BioSystems 99 (2010) 42–49

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

Scaling and fractal behaviour underlying meiotic recombination

D. Waxman*, N. Stoletzki

Centre for the Study of Evolution, School of Life Sciences, University of Sussex, Brighton BN1 9QG, Sussex, UK

ARTICLE INFO

Article history:
Received 14 June 2009
Received in revised form 12 August 2009
Accepted 15 August 2009

Keywords:
Recombination
Meiosis
Population genetics
Theory
Fractals

ABSTRACT

In this paper we investigate some of the mathematical properties of meiotic recombination. Working within the framework of a genetic model with n loci, where α alleles are possible at each locus, we find that the proportion of all possible diploid parental genotypes that can produce a particular haploid gamete is $\exp[-n \log(\alpha^2/[2\alpha-1])]$. We show that this proportion connects recombination with a fractal geometry of dimension $\log(2\alpha-1)/\log(\alpha)$. The fractal dimension of a geometric object manifests itself when it is measured at increasingly smaller length scales. Decreasing the length scale of a geometric object is found to be directly analogous, in a genetics problem, to specifying a multilocus haplotype at a larger number of loci, and it is here that the fractal dimension reveals itself.

1. Introduction

The process of meiotic recombination occurs during the production of gametes by diploid organisms. It generally results in new combinations of genes arising in gametes. Meiotic recombination allows genes to experience new genetic backgrounds and gives sexual populations the ability to respond to changing environments at a significantly faster rate, and with a lower fitness cost, than asexual populations (Crow, 1994; Waxman and Peck, 1999). An analogue of meiotic recombination is also a key aspect of some evolutionarily inspired numerical optimisation procedures, namely genetic algorithms, where novel solutions to computationally complex problems can be rapidly located because of the shuffling of the "genes" in parental bit strings (Goldberg, 1989).

In the present work we investigate some of the mathematical properties of recombination that occur within one generation. We shall omit mutation, on the assumption it occurs at a sufficiently low rate that over a single generation it may be neglected.

To carry out this investigation, we could focus on informative sites within the genome, such as a particular set of SNPs (single nucleotide polymorphisms) and consider recombination events between such sites. Alternatively, we could focus on recombination between genes. For definiteness, shall couch the analysis we present in the language of *genes* and this gives us the freedom to devote some space to cases where an arbitrary number of alleles exist at a locus, rather than being limited to the maximum number of 4—for a SNP. For simplicity, we shall neglect intragenic recombination, so the haploid products of meiosis are chromosomes consisting solely of parental genes. The neglect of intragenic recombination corresponds to omission of a process of relatively low probability, while for SNPs there is no notion of intragenic recombination.

We first address the following rather general question.

> "What proportion of all possible parental genotypes can give rise to a specific gamete-type?"

Specifically, this question is concerned with the *set* of genotypes, out of *all possible* parental genotypes, that have a chance of producing a particular gamete-type. The question does not enquire into the value of the probability with which a specific gamete-type is produced by parents of a particular genotype. As a consequence, the answer does not depend on details of recombination such as the linkage map, providing crossover can occur between all loci with a non-zero probability—as we shall henceforth assume. The question enquires into a fundamental aspect of the nature of the transmission of genetic information from parents to gametes, and the degree to which the specification of a particular type of gamete constrains the genotypes of parents.

2. Answer

The haploid products of meiosis are gametes. Their haploid genotype (haplotype) specifies the set of alleles present in a gamete. We consider gametes with n loci, with α alleles possible at each locus. We can characterise haplotypes by a list, such as $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, where the allele at locus i is represented by the variable x_i which can take the values $0, 1, \ldots, \alpha-1$. Given n loci

* Corresponding author. Tel.: +44 01273 678559.
E-mail address: D.Waxman@sussex.ac.uk (D. Waxman).

0303-2647/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2009.08.006

that all have α alleles, there are a total of α^n different $\mathbf{x}$'s, i.e., α^n different haplotypes.

Consider a diploid individual whose haplotypes of maternal and paternal origin are $\mathbf{x}$ and $\mathbf{y}$, respectively. We write the *ordered genotype* of this individual as $(\mathbf{x}, \mathbf{y})$. Let $f(n, \alpha)$ denote the proportion of all *ordered parental genotypes* that can give rise to a particular gamete-type. It may be verified that $f(n, \alpha)$ is independent of the particular gamete-type selected. We calculate $f(n, \alpha)$ by allowing $\mathbf{x}$ and $\mathbf{y}$ to independently range over all α^n possible haplotypes, so $(\mathbf{x}, \mathbf{y})$ covers all possible ordered genotypes. We note that:

(i) There are a total of $\alpha^n \times \alpha^n = \alpha^{2n}$ different ordered parental genotypes.
(ii) In the absence of mutation, it follows, from a general counting argument (see Appendix A), that there are a total of $(2\alpha - 1)^n$ ordered parental genotypes that have a finite probability of producing a single gamete-type.

It directly follows from (i) and (ii) that the proportion of all ordered parental genotypes that can give rise to a particular gamete-type is

$$f(n, \alpha) = \left(\frac{2\alpha - 1}{\alpha^2}\right)^n = \exp\left[-n \log\left(\frac{\alpha^2}{2\alpha - 1}\right)\right]. \quad (1)$$

The function $f(n, \alpha)$ indicates the way the proportion of parental genotypes (that can produce a given gamete-type) *scales* with the number of loci, n. As the number of loci increases, the proportion of all such parental genotypes decreases exponentially with n. For example, if there are just 2 alleles at each locus ($\alpha = 2$), then the relevant proportion of parental genotypes is $f(n, 2) = (3/4)^n \simeq \exp(-0.28768n)$.

