to analysis of DNA and other biochemistry topics.	University of Michigan	Ph.D.
Biostatistical Inference	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Biotechnology Laws and Ethics	This course deals with all the ethical and legal issues surrounding bio-technical research.	Boston University	Ph.D. / M.S.
Cell and Biomolecular Mechanics Laboratory	Please see university web site referenced in Appendix A for additional information.	Boston University	Ph.D. / M.S.
Cell Biology	This is a standard Biology course.	University of Michigan	Ph.D.
Cellular and Molecular System Analysis	Please see university web site referenced in Appendix A for additional information.	Boston University	Ph.D.
Cellular and Systems Neuroscience	Please see university web site referenced in Appendix A for additional information.	Boston University	Ph.D. / M.S.
Cellular Aspects of Development and Differentiation	Please see university web site referenced in Appendix A for additional information.	Boston University	Ph.D. / M.S.
Cellular Biotechnology	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Comparitive Microbial Genomics: Bioinformatics Approach	Please see university web site referenced in Appendix A for additional information.	Technical University of Denmark	Ph.D. / M.S.
Computational Chemistry and Biophysics	Please see university web site referenced in Appendix A for additional information.	Boston University	Ph.D. / M.S.
Computational Genetics and Evolution	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Computational Genomics	This course applies what is taught in the Sequence Analysis and Database Systems courses.	Boston University	Ph.D. / M.S.
Computational Models of Learning	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Computational Molecular Biology	Hands-on introduction to computer based molecular biology. This course has computer science and statistics prerequisites.	Iowa State	Ph.D. / M.S.
Computational Techniques for Genome Assembly and	The course description for this course is similar to that of the Boston University Computational Genomics course.	Iowa State	Ph.D. / M.S.

Course Name	Comment	Institution	Degree Program
Analysis			
Computations in Probabilistic Modeling in Bioinformatics	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Computer Modeling of Complex Systems	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
COOP in Bioinformatics (M.S. only)	This course indicates that Boston U. offers local field opportunities for its Ph.D. students.	Boston University	M.S.
Critical Discussion in Bioinformatics	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Current Topics in Databases	This is an EECS course made available to give the Bioinformatics doctoral candidate a current background in database.	University of Michigan	Ph.D.
Database Management Systems	This is an EECS course made available to give the Bioinformatics doctoral candidate a current background in database.	University of Michigan	Ph.D.
Design and Analysis of Algorithms	This is a computer science course offered to give ISU candidates a background in algorithm design.	Iowa State	Ph.D. / M.S.
Directed Study in Bioinformatics	This is an independent research course.	Boston University	Ph.D. / M.S.
Discrete Stochastic Models	This course teaches some of the statistical analysis techniques used in applied Bioinformatics research including Markov chains, Chapman-Kolmogorov equation. Classification of states, limiting probabilities, Poisson process and its generalization, continuous-time Markov chains, queuing theory, reliability.	Boston University	Ph.D. / M.S.
DNA Microarray Bioinformatics	This course trains the Ph.D. candidate in DNA Microarray analysis techniques.	Technical University of Denmark	Ph.D. / M.S.
DNA Structure and Function	This is a molecular biology course that would as likely be taken by a Biology Ph.D. as a candidate for Bioinformatics.	Boston University	Ph.D.
Elements of Neural Computation	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Evolutionary Genetics	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Evolutionary Problems for Computational Biologists	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Exploratory Data Analysis (offered irregularly)	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Fundamental Algorithms in Computational Biology	This course is similar to Boston U.'s Information – Theoretical Design of Algorithms course.	Iowa State	Ph.D. / M.S.
Gene Structure Regulation	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
General Biochemistry I & II	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Genetic Statistics	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.

