2.4.1 The Equations of Evolution

The previous chapters explained how the need of genes to exactly copy sequence results in a cost or directionality to change. Yet, however plausible the explanation seems, equations now used in evolution do not reveal any such effects. It is not that the effects are not there physically. They simply do not appear on the 'radar screen' of mathematical biology. So, if somebody proposes that these effects occur, that person needs to explain why they are not appearing in any equations.

Needless to say, this is very difficult. The topic is highly specialized, plus much of the subject concerns how diploid alleles will distribute in a mostly stable population. The equations are variations to the equality;

$(p + q)^2 = p^2 + q^2 + 2pq$ (The Hardy-Weinberg Equation)

The problem is that diploid organisms in a stable population represent an already highly directional form of change. (The direction is towards the wealth of allele variety.) Yet, the first 2-3 billion years of life were haploid, or even among diploid species, 70-90% of loci are *homozygous* (with little allele variety). So, although it is often derived from a Hardy-Weinberg equation, broader evolution can better analyzed via a so-called Fisher/Wright model (after R. A. Fisher and Sewall Wright).

A very simple Fisher/Wright model can be shown here. Suppose an individual has a locus X on a chromosome, which can be occupied by a range of genes or alleles x_i, where i = 1, 2, 3... (x_i means the distribution value, so 1 in 100 is 0.01. Here it also refers to the gene 'x_i'.) If each variation of x_i has a fitness w_i, the population has a mean fitness $\bar{w}$ (w bar) of $\Sigma w_i x_i$ about X. R. A. Fisher showed that if w_i of any gene x_i was greater than the mean, $\bar{w}$, rate of spread Δx_i of x_i in a natural population (for a haploid, at t= 0) would be;

$\Delta x_i = x_i (w_i - \bar{w})/\bar{w}$ (Call it Fisher's equation.)

This shows that the fitter w_i makes an individual above $\bar{w}$, the greater is $w_i - \bar{w}$, so the faster x_i spreads until $\bar{w}$ rises to w_i.

For example, suppose a mosquito population has 1,000 individuals. One individual has an allele, x_2, resistant to DDT (w_2 =1) and the other 999 x_1 individuals have only 50% resistance (w_1 = 0.5) With no DDT, each generation ($w_i - \bar{w}$) = 0, so x_2 does not increase ($\Delta x_2 = 0$). But once DDT is present, x_1 halves each generation, while x_2 quickly increases its frequency from x_2 = 0.001 to x_2 = 1.0 by about the 18th generation. So although favorable mutations may be small, they can spread very fast. And the spread can be traced by the history, or frequency, of the allele or gene causing the change. This principle is so central that Fisher called it the *fundamental theorem* of natural selection.

However, by 'mean fitness' $\overline{w}$, Fisher's theorem refers to the mean about a single locus X. But genomes contain many loci, X, Y, Z. If one takes a different mean fitness over all genome loci X, Y, Z... as w_G, then most equations (not just Fisher's one) rely on a condition $w_G = w_i$. This assumption simplifies calculations by replacing fitness of the thousands of genes in the genome, w_G, by the fitness of just the one gene or allele, w_i, causing the change. Even so, assumption $w_G = w_i$ "throws away the organism", so it needs to be shown exactly which role the organism does play, in how genes spread.

2.4.2 Gene Trajectory

Earlier, Section 2.2.3 introduced the concept of gene *trajectory*. As explained, physically there are many reasons why genes alter or mutate at different rates over the history of life. But whatever the physical cause, one might liken fidelity of copy of a gene or DNA sequence to a force, call it ε_i (eta i). If say, a gene did not alter copy by even 1 bp for eternity, then $\varepsilon_i = \infty$. If a gene could alter each reproduction, then $\varepsilon_i = 0$. No gene can obtain these extremes (life has not existed for eternity) so we assume that there is an average $\overline{\varepsilon}$ (eta bar) for all genes, that can be normalized such that $\overline{\varepsilon} = 1$, for any typical gene.

