An Introduction to Genetics for Mathematicians

Emily Scheele

Preface

Beyond the shadow of the Alps lies a historic stretch of Austrian farmland. It was here that, just four years after the discovery that all life is composed of cells and a mere sixteen before the proposition of evolutionary theory, Gregor Mendel renounced his responsibilities to the family farm. A brilliant physicist and mathematician, Mendel was drawn to life as a monk – and to the vast library promised to those who lived in the monastery. He continued to farm upon arrival, but the meticulous methods he chose for breeding pea plants suggested a purpose outside practical application.

Mendel was researching patterns of inheritance and, although his laws of Segregation and Independent Assortment never became popular in his lifetime, they are now foundational components of modern genetics.

Why research genetics? For Mendel, the appeal stemmed from an innate desire to *know* and to attain the "great gratification … afforded" him by his studies.[1] For James Watson, co-discoverer of the structure of DNA, the goal wasn't "simply to win, [but rather] to win at something really difficult."[2] For you, the appeal may be to acquire a finer understanding and appreciation of an avenue of science that – as you may or may not be aware – desperately calls for aid from skillful mathematicians.

Research into the genetic variation of living organisms has made leaps and bounds in the past few decades, but is beginning to leap and bound in the direction of a more computational and algorithmic approach than the chemical and physiological approach of years gone by.

The advent of genomic sequencing and its dependence upon computers demands that geneticists now take on additional roles as statisticians, programmers, and analysts. How much easier would it be for mathematicians – whose training already encompasses many of the responsibilities held by the aforementioned roles – to learn a bit of biology and to assist geneticists in their endeavors?

We will hereby discuss the vital yet seldom-recognized part that mathematics has played in the development of modern genetics. We will then turn our attention to differences that led to a separation of fields and how this has affected collaborative efforts. Our focus will then be to introduce interested mathematicians to the basics of genetics and to the technology that demands their attention today, as well as to suggest areas of future involvement beyond the scope of this reading.

---

[1] Biography.com Editors. "Gregor Mendel Biography." Accessed from: http://www.biography.com/people/gregor-mendel-39282.

[2] James Watson. "Succeeding in Science: Some Rules of Thumb," Science, 261, 24 (September 1993): 1812. September 1993.

Historical Relationship

Mendel was a mathematician by trade and he grew pea plants accordingly. Although the traits he chose to follow were qualitative in nature, his methods for keeping track of trait changes between each generation were purely quantitative. Mendel began with purebred seeds of two colors: green and yellow. He reasoned that, if traits were hereditary, one part must come from the mother and one from the father (stamen and pistil). As such, he labeled the purely yellow seeds *YY* (each *Y* representing the yellow trait passed down from mother or father) and the green seeds *yy* (each lower-case *y* representing the non-yellow (green) trait passed down from mother or father). When he crossbred these purebred plants, he saw that the offspring were *all yellow*. If one trait had indeed come from mother and another from father, the offspring should therefore all be labeled *Yy* (or *yY*). Working with a tentative hypothesis, Mendel then crossed two plants from the same generation of these offspring: *Yy* crossed with *Yy*. He meticulously counted the results and concluded that ¾ of this second generation were yellow and ¼ were *green*.

We can envision the possible combination of mother-father traits (*Y* and *y*) with a simple two-by-two matrix:

	Y	*y*
Y	*YY*	*Yy*
y	*yY*	*yy*

This is a standard Punnett square, named after Reginald C. Punnett, whose 1905 *Mendelism* is thought to have been the first popular introduction of genetics to the public.

Mendel concluded that the yellow trait must be *dominant* to green, and that any pair containing a *Y* trait (*YY, Yy,* or *yY*) would express itself as yellow, while only two copies of the *recessive y* trait (*yy)* would express a green seed. The ratio of 3:1 dominant to recessive displays in the second generation was strengthened by two factors: the number of plants included in the study and the repeatability of this ratio when studying other traits. Mendel counted 6,022 yellow and 2,001 green-seeded plants in the second generation, which led to a 3.01:1 initial ratio; he also counted the number of purple vs. white flowers in axial vs. terminal position, round vs. wrinkled seeds, inflated vs. constricted pod shape, green vs. yellow pod color, and tall vs. dwarf stem length in similar experiments. These ratios all converged upon the persistent 3:1.