Course Name	Comment	Institution	Degree Program
Genomic Data Processing	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Graduate Biochemistry	Please see university web site referenced in Appendix A for additional information.	Boston University	Ph.D. / M.S.
Graphs and Networks	This is a CS course and shows that ISU thinks that its Ph.D.'s should have the opportunity to learn about graph theory and network analysis.	Iowa State	Ph.D. / M.S.
Immunological Bioinformatics	This course focuses on bioinformatical research on viruses and other biological entities that affect the human immune system.	Technical University of Denmark	Ph.D. / M.S.
Information Architecture	This is an EECS course.	University of Michigan	Ph.D.
Information Theory and Coding	Introduction to information theory; entropy and information; discrete sources and much more.	Boston University	Ph.D. / M.S.
Information-Theoretical Design of Algorithms	This course involves a theoretical approach to the analysis and design of computer algorithms.	Boston University	Ph.D. / M.S.
Integrative Genomics	This is a Bioinformatics course that would be available and of interest to a geneticist.	University of Michigan	Ph.D.
Intro to parallel Algorithms and Programming	This is a CS course that can help a Bioinformatics Ph.D. determine evaluate the best tools and techniques to use in analysis.	Iowa State	Ph.D. / M.S.
Introduction to Algorithms	This is an EECS course that can help the Ph.D. candidate with algorithm design.	University of Michigan	Ph.D.
Introduction to Bioinformatics	This course gives the Ph.D. candidates the basic overview of the field, tools and techniques.	Technical University of Denmark	Ph.D. / M.S.
Introduction to Bioinformatics	This course gives the Ph.D. candidates the basic overview of the field, tools and techniques.	University of Michigan	Ph.D.
Introduction to Biostatistics	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Introduction to Database Management Systems	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Introduction to Database Systems	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Introduction to Dynamical Systems for Biocomplexity	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Introduction to Molecular Biology Techniques	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Introduction to Probability	This is from the mathematics and statistics department. This course can help make the Ph.D. candidates conversant with probability rules, techniques and methods.	University of Michigan	Ph.D.
Introduction to Probability and Statistics	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Introduction to Protein Structure & Function	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Introduction to Theoretical	Please see university web site referenced in Appendix A for	University of	Ph.D.

Course Name	Comment	Institution	Degree Program
Statistics	additional information.	Michigan	
Introductory Biochemistry	This is a standard Biology and Chemistry course.	University of Michigan	Ph.D.
Introductory Biochemistry Lab	This is a standard Biology and Chemistry course.	University of Michigan	Ph.D.
Introductory Combinatorics	Stat/Math	Iowa State	Ph.D. / M.S.
Introductory Computational Structural Biology	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Journal Club	This is a discussion course.	University of Michigan	Ph.D.
Knowledge-Based Systems (Even years)	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Machine Learning	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Masters Project	This is the course a Boston U. masters candidate must complete in order to attain the Master of Science degree.	Boston University	M.S.
Mathematical Models of Infectious Diseases (Even years)	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Membrane Biochemistry	Please see university web site referenced in Appendix A for additional information.	Boston University	Ph.D. / M.S.
Methods of Computational Chemistry	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Modeling Discrete State Stochastic Processes	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Molecular Biology	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Molecular Biology I & II	Research, application and discussion of molecular biological techniques, including genetics and recombinant DNA techniques.	Boston University	Ph.D. / M.S.
Molecular Biophysics	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Molecular Evolution	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Molecular Genetics	This course covers the principles of molecular genetics: "gene structure and function at the molecular level, including regulation of gene expression, genetic rearrangement, and the organization of genetic information in prokaryotes and eukaryotes."	Iowa State	Ph.D. / M.S.
Molecular Phylogenetics	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Molecular Physical Chemistry	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Molecular Systematics and Evolution	Please see university web site referenced in Appendix A for additional information.	Boston University	Ph.D. / M.S.
Molecular, Cellular and	Please see university web site referenced in Appendix A for	University of	Ph.D.

