Modeling Evolution

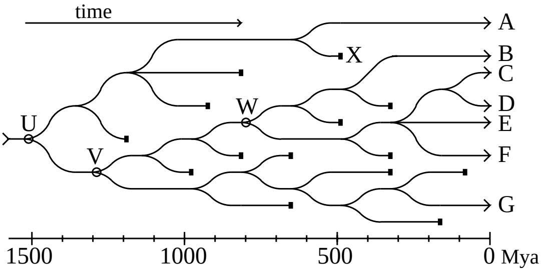
John Milnor

Stony Brook University

DAVID EPSTEIN CELEBRATION

Warwick University, July 14, 2007

A First Model for Evolution: The Phylogenetic Tree,



showing how a few of the millions of species have evolved.

At some point in time, one species may split up into two or more distinct species. (*U*, *V*, *W* in the diagram.)

Most species eventually become extinct. (*X* in the diagram.)

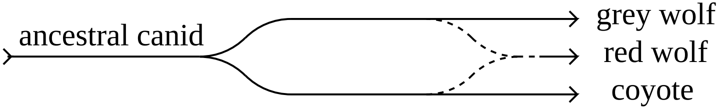
Tree Properties: Lines can only separate —never come back together. Any two distinct species at two points in time have a unique most recent common ancestor. (Thus for *F* and *G*, the most recent common ancestor is *V*.)

Common Ancestry

Such diagrams are based on the fundamental hypothesis that all **existing** forms of life are descended from a common ancestor.

One minor difficulty:

The tree structure can be rather fuzzy.



There is a similar fuzziness in connection with the separation between humans and chimps. See Patterson et al. [2006]. (References will be listed at the end.)

A more serious difficulty:

How do we account for symbiotic species?

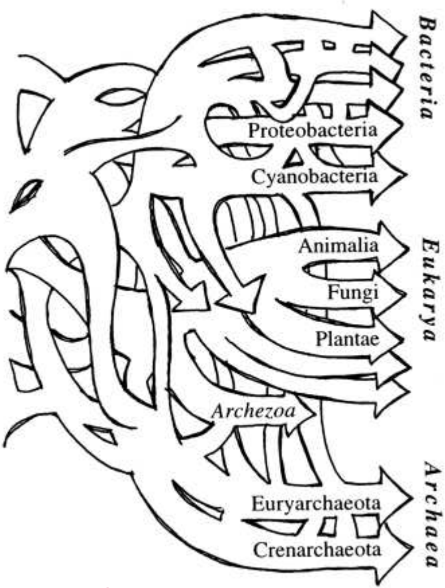
Lynn Margulis (1997): "··· a more important source of Darwinian evolutionary novelty in living beings is symbiogenesis, evolutionary change through long-term physical contact between members of different species. ···

"Random mutations in DNA ··· lead to small, mostly harmful changes. Mergers of symbionts lead to large, functional evolutionary jumps: new organs or major new groups of organisms."

To the extent that this is true, it contradicts the
hypothesis that separate species can never join together!

(This difficulty is most important for single-celled organisms.)

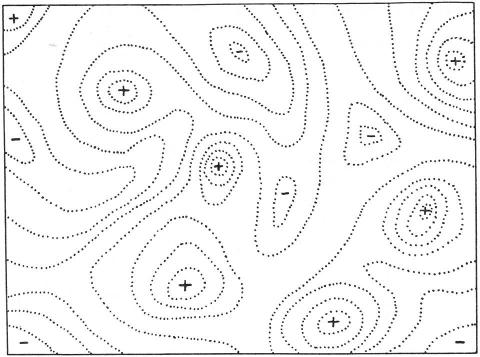
ALL life (Doolittle 1999):



A “Phylogenetic Tangle” ?

The Fitness Landscape Model.

Sewall Wright (1932) described the evolution of the gene pool within a population as a walk on a **fitness landscape**.



