Crossing Over…Markov Meets Mendel

Saad Mneimneh*

Department of Computer Science, Hunter College of CUNY, New York, New York, United States of America

Abstract: Chromosomal crossover is a biological mechanism to combine parental traits. It is perhaps the first mechanism ever taught in any introductory biology class. The formulation of crossover, and resulting recombination, came about 100 years after Mendel's famous experiments. To a great extent, this formulation is consistent with the basic genetic findings of Mendel. More importantly, it provides a mathematical insight for his two laws (and corrects them). From a mathematical perspective, and while it retains similarities, genetic recombination guarantees diversity so that we do not rapidly converge to the same being. It is this diversity that made the study of biology possible. In particular, the problem of genetic mapping and linkage—one of the first efforts towards a computational approach to biology—relies heavily on the mathematical foundation of crossover and recombination. Nevertheless, as students we often overlook the mathematics of these phenomena. Emphasizing the mathematical aspect of Mendel's laws through crossover and recombination will prepare the students to make an **early** realization that biology, in addition to being experimental, IS a computational science. This can serve as a first step towards a broader curricular transformation in teaching biological sciences. I will show that a simple and modern treatment of Mendel's laws using a Markov chain will make this step possible, and it will only require basic college-level probability and calculus. My personal teaching experience confirms that students WANT to know Markov chains because they hear about them from bioinformaticists all the time. This entire exposition is based on three homework problems that I designed for a course in computational biology. A typical reader is, therefore, an instructional staff member or a student in a computational field (e.g., computer science, mathematics, statistics, computational biology, bioinformatics). However, other students may easily follow by omitting the mathematically more elaborate parts. I kept those as separate sections in the exposition.

Introduction

Mendel and High School Biology

Sexually reproducing organisms generally combine heritable traits from two parents. The biological process that combines those traits is called meiosis. While mutations could occur during meiosis, most of the variation arises from the combinations of parental traits. How do these parental traits combine? The dominant theory was that some sort of blending or averaging took place. However, such a mode of inheritance would result in an average of all ancestors after only a modest number of generations (imagine repeatedly mixing colors). Instead, by performing experiments on plants, Mendel pointed out the existence of discrete elements that combine but do not mix. Figure 1 shows the simulated number of types of individuals as a function of time. Averaging, with traits taking real values in $[1,10]$, is used on one population, and the model described in the section "A Simple Model", with elements (later called alleles) taking discrete values in $\{0,1\}$, is used on another. Mutations are ignored. In both cases, a population size of 100 is kept constant for the entire duration of the simulation (100 time steps). The simulation is repeated 1,000 times to obtain an average for each time step.

Mendel formulated the concept of a *gene* (unit of inheritance), and hypothesized that inheritance is governed by the following two laws of *gene*tics:

1. **Segregation**: Each sexually reproducing organism has two *alleles* (copies) for each gene, one inherited from each parent; and in turn will contribute, **with equal probability** (1/2), only one of these two alleles.
2. **Independent assortment**: Alleles of different genes are inherited **independently** (later deemed not so accurate).

The state of a gene, the *genotype*, is determined by the two alleles. The resulting trait, the *phenotype*, is then a function of this state. When the alleles are the same, the gene, or equivalently the genotype, is *homozygous*; otherwise, it is *heterozygous*. For example, if an allele can be either a or A, then the possible genotypes are aa, aA, Aa, and AA. Table 1 shows the possible segregations of parental genotypes when at least one of them is heterozygous.

In a dominant/recessive mode where A is dominant, the corresponding phenotype is obtained as a function of the genotype as shown in Table 2, leading to a 3:1 ratio, a 1:1 ratio, and a 1:0 ratio of dominant to recessive phenotypes, respectively.

Students often overlook that these ratios are not simply based on counting the entries, but the result of the segregation law: each allele is contributed with equal probability, i.e., 1/2, resulting in a probability of $1/2{\cdot}1/2=1/4$ for each entry in the tables. Table 3 shows another example involving two heterozygous dominant/recessive genotypes that lead to a 9:3:3:1 ratio of phenotypes. In addition to the segregation law, students should be reminded that this ratio assumes that the law of independent assortment holds: alleles of different genes are inherited independently, resulting in a probability of $1/2{\cdot}1/2=1/4$ for each assortment (refer to the next section for a mathematical definition of independence),

Citation: Mneimneh S (2012) Crossing Over…Markov Meets Mendel. PLoS Comput Biol 8(5): e1002462. doi:10.1371/journal.pcbi.1002462

Editor: Fran Lewitter, Whitehead Institute, United States of America

Published May 17, 2012

Funding: The author received no specific funding for this article.

Competing Interests: The author has declared that no competing interests exist.

* E-mail: saad@hunter.cuny.edu

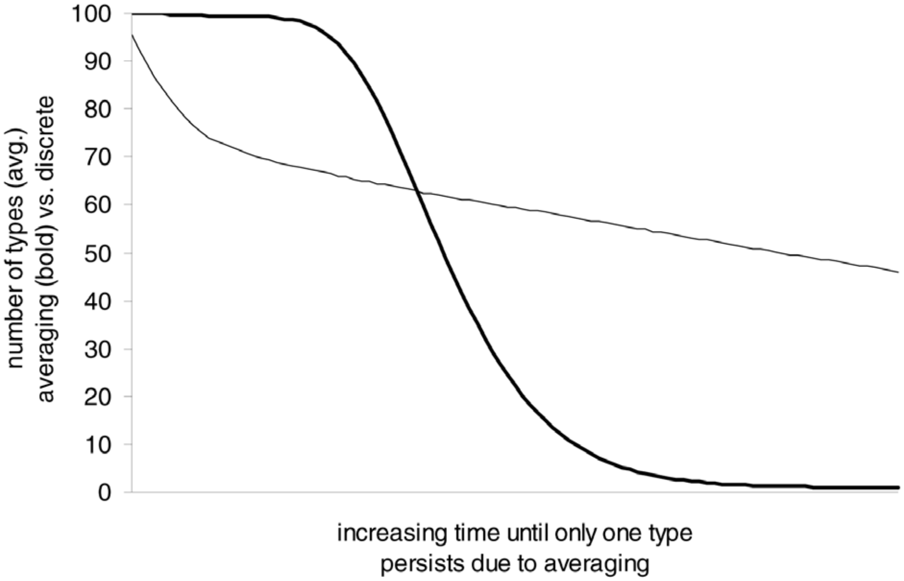

Figure 1. Fast convergence of inheritance by averaging.
doi:10.1371/journal.pcbi.1002462.g001

thus a probability of $1/4 \cdot 1/4 = 1/16$ for each entry in the table.

Chromosome, Crossover, and Recombination

About 100 years later, it was established that the physical structure underlying Mendel's laws is the chromosome (for simplicity, a long molecule of DNA). This discovery matched Mendel's experiments really well: In diploid organisms like us chromosomes come in pairs (thus the name diploid), one from each parent! With few exceptions, each chromosome of the pair has copies of the same genes (special stretches of DNA) arranged in the same order: the alleles! In an attempt to explain experimental results and confirm Mendel's laws, chromosomal *crossover* was formulated and described by Thomas Morgan (coincidentally, his student John Northrop was a teacher of botany at Hunter College, the author's institution), but demonstrated only about 20 years later. Crossover is a mechanism that occurs at the early stages of the meiotic prophase, and combines the two chromosomes of the pair into one, a process called *genetic recombination*. During this process, the chromosome of the pair that is the source of the allele alternates every so often. Exactly when the switch—the crossover—happens is almost arbitrary.

When two alleles come from different chromosomes of the pair, their corresponding genes are said to recombine (can you identify the recombinations in Table 3?). Figure 2 illustrates a genetic recombination with one crossover.

