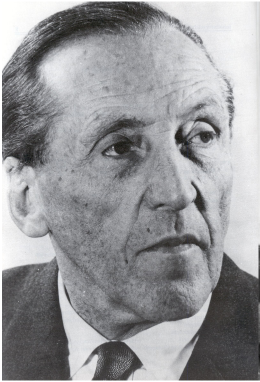
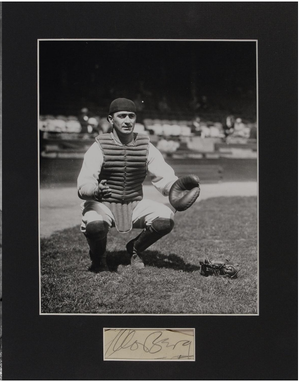
Die Eva der Mitochondrien

Christoph Leuenberger

Mathematikdepartement und
Institut für quantitative Wirtschaftsforschung,
Universität Fribourg

Villigen, 30. März 2011



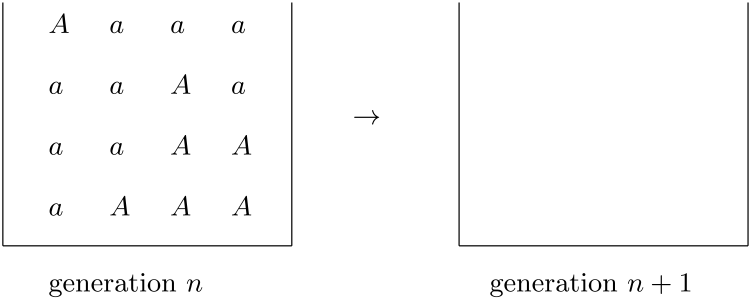


Darwin an Wallace, 1866

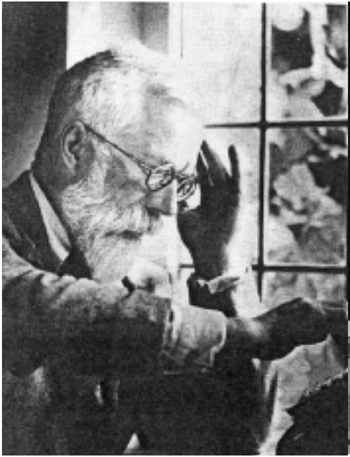
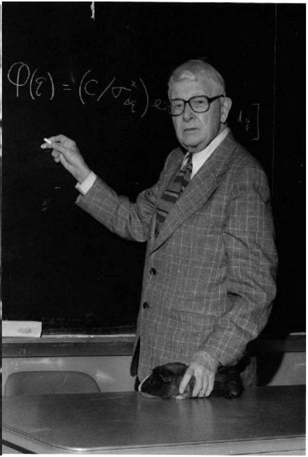
«I do not think you understand what I mean by the non-blending of certain varieties. It does not refer to fertility ; an instance I will explain. I crossed the Painted Lady and Purple sweetpeas, which are very differently coloured varieties, and got, even out of the same pod, both varieties perfect but not intermediate. Something of this kind I should think must occur at least with your butterflies ...»

G. H. Hardy (1877 – 1947) und Wilhelm Weinberg (1862 – 1937)


Fisher-Wright Modell



Ronald A. Fisher (1890 – 1962) und Sewall Wright (1889 – 1988)

Hardy-Weinberg Gleichgewicht

Allele : A rote Blätter
a weisse Blätter

Genotyp	Häufigkeit
AA	p^2
Aa	$2pq$
aa	q^2

Hardy, 1908

«To the Editor of Science : I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists. However, some remarks of Mr. Udny Yule, to which Mr. R. C. Punnett has called my attention, suggest that it may still be worth making ...»