The concept of ε_i, allows one to investigate the assumption $w_G = w_i$. Basically, when gene x_i increases its frequency, say $0.1 \rightarrow 0.9$, it does so in a certain "direction", in which every gene in the host genome, G, also increases frequency, say $0.1 \rightarrow 0.9$, by the same amount. So it seems safe to set $w_G = w_i$, because for any selective event every gene in G alters its frequency by the same amount anyway. However, the value of ε_i, if it exist, will be very different for each gene in the genome. And while x_i in small populations alters rapidly, ε_i alters slowly over the history of life, and is unlikely to be affected by small changes.

In fact, while it is not the same, ε_i can be derived from mutation rate μ_i (mu i). To be sure, μ_i, is a *scalar*. It measures statistical change in the present time, such as the rate by which an allele $x_1 \rightarrow x_2$ mutates to enter the gene pool of a modern population. On the other hand, ε_i is a *vector*. It is the retentive force that holds a gene within a copy *trajectory*, over the history of life. Genes also mutate for many reasons. Instability at a single region of a gene could cause high μ_i, but low ε_i if the rest of the gene was stable. Or an opportunistic gene can have low ε_i, but medium μ_i. Still, data for μ_i is available. If average normalized $\overline{\varepsilon}$ is a function of average mutation rate, $\overline{\mu}$ (mu bar) such that;

$\overline{\varepsilon} = f(\overline{\mu}) = 1$, then for any gene mutating at rate μ_i, approximately;
$\varepsilon_i = \sqrt[7]{(\overline{\mu}/\mu_i)}$

This formula gives a rough value of ε_i, against a measure (mutation rate) that is familiar. The term is reduced a 7^{th} root because ε_i is a weak force, acting about 1 in 10^7 against x_i. Highly conserved genes mutate at about $\mu_i = 10^{-13}$ which for $\bar{\mu} = 10^{-7}$ gives $\varepsilon_i = 7.2$; a fast mutating gene $\mu_i = 10^{-5}$ will have $\varepsilon_i = 0.52$. In fact, ε_i is never that accurate, and $\varepsilon_i = 1$ would cover a range $10^{-6} < \mu_i < 10^{-8}$. This is to give a broad idea of ε_i. Its precise values are not required here.

Having broadly defined ε_i, its relationship to x_i must be formulated in ways that conserve standard theory. Take distribution D_i of a gene x_i as $D_i = x_i$. Then $D_i = x_i$ for ε_i, is conserved if $D_i = x_i(1 + j\varepsilon_i)$ where $j = \sqrt{-1}$. However, because $0 < \varepsilon_i < \infty$, this must be normalized to keep $D_i \leq 1$, so the full expression becomes;

$$D_i = x_i(1 + j\varepsilon_i)/\sqrt{(\varepsilon_i^2 + 1)}$$

It looks complicated, but notice that the value of ε_i does not alter the value of D_i as x_i, but only varies its complex sign. (If $\varepsilon_i = 0$, $D_i = x_i$. Yet if $\varepsilon_i = \infty$, $D_i = jx_i$.) It is harder to show, but if ε_i was the same for all genes then again $D_i = x_i$. (If for two alleles $D_1 = kD_2$, if $\varepsilon_1 = \varepsilon_2$, then $x_1 = kx_2$.) So, the new notation is not that different from standard theory. If $\varepsilon_i = 0$, is the same for all genes, or has no effect, standard theory is conserved. Just that if ε_i does exist, or is not the same for all genes, one can now examine what is lost when setting $w_G = w_i$.