The plane of the screen is a 2-dimensional substitute for the space of all possible mixtures of genotypes within a population. The contours are loci of constant **Darwinian fitness**, and the course of evolution is described as a random path **which tends to climb towards the nearest peak.**

Goal of this talk:

To describe a *mathematically* precise model, compatible with Wright's picture, thus interpreting evolution within a species as a random dynamical system.

(based on Shahshahani 1979, Akin 1980, etc.)

First some standard definitions:

A **gene** is a stretch of DNA at a specific location (or **locus**) on a specific chromosome with some biologically meaningful function, for example a recipe for manufacturing some protein. (In mathematical terms, a gene is a word in an alphabet which has four letters.)

Most genes have several alternative versions called **alleles**.

Each human (or each mammal) has exactly two copies of each chromosome in each cell (excluding the X and Y chromosomes), hence two copies of each gene.

Genotype:

The two copies of this gene will correspond to two (not necessarily distinct) alleles.

If these alleles are A_j and A_k, then the individual is said to have **genotype** $A_jA_k = A_kA_j$ **for this gene.**

If a gene has n distinct alleles $A_1, \ldots, A_n$, then there are $n(n+1)/2$ possible genotypes for this gene.

Example: Two alleles $\Longrightarrow$ three possible genotypes.

Now consider 25 different genes, with two alleles each. Then there are

$$3^{25} \;=\; 847{,}288{,}609{,}443$$

possible genotypes, if we consider **only** these 25 genes.

This is much larger that the number of human beings who have ever lived! Yet there are actually **many thousands** of genes. How can we deal with such numbers?

G.H. Hardy to the rescue!

Let N_j be the number of copies of the allele A_j within a large population; and let

$$a_j = \frac{N_j}{N_1 + \cdots + N_n} = \frac{N_j}{2P}$$

be its **statistical frequency**.

Thus $a_1 + \cdots + a_n = 1$ with $a_j \geq 0$.

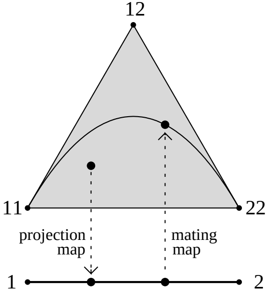
In other words the **frequency vector** $\vec{a} = (a_1, \ldots, a_n)$ belongs to the **standard simplex** $\Delta^{n-1} \subset \mathbb{R}^n$.

Hardy-Weinberg Law (1908): Under “suitable conditions,” allele frequencies determine genotype frequencies: **The frequency $f_{j,k}$ of the genotype A_jA_k within the population is approximately**

$$f_{j,k} = \begin{cases} 2a_ja_k & \text{if} \quad j \neq k, \\ a_ja_k & \text{if} \quad j = k. \end{cases}$$

(Unfortunately, these “suitable conditions” include the rather unrealistic hypothesis of **completely random mating**.)

Hardy-Weinberg for a gene with two alleles



The triangle of possible genotype frequencies for a gene with two alleles (above), projected to the interval of possible allele frequencies (below). Each genotype frequency vector, represented by a point in the triangle, corresponds to an allele frequency represented by the point on the interval directly below it. Conversely, under random mating, each point on the interval below corresponds to the point directly above it on the Hardy-Weinberg parabola which is defined by the equation $f_{1,2} = 2a_1 a_2 = 2a_2(1 - a_2)$.

Changes with time.

Two basic mechanisms make the gene pool gradually change:

- **Selection.** Those genotypes which are most successful in surviving and reproducing will tend to predominate in future generations.
- **Genetic Drift.** Due to the random nature of reproduction, the frequencies of different alleles in the gene pool will vary in a chaotic fashion (even without selective pressure).

Other important mechanisms that I will **not** discuss:

- **Mutation and copying errors.** Accidental errors in reproducing a gene are usually bad, and often fatal. However, accidental changes which are actually viable could be a major source of new genetic possibilities.
- **Lateral transfer of genes.** A symbiotic relationship may lead to sharing or transfer of genes. A retroviral infection (such as HIV) can inject **new genes** into the genotype (.03% of the human genome)!

