Duchenne Muscular Dystrophy (DMD) Genetics Incorporating Mutation Probabilities

Katie Collins and Bhavik Nagda

June 16, 2016

Abstract

Often diagnosed in childhood, Duchenne Muscular Dystrophy (DMD) is a fatal, sex-linked disorder. The disorder reaks havoc on the muscular fortitude of affected individuals, hampering both quality of life and life expectancy. DMD disproportionately affects males, as the mutation occurs on the X-chromosome. The cruel affliction, however, is acquired not only through genetic transmission, but also through *de novo* mutations within the egg prior to fertilization. Hardy-Weinberg genetics models DMD population prevalence, but fails to take into account *de novo* mutation probabilities. The following paper will first introduce a mathematical tool for modeling genetics, and will then apply this tool to incorporate *de novo* mutation probabilities. As such, this paper proposes a mathematical model to aid in predictive analytics for DMD and other sex-linked genetic disorders.

Contents

1 Introduction

Duchenne Muscular Dystrophy (DMD) is a debilitating genetic disorder that manifests through the progressive wasting away, or atrophying, of the muscles. The disorder arises from a mutation in the DMD gene on the X chromosome, in which the gene can no longer encode for the dystrophin protein. Dystrophin is essential in maintaining the integrity of muscle fibers, anchoring the muscle cell to the encircling extracellular matrix. However, without this protein, muscle fibers begin to leak kinase proteins and take on excess calcium, ultimately culminating in the complete degradation and death of the muscle cells. Such is the root cause of the incapacitating symptoms of the disorder: breathing and swallowing difficulties, large calf muscles (from the buildup of scar tissue), bone-thinning, heart problems, and the eventual inability to walk or stand without support.

Dystrophin is vital to the body and the disorder thus materializes itself quickly; initial diagnosis generally occurs around the age of two, and it is rare for a child with DMD not to be wheel-chair ridden by the age of 12. Like all genetic disorders, there is no cure; only an effort to increase the quality of life for the affected. Death frequently strikes in the form of respiratory or heart problems before the age of 30.

Despite being the most common and severe form of the muscular dystrophy-related disorders, DMD became a focus starting in the 1980s. As such, new research identified the pattern of inheritance of the disorder as X-linked recessive. Thus, seeing as men have only one X chromosome, a single mutation in the DMD gene results in an affected male. A female needs two affected X chromosomes (one from her mother and one from her father) in order to express the phenotype; nonetheless, even carriers – women who hold only one affected chromosome – can show symptoms of fatigue and heart muscle weakness.

The vast majority (80-90%) of carriers, however, show no visible symptoms, and thus pass on their diseased chromosome to their kin. Equally disturbing, random mutations in the egg pre-fertilization can negatively alter the DMD gene; thus, women with no family history of Duchenne's can pass the condition on to their children simply due to a fateful, random mutation. It is through mathematical modeling, in particular generating functions, that this paper attempts to tease out the relationship between carriers and affected individuals.

2 Generating Functions

2.1 Introduction

Generating functions are a discrete mathematical tool typically used to solve and prove otherwise complex counting problems. From partition theory conducted by Harding and Ramanujan in the 20th century to special relativity in classical physics, generating functions have become yet another tool to augment mathematical inquiry.

Essentially, generating problems translate problems about counting and probability into problems with functions. These functions generally have two characteristic values necessary for consideration: the coefficients of the functions, and the exponents of the functions.

The general form of the generating function for a sequence $< a_0, a_1 \dots a_n >$ is:

$$< a_0, a_1, a_2 \dots a_n > \Rightarrow a_0 + a_1 x + a_2 x^2 + \dots a_n x^n \tag{1}$$

2.2 Generating Functions and Punnett Squares

Let's look at a different method of modeling Punnett squares using Gregory Mendel's own experiments, specifically regarding pea seed color. The yellow seed color (denoted by the allele S) is dominant over the green seed color (denoted by the allele s).

Indeed, the principles of independent assortment and segregation lead to the following representation for the pea seed color Punnett square crossing two heterozygous yellow peas:

	S	s
S	SS	Ss
s	Ss	ss

Both the father and the mother have genes coding for Ss, and solving Punnett squares essentially involves 'multiplying' the genes together to codify the final phenotypes and genotypes.