Albeit simple, the numerical approach taken by Mendel was new to biology. Its irrefutability left a mark on the field that was popularized thirty-five years later by William Bateson. However, his ideas remained controversial given that it seemed dominant traits would eventually overtake the population and that recessive traits

would become increasingly rare – something that was not observed in reality. Stumped by this seeming dilemma and determined to unravel the mystery, Reginald Punnett (who had begun working with Bateson to popularize Mendelian ideas) introduced the problem to British mathematician and cricket partner G. H. Hardy. The sentiment he expressed in a 1908 paper is unfortunately echoed today – over a century later – by many mathematicians who find themselves similarly equipped to demystify genetics:

> To the Editor of Science: I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists. However, some remarks of Mr. Udny Yule, to which Mr. R. C. Punnett has called my attention, suggest that it may still be worth making...[3]

Hardy went on to derive an equation that proves the frequency of a given trait will remain constant in a population over time under certain conditions. This "simple point" is now known as the Hardy-Weinberg principle and is taught to budding geneticists in introductory biology classes across the globe. Its publication was instrumental to the eventual acceptance of Mendelian genetics, paving the way for incredible feats – not least of which is treatment for countless genetic diseases that might otherwise never have been properly understood.

The contribution of mathematicians is essential. If the only barrier keeping a competent mathematician from lending his or her input to geneticists is a reluctance to "intrude in a discussion concerning matters of which [they] have no expert knowledge," then perhaps we ought to inform these mathematicians that an elementary knowledge is all that is needed, and is easily attainable. We will set out to do so here, but must first recognize that there is another barrier: Time has drawn apart the two fields and has built a wall rife with jargon. Just as a college freshman cringes at the mention of a second order partial derivative, so does an experienced mathematician recoil at the idea of learning Polymerase Chain Reaction (PCR) technique. Both are relatively simple to comprehend, but require a baseline confidence before they can be tackled.

Genetics and mathematics need not be foreign to one another. In fact, many familiar concepts are at the heart of each, but are called different names. For example, x-ray crystallography is merely a practical application of group theory – with much tighter constraints than those afforded to a student of abstract algebra. Similarly, these students will learn that the elements of a Cayley Table are written as *row* followed by *column*, whereas a peek at any Cayley Table used in chemistry classes will show *columns* followed by *rows* – an important distinction when performing operations.

It is important to note that, despite the seemingly important differences in notation and the false air of significance they waft, group theory is no more a presence in the

---

[3] Hardy, G. H. 1908. "Mendelian proportions in a mixed population," *Science*, N. S. Vol. XXVIII: 49–50. (letter to the editor)

life of an organic chemist than is football in the life of a mathematician. This is made blatantly clear in an article published by Southern Illinois University in 1967, more than fifty years after the revelation of the Hardy-Weinberg principle: "It has been said that group theory at present is 'a tool not at the fingertips of most chemists.'" We should not be surprised if we were to find that the same holds true today.

Fortunately, the disparity between mathematics and genetics is being slowly but surely addressed from both sides. The American Mathematical Society (AMS) published a feature column in April of 2002 entitled *Mathematics and the Genome*, complete with a discussion of Classical Genetics, a history from Mendelian days to present, Population Genetics, and a discussion of present and future trends.[4] The introduction links to a *Molecular Biology and Genetics Primer for Mathematicians* that comes highly recommended and would serve as an excellent supplement to this reading. The AMS also openly draws attention to the way that "mathematical science has played an accelerating role in speeding up the developments in understanding the genome" and points to the "ongoing partnership between mathematics and biology in general."

The Genetic Science Learning Center (GSLC) has partially reciprocated this sentiment through publication of a set of learning materials under the title of *Mathematics*. Its content includes video introductions to probability, statistics, and growth and decay – meant to elucidate the process of mathematical modeling. The GSLC websites received 20 million visits in 2013 from nearly every country in the world, making it an ideal resource for geneticists and laypersons alike to learn about this enduring and necessary relationship. With this in mind, we turn our attention to the basics of genetics.

