Sequence Space

All possible sequences of a given length composed of a given alphabet

- binary sequences 0 and 1
- RNA — A C G U
- DNA — A C G T
- proteins — 20 amino acids

Sequence Space is Huge :

In a K letter alphabet, the number of sequences of length L is $\Omega = K^L$

L	Ω (K=2) binary	Ω (K=4) RNA	Ω (K=20) proteins
4	16	64	160000
12	4096	16.7×10^6	4×10^{15}
100	1.27×10^{30}	1.61×10^{60}	1.27×10^{130}

cf total human population 5×10^9

Avogadro's number 6×10^{23}

Hamming distance = number of positions at which two sequences differ.

Measures distance in sequence space.

Number of sequences at a Hamming distance d from any given sequence is

$$\omega(d) = (K-1)^d \frac{L!}{d!(L-d)!}$$

Example with L = 12

d	ω(d) (K=2)	ω(d) (K=4)
0	1	1
1	12	36
2	66	594
3	220	5940
4	495	40095
5	792	192456
6	924	673596
7	792	1732104
8	495	3247695
9	220	4330260
10	66	3897234
11	12	2125764
12	1	531441

Mean Hamming distance between two random sequences is

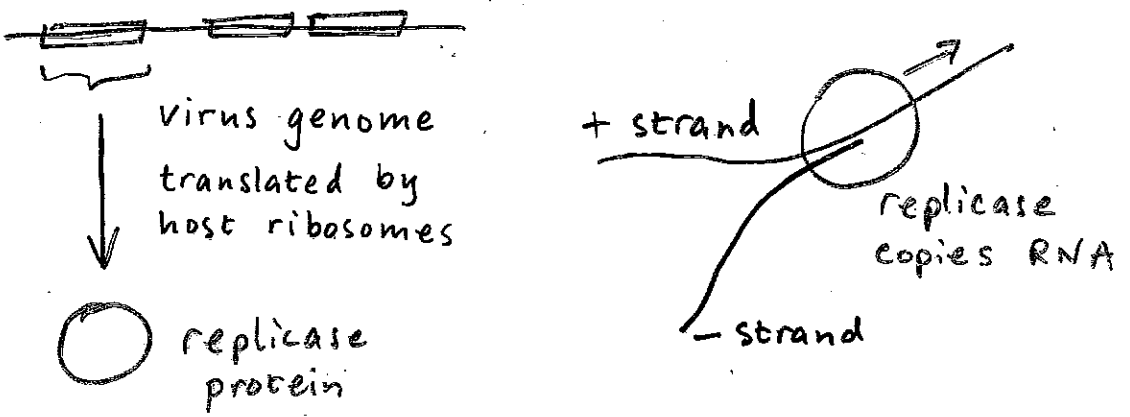
$\langle d \rangle = (K-1)\, L/K$

Binary sequence space - L = 4

RNA sequence space - L = 2

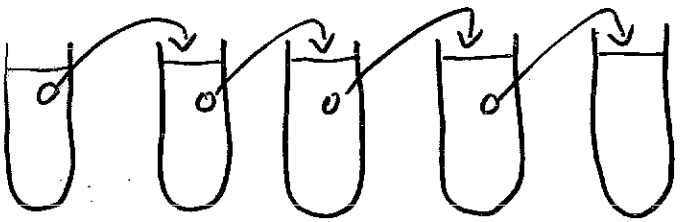
Qβ system

Qβ is an RNA virus which infects bacteria. Single strand ~ 4500 bases.
Codes for replicase enzyme (RNA dependent RNA polymerase).
Copies + strand to -, and - to +. Two step replication.



Error rate $u = 5 \times 10^{-4}$. Fidelity per base $q = 1-u$.
Overall fidelity $Q = q^L = 0.1$

In vitro experiment - replicase + activated ribonucleotides + RNA template



Sequence evolves so as to be most rapidly replicated.
Sequence length tends to decrease by elimination of unnecessary parts -
eg midi-variant $L = 218$
Short sequences can also be formed de novo which can then act as templates.

Sequence analysis indicates there is a distribution of related sequences, not just a single one. This is the quasispecies.

Continuous equations — Chemist's Way

Binary sequences of length L.

x_i = conc of sequence i $\quad i = 1 \ldots 2^L$

A_i = replication rate
D_i = death / breakdown rate } sequence specific

Step 1 $\quad \dot{x}_i = (A_i - D_i)\, x_i$

excess production of seq i is $E_i = A_i - D_i$

→ what happens ?