Genetic Drift. The Wright-Fisher Model:

Consider a single gene with n alleles, in a population of P individuals. Then there are $2P$ copies of this gene altogether.

Each of the $2P$ copies of this gene in the next generation is to be randomly chosen, according to the probability distribution $(a_1, \ldots, a_n)$. These $2P$ choices are to be independent random variables.

Thus the frequency vector $(a_1, \ldots, a_n) \in \Delta^{n-1}$ will vary from generation to generation, in a random manner (Markov chain).

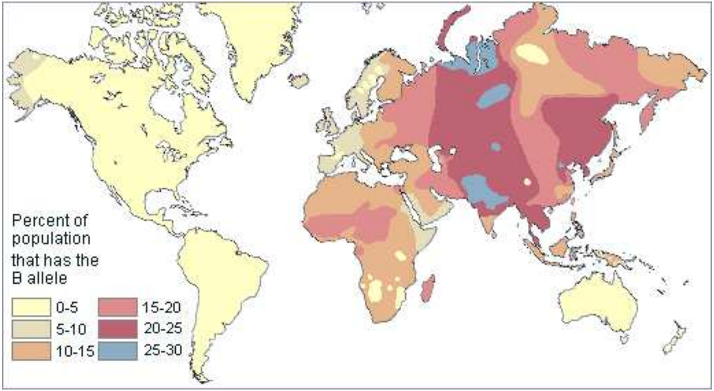
Theorem. If this process is continued long enough, then with probability one, all but one of the alleles will eventually become extinct.

The probability that A_i will be the unique eventual survivor is precisely equal to its initial frequency a_i.

Example: One mutant allele in a large population is very unlikely to survive **(unless it has a big selective advantage)!**

Allele extinction

VERY ROUGH ESTIMATE: The number of generations until the first extinction among significant alleles of a specified gene has the same order of magnitude as the total population size.



Example: Blood type B is almost totally missing among Native Americans. Presumably the corresponding allele became extinct twelve thousand years ago when small populations wandered from Asia over many generations.

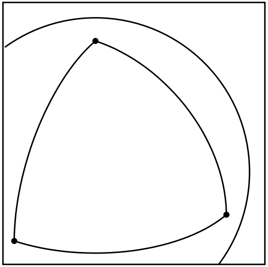
The Spherical Geometry for a Fitness Landscape.

Again consider a single gene with alleles $A_1, \ldots, A_n$.
Its frequency vector $(a_1, \ldots, a_n)$ within a given population belongs to the standard simplex $\Delta^{n-1} \subset \mathbb{R}^n$.

But the flat geometry for Δ^{n-1} is misleading!
Whether studying genetic drift or selection, it is often more useful to work with the **root-frequency vector**

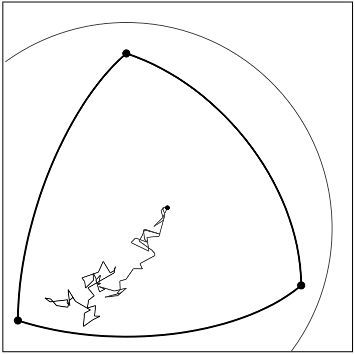
$$\vec{u} = \left(\sqrt{a_1}, \ldots, \sqrt{a_n}\right) \quad \in \quad S^{n-1} \subset \mathbb{R}^n .$$

This belongs to the standard **spherical simplex**,
with the n standard basis vectors for $\mathbb{R}^n$ as vertices.



Genetic Drift as a Random Walk on the Sphere.

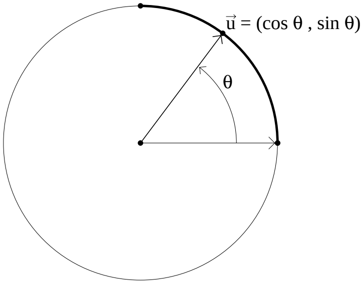
Using this spherical model for allele frequencies, the course of genetic drift can be described as a (nearly) unbiased random walk on the standard spherical simplex.