2.4.3 The Use of Angle Notation

The effects of change in a genome, where different forces of ε_i act on different genes, can be best visualized using an angle notation. When people are told that there is an angle, they expect to see a physical angle, like angles forming the DNA helix. However, the term $(1 + j\varepsilon_i)/\sqrt{(\varepsilon_i^2 + 1)}$ is also an angle, where $\theta_i = \tan^{-1}(\varepsilon_i)$. So;

$D_i = x_i (1 + j\varepsilon_i)/\sqrt{(\varepsilon_i^2 + 1)}$ is the equivalent of;
$D_i = x_i (\cos \theta_i + j \sin \theta_i)$ or, $D_i = (x_i, \theta_i)$

Further, for any value of ε_i, broadly;

$\varepsilon_i = \infty$, $\mu_i \approx 0$ ("forever"), $\theta_i = 90^O$
$\varepsilon_i = 1$, $\mu_i \approx \bar{\mu}$ ("average"), $\theta_i = 45^O$
$\varepsilon_i = 0$, $\mu_i \approx 1$ ("each reproduction"), $\theta_i = 0^O$

This is shown in Fig. 2.4.1. There is no physical angle, but the notation helps visualize how genes, genomes and DNA segments interact over the

history of life. Highly conserved genes barely alter over huge times, so they are at high angles. Because evolution is adaptation to change, genes will only be able to stay unaltered while adapting into a huge variety of types if other DNA in the genome bears the cost of change. This will appear on the diagram as though, over time, conserved genes 'rotate' higher, but genomes 'rotate' to lower angles.

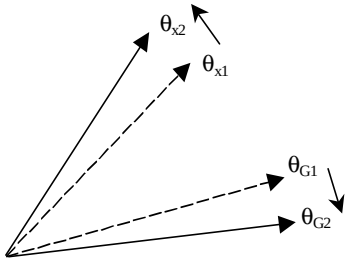

Fig 2.4.1 Genes, DNA, and genomes appear to spread together in single organisms. But over the history of life, individual genes try to avoid altering sequence, by forcing host genomes to bear the cost of change. On an angle diagram, it would appear that genes rotated 'higher' while forcing host genomes into a lower angle.

Still, Fig 2.4.1 only shows how genes or DNA distribute over time when mapped on a diagram of this type. The angles appear to change because the DNA does. Yet if genes really do try to rotate to higher angles, there must be some "force" driving them to do so. True, that force is natural selection, but in a Fisher equation it is the pressure (w_i - $\overline{w}$) that drives the value Δx_i to increase. So how would selection drive $\Delta\theta_i$ to increase in the new formulation?

Well, the equation is not fully derived yet, but to see how it works requires setting a "goal" that all genes try to achieve. In standard theory the gene seeks maximum probability P_i of survival in the next generation. Take distribution of any gene as D_i, and fitness of a host as F_i. The gene has a probability of existing of $P_i = D_i F_i$, with maximum of $P_i = 1$ (when $D_i = 1$ and $F_i = 1$). This can be written (in standard theory) as;

$$P_i = x_i w_i$$

The new equations, though, would involve two new terms; ε_i, (the exact copy), and w_G, (the fitness of the organism). The relationship of these new terms to w_i and x_i is not known. However, it is likely that that genes at high ε_i (highly conserved genes) tend to spread anyway, regardless of which host genome they happen to be in. Using this principle, one can approximate the new equation of P_i to be something like;

$P_i = x_i (w_G + w_i \varepsilon_i^2)/(\varepsilon_i^2 + 1)$, or in angle notation;
$P_i = x_i (w_G \cos^2 \theta_i + w_i \sin^2 \theta_i)$

Note that when ε_i (viz. θ_i) is low, x_i must be inside a fit genome in order to propagate. Yet when ε_i is high, the gene relies on its inherent fitness. And strange as this equation appears, it fully conserves standard theory. For the condition $w_G = w_i$ for any ε_i, or for $\varepsilon_i = 0$, the equation will revert to $P_i = x_i w_i$. (Note that $\cos^2\theta_i + \sin^2\theta_i = 1$.)