A Brief Genetic Primer: Terminology

Human Genome – the complete instruction manual contained within every cell for production of various proteins necessary for structure, movement, metabolism, and overall function of the human body

Protein – a string of between 20 and 34,350 amino acids that folds in upon itself to create one of an estimated 250,000-1,000,000 different structures (in the human body). These structures will dictate the function of the protein. Examples:

- **Hemoglobin** - a transport protein responsible for transporting oxygen from lungs to tissues
- **Immunoglobulin** – an antibody serving to protect the body from foreign particles (e.g. bacteria and viruses)
- **Collagen** – a structural protein abundant in connective tissues such as skin, bone, muscle, and tendons

[4] Malkevitch, Joseph. "Mathematics and the Genome: Introduction," *American Mathematical Society*, April 2002. Accessed from: http://www.ams.org/samplings/feature-column/fcarc-genome1.

Amino Acid – a building block for proteins, consisting on average of 19.2 atoms. Twenty different amino acids exist in the human body and they are strung together in an order dictated by codons.

Codon – a string of three nucleotides that codes for a specific amino acid. The code that translates mRNA (a string of codons) into a corresponding sequence of amino acids is *onto* (surjective), but not *one-to-one* (injective). Geneticists call this phenomenon *degeneracy* and it is one of the remaining mysteries of evolution. The code is represented in the table below, where each letter corresponds to one of four possible nucleotides.[5]

first base in codon	second base in codon: T	second base in codon: C	second base in codon: A	second base in codon: G	third base in codon
T	TTT Phe	TCT Ser	TAT Tyr	TGT Cys	T
	TTC Phe	TCC Ser	TAC Tyr	TGC Cys	C
	TTA Leu	TCA Ser	TAA stop	TGA stop	A
	TTG Leu	TCG Ser	TAG stop	TGG Trp	G
C	CTT Leu	CCT Pro	CAT His	CGT Arg	T
	CTC Leu	CCC Pro	CAC His	CGC Arg	C
	CTA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CTG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	ATT Ile	ACT Thr	AAT Asn	AGT Ser	T
	ATC Ile	ACC Thr	AAC Asn	AGC Ser	C
	ATA Ile	ACA Thr	AAA Lys	AGA Arg	A
	ATG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GTT Val	GCT Ala	GAT Asp	GGT Gly	T
	GTC Val	GCC Ala	GAC Asp	GGC Gly	C
	GTA Val	GCA Ala	GAA Glu	GGA Gly	A
	GTG Val	GCG Ala	GAG Glu	GGG Gly	G

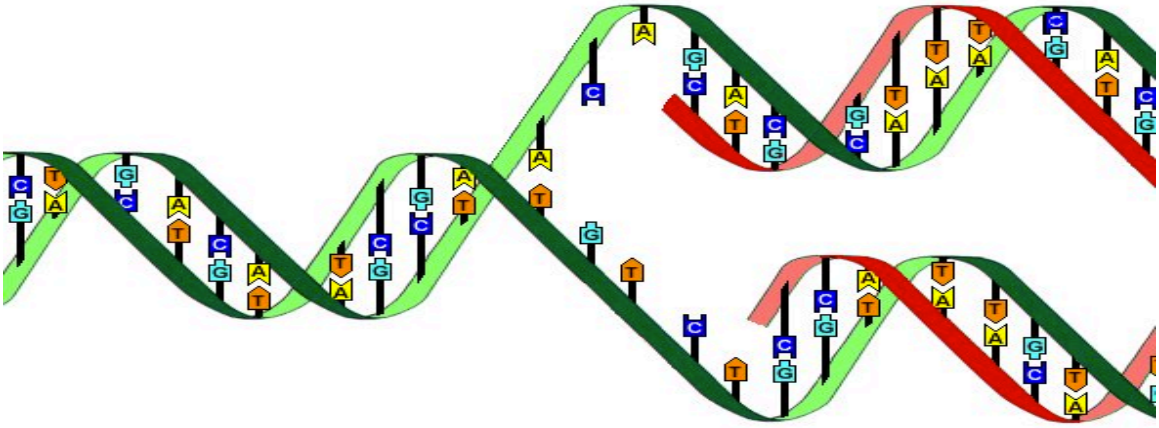
Nucleotide – A molecule consisting of three units (sugar, phosphate, and nitrogenous base) and comprising the backbone of DNA. The four bases are adenine, guanine, cytosine, and thymine (represented by A, G, C, and T above). The ordering of these bases along a string of DNA ultimately serves as a list of instructions for creating proteins that will manage the structure and function of the human body.