To a first approximation, all directions are equally probable, and the likely step size is independent of position, until the walk hits some boundary simplex. At that point one allele becomes permanently extinct, and the walk continues on this boundary simplex.

The Two Allele Case

For a gene with only two alleles, the geometry is much simpler.

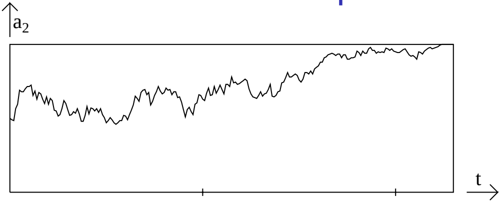


In this case, the standard spherical simplex becomes a **quarter-circle.** The alleles frequencies are now described by the **root-frequency angle** $\theta = \theta_A$, defined by the equation

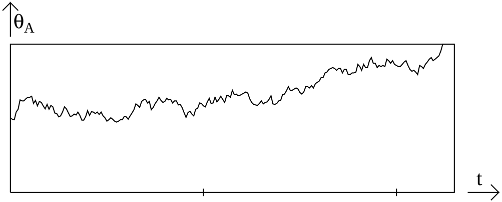
$$\vec{u} = (\cos\theta\,,\,\sin\theta)\;, \quad \text{or equivalently} \quad \vec{a} = (\cos^2\theta\,,\,\sin^2\theta)\;,$$

where θ measures arclength along the quarter-circle.

The Two Allele Case: A sample run



plotting the allele frequency a_2 as a function of time over 230 generations, with a population of fixed size $P = 100$.
Here A_2 becomes the sole survivor after 223 generations.



The same run, but plotting the angle θ_A rather than the frequency $a_2 = \sin^2 \theta_A$ as a function of time. The “roughness” of the graph is more uniform in this version.

Selection: Genotype Fitness

Let $P_{jk}(t)$ be the total number of individuals of genotype A_jA_k at time t. Then the number $N_j(t)$ of copies of allele A_j can be computed as

$$N_j = P_{jj} + \sum_k P_{jk} .$$

Hypothesis: To each genotype A_jA_k there corresponds a constant $\Phi(A_jA_k)$, or briefly $\Phi_{jk} = \Phi_{kj}$, which I will call its **fitness**, so that the following differential equation is satisfied:

$$\frac{dN_i}{dt} = \Phi_{jj}P_{jj} + \sum_k \Phi_{jk}P_{jk} .$$

Intuitively we can interpret this fitness as the difference

$$\Phi_{jk} = \beta_{jk}/2 - \delta_{jk} ,$$

where β_{jk} is the number of children per unit time born to a parent of genotype A_jA_k, and where δ_{jk} is the number of deaths per unit time for this genotype. (The factor $1/2$ is inserted since each parent contributes only half of the genotype of the child.)

Allele Fitness

Now suppose that the frequencies P_{jk}/P of the various genotypes are determined by the allele frequencies a_j, according to the Hardy-Weinberg formula.

Assertion. The exponential growth rate for N_j, the number of copies of the allele A_j in the population, is then given by

$$\frac{d \log N_j}{dt} = \sum_k \Phi_{jk} a_k .$$

Proof. Just substitute the Hardy-Weinberg expression into the differential equation for dN_j/dt and simplify. $\square$

This weighted average $\sum_k \Phi_{jk} a_k$ will be called the **fitness** Φ_j **of the allele** A_j**.** Thus

$$\Phi_j = \frac{d \log N_j}{dt} \quad \Longleftrightarrow \quad N_j \Phi_j = \frac{dN_j}{dt} .$$

Total Population Fitness.