Step 2 Add dilution term in order to keep total concentration constant.

$$\dot{x}_i = (A_i - D_i)\, x_i - \bar{E}\, x_i$$

$\bar{E}$ ↑ dilution rate

Define $\sum_i x_i = 1$ $\quad \therefore \sum_i \dot{x}_i = 0$

$\therefore$ Must have $\quad 0 = \sum_i (A_i - D_i)\, x_i - \bar{E} \sum_i x_i$

$$\therefore \bar{E} = \sum_i (A_i - D_i)\, x_i$$

mean excess production

→ what happens ?

Need to account for replication error

q = fidelity per base

$u = 1 - q$ = error rate per base

Q_{ij} = probability that sequence i is produced by replication of sequence j

$$Q_{ij} = (1-u)^{L-d_{ij}} u^{d_{ij}}$$

where d_{ij} = Hamming dist

Prob of correct replication of whole sequence is

$$Q_{ii} = (1-u)^L \approx e^{-uL} \quad \text{if } u \ll 1, \; L \gg 1$$

uL is average number of mistakes per replication.

$$\dot{x}_i = (Q_{ii} A_i - D_i) x_i + \sum_{j \neq i} Q_{ij} A_j x_j - \bar{E} x_i$$

note $\sum_i \dot{x}_i = 0$ still

what happens ?

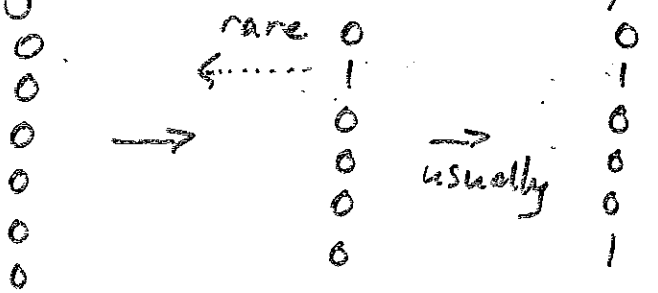
Look at special case.
Assume D_i is independent of i.

Simplifies to $\dot{x}_i = Q_{ii} A_i x_i + \sum_{j \neq i} Q_{ij} A_j x_j - \bar{A} x_i$

Assume a Master sequence "0" with replication rate A_0 and let all other sequences have equal replication rate A_1 $\quad A_0 > A_1$

Can 1 good sequence beat many bad ones?

Can neglect "back mutations" if $L \gg 1$

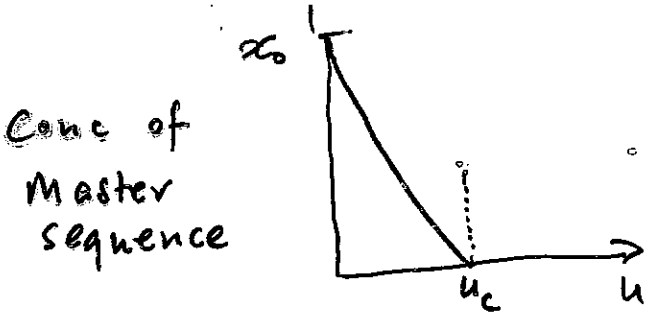


$$\therefore \dot{x}_0 = e^{-uL} A_0 x_0 - \bar{A} x_0$$

At equil $\dot{x}_0 = 0$ $\qquad \bar{A} = x_0 A_0 + (1 - x_0) A_1$

$$0 = e^{-uL} A_0 x_0 - (x_0(A_0 - A_1) + A_1) x_0$$

$x_0 = 0$ $\quad$ OR $\quad x_0 = \dfrac{e^{-uL} A_0 - A_1}{A_0 - A_1}$



$x_0 \to 0$ when $u \to u_c$ error threshold

$$e^{-u_c L} A_0 = A_1$$

$$u_c = \frac{1}{L} \ln\left(\frac{A_0}{A_1}\right)$$

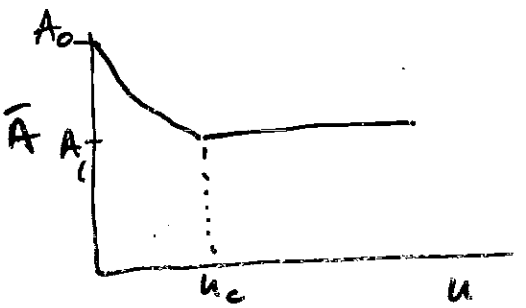
A higher fitness sequence can survive more errors.

Longer sequences require fewer errors

Issue of self replicating molecules at the Origin of life . . .

$$\bar{A} = A_0 e^{-uL} \quad \text{if } u < u_c$$

$$= A_1 \quad \text{if } u > u_c$$



Mutation Rates

RNA viruses – usually $10^{-3} - 10^{-5}$ per nucleotide

longer genomes have lower mutation rates

DNA based micro-organisms (reviewed by JW Drake (1991, 1993))

range $10^{-7} - 10^{-10}$

~1 million times more accurate than RNA

	genome size b.p.	mutation rate per b.p.
Bacteriophage M13	6.4×10^3	7.2×10^{-7}
" T2	1.6×10^5	2.7×10^{-8}
E. coli	4.7×10^6	4.1×10^{-10}
S. cerevisiae	1.4×10^7	2.8×10^{-10}
N. crassa	4.2×10^7	4.5×10^{-11}

mean number per genome ~ 0.03

cf 0.1 – 1.0 for RNA viruses.