Now for the analogous generating function representation. Denote $M(S, s)$ as the function representing the allele configurations for the mother's genes, and $F(S, s)$ as the function representing the allele configurations for the father's genes. Note that both the mother and father are hybrids with the Ss genotype. The law of independent assortment dictates that for both the mother and the father, the chance of selecting the dominant S allele equals $1/2$ and the chance of selecting the recessive s allele equals $1/2$. The mother and father generating functions can thus be represented as follows:

$$M(S, s) = 1/2S + 1/2s \tag{2}$$

$$F(S, s) = 1/2S + 1/2s \tag{3}$$

Note that the coefficients of each of the mother and father generating functions sum to one, since all possibilities have been taken into account. Breeding between the mother and father is represented as follows (for ease of interpretation, S^2 and s^2 have been represented as SS and ss respectively):

$$M(S, s) * F(S, s) = (1/2S + 1/2s)(1/2S + 1/2s) = 1/4SS + 1/2Ss + 1/4ss \tag{4}$$

Note that the coefficients of the resulting terms decree the exact probabilities of specific genes represented by variable terms in the resulting generation. Also, since the coefficients of the initial functions sum to one, the coefficients of the final function sum to one as well (via the multinomial theorem).

2.3 Similarity to the Hardy-Weinberg Theory

The generating functions proposed above bear striking resemblance to G. H. Hardy and Wilhelm Weinberg's equilibrium theory. Indeed, suppose the allele frequencies of two alleles on a single locus, A and a, are p and q respectively. Thus $P(A) = p$ and $P(a) = q$. Note that $p + q = 1$ since both alleles lie on the same locus. Thus, when we assume allele frequencies are equal in the sexes, as Hardy and Weinberg did, the parent functions can be generated as follows:

$$M(A,a) = F(A,a) = pA + qa = 1. \tag{5}$$

Following mating, genotype frequencies are generated:

$$M(A,a) * F(A,a) = (pA + qa)^2 = p^2AA + 2pqAa + q^2aa \tag{6}$$

The functions thus suggest that the genotype frequencies in the following generation will be p^2 for the AA genotype, $2pq$ for the Aa genotype, and q^2 for the aa genotype. As such, the generating functions model is consistent with the findings of Hardy and Weinberg. Generalizing now, we have that the Mendelian generating function for a parent (in this case, mother) can be represented by the following polynomial:

$$M(x_0, x_1, x_2....x_n) = a_0x_0 + a_1x_1 + ...a_nx_n \tag{7}$$

Where $< x_0, x_1...x_n >$ represents individual alleles and $< a_0, a_1...a_n >$ represents allele selection probabilities, respectively. In addition, $\sum_{i=0}^{n} a_i = 1$. Resulting cross probabilities are modeled by the following function:

$$M(x_0, x_1, x_2....x_n) * F(y_0, y_1, y_2...y_n) = \sum_{i,j,k=0}^{n} p_k x_i y_j \tag{8}$$

With the mathematical representation defined, the crux of the paper – modeling Duchenne Muscular Dystrophy probabilities – can now begin.

3 Modeling Duchenne Muscular Dystrophy

3.1 Assumptions and Justifications

- **Assumption:** Mating is random; the population is panmictic (i.e. all individuals are potential partners and equally likely to mate).
  Justification: While mating in a human population is seldom random and instead depends on a combination of geopolitical and socioeconomic factors, the assumption of random mating allows for greater simplification in this model, as mating probabilities are essentially equal. To further incorporate non-random mating in the future, Sewall Wright's F-Statistics models could be applied.

- **Assumption:** There is no gene flow through migration or selection.
  Justification: While gene flow certainly distorts mathematical outcomes, it is far too complicated for inclusion into this model. The numerical applications of this model use America as the considered population; gene flows within such a large population are often negligible.

- **Simplification:** Both individuals with and without Duchenne Muscular Dystrophy undergo mating, at equal probabilities, with one exception: affected females (both DMD sex alleles) will not exist.
  Justification: The MDA (Muscular Dystrophy Association) finds that "many young

adults with DMD attend college, have careers, get married and have children." Nevertheless, the Leiden University Medical Center reports that "Females carrying two mutated dystrophin genes have not been reported so far," eliminating all probabilities of affected females existing and mating.