Markov-Eigenschaft des Fisher-Wright Modells

Definiere : $X_t =$ Anzahl Allele A in Generation t

$$p_i(t) = \text{Prob}(X_t = i), \quad i = 0, 1, \ldots, 2N$$

$$P_{ij} = \text{Prob}(X_{t+1} = i | X_t = j) = \binom{2N}{i} p^i (1-p)^{2N-i}$$

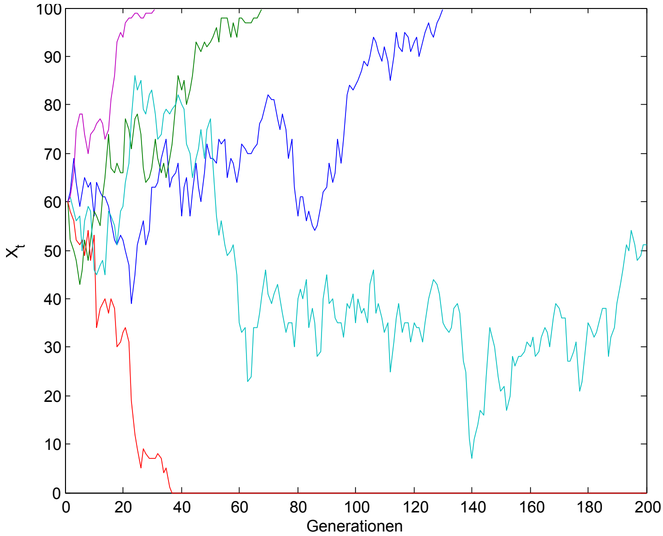
mit $p = j/2N$

Dann gilt : $$p_i(t+1) = \sum_{j=0}^{2N} P_{ij} p_j(t)$$

$$\mathbf{p}(t+1) = \mathbf{P}\,\mathbf{p}(t)$$

$$\mathbf{p}(t) = \mathbf{P}^t\,\mathbf{p}(0)$$

Matlab-Simulation des Fisher-Wright-Modells



Eigenvektoren

$$\mathbf{P}\mathbf{v}_k = \lambda_k \mathbf{v}_k, \quad k = 0, 1, \ldots, 2N,$$

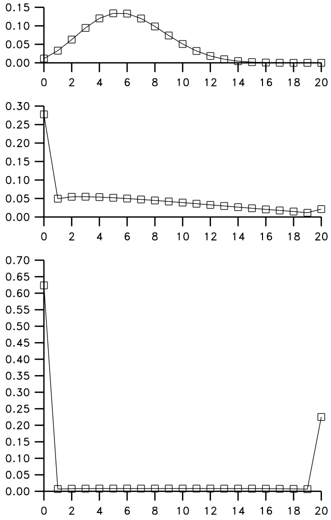
mit $\lambda_k = 1, 1, 1 - \frac{1}{2N}, (1 - \frac{1}{2N})(1 - \frac{2}{2N}), \ldots, \prod_{i=1}^{2N-1} (1 - \frac{i}{2N}).$

Setze $\mathbf{p}(0) = \sum_{k=0}^{2N} c_k \, \mathbf{v}_k.$

Verteilung nach t Generationen

$$\mathbf{p}(t) = \sum_{k=0}^{2N} c_k \, \lambda_k^t \, \mathbf{v}_k$$

Verteilung der Allele



Stationäre Verteilung

$$\mathbf{p}(\infty) = \begin{pmatrix} 1 - X_0/2N \\ 0 \\ \vdots \\ 0 \\ X_0/2N \end{pmatrix}$$

Martingal-Eigenschaft

Setze $x_t = X_t/2N$

$$\mathbb{E}(x_t|x_{t-1}) = x_{t-1}$$

$$\mathbb{E}[\mathbb{E}(x_t|x_{t-1})] = \mathbb{E}(x_{t-1})$$

Somit :

$$\mathbb{E}(x_t) = \mathbb{E}(x_{t-1}) = \ldots = \mathbb{E}(x_0) = x_0$$

Berechnung der Varianz von x_t

$$x_{t+1} = x_t + e_t$$

$$\begin{aligned}
\mathbb{E}(e_t|x_t) &= 0 \\
\mathbb{V}(e_t|x_t) &= \mathbb{E}(e_t^2|x_t) - [\mathbb{E}(e_t|x_t)]^2 = x_t(1-x_t)/2N \\
\mathbb{E}(x_{t+1}^2|x_t) &= \mathbb{E}((x_t+e_t)^2|x_t) \\
&= x_t^2 + 2x_t\mathbb{E}(e_t|x_t) + \mathbb{E}(e_t^2|x_t) \\
&= x_t^2 + x_t(1-x_t)/2N \\
\mathbb{E}(x_{t+1}^2) &= \mathbb{E}(x_t^2) + \mathbb{E}(x_t)/2N - \mathbb{E}(x_t^2)/2N \\
&= \mathbb{E}(x_t^2)\left(1 - \frac{1}{2N}\right) + \mathbb{E}(x_t)/2N
\end{aligned}$$

Berechnung der Varianz von x_t (Forts.)