A Slight Discrepancy and Genetic Linkage

Mendel's laws (segregation and independent assortment) dictate that genetic recombination occurs with a probability of 1/2. Let's re-examine why this holds true. Let a and A be the two alleles of gene i on the two chromosomes. Similarly, let b and B represent the same for gene j, respectively. Chromosomal crossover will result in recombination of gene i and gene j if one of the two assortments aB and Ab occurs. Since each allele is contributed with equal probability (segregation), both a and B are contributed with probability 1/2. Since alleles of different genes are inherited independently (independent assortment), the assortment aB occurs with probability $1/2 \cdot 1/2 = 1/4$ (refer to the next section for a mathematical definition of independence). The same analysis applies for the assortment Ab, leading to an overall recombination probability of $1/4 + 1/4 = 1/2$.

However, it has been observed that some pairs of genes show a correlation in their alleles, e.g., their probability of recombination is less than 1/2. In this case, there is a *linkage* between the genes. How can we now incorporate this notion into the mathematics of Mendel's laws, which so far have relied on the fact that genes are not correlated (assorted independently)? Fortunately, a simple probabilistic model based on Figure 2 (1 crossover) will capture the effect of linkage, and as a result, alleles that are near each other on a chromosome will tend to be inherited together. The inaccuracy of Mendel's law of independent assortment lies therein. Nevertheless, one should still expect that genes which are far from each

Table 1. Genotypes.

	a	*A*		*a*	*a*		*A*	*A*
a	*aa*	*aA*	*a*	*aa*	*aa*	*a*	*aA*	*aA*
A	*Aa*	*AA*	*A*	*Aa*	*Aa*	*A*	*AA*	*AA*

doi:10.1371/journal.pcbi.1002462.t001

Table 2. Phenotypes.

	a	*A*		*a*	*a*		*A*	*A*
a	*a*	*A*	*a*	*a*	*a*	*a*	*A*	*A*
A	*A*	*A*	*A*	*A*	*A*	*A*	*A*	*A*

doi:10.1371/journal.pcbi.1002462.t002

Table 3. Phenotypes for two heterozygous genotypes.

			aA	*bB*	
		ab	*aB*	*Ab*	*AB*
	ab	*ab*	*aB*	*Ab*	*AB*
aA	*aB*	*aB*	*aB*	*AB*	*AB*
bB	*Ab*	*Ab*	*AB*	*Ab*	*AB*
	AB	*AB*	*AB*	*AB*	*AB*

doi:10.1371/journal.pcbi.1002462.t003

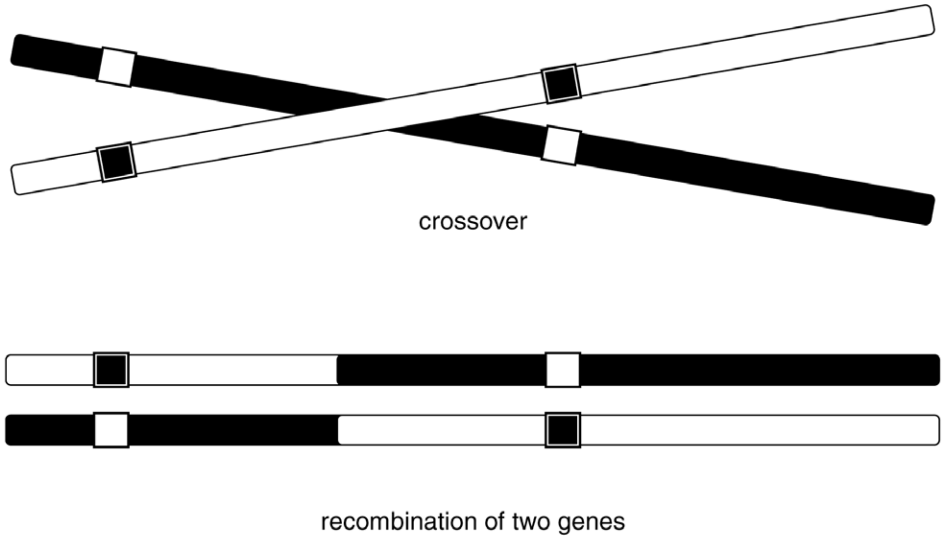

Figure 2. One chromosomal crossover and a genetic recombination.
doi:10.1371/journal.pcbi.1002462.g002

other on a chromosome (or on different chromosomes altogether) will assort independently, as Mendel once observed. It will require a better probabilistic model to reflect those two contradictory behaviors (genetic linkage and independence); the later introduction of the Markov chain will take care of this. But first, I will present a simple probabilistic model for genetic linkage. And before doing so, let's review some basic mathematics.

What Do We Need to Know?

Probability

Let $S=\{1,...,n\}$. A subset of S, $E\subseteq S$, is considered as an event (but not all events are subsets of S). Given a variable x, define the following probabilities of events:

$$P(x=i)=P(\{i\})=\frac{1}{n},\ 1\leq i\leq n$$
$$\text{(uniformly random)}$$

$$P(E)=|E|\frac{1}{n}=\frac{1}{n}+\ldots+\frac{1}{n}\ (|E|\text{ times})$$

where $|\ |$ denotes the size of a set. So $P(S)=1$. The negation of an event will always satisfy:

$$P(\text{not } E)=1-P(E)$$

Given two events E_1 and E_2, E_1 and E_2 are exclusive (cannot occur together) if and only if

$$P(E_1\text{ or }E_2)=P(E_1)+P(E_2)$$

Given two events E_1 and E_2, E_1 and E_2 are independent if and only if

$$P(E_1\text{ and }E_2)=P(E_1)P(E_2)$$

For instance, if E_1 is an event of probability q and $E_2\subseteq S$, then $P(E_1\text{ and }E_2)=q|E_2|/n$. In general, however, E_1 and E_2 may not be independent. So we define the probability of E_2 conditional on E_1, i.e., the probability of E_2 given that E_1 occurs.

$$P(E_2|E_1)=\frac{P(E_1\text{ and }E_2)}{P(E_1)}$$

For instance, let $E_1=\{i+1,...,m\}$ and $E_2=\{d+1,...,n\}$ with $i\leq d<m$. Note that $\begin{aligned}&P(E_1\text{ and }E_2)=P(\{d+1,...,m\})\\&=(m-d)/n\neq P(E_1)P(E_2)\end{aligned}$. Then,

$$\begin{aligned}P(E_2|E_1)&=\frac{P(E_1\text{ and }E_2)}{P(E_1)}\\&=\frac{(m-d)/n}{(m-i)/n}=\frac{m-d}{m-i}\end{aligned}$$

Matrix Multiplication

I will assume some familiarity with matrices. If, however, this notion is unfamiliar, the parts of the exposition that use matrices may be skipped. Only 2×2 matrices will be considered in this exposition. The multiplication of 2×2 matrices is defined below.

$$\begin{bmatrix}a&b\\c&d\end{bmatrix}\begin{bmatrix}e&f\\g&h\end{bmatrix}=\begin{bmatrix}ae+bg&af+bh\\ce+dg&cf+dh\end{bmatrix}$$

$$\begin{bmatrix}a&b\\c&d\end{bmatrix}^n=\begin{bmatrix}a&b\\c&d\end{bmatrix}\ldots\begin{bmatrix}a&b\\c&d\end{bmatrix}(n\text{ times})$$

Geometric Series

One of the series that is almost invariably covered in basic calculus is the geometric series.

$$1+a+a^2+\ldots+a^{n-1}=\begin{cases}\dfrac{1-a^n}{1-a},\ a\neq 1\\ n,\text{ otherwise}\end{cases}$$

Exponential Limit

This is one of the basic expressions covered when studying limits.

$$\lim_{n\to\infty}(1+\frac{a}{n})^n=e^a,\ e=2.71828183$$

Therefore, $(1+a/n)^n\approx e^a$ for large n.