Implication of error threshold for origin of life

Length 50

$1-q \equiv u$

$A_0/A_1 = 10 \Rightarrow u_c = 0.046$

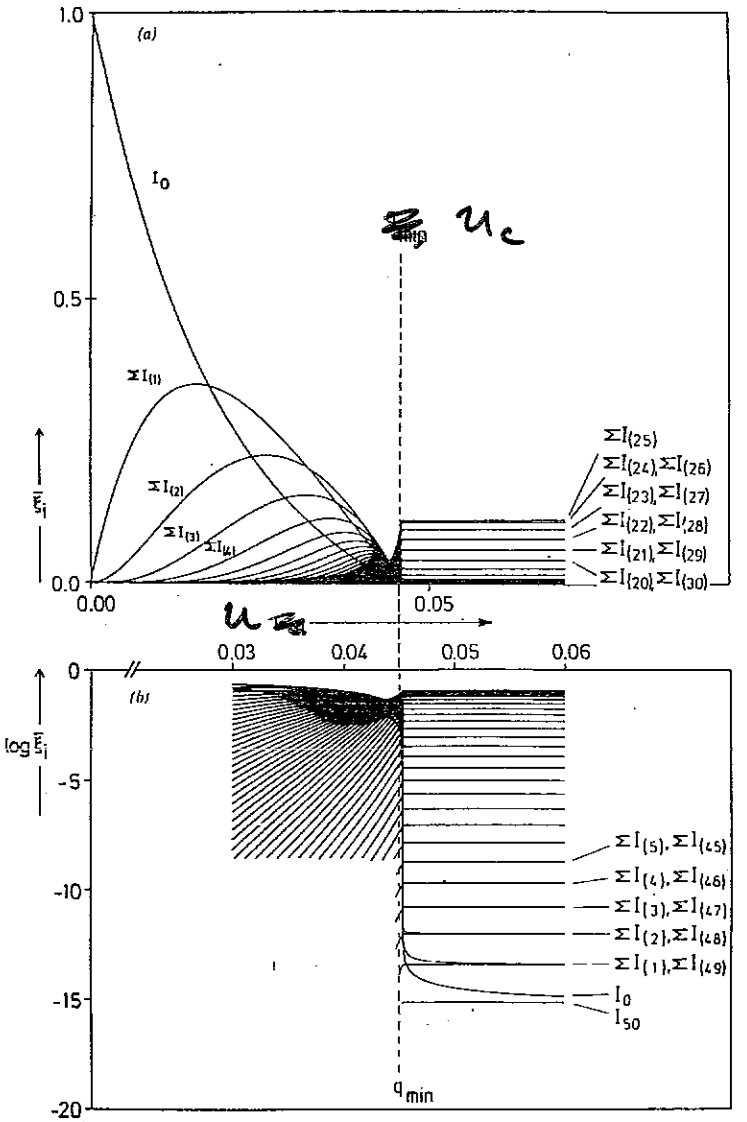


Figure 12. (*Caption overleaf*)

Discrete time equations – Biologist's way

Population of asexual organisms with discrete generati

Fitness w_i = average number of offspring of an individual with gene sequence i

Q_{ij} = prob of mutation from sequence j to i (ie that offspring of j is i)

$= u^d (1-u)^{L-d}$

x_i = frequency of sequence i in the population

$$x_i(t+1) = \frac{1}{\bar{w}} \sum_j Q_{ij} w_j x_j(t)$$

where mean fitness $\bar{w} = \sum_j w_j x_j(t)$

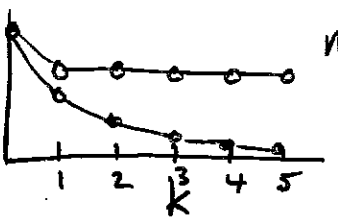
This ensures that population remains constant

$$\sum_i x_i(t+1) = \frac{1}{\bar{w}} \sum_i \sum_j Q_{ij} w_j x_j(t)$$

$$= \frac{1}{\bar{w}} \sum_j w_j x_j(t) = 1$$

Assume single peak with optimal sequence 0

Assume fitness depends on Hamming distance from optimum.