Course Name	Comment	Institution	Degree Program
Population Genetics	additional information.	Michigan	
Natural Language Processing	This is an EECS course that can provide the Ph.D. candidate with knowledge of Natural Language processing. This is relevant to Bioinformatics because of the preponderance of text strings in the electronic rendering of DNA and other biological entities.	University of Michigan	Ph.D.
Numerical Methods for Scientific Computing	This is an EECS course that can train the Ph.D. candidate to apply various numerical methods to problems in general and then extend them to the types of problems encountered in Bioinformatics research.	University of Michigan	Ph.D.
Object-Oriented Software Development	This is another EECS course and suggests that UM wants to equip its Ph.D's with the skills to design and build tools.	University of Michigan	Ph.D.
Optimization and Modeling with Artificial Life	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Perl and Unix for Bioinformaticians	Please see university web site referenced in Appendix A for additional information.	Technical University of Denmark	Ph.D. / M.S.
Physical BioChemistry	Train the researcher to "identify the physical events which underlie modern biomolecular research" rather than to teach the physical methods of research.	Boston University	Ph.D. / M.S.
Physical Biochemistry	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Physical Chemistry I	Please see university web site referenced in Appendix A for additional information.	Boston University	Ph.D. / M.S.
Physical Chemistry of Macromolecules	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Principles of Artificial Intelligence	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Principles of Database Systems	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Probabilistic Modeling in Bioinformatics	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Probability and Distribution Theory	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Programming I (Java)	This is an introductory programming course taught in Java.	University of Michigan	Ph.D.
Protein and DNA Sequence Analysis	This course teaches the researcher to execute some of the techniques described in section VI above.	Boston University	Ph.D. / M.S.
Protein Chemistry - Physical Methods	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Protein Structure	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Protein Structure and Function	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Recombinant DNA Laboratory	Please see university web site referenced in Appendix A for additional information.	Boston University	Ph.D. / M.S.

Course Name	Comment	Institution	Degree Program
Representation and Organization of Information Resrources	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Research in Bioinformatics	Please see university web site referenced in Appendix A for additional information.	Boston University	Ph.D. / M.S.
Search and Retrieval	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Special Topics - Practical Programming Concepts	This is another EECS course and suggests that UM wants to equip its Ph.D's with the skills to design and build tools.	University of Michigan	Ph.D.
Special Topics in Theoretical Statistics I	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Special Topics: Database Application Design	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Statistical Computing	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Statistical Mechanics	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Statistical Methods	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Statistical Models & Methods in Human Genetics	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Statistics for Molecular Genetics	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Structural Bioinformatics	Please see university web site referenced in Appendix A for additional information.	Boston University	Ph.D. / M.S.
Structure and Reactivity of Biomolecules	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Subcellular BioMechanics: I. Membranes and Interfaces	"The physical and chemical basis for the mechanical properties and activities of living cells will be considered from an engineering perspective."	Boston University	Ph.D. / M.S.
Theoretical Computer Science II: Foundations of Computer Security	Please see university web site referenced in Appendix A for additional information.	University of Michigan	Ph.D.
Theory of Probability and Statistics I & II	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.
Thesis/Research (Post-candidacy credit)	Please see university web site referenced in Appendix A for additional information.	Boston University	
User Interface Design and Analysis	This is another EECS course and suggests that UM wants to equip its Ph.D's with the skills to design and build tools.	University of Michigan	Ph.D.
Workshop in Bioinformatics and Computational Biology	Please see university web site referenced in Appendix A for additional information.	Iowa State	Ph.D. / M.S.

Appendix C – Amino Acids' 1- and 3-character codes

#	Name	1-character code	3 character code
1	Alanine	A	Ala
2	Arginine	R	Arg
3	Asparagine	N	Asn
4	Aspartic acid	D	Asp
5	Cysteine	C	Cys
6	Glutamine	Q	Gln
7	Glutamic acid	E	Glu
8	Glycine	G	Gly
9	Histidine	H	His
10	Isoleucine	I	Ile
11	Leucine	L	Leu
12	Lysine	K	Lys
13	Methionine	M	Met
14	Phenylalanine	F	Phe
15	Proline	P	Pro
16	Serine	S	Ser
17	Threonine	T	Thr
18	Tryptophan	W	Trp
19	Tyrosine	Y	Tyr
20	Valine	V	Val