Logarithm

Here's the definition of natural logarithm and some of its properties:

$$\ln a=b\Leftrightarrow a=e^b$$

$$\ln a^b=b\ln a$$

$$\ln ab = \ln a + \ln b$$

Harmonic Series

Another famous encounter is the harmonic series and its approximation.

$$1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n} \approx \ln n, \text{ for large } n$$

Derivatives

A function $f(x)$ reaches a local maximum or minimum when its derivative $f'(x) = 0$. Here are some examples of derivatives:

$$[ax + b]' = a$$

$$[f(x)g(x)]' = f'(x)g(x) + f(x)g'(x)$$

$$[\ln f(x)]' = \frac{f'(x)}{f(x)}$$

A Simple Model

Motivated by Figure 2, a uniform 1-crossover model can be constructed as follows: Consider a chromosome with n genes, i.e., n alleles on each chromosome of the pair. A crossover x is equal to i if it separates gene i and gene $i+1$, where gene $n+1$ is hypothetical when $x = n$, i.e., no crossover. Assume that x is uniform in $\{1, \ldots, n\}$ (thus the name of the model).

Linkage

Based on the above setting, x takes any value in $\{1, \ldots, n\}$ with probability $1/n$. Two genes at a distance $0 \le d < n$, say i and $i+d$, will recombine if x is in $\{i, \ldots, i+d-1\}$, i.e., with probability $1/n + \ldots + 1/n$ (d times),

$$p_d = \frac{d}{n}$$

This confirms that genes within a close distance (small d) on the chromosome are less likely to be subject to recombination (genetic linkage). Genes that are far apart (large d) have a high probability (up to $1 - 1/n$) of recombination, but are they independent (see "What Is Wrong" section)?

Segregation

To find the probability that a given allele of gene i is inherited, let E with probability q be the event that the recombination process starts on the given chromosome of the pair. This event and that genes 1 and i recombine (an event of probability $(i-1)/n$) are independent. The probability of inheriting the given allele is:

$$P(E \text{ and genes 1 and } i \text{ do not recombine or not } E \text{ and genes 1 and } i \text{ recombine})$$

$$= P(E \text{ and genes 1 and } i \text{ do not recombine}) + P(\text{not } E \text{ and genes 1 and } i \text{ recombine})$$

The addition is justified by the exclusivity of the events: a given allele is inherited when the process starts on the given chromosome and genes 1 and i do not recombine, or when the process starts on the other chromosome and genes 1 and i recombine. Due to the independence of E and recombination, the above becomes:

$$= q\left(1 - \frac{i-1}{n}\right) + (1-q)\frac{i-1}{n}$$

A reasonable assumption is that $q = 1/2$ and, in this case, the above evaluates to $1/2$ for every i, as predicted by the segregation law.

Genetic Mapping

Genetic mapping is the problem of placing the genes along the chromosome in their correct relative order. The bad news: It is hard! The good news: Genetic linkage can be used to infer genetic mapping. Though obsolete (it has been done), genetic mapping can be considered to be the first effort towards a computational approach to biology. How does it work?

In the uniform 1-crossover model, genetic linkage tells us that the probability of recombination of two genes is proportional to the distance between these genes.

Now consider the genotyping depicted in Table 4 where frequency of recombination can be used as a measure of distance. In a way analogous to Table 4, analyzing the frequency of different **pairs** of the phenotypes A, B, and C might reveal, for instance, that B and C recombine more often than A and B; therefore, we infer that B is closer to A than C. Such arguments help us to derive the gene order on the chromosome (relative order, not exact distances). While it may be hard to set up the experiment and obtain many offsprings to estimate probabilities, such arguments were definitely behind the construction of the early genetic maps, e.g., the first map of the human genome (all the chromosomes) in 1987.

Table 4. Frequency and distance.

			aA	*bB*	
		ab	*aB*	*A b*	*AB*
aa bb	*ab*	*ab*	*aB*	*Ab*	*AB*

The frequency of observing *aB* and *Ab* determines the probability of recombination of the two genes, thus a measure to reflect their distance.

doi:10.1371/journal.pcbi.1002462.t004

What Is Wrong?

The reader may choose to skip this section to the next. The uniform 1-crossover model is very insightful in explaining Mendel's law of segregation with independent assortment corrected to reflect genetic linkage. However, it suffers from a few deficiencies.

Linkage: OK But…

Nothing is seriously wrong about this aspect. By assigning lower probabilities of recombination for smaller distances, the distance between two genes justifies their linkage when they do not assort independently. However, the actual probability of recombination may not necessarily be **proportional** to distance or have a dependence on the chromosome length, as in $p_d = d/n$ (but more on this in the Markov section).

Segregation: Too Sensitive

The probability of inheriting a given allele is contingent on the probability that the recombination process starts on the given chromosome of the pair, previously called q. If $q = 1/2$, the probability of inheriting a given allele is $1/2$, as it should be by the segregation law. While this is a biologically reasonable assumption on q, the segregation law stands very sensitive to this particular choice. A slight deviation from $q = 1/2$ could result in a similar deviation in the probability of inheriting the given allele. Let $q = 1/2 - \epsilon$, then this probability for gene i is (from the "Segregation" section):

$$\left(\frac{1}{2}-\epsilon\right)\left(1-\frac{i-1}{n}\right)+\left(\frac{1}{2}+\epsilon\right)\frac{i-1}{n}$$

When $i=n$, i.e., $(i-1)/n\approx 1$, this is approximately $1/2+\epsilon$. If the starting of the recombination process favors one chromosome, ϵ can be large, say close to $1/2$ ($q\approx 0$). The above probability becomes arbitrarily close to 1. This means that the given allele will be inherited almost always.

Independent Assortment: Breaks

Despite genetic linkage, one should still expect that genes which are far from each other on the chromosome will assort independently. Because each chromosome can be treated separately, this independence is certainly true for genes that are on different chromosomes altogether. But on the same chromosome, the probability of recombination $p_d=d/n$ implies, for instance, that recombination of gene 1 and gene n occurs with a probability of $(n-1)/n\approx 1$ for large values of n. Therefore, gene 1 and gene n are highly correlated, and thus dependent (they will almost always recombine).

In retrospect, two genes i and j recombine when the alleles of the two genes are inherited from different chromosomes. Since the probability of inheriting a given allele is $1/2$ when the segregation law holds, independence then dictates that the probability of recombination of gene i and gene j must be equal to $1/2$. To see this, let E_i and E_j represent the events of inheriting a given allele for gene i and gene j, respectively, then:

$$P(\text{genes } i \text{ and } j \text{ recombine})$$

$$=P(E_i \text{ and not } E_j \text{ or } E_j \text{ and not } E_i)$$

$$=P(E_i \text{ and not } E_j)+P(E_j \text{ and not } E_i)$$

$$=P(E_i)[1-P(E_j)]+P(E_j)[1-P(E_i)]$$

where addition is justified by exclusivity of events, and the last equality follows from that gene i and gene j are independent. When the segregation law holds, $P(E_i)=P(E_j)=1/2$ and the above expression evaluates to $1/2\cdot 1/2+1/2\cdot 1/2=1/4+1/4=1/2$. Assuming q in the previous section is $1/2$, genes are independent if and only if $d=n/2$. Therefore, the law of independent assortment fails when genes are on the same chromosome.

