BIS: Statistics, Computer Science, Life Sciences

Life Science has been something that I have been interested since I was young. I became interested in research, and in my sophomore year I secured a research position in a genetics lab. I was interested in the conclusions that could come from this research, but found myself frustrated by the slow pace that the experiments took to pan out. I was spending weeks on experiments to potentially learn something about one gene out of the thousands that a *c. elegans* worm has. Surely there must be a way to do these experiments more quickly?

My question was answered when a professor mentioned the emerging field of bioinformatics during a lecture. My interest was piqued, and I went home after class and did all the research I could on the field. I found that Bioinformatics was the biology-focused subdomain of the recent data science and big data movement. When the Human Genome Project began, it signaled the beginning of a data problem for genetics. As the cost of genome sequencing decreased and more techniques became available to sequence RNA and protein being expressed in tissue or even cells, this problem grew. In talking to faculty here at the U that work with this data, I found that analyzing this data by hand is at best tedious, and at worst impossible. In many cases, a genetics scientist enlists a data science professional who would come in to write code to find relationships that the geneticist could base a hypothesis on. Unfortunately, the data scientist often lacks the genetics domain knowledge needed to know exactly what he or she should be looking for in their results.

I began to realize that having an understanding of all three of these fields would be of huge value if I wanted to go into this field. The ability to simultaneously understand what may be significant in a genetics experiment, think critically about statistical methods are used, and understanding the limitations of computers and how to harness their power would be a huge benefit for my future career. I originally sought a genetics degree with a small amount of computer science and statistics taken outside of the major. After hearing about the B.I.S. degree from my advisor however, I realized that a B.I.S. degree would be the perfect way to formalize
what I am studying and to gain not just a serviceable understanding of genetics, computer science, and statistics, but an expertise in each of these areas.

However, completing this degree was not without challenges on the way. In Spring 2012, taking six classes that included difficult classes like Organic Chemistry and Biochemistry lead to a frustrating semester and disappointing result at the end of the semester. After that difficult period, I rededicated myself to my studies and my future, and I am happy to report that all subsequent semesters have been very successful.

**Concentration One: Life Sciences**

Although the majority of the technical skills that I will be using in the workplace are in my Statistics and Computer Science, my academic interests are in the area of life sciences, and more specifically cancer genetics. Informatics are a good fit with cancer studies, and I have been involved in projects focusing on Ductal Carcinomas in Situ, Glioblastomas, and Acute Myeloid Leukemia, that have only reinforced my passion for studying cancers. When I was choosing my classes in this concentration, I focused on subjects where informatics analyses are commonly used in research settings. This results in a Life Sciences concentration that could be broadly defined as Genetics, and will result in a comprehensive understanding of genetics, cell biology, how the condition of cancer affects both of those subjects. In addition, I included health informatics to learn how hospitals use informatics to inform treatment.

*(BIOC 3021) - Biochemistry*

Biochemistry contains many of the knowledge necessary to understand cellular biology. After taking this class I knew how DNA, RNA, and proteins worked at a chemical level. Going beyond the conceptual level of genetics and into chemistry showed me connections between chemical structure and function that I would not have seen before.
(GCD 3022) - Genetics

Although I had a solid base of genetics before I took this class, the materials that were covered both strengthened and augmented my previous knowledge. The sections on building genetic maps, gene transfer and using statistics to find the probability of evolution within a population were the highlights of this class for me.

(GCD 3485) - Bioinformatic Analysis: Introduction to the Computational Characterization of Genes and Proteins

This class was my first real foray into Bioinformatics, and I wasn’t disappointed. The work in this class centered on a gene of our choosing that we researched for an entire semester. I loved gathering evidence about where my gene’s protein product would reside in the cell, and what function it would have. I used a wide variety of bioinformatic tools, and learned how to deal with conflicting evidence about our gene. This class gave me an expert understanding in bioinformatics, and if I had to choose one class as my favorite, this would be it.