$$\mathbb{E}(x_t) = x_0, \ \mathbb{V}(x_t) = \mathbb{E}(x_t^2) - x_0^2$$

$$\begin{aligned} \mathbb{E}(x_{t+1}^2) &= \mathbb{E}(x_t^2)\left(1 - \frac{1}{2N}\right) + \mathbb{E}(x_t)/2N \\ x_0^2 + \mathbb{V}(x_{t+1}) &= \left(x_0^2 + \mathbb{V}(x_t)\right)\left(1 - \frac{1}{2N}\right) + x_0/2N \end{aligned}$$

Berechnung der Varianz von x_t (Forts.)

$$
\begin{aligned}
x_0^2 + \mathbb{V}(x_{t+1}) &= \left(x_0^2 + \mathbb{V}(x_t)\right)\left(1 - \frac{1}{2N}\right) + x_0/2N \\
x_0(1-x_0) - \mathbb{V}(x_{t+1}) &= \left(x_0(1-x_0) - \mathbb{V}(x_t)\right)\left(1 - \frac{1}{2N}\right) \\
x_0(1-x_0) - \mathbb{V}(x_{t+1}) &= x_0(1-x_0)\left(1 - \frac{1}{2N}\right)^{t+1}
\end{aligned}
$$

Varianz von x_t

$$
\mathbb{V}(x_t) = x_0(1-x_0)\left[1 - \left(1 - \frac{1}{2N}\right)^t\right]
$$

Heterozygotie

$\mathcal{H}_t$:	Wahrscheinlichkeit, dass zwei Allele der Generation t verschienden sind
$\mathcal{G}_t = 1 - \mathcal{H}_t$:	Wahrscheinlichkeit, dass zwei Allele der Generation t identisch sind

$$\begin{aligned} \mathcal{H}_{t+1} &= \left(1 - \frac{1}{2N}\right) \mathcal{H}_t \\ \mathcal{H}_t &= \mathcal{H}_0 \left(1 - \frac{1}{2N}\right)^t \end{aligned}$$

Effektive Populationsgrösse

$$\mathcal{H}_0 \left(1 - \frac{1}{2N_{eff}}\right)^t = \mathcal{H}_0 \prod_{i=0}^{t-1} \left(1 - \frac{1}{2N_i}\right)$$

$$\exp(-\frac{t}{2N_{eff}}) \approx exp(-\sum_{i=0}^{t-1} \frac{1}{2N_i})$$

Harmonisches Mittel

$$N_{eff} \approx \left(\frac{1}{t} \sum_{i=0}^{t-1} \frac{1}{N_i}\right)^{-1}$$

Mutation versus Drift

$\mu =$ Wahrscheinlichkeit einer Mutation in einen neuen Zustand (*infinite alleles model*)

$$
\begin{aligned}
\mathcal{G}_{t+1} &= (1-\mu)^2 \left[\frac{1}{2N} + (1 - \frac{1}{2N})\mathcal{G}_t\right] \\
&\approx (1-2\mu) \left[\frac{1}{2N} + (1 - \frac{1}{2N})\mathcal{G}_t\right] \\
&\approx \frac{1}{2N} + (1 - \frac{1}{2N})\mathcal{G}_t - 2\mu\mathcal{G}_t
\end{aligned}
$$

Mutation versus Drift (Forts.)

$$\mathcal{H}_{t+1} \approx (1 - \frac{1}{2N})\mathcal{H}_t + 2\mu(1 - \mathcal{H}_t)$$

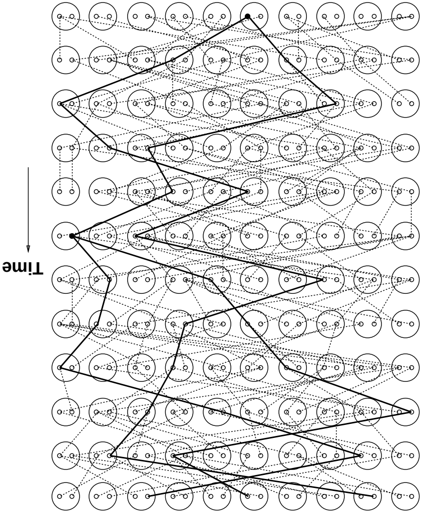
$$\begin{aligned} \Delta\mathcal{H}_t &\approx -\frac{1}{2N}\mathcal{H}_t + 2\mu(1 - \mathcal{H}_t) \\ &= \Delta_{drift}\mathcal{H}_t + \Delta_\mu\mathcal{H}_t \end{aligned}$$