Now, why do we insist that the model must satisfy, among other properties, the law of independent assortment? Well, first because it is a correct law for distant genes. And second, since the probability of recombination increases with distance due to genetic linkage, the law of independent assortment tells us that **the probability of recombination increases up to $1/2$, but cannot exceed** $1/2$ (this statement excludes *hotspots*, which are regions on the chromosome that experience a high probability of recombination even at small distances). It is important for students to make this realization, which will come in handy when solving genetic mapping problems, as illustrated in the section "A Computational Example of Genetic Mapping".

Generalization: Not Easy

One might consider extending the uniform 1-crossover model as an attempt of generalization to mimic the actual biological process. However, I will show that extending this model in the most natural way (mathematically, that is) will break the linkage property. For this purpose, consider a uniform 2-crossover model. Let x_1 be the first crossover which is uniform in $\{1,\ldots,n\}$ (as before), and x_2 be the second crossover which, conditional on x_1, is uniform in $\{x_1,\ldots,n\}$. Therefore, x_1 and x_2 are **not independent**, for x_2 cannot precede x_1. The choice of $x_2\geq x_1$ simplifies the math, but making $x_2>x_1$ does not change the results.

Now, why even bother to show that this model, which is more difficult to analyze than its predecessor, does not work? Well, my experience in teaching has been the following: While it is important to show students what works, it is equally important to show them what does not work.

With this in mind, all we need is a counter example, so consider gene 1 and gene $d+1$ (these two genes are at a distance d from each other). The probability of a recombination of gene 1 and gene $d+1$ is:

$$P(x_1\leq d \text{ and } x_2>d)$$

Using conditional probability and the harmonic series approximation, the "Uniform 2-crossover Model" section shows that when $n-d$ is large, this probability is approximately

$$\frac{n-d}{n}[\ln n-\ln(n-d)]$$

We can rewrite the above as:

$$-\frac{n-d}{n}\left[\ln\frac{1}{n}+\ln(n-d)\right]$$

$$=-\frac{n-d}{n}\ln\frac{n-d}{n}$$

This is not an increasing function of d. In fact, consider $f(x)=-x\ln x$. This function has a maximum of $1/e$ when $f'(x)=-\ln x-1=0\Rightarrow x=1/e$. Therefore, we have the highest probability of recombination when $(n-d)/n=1/e$, i.e., $d=n(1-1/e)$. Note that in this case $n-d=n/e$, which is large (as required above) when n is large. This means that gene 1 is most likely to recombine with a gene located at a distance approximately 63% of the chromosome length (see Figure 3). While this is an interesting result, it stands as a pure mathematical endeavor with no biological basis.

A Better Model: When Markov Meets Mendel

While the uniform 1-crossover model captures the essentials of segregation and linkage, it is lacking in some important aspects. First, the probability that a given allele is inherited (should be $1/2$) depends on an implicit parameter of the model (q in the "Segregation" section must be $1/2$). Second, genes exhibit the linkage property but they are almost never independent, as this would require a probability of recombination equal to $1/2$ (see "A Slight Discrepancy and Genetic Linkage" section). From the "Linkage" section, this probability is expressed as d/n, implying that only genes at a distance equal to half the chromosome length are independent. Moreover, the probability of recombination depends on the chromosome length and, therefore, two chromosomes that are locally similar but have different lengths exhibit different local recombination behavior. This is not biologically justifiable. Finally, a generalization (with uniformity maintained) to mimic the real biological process with multiple crossovers is not conceivable.

A better mathematical model is needed to rectify the above deficiencies. In principle, the model should satisfy the following three laws with multiple crossovers:

1. **Segregation**: The probability that a given allele of the gene is inherited is $1/2$.

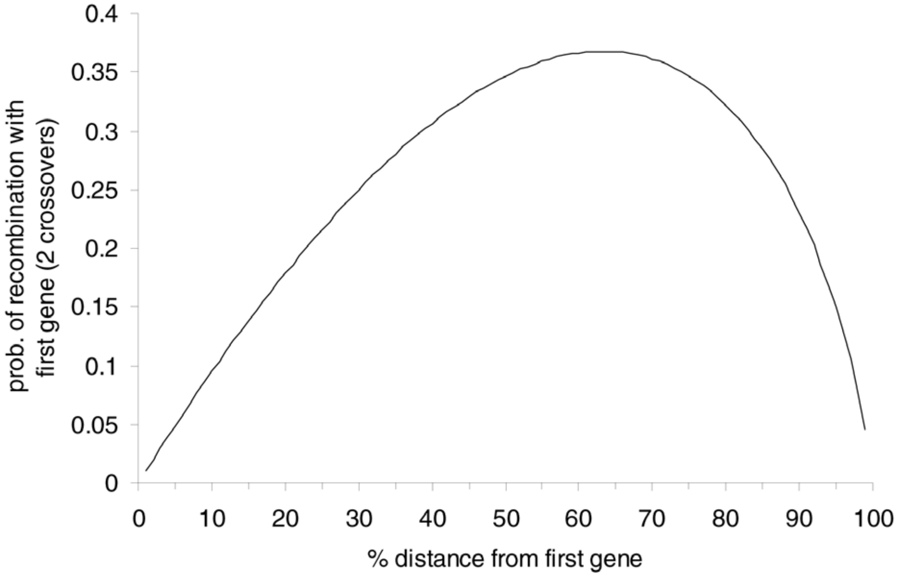

Figure 3. The uniform 2-crossover model. Probability of recombination of the first gene and a gene at a distance given as a percentage of the chromosome length. A maximum probability of $1/e = 0.367879$ occurs at $(1-1/e)\cdot 100 \approx 63\%$.
doi:10.1371/journal.pcbi.1002462.g003

2. **Linkage** (missed by Mendel): The probability of recombination of two genes is an increasing function of the distance between them, so it is higher for distant genes. Nevertheless, it should not depend on the chromosome length.
3. **Independent assortment**: This is impossible due to linkage where distance is a determining factor in the recombination. The alternative is to require genes to be asymptotically independent. As a result, the probability of recombination must approach $1/2$ when the distance between the two genes becomes large.

Being a computer scientist by training and not a biologist, when I first suggested to my students a model based on a Markov chain, I called it the *jumping model of recombination*. I also expressed to them my concern that it may not be *real*, but as it turned out, it made perfect sense. To be loyal to my first terminology, I will call it here the jumping model.

The Jumping Model

The jumping model is based on a Markov chain. A Markov chain consists of a set of states with probabilities of transition between them (thus the jumping term). For computer scientists, this is often illustrated as a directed weighted graph with vertices representing the states and directed edges representing the transitions between states. The weight of an edge is the probability of the corresponding transition. This is shown in Figure 4 for a Markov chain with two states. Operationally, one would start at a given state and follow transitions in discrete time steps as indicated by their probabilities, thus changing state from one step to another. Let a_{kl} be the probability of transition from state k to state l, and x_i be the state at time step i. Figure 4 shows a transition probability p between the two states (and $1-p$ to the same state, because the transition probabilities of a given state must sum up to 1). A generalized notion of a transition is captured by a conditional probability with the following property:

Markov property: For $j > i$,

$$P(x_j = l | x_i = k \text{ and } x_{i-1} = \dots)$$
$$= P(x_j = l | x_i = k)$$

When $j = i+1$, this probability is the transition probability $a_{kl} = P(x_{i+1} = l | x_i = k)$. In the event $(x_i = k \text{ and } x_{i-1} = \dots)$ only $x_i = k$ is relevant. In other words, the probability of a state at a given time depends only on the most recently known state.