(HINF 5430) - Health Informatics I

An important part of informatics is the organization and visualization of information. These elements are some of the most important problems in health informatics today, and this class taught me a great deal about each of these elements. After taking this class, I felt much more conscious of how data was organized, and I learned a lot about the thought process of a healthcare professional, which will be crucial for interfacing with them in the future.

(GCD 4141) - Molecular Biology of Cancer

The principles of molecular biology are turned on their head when cells become cancerous. Myriad changes occur in the genetic profile of the cell, and being aware of the molecular differences between a healthy cell and a cancerous cell is very important when studying different cancers.
(HINF 5531) - **Health Data Analytics and Data Science** (If I cannot enroll in GCD 4141)

While me taking HINF 5531 is contingent on me taking GCD 4141, it would still be a valuable course for me to take. This class combines elements all three of my concentration areas. The different subjects I would learn about are: methods of data collection in healthcare, critical evaluation of healthcare data, and data presentation. Difficulties with healthcare data is a problem that I have encountered during research, so this class would be very beneficial to me going forward.

**Concentration Two: Statistics**

An understanding of statistics is essential to any bioinformatic analysis. Statistics is what allows me to make sense of the data I am analyzing, and form conclusions. The statistical skills that I have learned and will learn are going to be essential in evaluating the piles of data that are available to study.

My Statistics concentration is focused on core statistics and probability, in addition to modeling skills. My goal was to be able to arrive at statistical conclusions with a solid explanation of why I did a certain analysis, and what caveats go along with that analysis. With that in mind, I chose classes that would give me a strong foundation in basic statistics and probability, and a strong understanding of modeling as well.

(STAT 3021) - **Intro to Probability and Statistics**

Intro to Probability and Statistics is all about learning the basics. I now know the basics of statistics and probability inside and out. Small but important details were also covered thoroughly. Internalizing these basic concepts eased my transition into the more difficult statistics classes, and has also made me feel more confident in applying them to real-life problems.
(STAT 3022) - **Data Analysis**

Data Analysis builds upon the foundation of STAT 3021, and explains where these methods should and should not be applied. My biggest takeaways from this class are the variety of assumptions that come with each statistical technique that you must always be conscious of, and that it is important not to simply accept a result without further analysis of how you arrived at that result.

(STAT 5201) – **Sampling Methodology in Finite Populations**

In Bioinformatics problems, there are often few subjects to draw from. For example, there often few cell lines that can be studied in rare forms of cancer, or a limited number of sequenced genomes for a species. The question that often arises is “What conclusions about this small group can we apply to the larger population?” This class will illuminate strategies to answer this question.

(STAT 4101) - **Theory of Statistics I**

In Statistics, it is very important to have a solid base in the mathematics of statistics. In comparison to STAT 3021, this class is more focused on mathematics, rather than using computer packages. The ability to use calculus and statistics together will further develop my understanding of the theories behind statistics.

**Concentration Three: Computer Science**

Computer Science is the mechanism that organizes the data and implements the analyses that I want to perform. Although bioinformatics is often accomplished through collaboration, knowing the strong suits and weak points of computers and different computing languages is very important when deciding how to attack a problem. For my Computer Science focus area, I wanted to focus on developing my ability to write fast and readable algorithms and to learn about data structures and data storage as well. This concentration could be summed up as back end development.
(CSCI 3003) - Introduction to Computing in Biology

This computer science class was tailored for the study of genetics, centered on the computing languages that could rapidly process DNA or Protein sequence data. This was my first computer science class, and it showed me how different computing languages are designed with different purposes in mind. Additionally, I learned the perl and Matlab languages, which I didn’t encounter in my other computer science concentration classes, but both are useful.

(CSCI 3081W) - Program Design and Development

To this point, most computer science that I have learned has been used to execute small, trivial problems, where a TA and myself were the only users. Program Design and Development will help me to understand the common mistakes and methods to use when developing large, multifaceted programs designed for production. This will help me to refine my process when building a program and will result in me becoming a more efficient and thoughtful worker.

(CSCI 4041) - Algorithms and Data Structures

Although the speed of computers is increasing at an extremely rapid rate, writing a bad algorithm for a large problem can still waste a huge amount of time. Evaluating and optimizing your algorithms is essential to avoid inefficiency. This class has given me a mathematical and logic based method of critique with which I can evaluate my algorithms.