Nitrogenous Base – Adenine (*A*), guanine (*G*), cytosine (*C*), and thymine (*T*). Each base is paired with another: *A* pairs with *T* and *G* pairs with *C*. **DNA** is comprised of two strings of nucleotides wound around each other in a helical conformation; these helices can be envisioned as twisted latters where each step is the bond between a nitrogenous base at the notch of one pole (e.g. *G*) and a nitrogenous base at the notch of the other (e.g. *C*). **Note:** When DNA is replicated, these strands are pulled

---

[5] Clark, Jim. "The Genetic Code," *chemguide*, 2007. Accessed from: http://www.chemguide.co.uk/organicprops/aminoacids/dna4.html

apart and an enzyme called *DNA polymerase* synthesizes a *new daughter strand* along the length of each "pole" by pairing nitrogenous bases.
The image below shows a *new daughter strand* (red) being synthesized as the *original parent strand* (green) serves as a template.[6]



Chromosome – the packaged form of double-stranded DNA. Chromosomes contain proteins around which DNA tightly coils so that it can fit and remain secure within the nucleus of a cell. Each human cell contains 46 chromosomes: 23 come from the mother and 23 from the father. Distinct instructions (in the form of varying base sequences) are written on each chromosome.

Gene – a segment of nucleotides along a given segment of DNA that will eventually be translated into a string of amino acids corresponding to the appropriate order of codons, forming a certain protein. A gene is informally referred to as the basic unit of hereditary passed from parent to offspring.

We are now equipped to discuss a relevant bit of technology that demands the attention of competent mathematicians.

DNA Sequencing Technology

The Human Genome Project began in 1990 at the recommendation of the National Institutes of Health (NIH). Among its primary goals was to map the complete human genome – in other words, to determine the sequence of all 3 billion base pairs residing within our 23 chromosome pairs. This incredible feat was accomplished in 2003 along with the discovery that distinct individuals share 99.9% of their genes. The remaining 0.1% of genes is all that accounts for *every difference* between independent human beings.

---

6 "Mitotic chromosome etc – DNA Replication.jpg."
Accessed from: http://imagesbiogeolfxm.free.fr/mitose/original/AND%20REPLICATION.html

Seeing as how we all consider ourselves to be rather unique, that relatively small number of genes is agreeably quite important. Even the subtlest of changes to a string of nucleotides, such as the replacement of a single adenine with thymine, can mean the difference between producing a functioning hemoglobin protein and having sickle cell anemia.[7] Unfortunately, the techniques for sequencing long segments of DNA are far from perfect and many still rely upon – for lack of a better word – *shady* statistics. Perfection of these techniques will elucidate the correct sequences for production of healthy, functional proteins that can be replicated and used for gene therapy in populations affected by malignant genetic mutations.

Before statistics can be incorporated, the basic method of sequencing DNA must be understood. The Sanger method takes advantage of the fact that strings of nucleotides are synthesized from *5'* – *3'* (pronounced five prime to three prime). This is fancy chemical jargon for saying that every string of nucleotides ends in a *free hydroxyl group* (the bumpy part of a Lego) to which a new nucleotide may be added (stacked on top). Imagine that the nucleotides (Legos) may not be stacked underneath one another; a new nucleotide can *only* be added to this *free hydroxyl* end of the chain.