What is the biological significance of the Markov chain in Figure 4? Each state represents a chromosome of the pair, and time in the Markov chain corresponds to

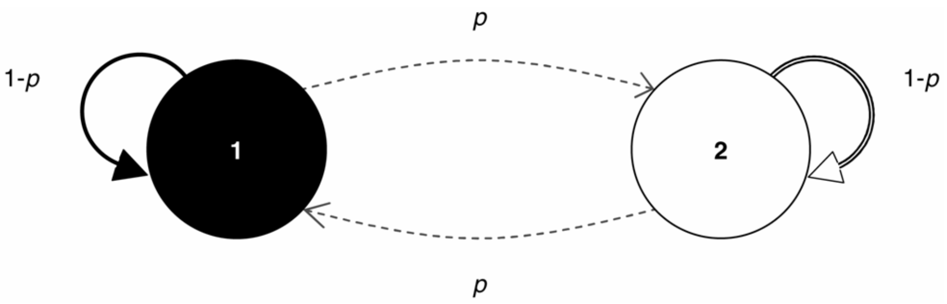

Figure 4. A simple Markov chain. Arguably the simplest Markov chain with two states, where each state represents one chromosome of the pair. Transitions between the two states (chromosomal crossovers) occur with probability p.
doi:10.1371/journal.pcbi.1002462.g004

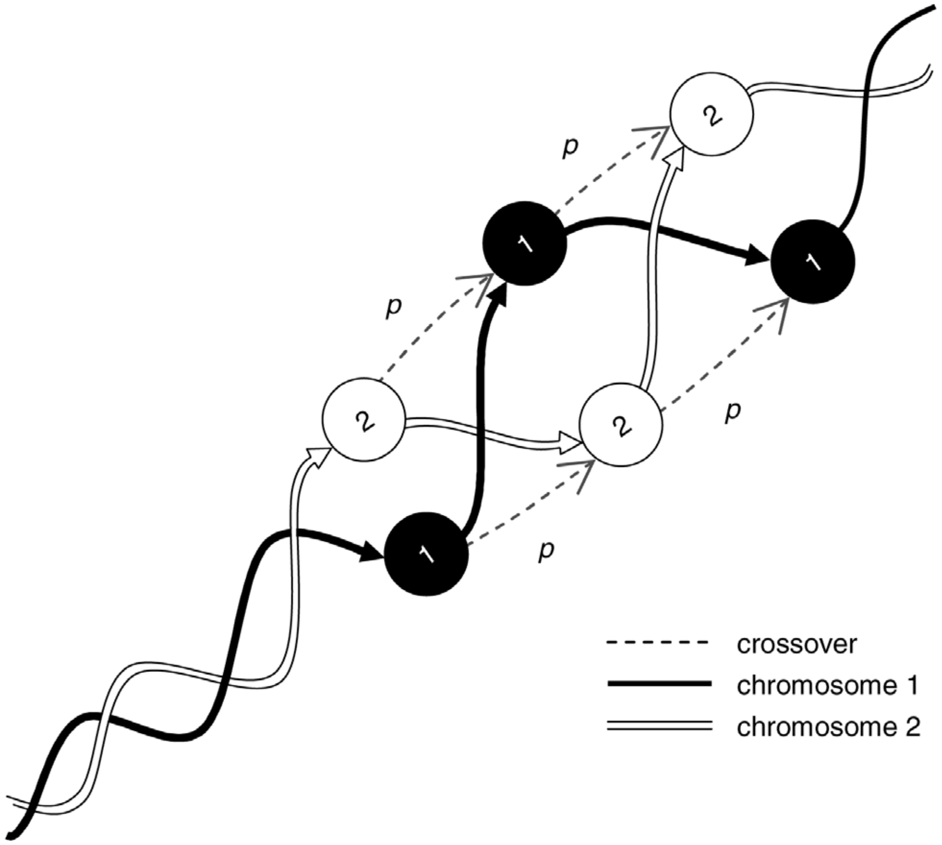

Figure 5. Crossover and recombination as a Markov chain. Dashed lines represent transitions (crossovers) with probability p, and solid lines (black and white) represent transitions (on the same chromosome) with probability $1-p$.
doi:10.1371/journal.pcbi.1002462.g005

genes on the chromosome. A transition between states in one time step signifies a crossover, and the probability of such a crossover is p. Therefore, x_i represents a crossover when $x_i \neq x_{i+1}$. One could then inquire about the probability of being in a given state at a given time. The event of being in a given state at time i parallels the event that the corresponding chromosome is the source of the allele for gene i. This is illustrated in Figure 5 by conceptually duplicating the chain for each gene to reflect the change of state over time.

A useful representation of a Markov chain is by a matrix P where P_{kl} (the term in the k^{th} row and l^{th} column of P) is the probability of transition from state k to state l; therefore, every row in P must add up to 1. If we call the states in Figure 4 state 1 and state 2, then our Markov chain can be expressed as:

$$P = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

In this matrix, P_{kl} can be interpreted as $P(x_{i+1}=l|x_i=k)=a_{kl}$. Why is this matrix representation useful? Let's multiply P by itself:

$$P^2 = \begin{bmatrix} 1-2p(1-p) & 2p(1-p) \\ 2p(1-p) & 1-2p(1-p) \end{bmatrix}$$

Note for instance that $P^2_{12}=2p(1-p)$ is equal to $P(x_{i+2}=2|x_i=1)$, because to transition from 1 to 2 in two time steps we can transition from 1 to 1 to 2 with probability $(1-p)p$ or from 1 to 2 to 2 with probability $p(1-p)$. As it turns out, $P(x_{i+d}=l|x_i=k)=P^d_{kl}$. The proof of this fact is in the "Markov Transitional Probabilities" section and uses conditional probability and the Markov property. Thus, every row in P^d must also add up to 1.

Because P is a symmetric matrix ($P_{12}=P_{21}$), a final note is that all powers of P are symmetric matrices. Therefore, $P^d_{kl}=P^d_{lk}$, which now implies that every column in P^d must also add up to 1. We can finally establish that the probability of recombination is

$$\begin{aligned} p_d = P(x_i=1 \text{ and } x_{i+d}=2 \\ \text{or } x_i=2 \text{ and } x_{i+d}=1) \\ = P(x_i=1)P(x_{i+d}=2|x_i=1) \\ + P(x_i=2)P(x_{i+d}=1|x_i=2) \\ = P(x_i=1)P^d_{12} + P(x_i=2)P^d_{21} \\ = P^d_{12}[P(x_i=1)+P(x_i=2)] \\ = P^d_{12} \cdot 1 = P^d_{12} = P^d_{21} \end{aligned}$$

Segregation and Independent Assortment

Following the logic of previous sections, the probability that a given allele of gene i is inherited is:

$$q(1-p_{i-1})+(1-q)p_{i-1}$$

$$P=\begin{bmatrix} 1/3 & 2/3 \\ 2/3 & 1/3 \end{bmatrix} \; P^2=\begin{bmatrix} 5/9 & 4/9 \\ 4/9 & 5/9 \end{bmatrix}$$

$$P^3=\begin{bmatrix} 13/27 & 14/27 \\ 14/27 & 13/27 \end{bmatrix} \; P^4=\begin{bmatrix} 41/81 & 40/81 \\ 40/81 & 41/81 \end{bmatrix}$$

$$P^5=\begin{bmatrix} 121/243 & 122/243 \\ 122/243 & 121/243 \end{bmatrix} \; P^6=\begin{bmatrix} 365/729 & 364/729 \\ 364/729 & 365/729 \end{bmatrix}$$

$$P^7=\begin{bmatrix} 0.499771 & 0.500229 \\ 0.500229 & 0.499771 \end{bmatrix} \; P^8=\begin{bmatrix} 0.500076 & 0.499924 \\ 0.499924 & 0.500076 \end{bmatrix}$$

$$P^9=\begin{bmatrix} 0.499975 & 0.500025 \\ 0.500025 & 0.499975 \end{bmatrix}$$

$$P^{10}=\begin{bmatrix} 0.500008 & 0.499992 \\ 0.499992 & 0.500008 \end{bmatrix}$$

$$P^{11}=\begin{bmatrix} 0.499997 & 0.500003 \\ 0.500003 & 0.499997 \end{bmatrix}$$

$$P^{12}=\begin{bmatrix} 0.500001 & 0.499999 \\ 0.499999 & 0.500001 \end{bmatrix} \; P^{13}=\begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$$

Figure 6. Convergence to steady state probabilities. Computation is performed with a rounding error $-5{\cdot}10^{-7} < \epsilon \leq 5{\cdot}10^{-7}$.

doi:10.1371/journal.pcbi.1002462.g006

Again, if $q=1/2$ the above probability is $1/2$, which makes the jumping model subject to the same sensitivity to q as the uniform 1-crossover model. However, this can now be alleviated. The theory of Markov chains tell us that P^d will converge for large values of d and all rows of P^d become identical. Therefore, the rows will define a *steady state* probability for each state. In other words, the effect of q will be washed out. This theory will not be presented here, but Figure 6 shows a few powers of a given matrix P.