(CSCI 4707) - Practice of Database Systems

Once completed, this class will teach me how a database works at a software level. I will learn what features are key to a relational database, and how to select what features to include if I were to design a database for a company. Not knowing the features and drawbacks of using a relational database and how they are structured was a problem I encountered during my distributed systems internship this summer. I am looking forward to learning all about this subject.
Data Mining is a field that will teach me general data classification and the mathematics behind evaluating the similarities and differences between different groups. By taking this class, I will gain insight into the cutting edge of general data analysis, which will serve me well on a variety of different problems.

Conclusion

The fields of bioinformatics, and its parent industry data science, are exploding. I believe my B.I.S. degree will not only help me to find work in this industry, but I will thrive due to the wide variety of subjects that I was able to tackle by pursuing this degree. The process of learning how these three disciplines work together has illuminated many different exciting new problems to tackle, and I cannot wait to start working on them.

This summer, I saw my work pay off when the Bioinformatics division of the chemical company Sigma-Aldrich hired me for an internship. My internship required me to use my computer science & statistics skills to work with professionals in the field of chemistry. I found that the knowledge and experience I have gained at the University of Minnesota was extremely effective and in demand in the workplace. In addition, the feedback I have received from prospective employers about my major and the subjects that I have been studying has also been very positive. I am very grateful for the unique opportunity that a B.I.S. degree has given me to pursue my interests.
**BIS/IDIM Program Course Worksheet**

**Total number of credits in proposed program:** 51

**Total number 3/4/5xxx level credits in proposed program:**

**Total credits in program left to complete:**

**IDIM Title:**

### Concentration One

**Title:** Life Sciences

<table>
<thead>
<tr>
<th>Dept.</th>
<th>Course Code</th>
<th>Course Title</th>
<th>Complete: Credits (Grade)</th>
<th>Not Complete: Credits (Term)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIOC</td>
<td>3021</td>
<td>Biochemistry</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>GCD</td>
<td>3022</td>
<td>Genetics</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>GCD</td>
<td>3485</td>
<td>Bioinformatic Analysis</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>HINF</td>
<td>5430</td>
<td>Health Informatics I</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>GCD</td>
<td>4141</td>
<td>Molecular Biology of Cancer</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>EEB</td>
<td>3409</td>
<td>Evolution</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

**3/4/5xxx level credits in area:** 15

**Total credits in area (IDIM only):**

### Concentration Two

**Title:** Statistics

<table>
<thead>
<tr>
<th>Dept.</th>
<th>Course Code</th>
<th>Course Title</th>
<th>Complete: Credits (Grade)</th>
<th>Not Complete: Credits (Term)</th>
</tr>
</thead>
<tbody>
<tr>
<td>STAT</td>
<td>3021</td>
<td>Intro to Probability and Statistics</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>STAT</td>
<td>3022</td>
<td>Data Analysis</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>STAT</td>
<td>5302</td>
<td>Applied Regression Analysis</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>STAT</td>
<td>5201</td>
<td>Sampling Methodology in Finite Populations</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>STAT</td>
<td>4101</td>
<td>Theory of Statistics</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

**3/4/5xxx level credits in area:** 18

**Total credits in area (IDIM only):**

### Concentration Three

**Title:** Computer Science

<table>
<thead>
<tr>
<th>Dept.</th>
<th>Course Code</th>
<th>Course Title</th>
<th>Complete: Credits (Grade)</th>
<th>Not Complete: Credits (Term)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSCI</td>
<td>3003</td>
<td>Computing in Biology</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>CSCI</td>
<td>3081W</td>
<td>Program Design and Development</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>CSCI</td>
<td>4041</td>
<td>Algorithms and Data Structures</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>CSCI</td>
<td>4707</td>
<td>Practice of Database Systems</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>CSCI</td>
<td>5523</td>
<td>Introduction to Data Mining</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

**3/4/5xxx level credits in area:** 18

**Total credits in area (IDIM only):**