The answer to the question posed in Section 1 also contains information about the *number* of gamete-types that different parental genotypes can produce. It is evident that there is variation in this number, since the two extreme cases, where parental genotypes that are either homozygotic at all loci or heterozygotic at all loci, lead to very different numbers of gamete-types that can be produced (1 or 2^n, respectively). If we consider only the *mean number* of different types of gamete that a parental genotype can produce, which we write as $\nu(n, \alpha)$, then there is a very simple result for this quantity. Since the proportion of all ordered parental genotypes that can give rise to a particular gamete-type, $f(n, \alpha)$, also equals the probability that a randomly picked parental genotype can produce a gamete of a specific type, summing $f(n, \alpha)$ over all α^n different gamete-types yields the mean number of gamete-types that a single parental genotype can produce, namely $\nu(n, \alpha) = f(n, \alpha)\alpha^n$, i.e.,

$$\nu(n, \alpha) = \left(\frac{2\alpha - 1}{\alpha}\right)^n = 2^n\left(1 - \frac{1}{2\alpha}\right)^n. \quad (2)$$

The final form for $\nu(n, \alpha)$ in Eq. (2) indicates the degree to which $\nu(n, \alpha)$ is suppressed below the value 2^n, which is the number of gamete-types that a parental genotype can produce, when heterozygotic at all loci.

More generally, we can calculate the probability that a randomly picked parental genotype is heterozygotic at just m loci, and hence can produce 2^m different gamete-types. From this probability distribution it is possible to calculate statistics, other than just the mean number, that are associated with the number of gamete-types that different parental genotypes can produce. For a randomly picked parental genotype we find there is a binomially distributed number of heterozygotic loci (see Appendix B):

$$\text{Prob}(m \text{ parental loci are heterozygotic}) = \binom{n}{m}(1 - \alpha^{-1})^m(\alpha^{-1})^{n-m} \quad (3)$$

where $\binom{n}{m}$ denotes a binomial coefficient. This binomial distribution has parameters n and $(1 - \alpha^{-1})$ and hence the expected number of heterozygotic loci is $n(1 - \alpha^{-1})$. It may be directly verified from Eq. (3) that the mean value of 2^m does coincide with the mean number of gamete-types, $\nu(n, \alpha)$, given in Eq. (2). Furthermore, the variance in the number of gamete-types produced is $4^n[(1 - \frac{3}{4\alpha})^n - (1 - \frac{1}{2\alpha})^{2n}]$.

Beyond the results of Eqs. (1)–(3), we note that there is additional mathematical structure associated with meiosis that underlies the question in Section 1. We now establish this additional structure.

3. Additional Mathematical Structure

The simplest way to see the additional structure is to produce a plot that indicates whether a given type of gamete can or cannot be produced by a particular parent.

Previously, we have characterised each haplotype by a list $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ where x_i indicates the allele at locus i (and x_i is an integer in the range 0 to $\alpha - 1$). By associating a unique numerical label with each haplotype, we are able to place the haplotypes in a definite order. One such ordering scheme is obtained by viewing $\mathbf{x}$ as the n digit representation, in base α, of an integer in the range 0 to $\alpha^n - 1$. The resulting numerical label, also an integer, is directly determined from the alleles present at all n loci. As an explicit example, consider the case of $n = 4$ loci where there are $\alpha = 2$ possible alleles at each locus. Because $\alpha = 2$, a particular haplotype, such as $(1, 1, 0, 1)$, is viewed as the *binary* representation of an "ordering label," here 13 (since $13 = 1 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0$) and this number can serve as the unique label of the haplotype. For general α, the numerical label associated with haplotype $\mathbf{x}$ is written X. It is given by $X = \sum_{m=1}^{n} \alpha^{n-m} x_m$ and runs from 0 to $\alpha^n - 1$.

Further aspects to the problem can then be expressed in terms of a new quantity $A_{X,Y}(G)$ where X, Y are the numerical labels of the parental haplotypes making up the ordered genotype $(\mathbf{x}, \mathbf{y})$ and G is the numerical label of the gamete-type, $\mathbf{g}$. We require $A_{X,Y}(G)$ to take the value 1 when a gamete of type $\mathbf{g}$ can be produced by a parent with ordered genotype $(\mathbf{x}, \mathbf{y})$, and to take the value 0 if a gamete of type $\mathbf{g}$ cannot be produced by such a parent. We give one possible mathematical representation of $A_{X,Y}(G)$ in Appendix C.

For a particular G, we view $A_{X,Y}(G)$ as the X, Y element of a matrix, $\mathbf{A}(G)$, of size $\alpha^n \times \alpha^n$, which is symmetric: $A_{X,Y}(G) = A_{Y,X}(G)$. When we wish to emphasise that the number of loci specifying a gamete haplotype is n, we will write $\mathbf{A}(G)$ as $\mathbf{A}(G; n)$.

Those elements of the matrix $\mathbf{A}(G)$ which are non-zero represent ordered parental genotypes that have a finite probability of giving rise to a gamete-type with numerical label G.

To illustrate the form of $\mathbf{A}(G)$, we consider the case of 2 loci with 2 alleles at each locus ($n = 2$ and $\alpha = 2$). There are then 4 possible haplotypes, which we write as $\mathbf{x} = (0, 0), (0, 1), (1, 0), (1, 1)$. These have the numerical labels $X = 0, 1, 2$ and 3 associated with them (i.e., the 4 $\mathbf{x}$'s are the binary representation of these numbers). Since there are 4 different haplotypes, there are $4 \times 4 = 16$ different ordered haplotype pairs i.e., 16 ordered parental genotypes. For the particular gamete-type, with numerical label $G = 0$ (corresponding to $\mathbf{x} = (0, 0)$), 9 of the 16 parental genotypes can give rise to gametes of this type, either by direct transmission of haplotypes or by recombination. However, those genotypes with labels $(X, Y) = (1, 1), (1, 3), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3)$ cannot

produce gametes with two 0 alleles and hence have *zero probability* of producing $G = 0$ type gametes. As a consequence, the matrix $\mathbf{A}(0)$ has zeros at the elements corresponding to these genotypes and is given by

$$\begin{pmatrix} A_{0,3}(0) & A_{1,3}(0) & A_{2,3}(0) & A_{3,3}(0) \\ A_{0,2}(0) & A_{1,2}(0) & A_{2,2}(0) & A_{3,2}(0) \\ A_{0,1}(0) & A_{1,1}(0) & A_{2,1}(0) & A_{3,1}(0) \\ A_{0,0}(0) & A_{1,0}(0) & A_{2,0}(0) & A_{3,0}(0) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}. \quad (4)$$

(we plot X and Y values along Cartesian axes).

An interesting pattern becomes apparent when we produce a plot of $\mathbf{A}(G)$ that is equivalent to Eq. (4), for an appreciable value of α^n. We plot dots only at the coordinate pairs $(X/\alpha^n, Y/\alpha^n)$ where the matrix elements $A_{X,Y}(G)$ are non-zero. Each such dot represents a parental genotype (composed of haplotypes with numerical labels X and Y), that has a finite probability of producing a gamete of type G. In Fig. 1 we give such plots for several different cases of interest.