Because P^d is symmetric in our case,

$$P^d_{11}=P^d_{21} \text{ (convergence)}$$

$$P^d_{21}=P^d_{12} \text{ (symmetry)}$$

$$P^d_{12}=P^d_{22} \text{ (convergence)}$$

Since rows and columns of P^d must both add up to 1, $p_{i-1}=P^{i-1}_{12}=P^{i-1}_{21}$ converges to $1/2$ for large enough i. By exchanging the roles of q and p_{i-1} in the top expression, we also get $1/2$, maintaining the segregation law for large enough distances when $q \neq 1/2$.

In addition, since both $P(x_{i+d}=l|x_i=k)$ and $P(x_{i+d}=l)$ approach $1/2$, we have that $P(x_{i+d}=l|x_i=k) \approx P(x_{i+d}=l)$ for large d. This makes $P(x_i=k \text{ and } x_{i+d}=l) = P(x_i=k)P(k_{i+d}=l|x_i=k) \approx P(x_i=k)$ $P(x_{i+d}=l)$ when d is large. Therefore, genes i and $i+d$ are asymptotically independent, confirming the law of independent assortment for large enough distances.

Linkage (and Hotspots!)

The previous sections show that $p_d=P^d_{12}$ and that P^d_{12} converges to $1/2$ for large values of d, thus establishing the laws of segregation and independent assortment. However, we wish to determine p_d for every value of d. This will re-establish the above results. This time, however, and instead of using the theory of matrices (e.g., eigen decomposition) to study how P^d evolves, I will revert to elementary mathematics. Two genes at a distance d from each other will recombine if and only if their chromosome experiences an odd number of crossovers along that distance. This is equivalent to the event of making an odd number of transitions between the two states of the Markov chain during d time steps. Let E_d be this event (thus $p_d=P(E_d)$). It is not hard to see that

$$\underbrace{E_d}_{\text{odd}} = (\underbrace{E_{d-1}}_{\text{odd}} \text{ and } \underbrace{\text{not } E_1}_{\text{even}})$$
$$\text{or } (\underbrace{\text{not } E_{d-1}}_{\text{even}} \text{ and } \underbrace{E_1}_{\text{odd}})$$

Observe that $p_1=P(E_1)=p$. Therefore, we can write:

$$p_d=p_{d-1}(1-p)+(1-p_{d-1})p$$

The Markov property is essential to justify the multiplication by $1-p$ and p in the above equation because it makes E_1 independent of the history E_{d-1}. Technically, $P(E_1|E_{d-1})$ does depend on the state at time step $d-1$, but given the symmetry in our Markov chain, it is always p. By rearranging and taking care of the special case when $d=1$ we get:

$$p_d=\begin{cases} (1-2p)p_{d-1}+p & d>1 \\ 1 & d=1 \end{cases}$$

It is easy to verify that the solution

$$p_d=\frac{1-(1-2p)^d}{2}$$

satisfies the above recurrence with a base case $p_1=1$ (following the pattern of the recurrence, we can retrieve the above expression if we replace d by $d-1$, multiply by $(1-2p)$, and add p).

While it is easy to verify the solution, obtaining it should not remain a wild guess. By working out a few iterations for p_d, the "Recurrence for p_d" section shows how to derive the solution using a geometric series.

The mathematically savvy could verify that $1-2p$ is an eigenvalue of P, and that the same expression could have been obtained using a technique called eigen decomposition. This expression for p_d reveals interesting properties (all can be verified from Figure 7):

- When d is large (and $p>0$), $(1-2p)^d$ goes to zero, causing p_d to converge to $1/2$. This convergence was discussed in the previous section, and should not be surprising by now.
- When $0<p<1/2$ ($1-2p$ is positive), $(1-2p)^d$ is greater than zero and less than one, causing p_d to increase with d (linkage). This increase, however, is not linear as in the uniform 1-crossover model; therefore, it is biologically more realistic.
- When $p>1/2$ ($1-2p$ is negative), the sign of $(1-2p)^d$ alternates, causing p_d to alternate between a typical value for d and high (hotspots, first time captured).

The jumping model captures the essential biology of crossover and recombination through the laws of segregation, linkage, and independent assortment. In addition, it reveals the non-typical high recombination probabilities of hotspots. Hotspots are regions on the chromosome that experience a high probability of recombination even at small distances. Therefore, depending on the parameter p, the jumping model embodies two modes of chromosomal recombination.

While a hotspot does not present a difficult concept, it is usually misinterpreted by students as a *region with high probability of recombination*. This is true if the region is too small (a peak in Figure 7), which is biologically typical of hotspots. However, if the region is large enough, there can be a high probability of recombination only if there is a corresponding low probability, as seen by the alternating pattern in Figure 7. What is interesting about the jumping model (which may not be true biologically) is that this low probability is the typical one for the given distance when p is replaced with $1-p$. This is also confirmed by the expression we derived for p_d, because when $p>1/2$ and $p_d<1/2$, d is even and, therefore, $(1-2p)^d=(2p-1)^d$:

$$\frac{1-(1-2p)^d}{2}=\frac{1-(2p-1)^d}{2}$$
$$=\frac{1-[1-2(1-p)]^d}{2}$$

The alternation itself should be intuitive because a high probability of recombination at a small distance must be driven by a high probability of crossover, which in turn means a high probability of crossing over back to the same chromosome. The jumping model captures this fact through the parameter p with a threshold of $1/2$ as a high probability of crossover.

Back to the Days of Morgan

Morgan established that the probability of recombination as a function of distance is the following:

$$p_d=\frac{1-e^{-2d}}{2}$$

which does not account for hotspots. In addition, the notion of distance in the above expression is not the same as ours. To see this, assume that p is close to zero in the jumping model (no hotspots) and, therefore, $1/p$ is large. Using the exponential limit,

$$(1-2p)^d=\left[(1-\frac{2}{1/p})^{\frac{1}{p}}\right]^{pd}$$
$$\approx[e^{-2}]^{pd}=e^{-2pd}$$

By making $\lambda=pd$, and replacing $(1-2p)^d$ with e^{-2pd} in the expression obtained for p_d, we get

$$p_\lambda=\frac{1-e^{-2\lambda}}{2}$$

which has the same form as Morgan's expression. So what is λ?

$$\lambda=\frac{d}{1/p}$$

where d is the distance and $1/p$ is the average distance until the next crossover (because a crossover occurs with probability p). So λ is the average number of crossovers between the two genes, and this is how Morgan defined his distance.

Why This Way?