However, now we have D_i and P_i, we can obtain F_i by dividing P_i/D_i. Note that P_i has a "real" (scalar) value, but once this is divided by the coordinate $D(x_i, \theta_i)$ this will result in a complex form of F_i, so we get;

$$P_i = x_i (w_G \cos^2 \theta_i + w_i \sin^2 \theta_i)$$
$$D_i = x_i (\cos \theta_i + j \sin \theta_i) \text{ or, } D_i = (x_i, \theta_i)$$

Dividing P_i/D_i gives;

$$F_i \approx w_G \cos \theta_i - j\, w_i \sin \theta_i$$

Note that F_i is approximate. (Following division there is an extra term in F_i that mostly reduces to 0, but might concern "past" or "future" events.) Again though, for the condition $w_G = w_i$ or $\theta_i = 0$, then $P_i = x_i w_i$. Or multiplying complex F_i by the complex D_i will also give $P_i = x_i w_i$ (the 'j' terms cancel) regardless of the value of θ_i (with some adjustments). So again, standard theory is conserved throughout.

Still, the equation is interesting. The first term shows that fitness of the organism, w_G, only acts on the real component of selection ($w_G \cos\theta_i$ is "real"). This infers that while organisms can evolve new designs by natural selection, they carry perfected designs into the next generation *without* selecting them out! This is the second term (with $j = \sqrt{-1}$). All genes were first selected in real genomes, but in the past (-ve sign on j). The deeper in the past (as $\theta_i >> 0^O$) the further the chance of selection is rotated away from the effects of modern events.

This is why genes 'want' to rotate to higher angles. They are trying to avoid selection! Selection is costly, for genes and nature. If a gene is already perfected in function, it is inefficient to re-design it by selection each time. It took billions of years to perfect the eukaryotic cell, and hundreds of millions of years to evolve large animals. Yet an intelligent being can evolve in a few million years by reincorporating earlier designs perfected over billions of years past. 'Selfish' gene theory has said that the organism is a way for genes to spread. The new formulation shows how it works. Nature conserves perfected designs by its own processes. But when humans model those processes with the mathematical tools available, it appears as though genes try to avoid selection by rotating deeper into an imaginary plane.

2.4.4 The Fall of Fitness

One of the conditions of the Fisher equation is that mean fitness can only rise as x_i spreads. But in life, this is often violated. Suppose that a genome G_1, consisted of two genes, x_i, y_i at loci X, Y. Suppose that gene x_i could double the total individuals in an area by splitting y_i into two new genes τ_i and ψ_i, then evolving G_1 into two new genomes G_2 and G_3. Suppose now G_1, G_2, G_3 each have 1,000 copies. We get;

$$G_1(x_i, y_i) \rightarrow [G_2(x_i, \tau_i) + G_3(x_i, \psi_i)] \text{ (For this case count the copies.)}$$

Here x_i has doubled but y_i has decreased. Plus if G_1 has gone extinct, its fitness decreased despite that x_i has increased. So fitness fell, but the condition $w_i = w_G$ of the Fisher equation was violated, by the case that;

$$w_{xi} > w_{G1} \text{ but } w_{yi} < w_{G1} \text{ (Again, just count copies.)}$$

Still, what happens when fitness falls but x_i increases is that the angle of the host genome, θ_G, falls. In a Fisher equation, mean fitness is a single scalar quantity, $\bar{w}$. It has not been derived, but in the new model mean fitness would be a complex sum, $(\bar{w} - j\bar{\varepsilon})$. (The j sign is –ve, because broadly, the population evolved in the past.) It would be difficult to sum this over thousands of small changes, but "pressure" about a locus X for change would be $(w_i - \bar{w}) + j(\bar{\varepsilon} - \varepsilon_i)$. The accumulated affects of these tiny decrements in $\bar{\varepsilon}$ over thousands of loci X, Y, Z, would be an eventual fall in the ε_G (or θ_G) of the entire genome.

If anything, one suspects that rather than sum θ_G over thousands of genes and billions of bp, one might assume that for a haploid $\theta_G \approx 45^O$, and a diploid $\theta_G \approx 0^O$. (When a new form of reproduction evolves, θ_G falls slightly. Evolution of sex was the 'great θ_G crash' from 45^O to 0^O, dwarfing all other decreases in θ_G.)