The nested set of triangular "holes" that is present in Fig. 1a and b or the related structure in (c), are highly suggestive of self-similar fractal structures (Mandelbrot, 1982). Indeed, when the number of loci, n, tends to infinity, the number of elements of $\mathbf{A}(G)$ tends to infinity and the quantities X/α^n and Y/α^n become continuous variables running from 0 to 1. Fig. 1a–c then becomes fractals on the unit square and for general α the fractal dimension is

$$D(\alpha) = \frac{\log(2\alpha - 1)}{\log(\alpha)}. \quad (5)$$

For Fig. 1a, where $\alpha = 2$, the fractal dimension is $\log(3)/\log(2) \simeq 1.5849$. This result may be verified by determining how the filled area of the square of Fig. 1a changes, when first measured at a given linear spatial scale, and then measured-again on a smaller linear spatial scale. Thus to determine the fractal dimension associated with Fig. 1a, we first pave the figure with 45°, 45°, 90° triangles, of a given linear scale, such as the length of the two shorter sides of the triangle. As a concrete example we take this length to be 1/2 and the paving of Fig. 1a with triangles of this linear scale is illustrated in Fig. 2a.

Measuring the filled area of Fig. 1a on this linear spatial scale means:

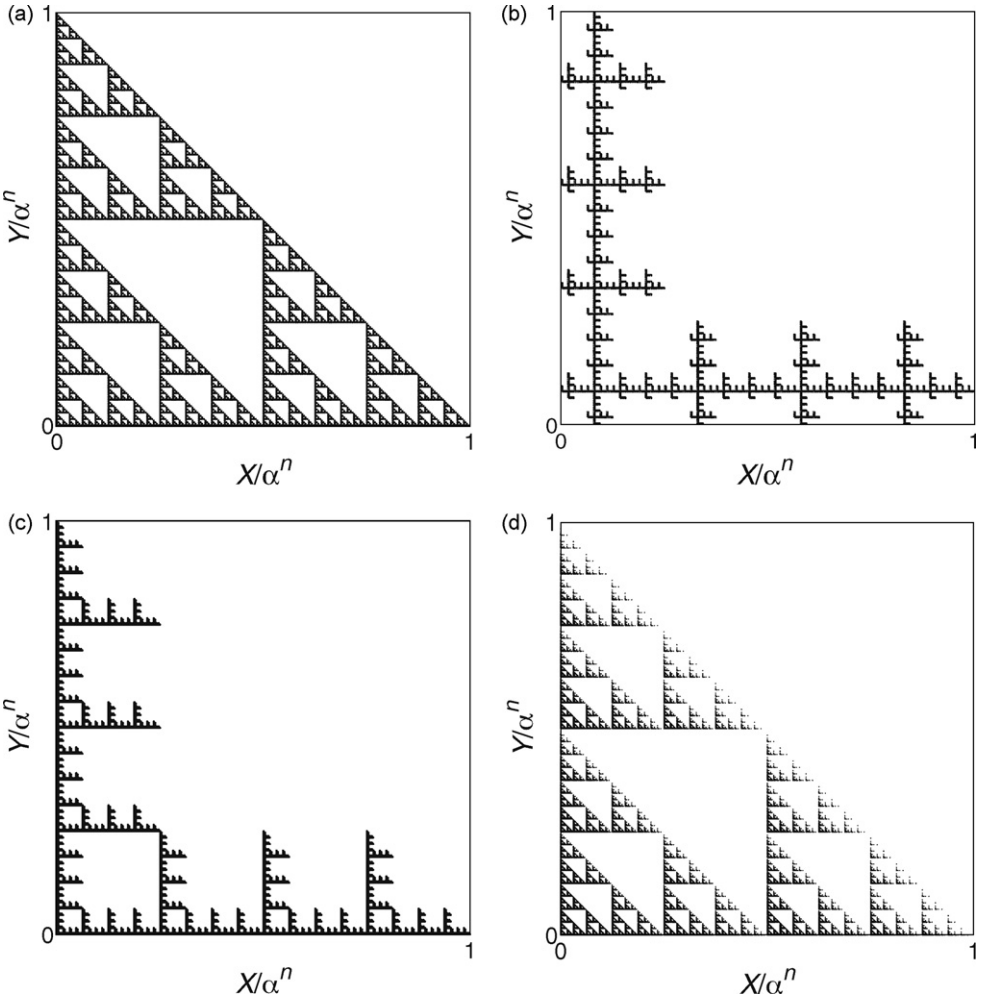

Fig. 1. The quantity $A_{X,Y}(G)$ takes the value 0 or 1, depending whether an individual, composed of haplotypes with numerical labels X and Y, has a zero or non-zero probability of producing gametes of type G. For n loci, with α alleles at each locus, the labels X, Y and G are integers running from 0 to $\alpha^n - 1$. We plot dots at the scaled coordinate values $(X/\alpha^n, Y/\alpha^n)$ where $A_{X,Y}(G)$ has the value of unity. Each such dot corresponds to a parental genotype that has a non-zero probability of giving rise to gamete-types with numerical label G. In both (a) and (b) we have taken $n = 10$ and $\alpha = 2$. The fraction of all parental genotypes that can produce any particular gamete-type is $(2\alpha - 1)^n/\alpha^{2n} = \exp(-n \log[\alpha^2/(2\alpha - 1)])$ and for (a) and (b) this fraction is $59,049/1,048,576 \simeq 5.6 \times 10^{-2}$. In (a) the specified gamete-type has numerical label $G = 0$, while in (b) we have chosen $G = 333$. The differences between (a) and (b) illustrate that the pattern obtained depends on the target gamete. For (c) we have used $n = 6$, $\alpha = 4$ and $G = 0$. It is important to ensure that the structure present e.g., in (a) possess a degree of robustness, and hence persists when recombination events of low probability are neglected. To establish this, we have produced (d), where all parameters are identical to those of (a), but dots are only plotted in the figure if the corresponding gametes can, under free recombination, be produced with probability larger than 1%. The structure produced is evidently a very similar to that of (a).