The first step to the Sanger method involves placing many copies of the same small segment of DNA in an environment prepped for replication. This environment will contain free nucleotides available for use by DNA Polymerase as it conducts the synthesis of a new daughter strand. However, a small proportion of these free nucleotides *will not have a free hydroxyl group at the 3' end*. In other words, no new Legos can be added once these "dideoxynucleotides" are incorporated into the daughter strand of a partially replicated DNA segment.

This sudden termination of the chain will occur at every spot along the original strand, given a sufficient number of copies and available nucleotides. The environment will now contain newly synthesized strands of lengths one, two, three, etc. up to the length of the original DNA segment. Using a technique called *gel electrophoresis*, these smaller segments can be separated according to size.

In newer sequencing models, these smaller segments can be scanned by a computer in order from smallest to largest. The dideoxynucleotides at the terminal end of each chain are fluorescently labeled according to whether they are an adenine, guanine, cytosine, or thymine derivative. In this way, a sequencing machine can detect which letter comes next until it has reached the largest fragment, which is capped with the final letter complementary to the original strand.

One important limitation to Sanger sequencing is that this method can only read between 800 and 1,000 base pairs at a time before physical constraints restrict further separation of fragments according to size; at any greater length, it becomes

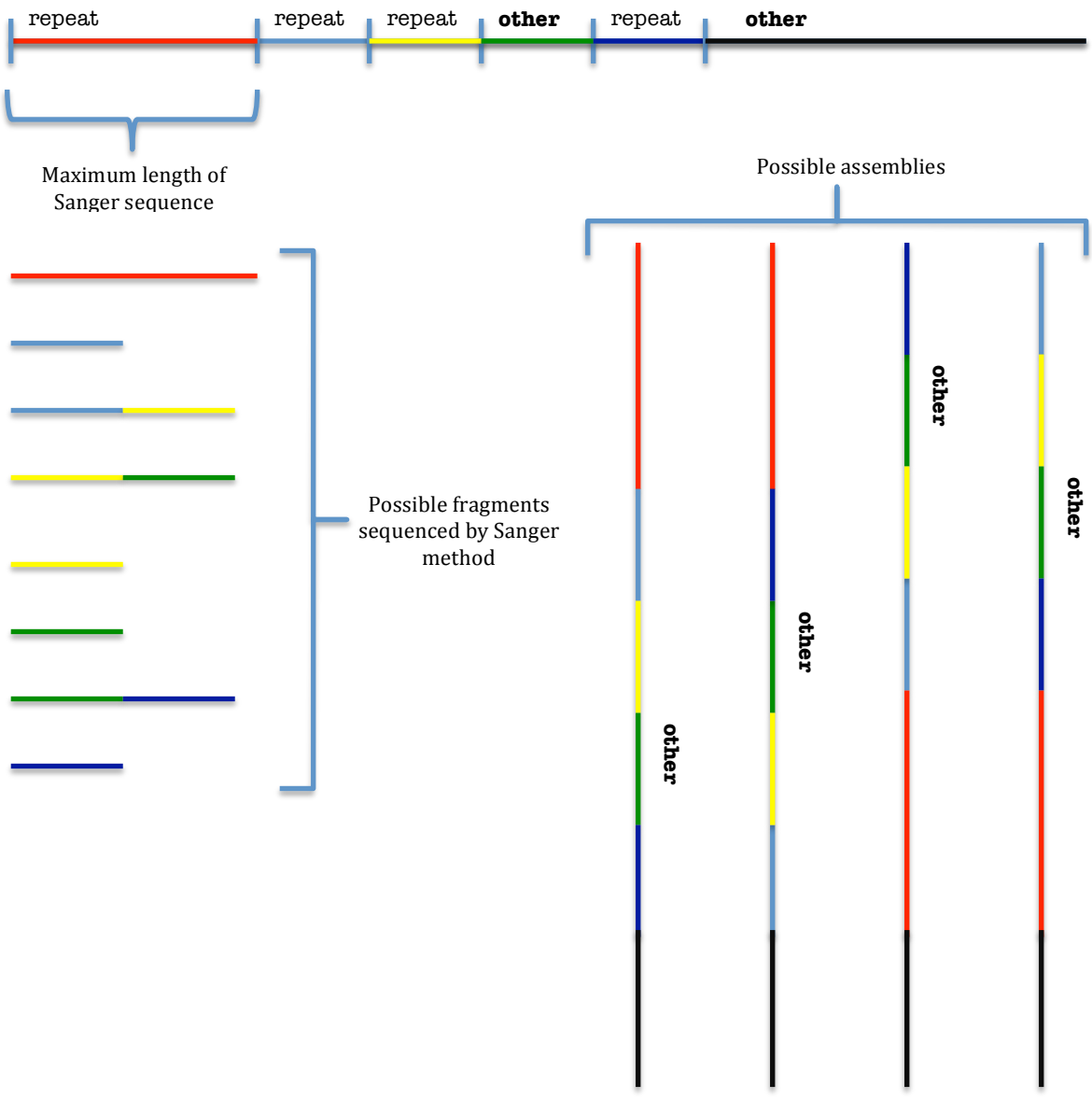
---

[7] Kimball, John. "Sickle Cell Disease," *Kimball's Biology Pages,* 2011.
Accessed from: http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/S/Sickle-cellDisease.html

impossible to distinguish various segments at a precision of just one nucleotide difference. Since the human genome contains 3 billion base pairs, a minimum of 3 million sequencing trials would have to be run. Even then, it would be impossible to know the order of these sequences as they are linked together in one chromosome because we cannot cut precisely at one and only one location.

The most precise method for cutting DNA into manageable chunks is to use *restriction endonucleases*, which are enzymes with the designated function of cutting after recognition of a given sequence of approximately six base pairs on average. Although this would imply an existence of 4^6 or 4,096 differently specified cutting enzymes, not all possibilities are accounted for and, furthermore, DNA is highly repetitive and may contain thousands of the same six-base sequence – in a row.