Im Gleichgewicht : $\Delta\mathcal{H}_t = 0$

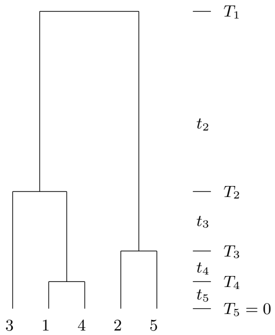
$$\hat{\mathcal{H}} = \frac{4N\mu}{1 + 4N\mu} = \frac{\theta}{1 + \theta} \quad \text{wobei} \quad \theta = 4N\mu$$

John F. Kingman (geb. 1939)

Locking Backward : Kingman's Coalescent



Coalescent-Baum von Kingman



$t_k =$ Anzahl Generationen, während denen k Abstammungslinien vorhanden sind

Erwartungswert von t_k

$$\text{W'keit, dass zwei Töchter die gleiche Mutter wählen} \approx \frac{k(k-1)}{2} \cdot \frac{1}{2N}$$

$$\begin{aligned}
\text{Prob}(t_k > m) &\approx \left(1 - \frac{k(k-1)}{2} \cdot \frac{1}{2N}\right)^m \\
&\approx \exp\left(-\frac{k(k-1)}{2} \cdot \frac{m}{2N}\right) \\
&= e^{-\lambda m} \quad \text{wobei} \quad \lambda = \frac{k(k-1)}{4N}
\end{aligned}$$

$$\mathbb{E}(t_k) = \frac{1}{\lambda} = \frac{4N}{k(k-1)}$$

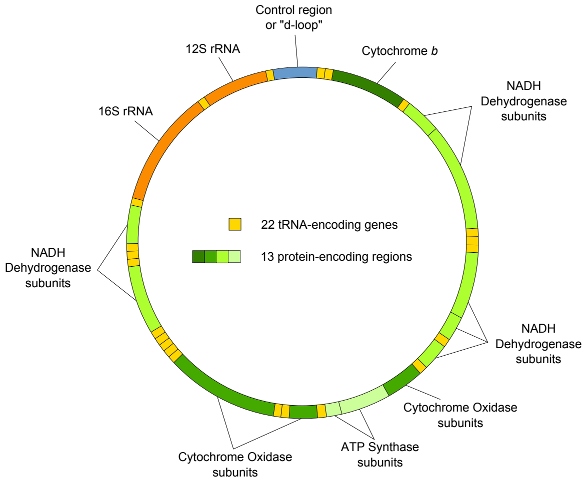
Zeit bis zum MRCA

MRCA	=	«Most Recent Common Ancestor»
T_{MRCA}	$=$	Anzahl Generationen bis zum MRCA für ein Sample von n Individuen

$$
\begin{aligned}
T_{MRCA} &= t_2 + t_3 + \ldots + t_n \\
\mathbb{E}(T_{MRCA}) &= \sum_{k=2}^{n} \mathbb{E}(t_k) = \sum_{k=2}^{n} \frac{4N}{k(k-1)} \\
&= 4N \sum_{k=2}^{n} \left(\frac{1}{k-1} - \frac{1}{k} \right) = 4N \left(1 - \frac{1}{n} \right)
\end{aligned}
$$

$\mathbb{E}(T_{MRCA}) \approx 4N = 2 \cdot$ Populationsgrösse

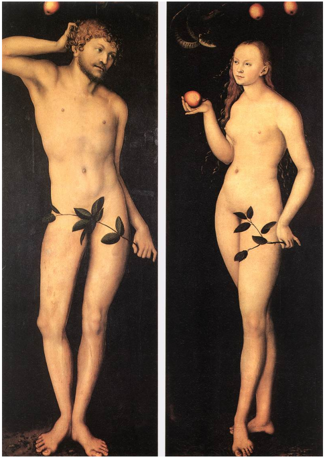
Mitochondriale DNS



Basensequenz der mtDNA (Beginn)