Appendix D – Biological Databases

Type	**Data**	Database	**Note/Link**
Info	Biomedical literature	PubMed	http://www3.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed
Sequence	Nucleotide Sequence	GenBank	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide
		SRS	http://srs.ebi.ac.uk/ European Molecular Biology Lab/European Bioinformatics Institute
	Protein Sequence	SWISS-PROT	Swiss Institute for Bioinformatics and European Bioinformatics Institute http://www.expasy.ch/sprot/sprot-top.html
	Genome sequence	TIGR	http://www.tigr.org
3D	Protein structure	PDB	Protein Data Bank http://www.rcsb.org/pdb/
Organisms	Human	UCSC database	http://genome.ucsc.edu/
		GenBank	http://www.ncbi.nlm.nih.gov/genome/guide/human/
	Biomedical Pathway	KEGG	http://www.genome.ad.jp/kegg/ Kyoto Encyclopedia of Genes and Genomes
		WITT	http://wit.mcs.anl.gov/WIT2/

Appendix E – Physics

University Physics [13]
Page 829 "Fig 37-5 A chart of the electromagnetic spectrum (to include visible light)"

Page 537 Example 1. " An alpha particle is a nucleus of doubly ionized helium. It has a mass m of 6.68 X 10^{-27}kgm and a charge q of +2e or 3.2 x 10^{-9} coul.
Compare the force of electrostatic repulsion between two alpha particles with the force of gravitational attraction between them.

The electrostatic force F_e is $Fe = k\ q^2/r^2$

And the gravitational force Fg is $F_g = G\ m^2/r^2$

The ratio of the electrostatic to the gravitational force is

$$\frac{F_e}{F_g} = \frac{k\ q^2}{G\ m^2} = 3.1 \times 10^{+35}$$

The gravitational force is negligible compared with the electrostatic force."

Appendix F – Genetic Code Table

First base	Second base: U	Second base: C	Second base: A	Second base: G
U	UUU, UUC: Phe UUA, UUG: Leu	UCU, UCC, UCA, UCG: Ser	UAU, UAC: Tyr UAA, UAG: TERM	UGU, UGC: Cys UGA: TERM UGG: Trp
C	CUU, CUC, CUA, CUG: Leu	CCU, CCC, CCA, CCG: Pro	AAU, CAC: His CAA, CAG: Gln	CGU, CGC, CGA, CGG: Arg
A	AUU, AUC, AUA: Ile AUG: Met	ACU, GCC, ACA, ACG: Thr	AAU, AAC: Asn AAA, AAG: Lys	AGU, AGC: Ser AGA, AGG: Arg
G	GUU, GUC, GUA, GUG: Val	GCU, GCC, GCA, GCG: Ala	GAU, GAC: Asp GAA, GAG: Glu	GGU, GGC, GGA, GGG: Gly

Figure 2 - Genetic Code Table [3]

Above you can see the sixty-four (64) combination of codons found in DNA used to cast the twenty amino Acids? The count is not important, how the codons are distributed is important. We find that three are used for "TERM" codons or stop (termination) codons used to stop and start the film of the protein, what part of the DNA, the gene, do we want to cast. They are the director of the "scene". So that leaves sixty-one (61) nucleotides. It turns out that the energy signature, the pencil shadow, of the first two nucleotides are the predominate energy signatures needed to define an amino acid type for thirty-two (32) of the codon combinations. That is, half of the sixty-four (64) combinations result in creating eight of the amino acids' types. The "pencil shadow" or energy signature of the first two nucleotides so overwhelms the shadow of the last that the last nucleotide's contribution can be ignored.

So far we've accounted for 35 of our 64 combinations, leaving twenty-nine (29) left. We find that the energy geometry, the signature, of all three nucleotides taken together only uniquely defines three combinations, one of which, TERM has already been counted. The other two result in amino acids.

Having accounted for 37 combinations, twenty-seven (27) remain. Twenty-four (24) combinations map to twelve (12) amino acids (plus an additional two for a TERM already counted) these share the first two nucleotides with one other TERM or amino acids. That is to say the first two shadows are predominate. However, the tiebreaker ends up being the third shadow signature.

That leaves three (3) amino acids to be defined. These have the first two amino acids in common and the third not equal to the third amino acid of Met. That means aside from making sure it is not Met's third amino acid, we don't particularly care which one it is.