Suppose though, that a gene maximizes spread if it replicates in a genome at an effective 'angle' of 45^O. This will occur at $w_G = w_i\varepsilon_i$. Then for genes at $\theta_i < 45^O$, the gene can afford a lower host fitness, $w_G < 1$, as this helps the gene increase effective angle. For conserved genes where $\theta_i > 45^O$ the gene could afford a lower w_i to get a 45^O effective angle. It is not clear physically what this means, but it vaguely infers how sex works. Highly conserved genes can accept a high fitness penalty for other genes in the host, because they are going to spread anyway.

On the other hand, while the case $w_i > w_G$ is hard to resolve, the case $w_i < w_G$ (the gene damages host fitness) becomes clearer. Note, w_G acts only on the real component of fitness. Broadly, any gene such as a rogue or parasite at $\theta_i < 45^O$ is losing copy at each reproduction at a faster rate than average. (A 10^2 bp long fragment that is mutating at $\mu_i = 10^{-4}$ will

destroy its copy in 10^6 reproductions.) The best strategy for such a gene is to "slow" its rate of reproduction, by damaging its host's fitness, at roughly $w_G \propto \varepsilon_i$. (At $\mu_i = 10^{-4}$ then the gene obtains equivalent copy of $\bar{\mu}$ at $w_G = 0.37$. The figures are not researched.) Note too that the equation of complex fitness is;

$$F_i = w_G \cos \theta_i - j\, w_i \sin \theta_i$$

Lowering w_G lowers the "real" part of the equation, so it pushes effective 'angle' of complex fitness higher. As a real process, rogue DNA damages host fitness because that is how it acts. But in the equation, the DNA is trying to increase its effective angle, hence its survivability, by rotating itself further away from the plane of real selection.

Generally, the gene, being "selfish", tries to manipulate a genome to its advantage, but the strategy will depend on the (w_i, θ_i) of the gene. A rogue gene with a low (w_i, θ_i) tries to replicate inside a strong genome with a high θ_G, despite that rogue genes might try to lower w_G of the host. (A low angle genome, like in sexual organisms, can alter rapidly, so it might quickly find a way to throw out the rogue gene.) Yet a very strong gene will, paradoxically, want to see life populated by highly variable (but low angle) genomes, so the strong gene can spread within a huge variety of types. (It is like the computer industry. If you make a part like the CPU needed in all computers, then the larger the variety of low cost computers built, the more parts you can sell.)

2.4.5 An Ongoing Debate

In summary, how is it that effects claimed here to be a major factor in evolution, do not appear in the math of standard theory?

Well, the math of standard theory is explicit. It is describing a well-understood physical process, in that a gene that is fit is also increasing its frequency in a population, say, from 1% to 99% distributed, relative to a rival. Moreover, the gene that is fit, spreading this way, is also contained "within" the equation modeling the process occurring. (The gene that is spreading, is the same gene that the equation is describing.) However, when a favored gene is spreading, say 1% to 99% distributed, other genes in the genome are also spreading, even though, perplexingly, they might be 100% distributed already for that population. The difference is that the gene causing the spreading, the "action" gene, was altered from an earlier sequence to gain the fitness to spread. Yet the genes that spread anyway, that were already distributed 100%, are now carried along by the "action" gene into a new adaptation, but are not themselves forced to alter their own sequence to adapt. These genes, able to adapt into new varieties without themselves being forced to alter, gain slight fitness over genes forced to bear the cost of change.

To model this process, requires capturing the effects of fitness from the perspective of any gene in the genome, not just the "action" gene. This is done using a second quality of gene distribution; the "exact copy" of a gene, here called ε_i. Genes that survived unaltered for billions of reproductions, or adapted into a huge variety of types at no alteration to their sequence are versatile designs, that inherently end up widely copied. And organisms that adapt proven genetic designs (by reshuffling existing genes, rather than evolving new ones) ultimately adapt at lower total cost of change. So although ε_i is the copy fidelity of a gene over the history of life, it approximates the cost and directionality of change.

