Inquiry Journal 2007 Inquiry Journal

Spring 2007

Bioinformatics: Merging Computer Science and Genetics

Lina Faller
University of New Hampshire - Main Campus

INQUIRY journal

UNDERGRADUATE RESEARCH JOURNAL SPRING 2007

research ARTICLE

Bioinformatics: Merging Computer Science and Genetics

—**Lina Faller** (Edited by Ashley Ward)

In the 1850's the ambitious scientist, Gregor Mendel, studied thousands of pea plants and their offspring. Collecting and analyzing data, such as the number and texture of seeds and the colors of the flowers, led him to define what scientists today call Mendel's Law of Inheritance. In Mendel's day data was collected by hand, and the findings were analyzed manually. Today, these tasks and many more, unimaginable in Mendel's time, can be accomplished by computers: Welcome to bioinformatics, an emerging interdisciplinary field that applies computer science skills to the field of life sciences!

Combining different fields, such as computer science and genetics, and acquiring skills from both disciplines can be a challenge. However, new technologies to support biological research are continuously being developed, and the field of life sciences calls for researchers with expertise in many disciplines. One of these sciences, comparative genomics, studies the relationships between species by analyzing their genomes for similarities and differences. By evaluating these factors, scientists may answer questions about concepts such as genome evolution and function. This, however, involves analyzing a vast amount of data.

As a computer science major, I wanted to learn more about the uses of bioinformatics outside the classroom and also get some hands-on experience of bioinformatics in the sciences. So I applied for and received a 2006 Summer Undergraduate Research Fellowship from the University of New Hampshire to investigate the evolution of microsatellites in fungal and bacterial genomes. To understand what I did, a little background is needed first.

The Genome and DNA

Every organism on earth has its own unique genome, which is the collection of genes that, in conjunction with the environment, ultimately determine such factors as its physical appearance or susceptibility to certain diseases. When a biologist wants to study an organism‘s genes, the usual first step is to collect and analyze the molecules that contains the genes called Deoxyribonucleic Acid, or DNA.

All cells in our body contain DNA molecules that carry our genetic heritage. Every DNA molecule is made up of smaller building blocks called *nucleotides*, or *bases*. Only four different kinds of bases are found: Adenine, Thymine, Guanine, and Cytosine, which are commonly shortened to A, T, G and C. A DNA molecule can be visualized as a long string of four different-colored beads. In the cell, two DNA strands tend to combine to form a molecule with a double helix shape. Due to chemical properties of the bases, Adenine molecules pair with Thymine molecules, whereas a Guanine molecule pairs with a Cytosine molecule. These molecules “stick” together due to chemical forces acting between them and thus form the helix. (Fig. 1)

The sequence of the nucleotides (bases) acts as a code that contains all the instructions needed for a cell to function and reproduce. As the whole DNA sequence is long, scientists divide it into shorter, functional regions of sequence, called genes. Genes contain the instructions for creating proteins, which are essential for our cells to operate.

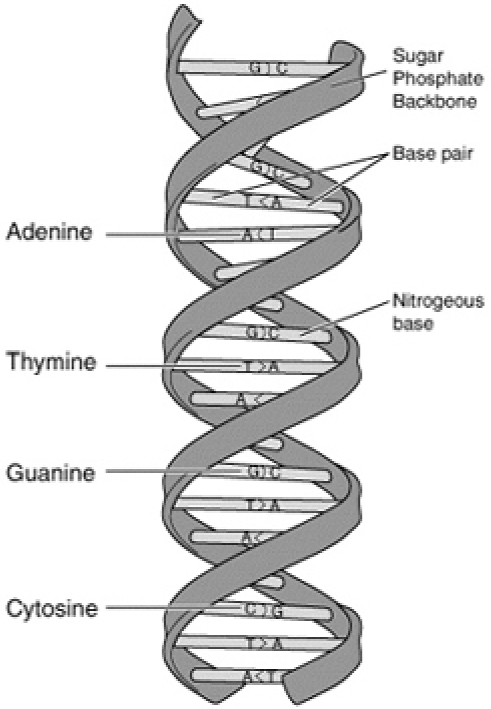


Fig. 1: A diagram of the DNA double helix illustrating the pairing of Adenine on one strand bonding with Thymine on the other, and in the same way Guanine with Cytosine

What are Microsatellites and Why Study Them?