Define the weighted average

$$\Phi \;=\; \sum_j \Phi_j a_j \;=\; \sum_{jk} \Phi_{jk} a_j a_k$$

to be the **fitness for the entire population.**

Corollary. Φ measures the exponential growth rate for the population size P. That is,

$$\frac{d \log P}{dt} \;=\; \Phi\,.$$

Proof. Sum over j in the equation

$$\frac{dN_j}{dt} \;=\; N_j\,\Phi_j \;=\; (2P\,a_j)\,\Phi_j \quad\Longrightarrow\quad \frac{d2P}{dt} \;=\; 2P\,\Phi\,.$$

Dividing by $2P$, the conclusion follows. □

The basic differential equation.

Using these equations, we can compute the exponential growth rate for the allele frequency

$$a_j = \frac{N_j}{2P}\,.$$

In fact, $$\frac{d\log N_j}{dt} = \Phi_j\,, \qquad \text{and} \qquad \frac{d\log P}{dt} = \Phi\,.$$

Thus the exponential growth rate for a_j is the difference

$$\frac{d\log a_j}{dt} = \Phi_j - \Phi \qquad \Longleftrightarrow \qquad \frac{da_j}{dt} = (\Phi_j - \Phi)a_j\,.$$

Thus $d\vec{a}/dt$ is a well defined smooth function of $\vec{a}$, and the course of evolution is described by an ODE on the standard simplex Δ^{n-1} .

However, this equation takes a more convenient form if we replace the Euclidean simplex of **allele frequencies** by the spherical simplex of **allele root-frequencies.**

The Spherical Gradient

We continue to consider the alleles of a single gene, with frequencies a_j. Set $u_j = \sqrt{a_j}$, and consider the population fitness Φ as a function of the u_j. In fact

$$\Phi \;=\; \Phi(\vec{u}) \;=\; \sum_{jk} \Phi_{jk}\, u_j^2 u_k^2 \,.$$

Theorem

Our differential equation $\quad d\log a_j/dt = \Phi_j - \Phi \quad$ *for evolution driven by selection corresponds to a* **gradient dynamical system** *on the sphere:*

The **velocity vector** $d\vec{u}/dt$ *satisfies the differential equation*

$$\frac{d\vec{u}}{dt} = \frac{1}{8}\,\overrightarrow{\mathrm{grad}}_{\,\mathrm{S}}\Phi(\vec{u})\,,$$ *where* $\overrightarrow{\mathrm{grad}}_{\,\mathrm{S}}$ *is the gradient* **on the sphere**.

In other words, the path $\vec{u}(t)$ always moves directly “uphill” on the sphere, in the direction of steepest ascent, with speed $\|d\vec{u}/dt\|$ which is directly proportional to the steepness.

There are no periodic cycles and no chaotic orbits!

Proof Outline

If we think of

$$\Phi \;=\; \sum_{jk} \Phi_{jk} a_j a_k \;=\; \sum_{jk} \Phi_{jk} u_j^2 u_k^2$$

as a function which is defined for all $\vec{u} \in \mathbb{R}^n$, then its **Euclidean** gradient has j-th component

$$\left(\overrightarrow{\mathrm{grad}}_{\mathrm{E}}\, \Phi(\vec{u})\right)_j \;=\; \frac{\partial \Phi}{\partial u_j} \;=\; \frac{\partial \Phi}{\partial a_j}\,\frac{da_j}{du_j} \;=\; (2\Phi_j)(2u_j) \;=\; 4\Phi_j u_j\,.$$

The component of this Euclidean gradient in the direction orthogonal to the sphere is given by the inner product $(\overrightarrow{\mathrm{grad}}_{\mathrm{E}}\Phi) \cdot \vec{u} = 4 \sum \Phi_j u_j^2 = 4\Phi$. Thus, in order to compute the component $\overrightarrow{\mathrm{grad}}_{\mathrm{S}}\Phi$ which is tangent to the sphere, we must subtract the normal vector $(4\Phi)\vec{u}$ from $\overrightarrow{\mathrm{grad}}_{\mathrm{E}}\Phi$. This yields

$$\left(\overrightarrow{\mathrm{grad}}_{\mathrm{S}}\, \Phi\right)_j \;=\; 4(\Phi_j - \Phi)u_j\,. \tag{1}$$