gatcacaggt ctatcaccct attaaccact cacgggagct ctccatgcat ttggtatttt cgtctggggg gtgtgcacgc gatagcattg cgagacgctg gagccggagc accctatgtc gcagtatctg tctttgattc ctgccccatt ccattattta tcgcacctac gttcaatatt acaggcgagc atacttactg aagtgtgtta attaattaat gcttgtagga cataataata acgactaaat gtctgcacag ctgctttcca cacagacatc ataacaaaaa atttc-cacca aacccccccct cccccgcttc tggccacagc acttaaacac atctctgcca aaccccaaaa acaaagaacc ctaacaccag cctaac-caga tttcaaattt tatcttttgg cggtatacac ttttaacagt caccccctaa ctaacacatt attttcccct cccactccca tactactaat ctcatcaata caacccccgc ccatcctacc cagcacacac cgctgctaac cccatacccc gagccaacca aaccccaaag acaccccccca cagtttatgt agct-tacctc ctcaaagcaa tacactgaaa atgtttagac gggctcacat caccccataa acaaataggt ttggtcctag cctttctatt agctcttagt aa-gattacac atgcaagcat ccccattcca gtgagttcac cctctaaatc accacgatca aaagggacaa gcatcaagca cgcaacaatg cagct-caaaa cgcttagcct agccacaccc ccacgggaaa cagcagtgat aagcctttag caataaacga aagtttaact aagctatact aaccccaggg ttggtcaatt tcgtgccagc caccgcggtc acacgattaa cccaagtcaa tagaagccgg cgtaaagagt gttttagatc accccctccc caa-taaagct aaaactcacc tgagttgtaa aaaactccag ttgacacaaa ataaactacg aaagtggctt taacatatct gaacacacaa tagctaa-gac ccaaactggg attagatacc ccactatgct tagccctaaa cctcaacagt taaatcaaca aaactgctcg ccagaacact acgagccaca gcttaaaact caaaggacct ggcggtgctt catatccctc tagaggagcc tgttctgtaa tcgataaacc ccgatcaacc tcaccacctc ttgctcagcc tatataccgc catcttcagc aaaccctgat gaaggctaca aagtaagcgc aagtacccac gtaaagacgt taggtcaagg tgtagcccat gaggtgg-caa gaaatgggct acattttcta ccccagaaaa ctacgatagc ccttatgaaa cctaagggtc gaaggtggat ttagcagtaa actgagagta gagtgcttag ttgaacaggg ccctgaagcg cgtacacacc gcccgtcacc ctcctcaagt atacttcaaa ggacatttaa ctaaaacccc tacg-cattta tatagaggag acaagtcgta acatggtaag tgtactggaa agtgcacttg gacgaaccag agtgtagctt aacacaaagc acccaactta

Y-Adam und Eva der Mitochondrien



Wie sah Eva aus ?

Anzahl segregierende Sites

$S_n =$ Anzahl segregierende Sites in einem Sample von n Individuen

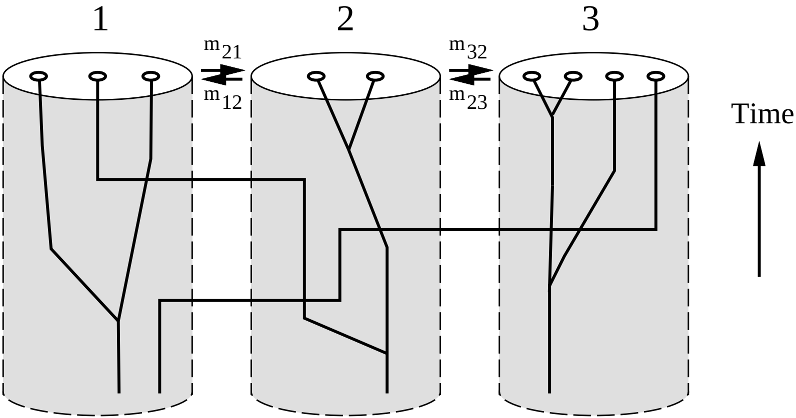
$T_{tot} =$ Gesamtlänge des Baumes

$$\begin{aligned} \mathbb{E}(T_{tot}) &= \mathbb{E}\left[\sum_{k=2}^{n} k t_k\right] = \sum_{k=2}^{n} k\mathbb{E}(t_k) = \sum_{k=2}^{n} k \, \frac{4N}{k(k-1)} \\ &= 4N \sum_{k=2}^{n} \frac{1}{k-1} \approx 4N \ln n \\ \mathbb{E}(S_n) &= \mu \cdot \mathbb{E}(T_{tot}) = \theta \ln n \quad \text{wobei} \quad \theta = 4N\mu \end{aligned}$$

Schätzer für θ

$$\hat{\theta} = \frac{S_n}{\ln n}$$

Coalescent mit Migration



aus *The Autobiography of Charles Darwin*

«I attempted mathematics, and even went during the summer of 1828 with a private tutor (a very dull man) to Barmouth, but I got on very slowly. The work was repugnant to me, chiefly from my not being able to see any meaning in the early steps in algebra. This impatience was very foolish, and in after years I have deeply regretted that I did not proceed far enough at least to understand something of the great leading principles of mathematics, for men thus endowed seem to have an extra sense.»