In most organisms, from bacteria to humans, not all bases in a DNA molecule have obvious functions such as encoding the information for a protein. Depending on the organism, there might well be a very large fraction of the genome that does not appear to serve any purpose at all. Scientists call these parts *non-coding* DNA. However, it is in these sequences that one can often observe repeating patterns of different bases called *satellite* DNA. If a stretch of DNA is made up of very short, repeated sequences, it is called a *microsatellite*. These sequence patterns occur most often in the non-functional sections of DNA; but scientists have linked some of these sequences, mainly with three-base repeating patterns, to genetic diseases such as Huntington's (“Trinucleotide Diseases,” Stanford University). These findings suggest that non-coding DNA is more functional than previously assumed. While scientists are still unsure about the exact functions of microsatellites, they have discovered useful applications for them. For example, the composition of these sequences changes so fast that each person has a unique set of them. Due to this feature, satellite sequences are the basis of DNA fingerprinting (Jeffreys, Wilson, Thein, 1985); and today microsatellites are often used as genetic markers in forensics and paternity testing.

My summer research project focused on how microsatellites appear in different fungal and bacterial species. I investigated how repeated sequences differ in number, length, and base composition. The genomes of bacteria and fungi provide ideal opportunities for studying microsatellites because data for many closely related genomes of individual species is available. If we can better understand the small genomes of closely related organisms, we might one day be able to explain larger, more complex genomes.

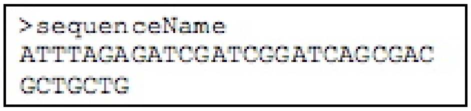
Bioinformatics

I needed first to decide on a computer language to use in the computational aspects of this project. Programming (or scripting) languages come in many varieties, and while most tasks can be accomplished with many different languages, there are often advantages to using one language over another. This project, for example, required manipulating text and searching for patterns in large text files. I decided on the “Practical Extraction and Report Language” (Perl), mainly because this programming language offers convenient tools to operate on text files. In addition, I had previous experience with Perl, so I could easily brush up my skills.

Next, I needed to obtain the data I wanted to analyze. Genetic information for many species is available in the public domain on the Internet. Scientists can publish their data on web sites set up by organizations like the National Center for Biotechnology Information (NCBI), which encourage global research collaboration. These organizations organize data in large databases, provide access to literature and papers on the research subjects, and develop software tools to access and analyze this data.

Research Methods and Results

After I decided on the species of fungi and bacteria I wanted to investigate, I searched the Internet for data on them. I was looking specifically for complete sequences of bases for each of the species. Usually, this data can be obtained as plain text files, so-called *fasta* files, where a DNA sequence is represented by a stretch of letters (A, T, G, or C, depending on the base).

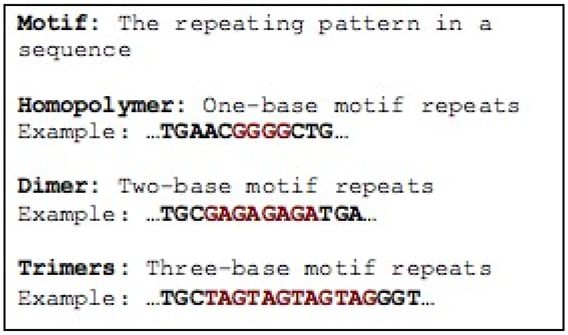


```
>sequenceName
ATTTAGAGATCGATCGGATCAGCGAC
GCTGCTG
```

Fig. 2: An example of a fasta file; the sequence name is preceded by > and the next line contains the nucleotide sequence

A fasta file is a simple text file whose contents follow a certain pattern. In every fasta file, the first line begins with an angle bracket (>), followed by a sequence name or identifier (Pearson, Lipman). (Fig. 2)

As my mentor, Dr. Thomas of the Hubbard Center for Genome Studies (HCGS) at UNH, explained to me: fasta files can be compared to rather boring textbooks without graphs or figures, where the text is written in an alphabet consisting of four letters and is without any kind of punctuation. The challenge of bioinformatics is to learn how to “read” this textbook and make sense of the information it contains.



Motif: The repeating pattern in a sequence

Homopolymer: One-base motif repeats
Example: ...TGAACGGGGCTG...

Dimer: Two-base motif repeats
Example: ...TGCGAGAGAGATGA...

Trimers: Three-base motif repeats
Example: ...TGCTAGTAGTAGTAGGGT...