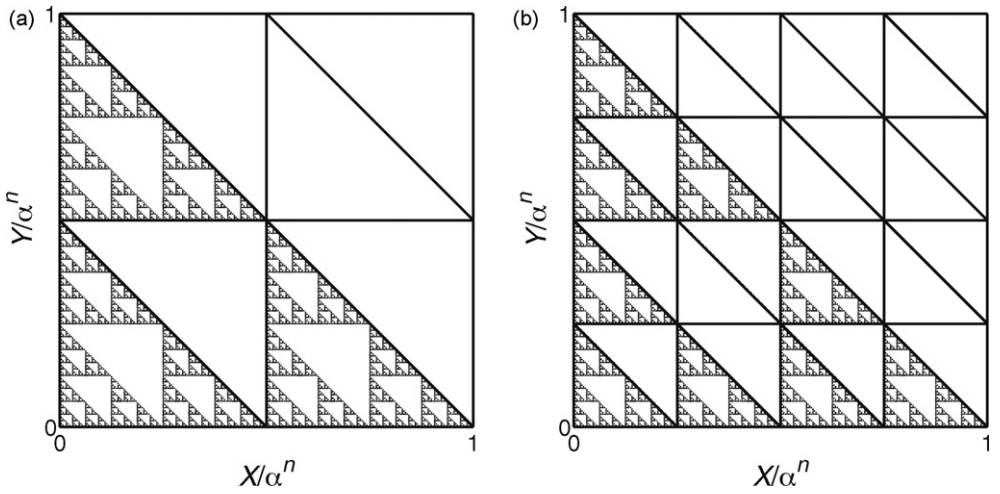

Fig. 2. In (a), the unit square, containing Fig. 1a, is paved with 45°, 45°, 90° triangles, whose shorter sides have length 1/2. At this linear spatial scale, a total of 3 out of 8 triangles cover any dots. The area covered by dots at the linear spatial scale of 1/2 is thus 3/8. In (b), the unit square, containing Fig. 1a, is paved with 45°, 45°, 90° triangles, whose shorter sides have length 1/4. At this linear spatial scale, a total of 9 out of 32 of the triangles cover any dots. The area covered by dots, at the linear spatial scale of 1/4 is thus 9/32.

(i) If a given triangle covers *any* dots, then the *entire area* of the triangle counts toward the filled area of Fig. 1a. This is illustrated in Fig. 2a.
(ii) If a triangle covers *no* dots, then it makes *zero* contribution to the filled area of Fig. 1a.

When we then halve the linear scale of the triangles—by taking the length of the two shorter sides to now be 1/4, we find this halving does not produce a 4-fold increase in the number of triangles required to cover the filled area, but only a 3-fold increase. This is due to the fact that the set of half-sized triangles that pave the "holes" in Fig. 1a, and cover no dots, make no contribution to the area. The paving of Fig. 1a with triangles whose shorter sides are of length 1/4 is illustrated in Fig. 2a. Thus the number of half-sized triangles that cover any dots in Fig. 2b is not a factor of 2^2 larger than those of Fig. 2a (as would follow in a normal, non-fractal, geometry) but rather $2^{D(2)}$, where $D(2)$ is the fractal dimension. Equating $2^{D(2)}$ to 3 leads to the result in Eq. (5) for the special case $\alpha = 2$.

This fractal nature is intimately related to the binomial sampling property of recombination and, indeed, in a mathematical context, the fractal nature of binomial coefficients and Pascal's triangle has been reported elsewhere (Wolfram, 1984), as it has the famous Sierpinski gasket (Mandelbrot, 1982), and both of these have the fractal dimension given in Eq. (5), when $\alpha = 2$. It appears there are very close mathematical connections with the result of Fig. 1a.

In Fig. 1c, where $\alpha = 4$, we use an identical procedure to determine the fractal dimension. That is, we first consider the number triangles that cover the filled area, when, e.g., their shorter sides have length 1/4, and then compare this with the number required when their shorter sides have length 1/16. This leads to a fractal dimension of $\log(7)/\log(4) \simeq 1.4037$.

The general result for the fractal dimension, given in Eq. (5) was inferred from considering figures analogous to Fig. 1a, for a number of different values of α. As we show below, in Section 5, a numerical determination of the fractal dimension fully agrees with the result in Eq. (5).

4. Ordering of Haplotypes

We note that the appearance and detailed properties of the fractal-like structure underlying meiosis (e.g., as in Fig. 1a) depends on the scheme used to *order* haplotypes. For the purposes of classification, we shall refer to the ordering scheme in Section 3, with $\alpha = 2$ as Scheme 1 and describe it as *deterministic*, for reasons that will shortly become clear.

4.1. Ordering Scheme 2: distance + deterministic

Consider an scheme of ordering haplotypes, where the numerical label of a haplotype contains *some* information about its genetic distance from a reference haplotype, but which also uses information about the particular alleles present at each locus. We consider such a scheme in the case of $\alpha = 2$ alleles per locus and let

$$\mathbf{x} = (0, 0, \ldots, 0) \tag{6}$$

be a reference n locus haplotype. Then the n haplotypes with a single "1" allele, which include $(1, 0, 0, \ldots, 0)$, $(0, 1, 0, \ldots, 0), \ldots (0, 0, 0, \ldots, 1)$, are all collected together into a group that is distance 1 from the reference haplotype. Next, the $n(n-1)/2$ haplotypes with two "1" alleles, such as $(1, 1, 0, \ldots, 0)$ are all collected together into a group that is distance 2 from the reference haplotype, etc. In general, the number of haplotypes a distance m from the reference haplotype are given by the binomial coefficient $\binom{n}{m}$. An ordering scheme that incorporates *both* a measure of genetic distance and *some* information about the particular allele residing at each locus, is defined by the following:

(i) Associate the label $X = 0$ with the reference haplotype, Eq. (6).
(ii) Associate the next n labels (i.e., $X = 1, 2, \ldots, n$) with the haplotypes a distance 1 from the reference haplotype, and place them in the order following from interpreting each haplotype, such as $(0, 0, \ldots, 0, 1, 0)$, as the *binary* representation of a number.
(iii) Associate the next $n(n-1)/2$ labels (i.e., $X = n+1, n+2, \ldots, n+n(n-1)/2$) with haplotypes a distance 2 from the reference haplotype, again placing them in the order following from interpreting each haplotype as the binary representation of a number.
(iv) ....