3.2 Notation

Producing a sound mathematical model of DMD requires sufficient understanding of DMD's modes of transfer. Indeed, DMD can be transferred in two distinct ways: (1) through genetic sex-linked inheritance and (2) through a random mutation in the egg prior to fertilization (in the female zygote).

Note that DMD is a sex-linked disease. For the purposes of the model, define our genetic notation as follows:

- $X \Rightarrow$ Normal X Chromosome
- $X^a \Rightarrow$ Affected Y Chromosome
- $Y \Rightarrow$ Normal Y Chromosome

Resulting genotypes and phenotypes would be:

Genotype	Phenotype
XX	Normal Female
X^aX	Carrier (Unaffected) Female
X^aX^a	Affected Female
XY	Normal Male
X^aY	Affected Male

Forming the Mendelian generating function for one pair of parents (see Section 2.3 for general definition) would result in a binomial for the father (one X or X^a and one Y chromosome) and a monomial/binomial for the mother (either two X chromosomes or one X chromosomes and one X^a chromosome or two X^a chromosomes). While this approach may hold validity, it lacks scalability; considering each individual mating probability for parents would limit the scale of the model. Rather, a single generating function can represent an entire population, as seen in the next section.

3.3 Modeling the Entire Population

Rather than observing resulting phenotype and genotype frequencies from individual mating pairs, we look at allele frequencies within an entire generation and then stipulate resulting genotype and phenotype frequencies in the resulting F_1 generation. Essentially, assuming random mating, we can use allele frequencies in the parent generation to calculate the fraction of F_1 individuals with a given phenotype.

For example, suppose the X^a allele made up 20% of the entire female gene pool, while the Y allele made up 50% of the entire female gene pool. The resulting frequency of the

X^aY genotype in the F_1 population would be calculated by multiplying the original gene frequencies: $20\% * 50\% = 10\%$. We thus conclude that 10% of the F_1 generation would be affected males with the X^aY genotype.

Unfortunately, allele frequencies are not numerically reported or tabulated by any sources. Genotypes and phenotype frequencies are reported, however, and we can thus calculate allele frequencies. For example, a male population that is 25% affected for DMD (X^aY) and 75% normal (XY) would have an allele frequency for the X^a gene of $\frac{1}{2} * 25\% = 12.5\%$, an allele frequency for the X gene of $\frac{1}{2} * 75\% = 37.5\%$, and an allele frequency for the Y gene of $\frac{1}{2} * 25\% + \frac{1}{2} * 75\% = 50\%$. Note that the sum of these allele frequencies ($12.5\% + 37.5\% + 50\%$) is 100%, or the entire allele pool.

Furthermore, denote the following coefficients of genotype frequencies in the parent generation:

Genotype	Phenotype	Frequency
XX	Normal Female	f_1
X^aX	Carrier (Unaffected) Female	f_2
X^aX^a	Affected Female	f_3
XY	Normal Male	m_1
X^aY	Affected Male	m_2

Whereby $f_1 + f_2 + f_3 = m_1 + m_2 = 100\%$, since we must separate between male and female gene pools to produce our parent generating functions. Now, with simple mathematics, we can produce the following chart of allele frequencies within male and female populations.

	Allele	Frequency
Female	X	$f_1 + \frac{1}{2}f_2$
	X^a	$\frac{1}{2}f_2 + f_3$
Male	X	$\frac{1}{2}m_1$
	X^a	$\frac{1}{2}m_2$
	Y	$\frac{1}{2}$

3.4 Incorporating Mutation Probabilities

Indeed, DMD can be inherited not only genetically, but also through a mutation within the egg prior to fertilization. Suppose we model our mutation probabilities such that the probability of mutating within the egg, or the probability of $X \Rightarrow X^a$ within the egg is a, and subsequently, the probability of no mutation within the egg, or the probability of $X \Rightarrow X$ within the egg is $1 - a$. While allele frequencies prior to mutation would be the same as represented in Section 3.4, post-mutation allele frequencies would change (only for the alleles from the mother).