Fig. 3: Microsatellite terminology

To identify the specific microsatellite patterns in fungal and bacterial genomes, I used a Perl program written by Way Sung, a graduate student in the HCGS. Sung's program analyzes a fasta file by searching for microsatellites. The user can specify how long the repeating pattern, or *motif*, should be and how many times it is repeated in a microsatellite. For a bacterial genome, for example, I looked for microsatellites whose motifs ranged from one base repeated at least four times to motifs of one hundred bases. (Fig. 3)

To facilitate comparisons of microsatellite data across several species (the comparative genomics aspect of this project), I devised a system to catalogue and organize the information into one file. I analyzed the microsatellites and ordered them first by length and then alphabetically for each species. I then compiled the data of all species into a large text file that can be opened by Microsoft's Excel program.

Comparing the microsatellite data across several species I discovered results that were to be expected, due to the close relationships among the species studied, as well as some unexpected insights on microsatellite abundance and distribution in select genomes.

When examining C- and G-homopolymers (repetitive strings of Cytosine and Guanine bases) in distinct bacterial species, I noticed a bias toward G-homopolymers on one strand of the genome and toward C on the other. Investigating these observations more closely, I realized that most of the completed genomes I studied, that is, the genomes whose DNA sequence is fully determined, showed such a bias (see genomes II-V in Figure 4). A Chi-square analysis revealed that these differences are statistically significant. However, there are also genomes that do not show a statistically significant strand bias (see genomes VI-VII, for example). If we assume microsatellites to be randomly distributed sequences, we would also assume them to be randomly distributed in the genome and on the two strands. No one knows why this asymmetry occurs in these bacteria, but my data suggests that this is not a random feature of microsatellites. Furthermore, I also observed this phenomenon when analyzing dimers and trimers (two- and three-nucleotide repeating patterns) in certain genomes.