In this way we arrive at a scheme of ordering haplotypes that fits with the intuitive notion that their labels reflect, to some degree, a real attribute of haplotypes, namely their genetic distance from a reference haplotype. We term this a *distance + deterministic* scheme.

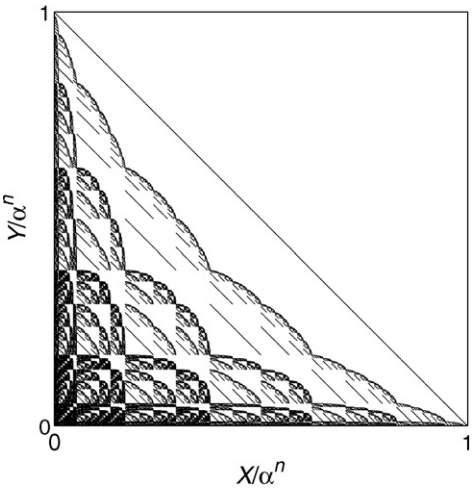

Fig. 3. The matrix $\mathbf{A}(0)$, which applies for $n = 10$ and $\alpha = 2$, and has elements $A_{X,Y}(0)$, is determined by using Scheme 2: a *distance + deterministic* scheme of ordering haplotypes, where they are grouped together, according to their distance from a reference haplotype, as measured in allelic changes. However, within a group of haplotypes that are a given genetic distance from the reference haplotype, the haplotypes are ordered by interpreting them as the binary representation of a number, as outlined in the text. In the figure, we have plotted dots at the scaled coordinate values X/α^n and Y/α^n where $A_{X,Y}(0)$ has the value of unity. There are the same number of dots as in Fig. 1a and b and there are some similarities of with both Fig. 1a and b.

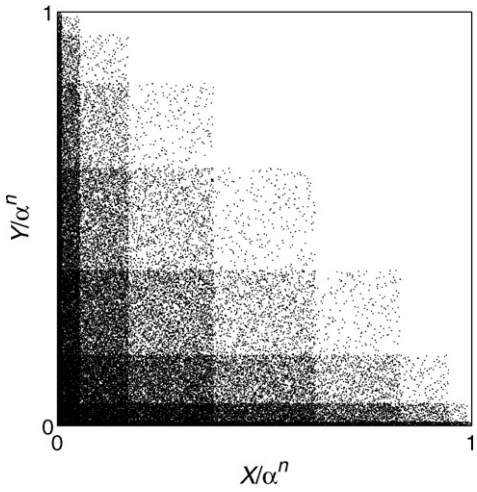

Fig. 4. The matrix $\mathbf{A}(0)$, which applies for $n = 10$ and $\alpha = 2$ and has elements $A_{X,Y}(0)$, is determined by using Scheme 3, a *distance + random* scheme of ordering, where haplotypes are grouped together according to their distance from a reference haplotype, as measured in allelic changes. However, within a group of haplotypes that are a given genetic distance from the reference haplotype, the haplotypes are randomly ordered, as outlined in the text. In the figure, we have plotted dots at the scaled coordinate values X/α^n and Y/α^n where the "distance labelled" $A_{X,Y}(0)$ has the value of unity. There are the same number of dots as in Fig. 1a and b. Although some structure is visible, it lacks the connectivity or organisation of a fractal, or fractal-like object.

The consequences of such a scheme is plotted in Fig. 3, with the resulting figure having some definite similarities to Fig. 1.

4.2. Ordering Scheme 3: distance + random

An alternative scheme for ordering haplotypes may be constructed that also contains information about the genetic distance of a haplotype from a reference haplotype, but unlike Scheme 2, discards information about the alleles present at each locus. Again we consider $\alpha = 2$ alleles per locus, and measure genetic distances from the n locus haplotype of Eq. (6). The n haplotypes with a single "1" allele are again all collected together into a group that is distance 1 from the reference haplotype, and similarly the $n(n-1)/2$ haplotypes with two "1" alleles, are all collected together into a group a distance 2 from the reference haplotype, etc. An ordering scheme that just incorporates a measure of distance and hence contains both ordered and random components is defined by the following:

(i) Associate the label $X = 0$ with the reference haplotype, Eq. (6)
(ii) Associate the next n labels (i.e., $X = 1, 2, \ldots, n$) with the haplotypes a distance 1 from the reference haplotype, and make a *random* assignment of the labels $1, 2, \ldots, n$ to the haplotypes *within* this group.
(iii) Associate the next $n(n-1)/2$ labels (i.e., $X = n+1, n+2, \ldots, n+n(n-1)/2$) with haplotypes a distance 2 from the reference haplotype, again making a *random assignment* of the labels with haplotypes within this group,
(iv) ....

The resultant labelling scheme, which incorporates only genetic distance, leads to Fig. 4. We term this a *distance + random* scheme.

4.3. Ordering Scheme 4: random

A scheme of ordering that does not reflect any information about genetic distance can be obtained by randomly permuting the labels associated with haplotypes. Thus if the haplotypes $(\mathbf{x}, \mathbf{y}, \mathbf{z}, \ldots)$ were, in the original binary scheme (Scheme 1), given the numerical labels $(X, Y, Z, \ldots)$ then under the random scheme proposed here, the haplotypes $(\mathbf{x}, \mathbf{y}, \mathbf{z}, \ldots)$ are given labels that are a random permutation of the numerical labels $(X, Y, Z, \ldots)$. Fig. 5 contains an example of one such ordering scheme, where the numerical labels of haplotypes used in Fig. 1a have been randomly permuted. We term this a *random* scheme. The resulting figure is significantly different to Fig. 1a.