What is the significance of the third nucleotides? At least half the time, they are irrelevant in determining which amino acid we have. This information would be important in design of an algorithm to read amino acids to define nucleotides.

Appendix G – RNA Image

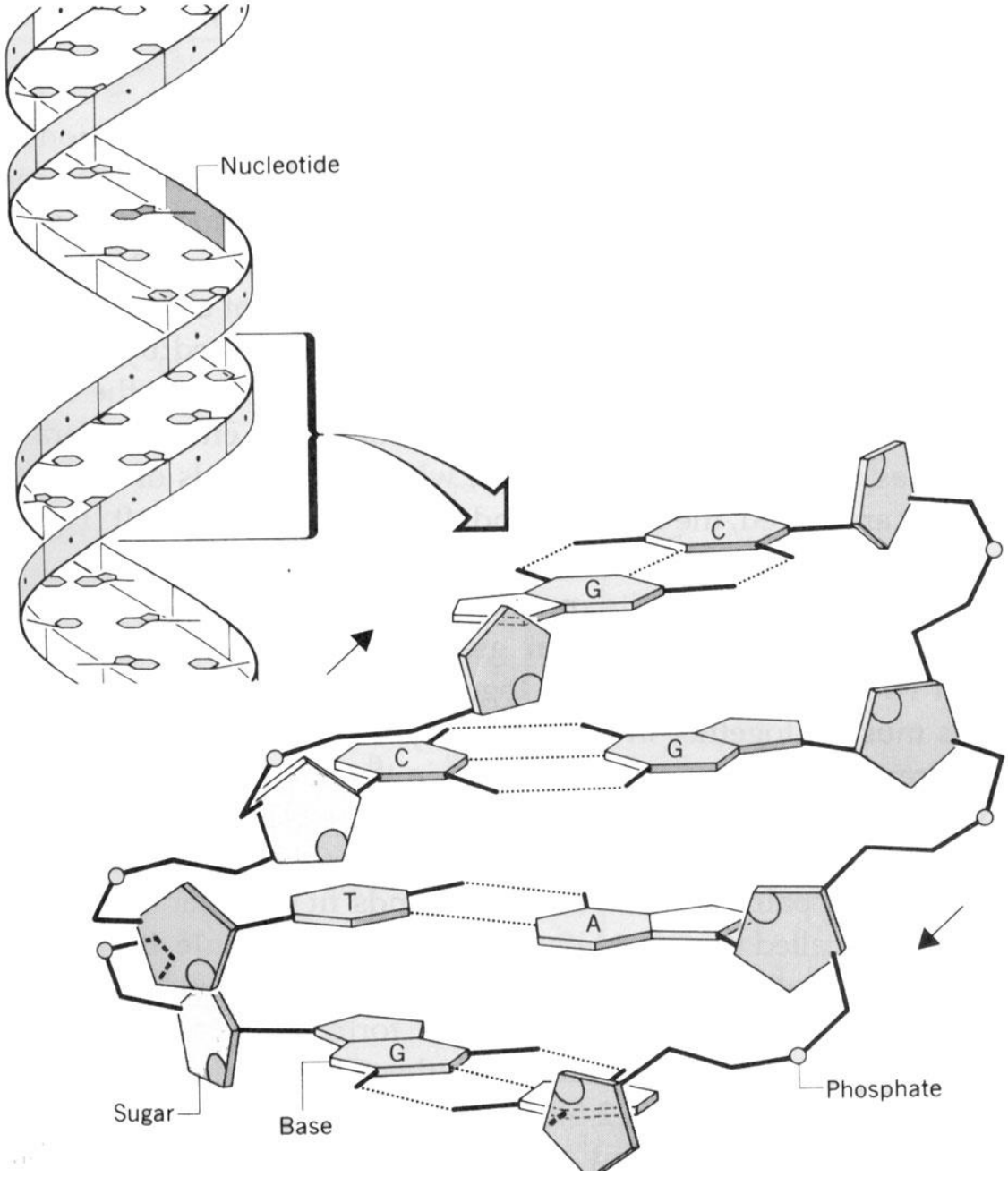

Figure 3 - [From Figure 2-3: Drlica, 1997]