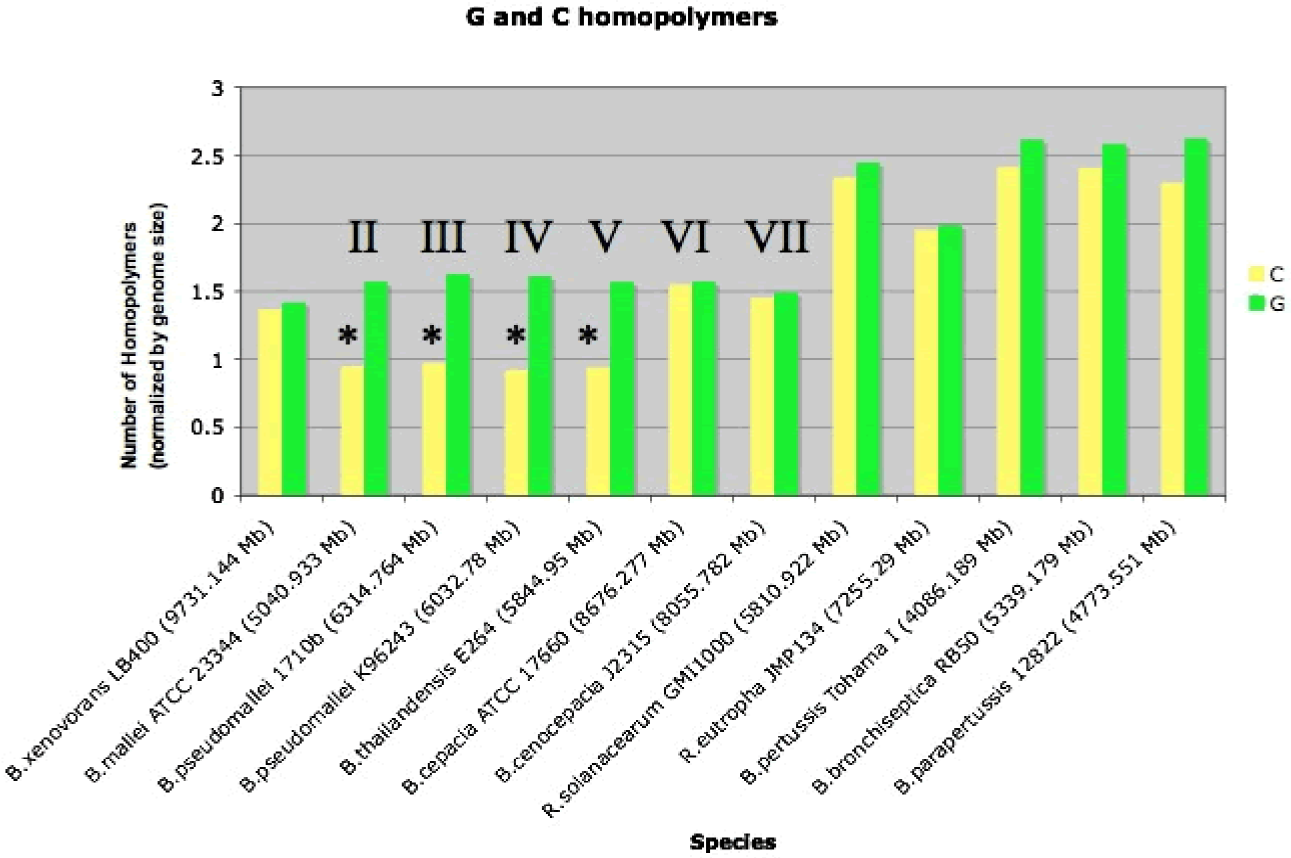


Fig 4: G an C homopolymers of various species within the genus *Burkholderia*, *Ralstonia*, and *Bordetella.*

This graph shows the number of C- and G-homopolymers (yellow and green respectively) of the completed bacterial genomes I investigated. The numbers are normalized with respect to the species' genome sizes, indicated in Megabases (Mb). Especially the first few genomes (II – V) exhibit a bias towards G-homopolymers. Genomes marked with an asterisk (*) showed statistically significant p-values.

Interdisciplinary Challenges and Future Research

One of the biggest challenges I faced when working on this project was communicating with the geneticists. Computer scientists and biologists speak almost different languages, each with its own extensive sets of acronyms, terms, and, of course, concepts. While I had no trouble picking up and using Perl, I found that most of the biological concepts went far beyond what my introductory genetics class had taught me. Sometimes, for example, I would look at my data (usually a whole lot of numbers compiled in a graph) and not see anything spectacular in it; but my genetics mentor, Dr. Thomas, would pick up on an interesting or unusual curve and point me in a new research direction or suggest a different way of visualizing the data to facilitate interpretation. Dr Bergeron, my computer science mentor, was also very helpful in bridging the gap between the two academic disciplines. He has been working with geneticists for a little longer than I have been, and he knows from experience what kind of difficulties a computer scientist might encounter when trying to understand concepts of genetics.

When doing research, as I discovered, one rarely answers a question without finding a new subject to investigate and more interesting problems to solve. For example, it would be interesting to look at longer microsatellite motifs. As the motifs increase in length, the data becomes more complex because there are more possibilities of base composition. Tri-nucleotide repeats (trimers) could be especially fascinating because the genetic code used by most organisms to translate DNA sequences into genes is composed of such three-base symbols. Studying the location of the microsatellites in the genome would also be intriguing. Other questions for further research could be whether some of these sequences are part of certain genes and if there are regions in the genome where microsatellites tend to cluster.

Through this project, I learned so much more about the field of genetics and how to apply computational tools to analyze some of the problems it presents than I could ever have learned in a classroom. I am fascinated by the number of unsolved questions about the process of life and by the prospect of working with geneticists to provide answers.

I would like to thank all the people who supported me and helped make this experience possible. Thank you to my UNH mentors, Drs. Bergeron and Thomas, for guiding me through this project. Also, thank you to the staff of the Undergraduate Research Opportunity Program at UNH and to the donors who made this grant available to me.

References

1. Picture of DNA: <http://encyclopedia.quickseek.com/images/DNA-structure-and-bases.png>
2. Jeffreys AJ, Wilson V, Thein SL. “Hypervariable ‘minisatellite' regions in human DNA.” *Nature* . 1985; **314** :67-73.
3. The National Center for Biotechnology Information (NCBI) <http://www.ncbi.nlm.nih.gov/>
4. “Trinucleotide Diseases.” Huntington's Outreach Project for Education at Stanford. 2002. Stanford University. 7 Jan. 2007 <http://www.stanford.edu/group/hopes/rltdsci/trinuc/f0.html>
5. Pearson WR, Lipman DJ.”Improved Tools for Biological Sequence Comparison.” *Proc Natl Acad Sci USA.* 1988; **85** :2444–2448.