Suppose that $F_i(X)$ is the frequency of the X allele pre-mutation, while $F_f(X)$ is the frequency of the X allele post-mutation. Each of the X alleles represented in the X allele pool has a $(1 - a)$ chance of remaining X, so post mutation, we have that:

$$F_f(X) = P(no\ mutation) * F_i(X) = (1 - a) * (f_1 + \frac{1}{2}f_2) \tag{9}$$

The X^a allele is a bit more tricky. We have to add the mutation probability with the original X^a as follows:

$$F_f(X^a) = P(mutation) * F_i(X) + F_i(X^a) = a * (f_1 + \frac{1}{2}f_2) + \frac{1}{2}f_2 + f_3 \tag{10}$$

We can now update our allele selection probabilities table:

	Allele	Selection Probability
Female	X	$(1-a) * (f_1 + \frac{1}{2}f_2)$
	X^a	$a * (f_1 + \frac{1}{2}f_2) + \frac{1}{2}f_2 + f_3$
Male	X	$\frac{1}{2}m_1$
	X^a	$\frac{1}{2}m_2$
	Y	$\frac{1}{2}$

With allele selection probabilities complete, we can now build our generating function that will provide the resulting genotype frequencies:

$$\begin{aligned} &M(X, X^a, Y) \ * \ F(X, X^a) \\ &= (\frac{1}{2}m_1X + \frac{1}{2}m_2X^a + \frac{1}{2}Y) \ * \ ((1-a)(f_1 + \frac{1}{2}f_2)X + (a(f_1 + \frac{1}{2}f_2) + \frac{1}{2}f_2 + f_3)X^a) \end{aligned} \tag{11}$$

3.5 The Numerical Implementation

Finally, the genetic probabilities and prevalence predicted by the equations calculated above can be applied with real numeric values to obtain applicable formulas. Essentially, coefficients of the generated functions provide the phenotype prevalence from one generation to another. Thus, to find the prevalence of affected male individuals in the second generation, we have the following equation:

$$P(X^aY) = \frac{1}{2} * (a * (f_1 + \frac{1}{2}f_2) + \frac{1}{2}f_2 + f_3) \tag{12}$$

While the above-calculated equations incorporate all probabilities, a more realistic representation would neglect f_3 values, or the prevalence of affected females (see 3.2-Assumptions and Justifications). Since so few affected females exist, we can simply neglect them and thus $f_3 = 0$. Furthermore, since $f_1 + f_2 + f_3 = 1$, $f_1 + f_2 = 1$, and thus $f_1 = 1 - f_2$, we can use these restrictions to simplify the above equation to the following:

$$P(X^aY) = \frac{1}{2}(a + (\frac{1}{2} - \frac{a}{2})f_2) \tag{13}$$

Garcia et al. found in 2014 that *de novo* mutations occurred in 16.7% of all DMD cases; given that DMD itself has on average a prevalence of 19.25 per 100,000 births (Mah et al.), the approximate prevalence of textitde novo mutations worldwide is .167 * 19.25 = 3.215 cases per 100,000 cases (prevalence of .00003215). Plugging in these numerical values, we find ourselves at the following formula:

$$P(X^aY) = .00001607 + .24999196f_2 \tag{14}$$

4 Conclusion

Generating functions provide simple yet complete models of genetics that pave the way for further study in the murky realm of genetic disorders. The brief introduction to this mathematical tool presented in this paper is merely a feet wetter, an appetizer to the true full course of population genetics.

While traditional genetics would indeed provide an accurate model, one of the requirements for the Hardy-Weinberg rule is that "no genetic mutations" exist. Reconciling Hardy-Weinberg genetics with probability theory provides a more convincing and complete image of the genetic mechanisms driving DMD.

The paper initially addressed nomenclature concerns, and then delved into genetic mathematics using generating functions. From there, equations neglecting *de novo* mutation probabilities were generated; brief revisions shown in chapter 3.5 incorporated *de novo* mutation probabilities. Finally, real world data was incorporated to produce a powerful equation that relates carrier prevalence in one generation to DMD incidence in the next generation. The equation is shown below:

$$P(X^aY) = .00001607 + .24999196f_2$$

The equation captures the relationship between carriers and affected individuals in a linear function. Note that as the number of carriers approaches zero (as in no carriers exist), the prevalence of DMD approaches a non-zero value – a result of the mutations. Furthermore, the coefficient on the carrier frequency would normally equal .25 since the chance of a carrier passing the X^a allele would be .5, and the chance of the male passing the Y allele would be .5. However, the calculated coefficient is slightly less than .25; as the number of carriers (f_2) increases, some carriers will pass on affected alleles that "mutate" back into normal alleles, and others will pass on unaffected alleles that will mutate into DMD alleles. The net effect results in a coefficient that is slightly below the classical .25 coefficient achieved when neglecting mutations.