If the repetitive region extends beyond the maximum readability of Sanger sequencing, we will not be able to assemble the original DNA chain, as shown below.



Clearly, the location of non-repetitive DNA is ambiguous. Customized sequence finishing is employed after such trials, wherein additional data is generated to help close the gaps and determine the precise order of fragments. While such measures are highly accurate, they are certainly not perfect.

Most telling of these imperfections is the existence of a plethora of software programs relying on different algorithms to generate a complete assembly. In a 2010 study published in *Genomics*, Miller et al. survey and review ten of these available packages: SSAKE, SHARCGS, VCAKE, Newbler, Celera Assembler, Euler, Velvet, ABySS, AllPaths, and SOAPdenovo. One enormous frustration for geneticists is that choice of program currently dictates the resulting sequence and assembly. The end of Miller's 2010 study points to an even bigger issue:

> No algorithm or implementation solves the WGS assembly problem. Each of the various software packages was published with claims about its own superiority… In our experience, the success of an assembler depends largely on the sophistication of its heuristics for real reads including error, real genomes including repeats, and the limitations of modern computers.[8]

One limitation that goes unmentioned is the scarcity of competent mathematicians working to address the problems with algorithmic assembly. Miller, himself a mathematician from New York University, would certainly agree.

Genetics and mathematics need not be foreign to one another. A partnership begun over a century ago by such prominent figures as Gregor Mendel and G. H. Hardy must be continued and fostered if we are to see advances the magnitude of genetic screening and therapy in our future.

<u>Ways to Get Involved</u>

1. Read *Mathematical and Statistical Methods for Genetic Analysis* by Kenneth Lange. This book was written specifically for students in the mathematical sciences and is infused with problems seeking attention.

2. Attend or organize a seminar for mathematicians and geneticists to share their research with one another. Plan to ask questions and offer insight.

3. Join the Society for Mathematical Biology as a *full* or *student* member to receive the Bulletin of Mathematical Biology, participate in society activities, attend annual meetings, and to generally stay informed of the contributions that mathematics and statistics continue to make to our better understanding of the fundamental units of life.

---

[8] Miller, Jason R., Sergey Koren, and Granger Sutton. "Assembly Algorithms for Next-Generation Sequencing Data." *Genomics* 95.6 (2010): 315–327. *PMC*. Web. 11 Sept. 2015.