I could have simply argued that the probability of recombination p_d is $(1-e^{-2d})/2$, and that this is consistent

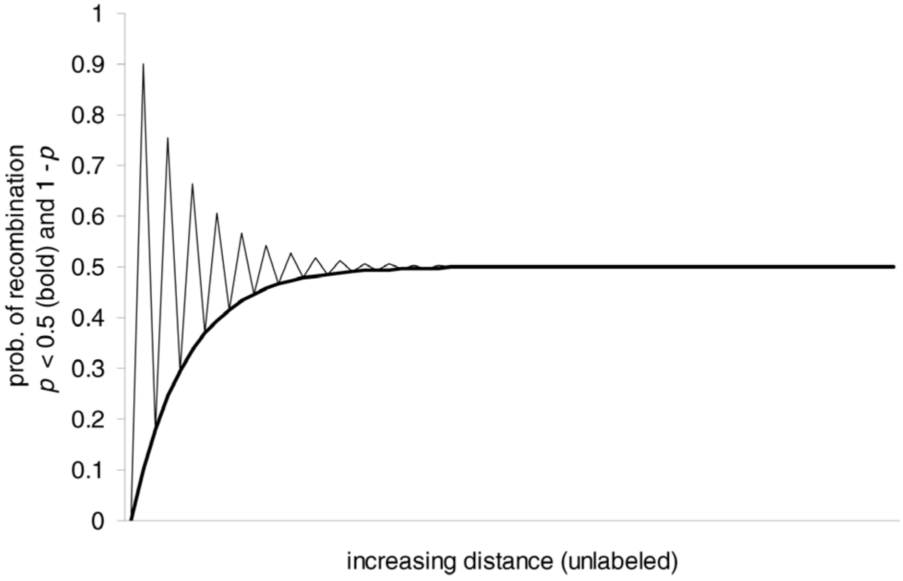

Figure 7. The jumping model, two modes of recombination, for $p<0.5$ and $1-p$.
doi:10.1371/journal.pcbi.1002462.g007

with the laws of inheritance. Therefore, I will list what I believe are important aspects of this exposition.

- There is a rapid prototyping with a simple uniform 1-crossover model that reflects the essential biological properties of crossover and recombination (though not perfectly). This allows the student to quickly make a connection between the biology and the mathematics.
- There is no need for advanced calculus or probability (e.g., no mention of Poisson processes or probability distributions other than uniform).
- To achieve a better understanding of the biological properties, the exposition proceeds by pointing out the deficiencies of the simple model.
- The simple model itself is a useful tool that is actually used for simulation, e.g., genetic algorithms.
- Having a model (whether mathematical or not) provides some operational sense, so the biology is made more concrete.
- Moving progressively through the models illustrates what it takes to make attempts, including wrong ones, in the modeling of biological systems.
- Multiple models reinforce the ideas by exposing them in different settings.
- Markov chains are useful as a tool for modern biological sciences and, therefore, introducing them in this context gives the student an early preparation.
- The jumping model captures two modes of recombination, normal and hotspots, and puts them in their biological context by means of the parameter p.
- The jumping model also provides the insight that the probability of crossover must be less than $1/2$ to observe the typical behavior of recombination (linkage), and hence giving the correct impression that p is rather small.
- The alternating behavior of the jumping model corrects one major misunderstanding of hotspots.
- Morgan's first result can be derived as a special case.
- The jumping model can be described (not necessarily analyzed) very easily and satisfies all the required biological properties of crossover and recombination. Therefore, a student can effectively retain and communicate the recombination process.

A Computational Example of Genetic Mapping

Consider the hypothetical family in Table 5 where alleles take values in $\{0,1\}$ (inspired by a homework assigned by Bonnie Berger at MIT).

To map the genes (genetic mapping), we count the number of recombinations, both paternal and maternal, for each pair of genes, AB, AC, and BC. Then we estimate the probabilities of recombination and relate them to distances.

There are $2n-i$ recombinations of A and B, $2n-i+x$ recombinations of A and C, and $2i-1+x$ recombinations of B and C. Therefore, A and B recombine with probability $(2n-i)/(2n)$, A and C with probability $(2n-i+x)/(2n)$, and B and C with probability $(2i-1+x)/(2n)$. Let's denote these probabilities by $P(AB)$, $P(AC)$, and $P(BC)$, respectively. If n is large enough, $P(AB)\approx P(AC)\approx 1-\alpha/2$ and $P(BC)\approx\alpha$.

First Attempt

Since $1-\alpha/2>1/2$ (for AB and AC), and it is not generally assumed that genes represent hotspots, we might suspect that our knowledge of the alleles of gene A is wrong. It is more plausible that the alleles of gene A are 1,0 for the father and mother, as shown in Table 6.

Table 5. A hypothetical family and three genes A, B, and C shown with their alleles.

	Genes		
	A	*B*	*C*
Father	0,1	0,1	0,1
Mother	0,1	0,1	0,1
Offspring 1	0,0	1,0	0,1
⋮	⋮	⋮	⋮
Offspring $i-1$	0,0	1,0	0,1
Offspring i	0,0	0,1	x,0
Offspring $i+1$	0,0	1,1	1,1
⋮	⋮	⋮	⋮
Offspring n	0,0	1,1	1,1

For simplicity of illustration, the chromosome of the pair with allele 0 inherited for gene A (both parental and maternal) is chosen for the offsprings, so this is not to be interpreted as if allele 0 is always inherited for gene A. Offsprings 1 to $i-1$ are identical, and similarly, offsprings $i+1$ to n are identical. Allele x is either 0 or 1, and $i=\alpha n$ for some $0<\alpha<1/2$.

doi:10.1371/journal.pcbi.1002462.t005

Table 6. The same hypothetical family after the alleles of gene A have been switched.

	Genes		
	A	*B*	*C*
Father	**1, 0**	0,1	0,1
Mother	**1, 0**	0,1	0,1
Offspring 1	0,0	1,0	0,1
⋮	⋮	⋮	⋮
Offspring $i-1$	0,0	1,0	0,1
Offspring i	0,0	0,1	x,0
Offspring $i+1$	0,0	1,1	1,1
⋮	⋮	⋮	⋮
Offspring n	0,0	1,1	1,1

doi:10.1371/journal.pcbi.1002462.t006

This will make $P(AB)\approx P(AC)\approx\alpha/2$ and will keep $P(BC)\approx\alpha$. Since the probability of recombination of distant genes is higher, the order of genes is B, A, C or C, A, B.

This solution puts B and C at equal distances from A and, therefore, makes the distance from B to C twice the distance from A to B (and that from A to C). However, doubling the distance should not double the probability of recombination unless the probability is a linear function of distance like in the uniform 1-crossover model. We may adopt this model here if we know in advance that only one crossover occurs; this conditioning makes the crossover uniform even when the underlying model is the jumping one (because of the symmetry in the Markov chain). For this argument to work we will also need $x=1$; otherwise, we observe a double crossover for Offspring i in Table 6.

Second Attempt

If we believe that our knowledge of the alleles in Table 5 is correct, then the genes are in a hotspot region. The obtained probabilities $1-\alpha/2$ and α must correspond to the alternating pattern in Figure 7. Therefore, the order is again B, A, C or C, A, B, with A situated at equal distances from B and C. But are the probabilities consistent? In the jumping model, one could easily show that $(1-2p_d)^2=1-2p_{2d}$. Therefore, we must verify that $[1-2(1-\alpha/2)]^2=1-2\alpha+\alpha^2\approx 1-2\alpha$, so we will need α to be small enough. Note also that if α is small enough, the probability that B and C recombine is $P(AB)[1-P$

$(AC)]+[1-P(AB)]P(AC)\approx\alpha(1-\alpha/2)$ $=\alpha-\alpha^2/2\approx\alpha$, which is consistent. Moreover, the probability of a double crossover is $P(AB)P(AC)\approx(1-\alpha/2)^2=1-\alpha+\alpha^2/4$ $\approx 1-\alpha$, which is the proportion of offsprings in Table 5 that exhibit the double crossover.