5. Numerical Investigation of the Fractal Character

The fractal dimension of a structure can be determined by investigating its "box counting dimension" (see e.g., Falconer, 1990). This

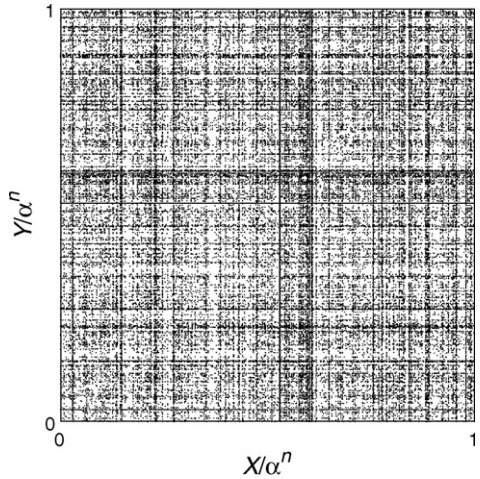

Fig. 5. The matrix $\mathbf{A}(0)$, with elements $A_{X,Y}(0)$, is obtained by assigning numerical labels in an arbitrary order to different haplotypes. This is Scheme 4, a *random* scheme of ordering. For the case $n = 10$ and $\alpha = 2$, the numerical labels used in Fig. 1a have been randomly permuted. In the figure, we have plotted dots at the scaled coordinate values X/α^n and Y/α^n where $A_{X,Y}(0)$ has the value of unity. While there are the same number of dots as in Fig. 1a, the dots plotted in this figure are not clustered together. Indeed, the various dots in Fig. 1a lie in a connected cluster, while in this figure the dots are approximately uniformly distributed, with an element of the matrix being non-zero with probability $f(n, \alpha) = [(2\alpha - 1)/\alpha^2]^n$. For the figures in question, $n = 10$ and $\alpha = 2$, hence $f(n, \alpha) \simeq 0.06$ in each case.

is determined from the number of elementary "boxes" of side ε, namely $N(\varepsilon)$, that cover the fractal. If the structure has fractal properties, possibly only over a range of box sizes, then over this range we have the relation

$$N(\varepsilon) = \varepsilon^{-D_{BC}} N(1) \tag{7}$$

where D_{BC} is the "box counting" dimension. If the relation of Eq. (7) holds for arbitrarily small ε, the structure is a fractal and the fractal dimension coincides with D_{BC}.

In the case at hand, we consider structures that may have fractal properties on small length scales (small ε) but not necessarily indefinitely small values. In the figures of this paper, such as Fig. 1, they have been plotted so that the smallest length scale (i.e., the closest separation of adjacent dots) is α^{-n}. In this case we would expect that if a fractal-like structure is present, then $\log(N(\varepsilon))/\log(1/\varepsilon)$ would approach a constant value (namely D_{BC}) as ε becomes small, but not smaller than α^{-n}. We have numerically investigated the degree to which $\log(N(\varepsilon))/\log(1/\varepsilon)$ does approach a constant as the length, ε, is allowed to become small, and hence the extent to which a fractal-like structure is present. We used Matlab software to perform this numerical investigation (Moisy, 2008).

In Fig. 6 we present a plot of the quantity

$$R(\varepsilon) = \frac{\log_2 N(\varepsilon) - \log_2 N(1)}{\log_2(1/\varepsilon)} \tag{8}$$

against $\log_2(\varepsilon)$. If the relation in Eq. (7) holds for small ε then $R(\varepsilon)$ will approach a constant value at large negative values of $\log_2(\varepsilon)$. We observe, in Fig. 6, that schemes of ordering haplotypes with a so-called deterministic component (i.e., schemes that take into account the particular alleles present at a locus), do lead to $R(\varepsilon)$ approaching a constant value at large negative $\log_2(\varepsilon)$. By contrast, other schemes do *not* lead to $R(\varepsilon)$ approaching a constant value at large negative $\log_2(\varepsilon)$. It is plausible that only haplotype ordering schemes with a deterministic component will expose any fractal-like structures underlying meiotic recombination.

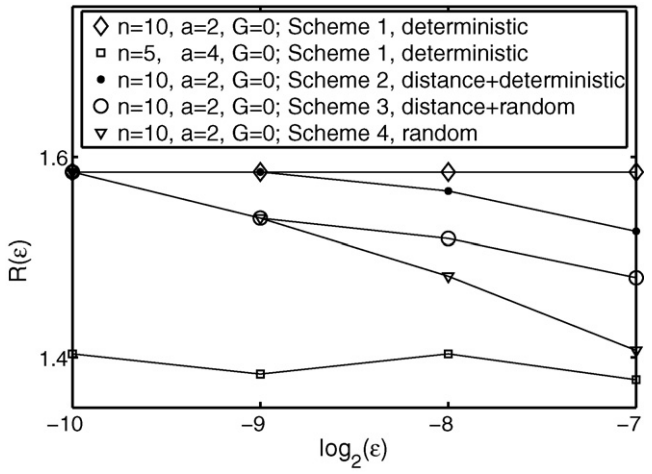

Fig. 6. In this figure we present results of a numerical investigation of the "box counting dimension" of different haplotype labelling schemes. The box counting dimension associated with non-zero values of $A_{X,Y}(G)$ was determined for the four labelling schemes considered in this work. With ε the length scale associated with an elementary box, the quantity $R(\varepsilon)$ of Eq. (8) was plotted against $\log_2(\varepsilon)$. Only ordering Schemes 1 and 2 lead to $R(\varepsilon)$ approaching a constant value for small length scales, i.e., large negative values of $\log_2(\varepsilon)$. We note that the values of $R(\varepsilon)$ for these schemes, at large negative $\log_2(\varepsilon)$, are very close to the values of the fractal dimension $D(\alpha)$ of Eq. (5).

Table 1
Connects notions associated with recombinant genetics and fractals in the plane.

Recombinant genetics	Fractal in the plane
Using alleles at n loci to specify a gamete	Using a linear scale, L, to measure a fractal
Determining the proportion of all parental genotypes that can produce a given n locus gamete haplotype	Determining the area covered by a fractal, when the fractal is measured on a linear scale of L
Determining the proportion of all parental genotypes that can produce a given $n+1$ locus gamete haplotype	Determining the area covered by a fractal, when the fractal is measured on a linear scale of L/α

6. Analogy Between Recombinant Genetics and Fractals

A fractal is a geometric object composed of an infinite number of points. We would, apparently, have a fractal structure in e.g., Fig. 1 if there were an infinite number of points in the figure. However, this would require an infinite number of loci, which does not hold in our problem or, indeed, in any problem except idealised, but useful theoretical models, such as the Fisher–Bulmer infinitesimal model (Fisher, 1918; Bulmer, 1980). How can the connection be made between fractal results such as Eq. (7) and a genetics problem with a finite number of loci, $n\,(<\infty)$? We argue that the connection arises from the scaling property given in Eq. (1).