Yet, using ε_i must conserve the equations of standard theory where these are correct. This is done by adding ε_i to x_i as a complex sum, so normalized distribution D_i, becomes $D_i = x_i(1 + j\varepsilon_i)/\sqrt{(\varepsilon_i^2 + 1)}$. This form conserves standard theory (say, by setting $\varepsilon_i = 0$). Still, manipulating this further provides a new equation, showing how w_i (gene fitness) relates to w_G, (fitness of the host genome in which the gene is resident). This is;

$F_i = (w_G - jw_i\varepsilon_i)/\sqrt{(\varepsilon_i^2 + 1)}$, or in angle notation;
$F_i = w_G \cos \theta_i - j\, w_i \sin \theta_i$

This equation is incomplete. There are missing terms, and it does not show angle, θ_G, of the host genome (which might differ between diploid to haploid organisms). The equation also does not show the time variant conditions, or effective angle for F_i for a gene to maximize propagation. (Though one suspects it is 45^O.) Even so, the equation does confirm how life works! Succinctly, it shows that fitness, w_G, of the host genome acts on the "real" part of the equation, so as is the case, it is the organism (not the gene) that is selected at each fitness event.

Genes reproduce physically, inside organisms. And they pass on to offspring physically, like passing a baton in a relay. Yet genes still only reproduce information. In the famous polymerase chain reaction (PCR) humans provide the chemical ingredients. It is the "information" in the DNA snippet, not the chemicals, that is multiplied millions of times. So, organisms play two roles in transmitting DNA. By physical reproduction they are a chemical relay station. By mutation and selection, they are a way to modify DNA information. DNA as molecules is copied as a "real" physical process, and change of sequence occurs at real physical events, even for events in the past. Even so, when modern organisms are selected for changes of allele frequencies, 99% of the stable sequences in those organisms are being copied in other organisms, in other times, over the biota of life. If one models this among a small population from which the gene has already radiated, it should show as 'imaginary' selection in a correctly formulated equation.

Yet if this equation is correct, it means that any gene at any locus in the genome, tries to increase not just its distribution fitness, x_i, but its total fitness, $x_i(1 + j\varepsilon_i)/\sqrt{(\varepsilon_i^2 + 1)}$, where ε_i is the "exact copy" of the gene. When a gene first comes into existence, at $\varepsilon_i = 0$, the gene relies on the fitness of its host, w_G, to spread. Here, $w_G = w_i$ for that gene, which applies as in standard theory. But as the gene matures and radiates into many types, it will become less dependent on its host to avoid sequence death. Broadly, as ε_i increases the gene sequence *radiates* out from the point of origin of the sequence much like a wave, through millions of descendant reproductions. (When $\varepsilon_i = 0$, the gene is like a particle. When $\varepsilon_i = \infty$, it is like a wave.) It has not yet been modeled, but it is hoped that some fast mutating DNA will exhibit this wave-like effect as a concerted synchronism across physically separate organisms.

Modern evolutionary theory has become divided between so-called gene-centric or reductionist models, focused on genes and equations, and a more holistic, observational approach. The assertions of this chapter seem to take the division to an extreme. Just when the reductionist school is conceding that genes might also be cooperative or parliamentary, this chapter argues why DNA is consistently selfish. Genes might cooperate to spread in unison, but each gene also competes to preserve its own copy unaltered, and force other genes in the genome to bear the cost of change. Just that genes compete for spread over tens of generations, but compete for exactness of copy over millions of generations, and this difference of scale is hard to model. This is the second contention. All the processes of life are real physical events at the instant when they occur. But within equations, humans try to capture events from billions of years past into single events of the present. Within this restriction equations will show strange effects, such as genes radiating like waves of information, rather than processes normally associated with life.

Even so, the math explained here is more a notational argument than proven equations, and no one equation anyway will ever fully capture the vast processes of life. Yet incomplete as it is, the argument here can still challenge existing models of how large-scale evolution works, or how genes and organisms do interact. Also, despite the reductionist approach inherent to equations, there is a cautious optimism. Even from a model of gene selfishness, these equations illuminate the one result that everybody suspected was the case all along. Evolution of complex new creatures, or complex new adaptations such as thought and emotion, will take more than just a few changes in allele frequencies. It is the combined effects of all evolution, accumulating over the history of life.