A Possible Delivery Method

Here's a possible method for delivering the content of this exposition to students:

1. Describe the recombination process and genetic linkage with the uniform 1-crossover model as a hypothetical prototype, and explain how genetic mapping can be done based on observed probabilities. Introduce hotspots as an exception to the normal behavior of recombination.
2. As part of a homework assignment, ask which biological properties are satisfied by the uniform 1-crossover model and which are not. Assume that q in the "Segregation" section is $1/2$. In addition, ask the students to solve a genetic mapping problem with the biological properties in mind and determine whether hotspots are involved or not.
3. (optional) As an advanced question, ask to prove that a uniform 2-crossover model breaks the linkage property.
4. Provide solutions and briefly go over them in class. Introduce Markov chains and the jumping model.
5. As a programming assignment, ask to simulate the jumping model with various values of the parameter p and observe how the probability of recombination changes with distance. Assume that q in the "Segregation" section is $1/2$.
6. Provide solutions and wrap up by explaining some of the properties of a Markov chain through the jumping model, including the ability to model hotspots.

Uniform 2-Crossover Model

The derivation of the result is as follows:

$$P(x_1\le d \text{ and } x_2>d)$$

$$=P(x_1=1 \text{ and } x_2>d$$

$$\text{or } x_1=2 \text{ and } x_2>d$$

$$\text{or} \ldots \text{or } x_1=d \text{ and } x_2>d)$$

By the exclusivity of events, this is

$$P(x_1=1 \text{ and } x_2>d)$$

$$+P(x_1=2 \text{ and } x_2>d)$$

$$+\ldots+P(x_1=d \text{ and } x_2>d)$$

$$=P(x_1=1)P(x_2>d|x_1=1)$$

$$+P(x_1=2)P(x_2>d|x_1=2)$$

$$+\ldots+P(x_1=d)P(x_2>d|x_1=d)$$

and since $x_1=i$ means x_2 is in $\{i,\ldots,n\}$, this is

$$\frac{1}{n}P(\{d+1,\ldots,n\}|\{1,\ldots,n\})$$

$$+\frac{1}{n}P(\{d+1,\ldots,n\}|\{2,\ldots,n\})$$

$$+\ldots+\frac{1}{n}P(\{d+1,\ldots,n\}|\{d,\ldots,n\})$$

$$=\frac{1}{n}\left(\frac{n-d}{n}+\frac{n-d}{n-1}+\ldots+\frac{n-d}{n-d+1}\right)$$

$$=\frac{n-d}{n}\left(\frac{1}{n}+\frac{1}{n-1}+\ldots+\frac{1}{n-d+1}\right)$$

$$\approx\frac{n-d}{n}[\ln n-\ln(n-d)]$$

when $n-d$ is large.

Markov Transitional Probabilities

The proof is by induction where $P(x_{i+1}=l|x_i=k)=a_{kl}=P^1_{kl}$ is the base case.

$$P(x_{i+d}=l|x_i=k)=$$

$$P(x_{i+d-1}=1 \text{ and } x_{i+d}=l$$

$$\text{or } x_{i+d-1}=2 \text{ and } x_{i+d}=l|x_i=k)$$

By exclusivity of the two events, this is:

$$P(x_{i+d-1}=1 \text{ and } x_{i+d}=l|x_i=k)$$

$$+P(x_{i+d-1}=2 \text{ and } x_{i+d}=l|x_i=k)$$

Note that

$$P(E_1 \text{ and } E_2|E_3)=$$
$$P(E_1|E_3)P(E_2|E_1 \text{ and } E_3)$$

which can be derived from the definition of conditional probability. Therefore, we can rewrite the above as:

$$P(x_{i+d-1}=1|x_i=k)$$
$$P(x_{i+d}=l|x_{i+d-1}=1 \text{ and } x_i=k)$$

$$+P(x_{i+d-1}=2|x_i=k)$$
$$P(x_{i+d}=l|x_{i+d-1}=2 \text{ and } x_i=k)$$

By the Markov property this is:

$$P(x_{i+d-1}=1|x_i=k)$$
$$P(x_{i+d}=l|x_{i+d-1}=1)$$

$$+P(x_{i+d-1}=2|x_i=k)$$
$$P(x_{i+d}=l|x_{i+d-1}=2)$$

$$=P^{d-1}_{k1}P_{1l}+P^{d-1}_{k2}P_{2l}=P^d_{kl}$$

The equality before last represents the inductive step of the proof. The last equality follows immediately from the definition of matrix multiplication.

Recurrence for p_d

Knowing that $p_1=p$, we have a recurrence for p_d that we can solve, $p_d=(1-2p)p_{d-1}+p$. To obtain p_d we multiply p_{d-1} by $(1-2p)$ and add p. Here are a few attempts:

$$p_1=p$$

$$p_2 = (1-2p)p_1 + p = (1-2p)p + p = p[1+(1-2p)]$$

$$p_3 = (1-2p)p_2 + p$$

$$= (1-2p)p[1+(1-2p)] + p$$

$$= p[1+(1-2p)+(1-2p)^2]$$

$$p_4 = (1-2p)p_3 + p$$

$$= (1-2p)p[1+(1-2p)+(1-2p)^2] + p$$

$$= p[(1+(1-2p)+(1-2p)^2+(1-2p)^3]$$

We can easily generalize those attempts to obtain a geometric series:

$$p_d = p[1+(1-2p)+ (1-2p)^2 + \ldots + (1-2p)^{d-1}]$$

$$p_d = p\frac{1-(1-2p)^d}{1-(1-2p)} = \frac{1-(1-2p)^d}{2}$$

Conclusion

I am not aware of any other exposition of chromosomal crossover, recombination, genetic linkage, hotspots, and genetic mapping that takes the approach outlined herein. The approach represents a simple and modern treatment of an ancient subject, without a compromise of its scientific and mathematical integrity.

The reader should find an insightful explanation with a focus on reinforcing the ideas by exposing them in different settings. In addition, there is an attempt to introduce the reader to the process of modeling by showing what works and what doesn't. Most importantly, this should provide an early chance to convey to our students that biology is a computational science.

Disclaimer

I ignored some of the biological detail in favor of simplicity and consistency. Keep in mind, however, that in biology there is always an exception to the rule!

Further Readings

There is no explicit referencing in the text. This is intentional. I used what everyone would now consider folklore from biology, probability, and calculus. All can be found in textbooks, even elementary ones. For the interested reader, however, and in addition to any introductory texts on probability and calculus, here is a list (in alphabetical order by author) of book chapters that will provide enough background for further endeavors.

1. Gallager RG (1996) Finite State Markov Chains. In: Discrete Stochastic Processes (pp. 103–112). Norwell, MA: Kluwer Academic Publishers.
2. Hunter LE (2009) Evolution. In: The Process of Life: An Introduction to Molecular Biology (pp. 19–47). Cambridge, MA: The MIT Press.
3. Lovász L, Pelikán J, Vesztergombi K (2003) Combinatorial Probability. In: Discrete Mathematics: Elementary and Beyond (pp. 77–80, Uniform Probability). New York, NY: Springer.
4. Stein C, Drysdale RL, Bogart K (2011) Probability. In: Discrete Mathematics for Computer Scientists (pp. 276–279, Conditional Probability). Boston, MA: Pearson Education Inc. (Addison-Wesley).
5. Pevzner PA (2001) Computational Gene Hunting. In: Computational Molecular Biology: An Algorithmic Approach (pp. 1–18). Cambridge, MA: The MIT Press.

Acknowledgments

I would like to thank the QuBi (QUantitative BIology) program committee at Hunter College for their encouragement to write this article, and the reviewers for their valuable suggestions to improve it.