Proof Outline (continued)

On the other hand, since $u_j = \sqrt{a_j}$, we have

$$\frac{d \log u_j}{dt} = \frac{1}{2} \frac{d \log a_j}{dt} = \frac{1}{2} (\Phi_j - \Phi) ,$$

hence

$$\frac{du_j}{dt} = \frac{1}{2} (\Phi_j - \Phi) u_j . \qquad (2)$$

Comparing Equation (1):

$$\left(\overrightarrow{\mathrm{grad}}_S \, \Phi \right)_j = 4(\Phi_j - \Phi) u_j ,$$

it follows that

$$\frac{d\vec{u}}{dt} = \frac{1}{8} \overrightarrow{\mathrm{grad}}_S \Phi ,$$

as required. □

The “Fundamental Theorem of Natural Selection.”

R. A. Fisher’s **fundamental theorem** was stated as follows in 1930:

> *"The rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time."*

In our terminology, this statement (for the single gene case) would be expressed by the equation

$$\frac{d\Phi}{dt} \quad = \quad \overrightarrow{\mathrm{grad}}_S(\Phi)\cdot\frac{d\vec{u}}{dt} \quad = \quad \frac{1}{8}\left\|\overrightarrow{\mathrm{grad}}_S(\Phi)\right\|^2 \ \geq\ 0\,.$$

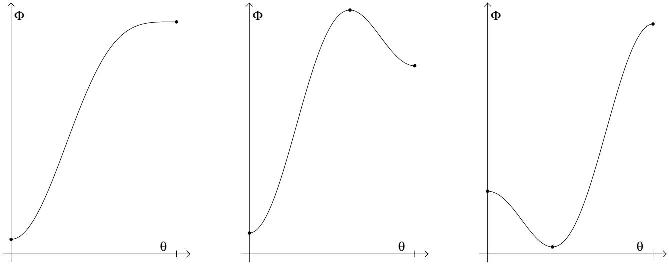
Thus the fitness Φ strictly increases with time, except at a stationary point where $\overrightarrow{\mathrm{grad}}_{\mathrm{S}}\Phi = 0$.

Caution: This is a theorem about a highly simplified mathematical model. It is ***not*** *a statement about the real world.*

The Two Allele case

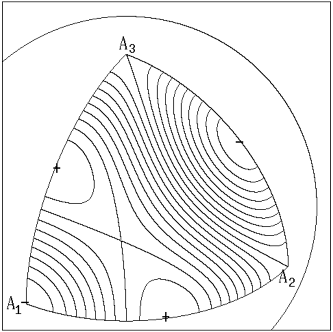
is governed by the equation

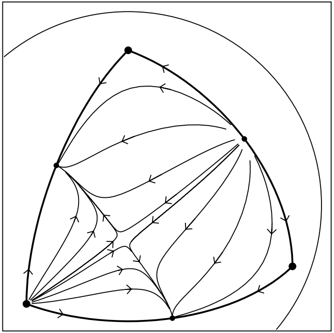
$$\frac{d\theta}{dt} \; = \; \frac{1}{8}\,\frac{d\Phi}{d\theta}\,.$$



Three possible shapes for the graph of Φ as a function of the root-frequency angle θ.

Example with 3 Alleles: Fitness and Gradient Flow





On the left, a contour plot for a possible fitness function $\Phi(\vec{u})$ associated with a gene with three alleles, where $\vec{u}$ ranges over the standard spherical 2-simplex. On the right, the arrows indicate the direction of the associated gradient flow.

Each of the flow curves travels straight uphill, and nearly all of them start at one of the two pits and converge towards one of the two peaks.

Two or More Genes

If we consider two genes A and B at the same time, we must be more careful.

Case 1. Separate Chromosomes. Two genes A and B are said to be **unlinked** if they lie on different chromosomes.