Opportunities for further study remain. Markov chain theory can also be applied to DMD genetics by separating genotypes into individual states and solving for transition probabilities. While this paper synthesizes realized data to create an equation, future statistical analysis could be conducted to validate the formulas. Furthermore, the model fails to incorporate DMD symptoms; indeed manifesting carriers of the disease – those who are slightly symptomatic – would reproduce at a diminished rate. As of the early 2000s, the US Center for Disease Control (CDC) began collecting DMD data through the Muscular Dystrophy Surveillance, Tracking, and Research Network, known as MD STARnet. This systematic statistical tracking will hopefully someday reveal truths that will engender greater discovery. However, a method of tracking and recognizing DMD carriers still remains to be seen (most DMD carriers are not aware of their state). As DMD science emerges in the coming years, the prediction formulas produced above should be tailored for greater accuracy.

Likewise, the augmentation of prediction formulas for genetic disorders coupled with big data will allow for enhanced analytics regarding genetic diseases. Medical centers can track the spread and progression of specific diseases; parents can receive more specialized genetic counseling; treatments and therapeutics can be more robust and targeted. The field is ripe

for a mathematical and technical makeover. Ultimately, nuanced mathematical models could revolutionize the field of genetics, not only providing fodder for predictive analytics, but also sparking the development of targeted, innovative therapeutics.

References

Bennett, J. H. "On The Theory Of Random Mating." Annals of Eugenics 17.1 (1952): 311-17. Web.

"Chapter 17: Generating Functions." MIT Open Courseware (2010): 385-408. Web. 16 June 2016. <db.math.ust.hk/notes_download/elementary/algebra/ae_A11.pdf>.

"Diseases - DMD - Causes/Inheritance." Muscular Dystrophy Association. N.p., 18 Dec. 2015. Web. 01 Apr. 2016.

"Diseases - DMD - Diagnosis." Muscular Dystrophy Association. N.p., 18 Dec. 2015. Web. 02 Apr. 2016.

"Diseases - DMD - Medical Management." Muscular Dystrophy Association. N.p., 18 Dec. 2015. Web. 02 Apr. 2016.

"Diseases - DMD - Signs & Symptoms." Muscular Dystrophy Association. MDA, 18 Dec. 2015. Web. 01 Apr. 2016.

"Diseases - DMD - Top Level." Muscular Dystrophy Association. N.p., 18 Dec. 2015. Web. 16 June 2016.

"DMD Dystrophin [Homo Sapiens (human)]." NCBI. N.p., N.d. Web. 2 Apr. 2016.

"DMD Gene." Genetics Home Reference. N.p., N.d. Web. 01 Apr. 2016.

"DMD/BMD in Females." Leiden Muscular Dystrophy Pages. Center for Human and Clinical Genetics, Leiden University Medical Center, 26 Oct. 2001. Web. 16 June 2016.

Garcia, Susana, Tomás De Haro, Mercedes Zafra-Ceres, Antonio Poyatos, Jose A. Gomez-Capilla, and Carolina Gomez-Llorente. "Identification of De Novo Mutations of Duchénnè/Becker Muscular Dystrophies in Southern Spain." International Journal of Medical Sciences Int. J. Med. Sci. 11.10 (2014): 988-93. Web.

"Learning About Duchenne Muscular Dystrophy." Learning About Duchenne Muscular Dystrophy. NIH - National Human Genome Institute. Web. 31 Mar. 2016.

Mah, Jean K., Lawrence Korngut, Jonathan Dykeman, Lundy Day, Tamara Pringsheim, and Nathalie Jette. "A Systematic Review and Meta-analysis on the Epidemiology of Duchenne and Becker Muscular Dystrophy." Neuromuscular Disorders 24.6 (2014): 482-91. Web.

"Muscular Dystrophy: Hope Through Research." National Institute of Neurological Disorders and Stroke. NIH, N.d. Web. 01 Apr. 2016.

Thomas, Jeffery. "Muscular Dystrophy." : Background, Pathophysiology, Etiology. Medscape, N.d. Web. 01 Apr. 2016.