Let us imagine that we have a population of organisms with n_{total} loci. Given the large number of loci in a gamete, it is assumed impractical to determine the alleles at all n_{total} loci. By determining the haplotype of a gamete at a smaller number of loci, say n, which is small compared with n_{total} ($n \ll n_{\text{total}}$), we have an incomplete description, i.e., we work at a lower level of genetic resolution. In Table 1 we show an analogy between recombinant genetics and fractals in the plane.

To see how this analogy ties together and employs the fractal dimension $D(\alpha)$ of Eq. (5), we adopt the following line of reasoning.

1. We begin by incompletely identifying gamete haplotypes, by specifying the alleles at just n loci (a haplotype has more than n loci).
2. For a particular gamete-type, G, the matrix $\mathbf{A}(G) \equiv \mathbf{A}(G;n)$, associated with haplotypes specified at n loci, is of size α^n by α^n and we view it as being composed of $\alpha^n \times \alpha^n$ small squares, with each square occupied by one element of the matrix. The squares are either filled or empty, depending whether the corresponding element of the matrix is 1 or 0 (we term this a "filling rule").
3. We regard the matrix $\mathbf{A}(G;n)$, for any n, as having a fixed area of unity. Given the *filling rule* of the previous point, the proportion of the area that is filled represents the proportion of all possible genotypes that can produce gametes of type G. Furthermore, from this viewpoint, each element of $\mathbf{A}(G;n)$ takes up an area of $1/(\alpha^n \times \alpha^n)$.
4. Increasing the number of loci to specify a haplotype, by going from n to $n+1$, corresponds to using the matrix $\mathbf{A}(G;n+1)$ in place of the matrix $\mathbf{A}(G;n)$. The total number of elements in the matrix $\mathbf{A}(G;n+1)$, is larger than that of $\mathbf{A}(G;n)$ by a factor of $(\alpha^{n+1} \times \alpha^{n+1})/(\alpha^n \times \alpha^n) = \alpha^2$. This factor is equivalent to *decreasing* the area of the square associated with any element of the matrix $\mathbf{A}(G;n+1)$, relative to that of $\mathbf{A}(G;n)$, by a factor of α^2. It is also equivalent to decreasing the *length* of the side of such squares associated with $\mathbf{A}(G;n)$ by a factor of α.
5. Additionally, changing n to $n+1$ also results in the number of *non-zero* elements of the matrix $\mathbf{A}(G;n+1)$ being larger, by a factor of $2\alpha - 1$, than the number of *non-zero* elements of $\mathbf{A}(G;n)$ (see Section 2).
6. It follows that when the number of loci specifying a haplotype is changed from n to $n+1$, the ratio of "the number of non-zero

elements of $\mathbf{A}(G; n+1)$" to "the total number of elements of $\mathbf{A}(G; n+1)$" differs from the corresponding ratio for $\mathbf{A}(G; n)$ by a factor of $(2\alpha-1)/\alpha^2 = \alpha^{D(\alpha)-2}$, where $D(\alpha)$ is the fractal dimension given in Eq. (5). The quantity $\alpha^{D(\alpha)-2}$ is precisely the factor by which the area of a fractal of dimension $D(\alpha)$ changes when the length scale, on which the fractal is measured, is decreased by a factor of α (i.e., when the original length scale of measurement, say L, is decreased to the length scale L/α).

7. Lastly, the ratio of non-zero elements of $\mathbf{A}(G; n)$ to the total number elements of this matrix is the proportion of all parental genotypes that can give rise to a particular gamete-type, and has already been denoted by $f(n, \alpha)$. From point (6) it follows that this proportion obeys the difference equation $f(n+1, \alpha) = \alpha^{D(\alpha)-2} f(n, \alpha)$. It may be explicitly verified that the result for $f(n, \alpha)$ can be written $f(n, \alpha) = \alpha^{(D(\alpha)-2)n}$ and is a solution of this difference equation.

We thus see how the intimate relation between the fractal dimension associated with the genetics problem when the number of loci is infinite, and the proportion of n-locus genotypes that can produce a specific gamete-type.

7. Discussion

In this work, we have investigated some of the mathematical properties of meiotic recombination. We have determined the proportion of all parental genotypes which can give rise to a particular gamete-type. This required classifying all parental genotypes into one of two categories; those that are capable of giving rise to the particular gamete-type and those that are not. Plots showing which parents can produce a certain type of gamete reveal a fractal-like structure that underlies meiotic recombination (see Figs. 1 and 3). Recombination is a well known phenomenon of extremely wide occurrence. This work provides a non standard (i.e., fractal) viewpoint of recombination. It is possible this different viewpoint and the associated scaling behaviour, such as that exhibited in Eq. (1), has implications for subjects where recombination plays an important role, such as evolutionary dynamics, genetics or genetic algorithms.

Applications of the results of the present work may be an intermediate stage in a calculation. In such a case, additional information, such as recombination fractions, mutation rates and a description of a population, may be required. However the results may also be used to directly address questions of the "yes–no" type, i.e., questions concerned with whether a class of genotypes is present or absent in a population. This is distinct from questions concerned with actual numbers or frequencies.

For questions of the "yes–no" type, consider a diploid sexual population, where a set of n loci are informative (i.e., variable across the population). If the n loci are in linkage equilibrium, and sufficiently polymorphic, then in a very large population, virtually all possible n-locus genotypes will be present in the population. If we are in possession of a particular gamete whose n-locus haplotype has been determined, then the result given in Eq. (1), applies for the proportion of all distinct genotypes that are present in the population, that could, in principle, have given rise to this gamete. Such information restricts the parentage of the gamete to a subclass of the full population. This information gives the knowledge that a major fraction of all genotypes are excluded from giving rise to the gamete and is especially significant if the gamete is distinguished or significant in some way.

We can also consider the situation where n recognition proteins on an egg have to be matched by proteins in a sperm, in order that the sperm can penetrate the egg membrane and fertilize the egg. If the matching mechanism restricts the genotype of the sperm to match the genotype of the egg, at n loci, then only a fraction $f(\alpha, n)$ (Eq. (1)) of all male genotypes can provide sperm of a given type and are "compatible" with the egg (cf. Gavrilets and Waxman, 2002), i.e., are capable of fertilizing it. Furthermore, increasing the number of loci involved in the recognition mechanism by unity: $n \to n+1$, results in a reduction in the fraction of "compatible" male genotypes by a factor of $(2\alpha-1)/\alpha^2$. Thus specificity of the recognition mechanism puts constraints on the genotypes that are compatible with a given egg and the proportion $f(\alpha, n)$ may be a significant component of fitness associated with an egg. Implications of modifications to this recognition mechanism, by changing the number of loci, or alleles possible, can be explicitly seen by the form of $f(\alpha, n)$ given in Eq. (1).