In this case. the allele frequencies a_i for A and b_k for B can be considered as independent random variables, and the theory goes through much as before:

Let $\Phi(A_{i_1} A_{i_2} B_{k_1} B_{k_2})$ be the "fitness" for genotype $A_{i_1} A_{i_2} B_{k_1} B_{k_2}$. Then we can define a corresponding fitness function

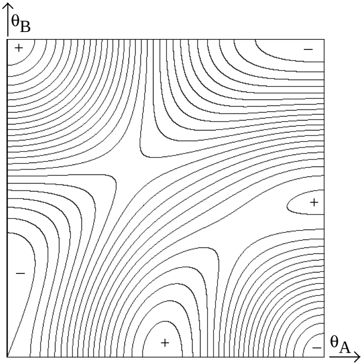
$$\Phi = \sum \Phi(A_{i_1} A_{i_2} B_{k_1} B_{k_2})\, a_{i_1} a_{i_2} b_{k_1} b_{k_2} \,.$$

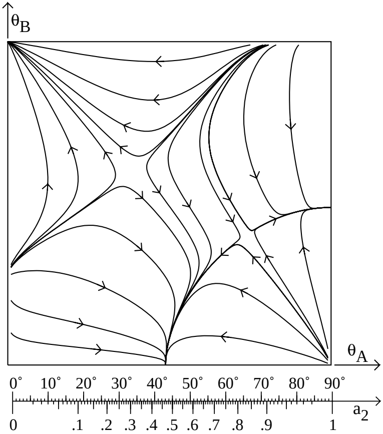
The appropriate phase space is now the cartesian product $S_A \times S_B$ of standard spherical simplexes for A and B,

and we again get a gradient dynamical system:

$$\frac{d\vec{u}}{dt} = \frac{1}{8}\, \overrightarrow{\mathrm{grad}}_{S_A \times S_B} \Phi(\vec{u})\,, \quad \text{so that} \quad \frac{d\Phi}{dt} = \frac{1}{8}\, \|\overrightarrow{\mathrm{grad}}_{S_A \times S_B} \Phi\|^2 \,.$$

An Example





On the left: Contour levels for one possible fitness landscape for two unlinked genes *A* and *B* with two alleles each, plotted as a function of the root frequency angles θ_A and θ_B.

On the right: The associated gradient flow.

Case 2: Genes on the Same Chromosome

In the case of two genes A and B which belong to the same chromosome, the allele probabilities a_i and b_k are no longer independent random variables.

Such genes are said to be **linked**.

Theorem (Akin 1981, Hastings 1981). The appropriate differential equation for $d\vec{u}/dt$ in the case of two linked genes is **not** a gradient dynamical system.

There can be periodic orbits.

Fitness does **not** always increase.

Hence Fisher's Fundamental Theorem also fails.

One can ask whether such models with linked genes can lead to chaotic dynamical systems, with sensitive dependence on initial conditions. (The Poincaré-Bendixson Theorem says that this cannot happen in dimension two; but higher dimensional examples could well be chaotic.)

The End: Some References

- R. A. Fisher [1930], “The Genetical Theory of Natural Selection” (also 1958, 1999).
- Sewall Wright [1932], The roles of mutation, inbreeding, crossbreeding and selection in evolution. (See: “Evolution, Selected Papers” (1986), p. 163.)
- S. Shahshahani [1979], A new mathematical framework for the study of linkage and selection, Memoirs AMS **17**, no. 211.
- E. Akin [1980], “The Geometry of Population Genetics,” Springer.
- ——— [1981/82], Cycling in simple genetic systems, J. Math. Biol. **13**, 305-324.
- A. Hastings [1981], Stable cycling in simple genetic models, Proc. Nat. Acad. Sci. USA **78**, 7224-7225.
- L. Margulis and D. Sagan [1997], “Slanted Truths,” Springer.
- W.F. Doolittle [1999], Phylogenetic classification and the universal tree, Science **284** (25 Jun) 2124-2128.
- N. Patterson et al. [2006], Genetic evidence for complex speciation of humans and chimpanzees, Nature, 29 June.