Overall, the results obtained in this work indicate a general property of the transmission of genes in meiotic recombination, and are independent of essentially all details of: recombination, the distribution of parental genotypes in a population, and the distribution of gamete numbers produced. The results thus give an insight into some general properties of meiotic recombination.

Acknowledgements

We thank H.-P. Yang for helpful discussions on this work and an anonymous reviewer for suggestions that have improved the manuscript. D.W. acknowledges support from the Leverhulme Trust and NS from the Biotechnology and Biological Sciences Research Council.

Appendix A.

In this Appendix we show that a total of $(2\alpha-1)^n$ ordered parental genotypes have a non-zero probability of producing any particular gamete haplotype.

To establish this result, we first note that a haplotype is determined by specifying which of the α alleles are present at each of the n loci. Suppose that at a particular locus, in a specified gamete haplotype, the particular allele is 0 (identical results follow if 0 is replaced by any other allele, i.e., by $1, 2, \ldots, \alpha-1$). The possible genotypes *associated with any locus*, including the particular one of interest, can be written (x, y) where both x and y are able to independently take the values $0, 1, 2, \ldots, \alpha-1$. At the particular locus of interest, only those genotypes that contain one or more 0 allele's can give rise to a 0 allele in a gamete. These genotypes are $(0, 0)$ and also $(0, a)$, $(a, 0)$ with $a = 1, 2, \ldots, \alpha-1$. It follows that there are $1 + 2(\alpha-1)$, i.e., $2\alpha-1$ such genotypes. Generally, it follows that at any locus, only $2\alpha-1$ of the α^2 genotypes contain a specific allele, and hence can produce a gamete-type containing the specific allele at the locus in question. Accordingly, the total number of ordered parental genotypes that have a non-zero probability of producing a particular gamete-type is $(2\alpha-1)^n$.

Appendix B.

In this Appendix we derive the probability that a randomly picked parental genotype has m heterozygotic loci.

For n loci, that each have α alleles, the total number of ordered parental genotypes is α^{2n}. For an ordered genotype $(\mathbf{x}, \mathbf{y})$, the number of heterozygotic loci is $M(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{n} [1 - \delta(x_j, y_j)]$ where $\delta(a, b)$ is a Kronecker delta ($\delta(a, b)$ has the value of unity when $a = b$ and is zero otherwise). Thus the probability of a randomly picked parental genotype having m heterozygotic loci is

$$\text{Prob}(m \text{ parental loci are heterozygotic}) = \frac{1}{\alpha^{2n}} \sum_{\mathbf{x}} \sum_{\mathbf{y}} \delta(m, M(\mathbf{x}, \mathbf{y})) \tag{9}$$

where the $\mathbf{x}$ and $\mathbf{y}$ sums separately cover all α^n haplotypes.

We can write $\delta(m, M(\mathbf{x}, \mathbf{y})) = \int_0^{2\pi} \frac{d\lambda}{2\pi} e^{-i\lambda m} e^{i\lambda M(\mathbf{x},\mathbf{y})}$ and hence

$$\begin{aligned} \sum_{\mathbf{x}} \sum_{\mathbf{y}} \delta(m, M(\mathbf{x}, \mathbf{y})) &= \int_0^{2\pi} \frac{d\lambda}{2\pi} e^{-i\lambda m} \left(\sum_{x=0}^{\alpha} \sum_{y=0}^{\alpha} e^{i\lambda[1-\delta(x,y)]} \right)^n \\ &= \int_0^{2\pi} \frac{d\lambda}{2\pi} e^{-i\lambda m} [\alpha + (\alpha^2 - \alpha) e^{i\lambda}]^n \\ &= \binom{n}{m} (\alpha^2 - \alpha)^m \alpha^{n-m} \end{aligned} \tag{10}$$

where we have used the binomial theorem. Substitution of Eq. (10) into Eq. (9) yields the result that Prob(m parental loci are heterozygotic) = $\binom{n}{m} (1 - \alpha^{-1})^m \alpha^{-(n-m)}$.

Appendix C.

In this appendix, we give an explicit mathematical form for $A_{X,Y}(G)$.

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ represent the haplotype with numerical label X, and similarly $\mathbf{y}$ and $\mathbf{g}$ represent the haplotypes with numerical label Y and G. Using a Kronecker delta $\delta(a, b)$ ($\delta(a, b) = 1$ when $a = b$ and zero otherwise), we can write

$$A_{X,Y}(G) = \prod_{m=1}^{n} [\delta(x_m, g_m) + \delta(y_m, g_m) - \delta(x_m, g_m)\delta(y_m, g_m)]$$

where the product is taken over all loci.

It may be verified that $A_{X,Y}(G)$ does have the required properties, i.e., if a gamete with haplotype $\mathbf{g}$ can (cannot) be produced by a parent with ordered genotype (x, y) then $A_{X,Y}(G)$ equals 1(0).

References

Bulmer, M.G., 1980. The Mathematical Theory of Quantitative Genetics. Oxford University Press, Oxford.

Crow, J.F., 1994. Advantages of sexual reproduction. Dev. Genet. 15, 205–213.

Fisher, R.A., 1918. The correlation between relatives under the supposition of Mendelian inheritance. Trans. R. Soc. Edinburgh 52, 399–433.

Falconer, K., 1990. Fractal Geometry. Wiley, New York.

Gavrilets, S., Waxman, D., 2002. Sympatric speciation by sexual conflict. PNAS 99, 10533–10538.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Kluwer Academic Publishers, Boston, MA.

Mandelbrot, B., 1982. The Fractal Geometry of Nature. W. H. Freeman, NY.

Moisy, F., 2008. http://www.fast.u-psud.fr/~moisy/ml/.

Waxman, D., Peck, J.R., 1999. Sex and adaptation in a changing environment. Genetics 153, 1041–1053.

Wolfram, S., 1984. Geometry of binomial coefficients. American Mathematical Monthly 91, 566–571.