Master sequence landscape
$w_k = w_1$ for $k > 1$

Multiplicative landscape
$w_k = (1-s)^k$

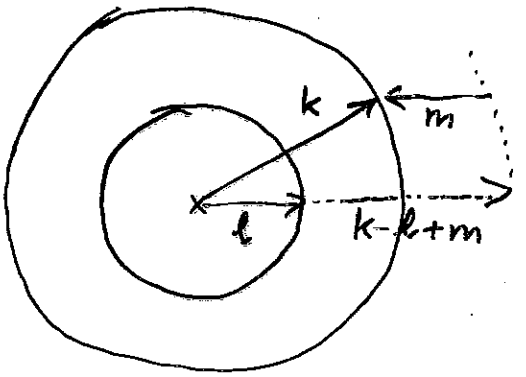
Frequency x_i depends only on distance k from peak.

Let $C_k = \Sigma x_i$ for all sequences at dist k

$$C_k(t+1) = \frac{1}{\bar{w}} \sum_{l=0}^{L} M_{kl}\, w_l\, C_l(t)$$

M_{kl}: prob that a sequence in class l mutates to a sequence in class k

$$\bar{w} = \sum_{l=0}^{L} w_l C_l$$



1 1 1 } l choose m

0 0 0 0 0 0 } $L-l$ choose $k-l+m$

m is the number of back mutations

$$M_{kl} = \sum_{m=m_{min}}^{m_{max}} \binom{l}{m} \binom{L-l}{k-l+m} u^{k-l+2m} (1-u)^{L-k+l-2m}$$

where $m_{min} = \max(0, l-k)$

$m_{max} = \min(l, L-k)$

When $L \gg 1$, $u \ll 1$ can neglect back mutations

i.e. $m_{kl} \approx 0$ if $k < l$

and m_{kl} is dominated by the $m=0$ term if $k \geq l$

$$\therefore \; m_{kl} \simeq \binom{L-l}{k-l} u^{k-l} (1-u)^{L-k+l}$$

$$\approx \frac{(uL)^{k-l}}{(k-l)!} e^{-uL}$$

So, at equil:

$$c_k = \frac{1}{\bar{w}} \sum_{l \leq k} \frac{(uL)^{k-l}}{(k-l)!} e^{-uL} w_l c_l \quad *$$

$k=0$: $\quad c_0 = \frac{1}{\bar{w}} w_0 e^{-uL} c_0$

$$\Rightarrow \quad c_0 = 0 \quad \text{or} \quad \underline{\bar{w} = w_0 e^{-uL}}$$

In master sequence landscape $\bar{w} = w_0 c_0 + w_1 (1-c_0)$

$$\therefore \; c_0 (w_0 - w_1) + w_1 = w_0 e^{-uL}$$

$$c_0 = \frac{w_0 e^{-uL} - w_1}{w_0 - w_1}$$

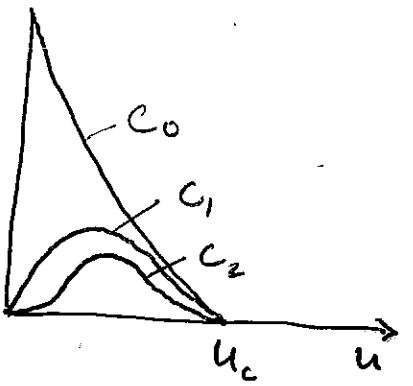
As before!

If $k=1$ in *

$$c_1 = \frac{1}{\bar{w}} e^{-uL} \left(uL\, w_0 c_0 + w_1 c_1 \right)$$

$$(w_0 - w_1)\, c_1 = uL\, w_0\, c_0$$

$$c_1 = \frac{uL\, w_0}{w_1 - w_0}\, c_0$$



Example 2 - Multiplicative landscape

$w_k = (1-s)^k$ each mutation reduces fitness by the same factor

$$c_k = \frac{1}{\bar{w}} \sum_{\ell \le k} \frac{(uL)^{k-\ell}}{(k-\ell)!} e^{-uL} (1-s)^{\ell} c_\ell$$

Solution is: $c_\ell = \frac{(uL/s)^{\ell}}{\ell!} e^{-uL/s}$

Poisson dist

Proove this:

we know that $\bar{w} = w_0 e^{-uL}$

↑ independent of landscape shape!

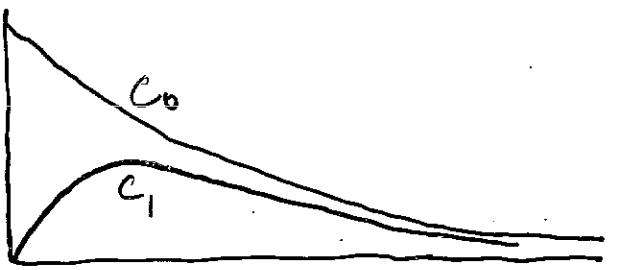
substitute soln.

$$\therefore c_k = \sum_{k \le l} \frac{(uL)^{k-l}}{(k-l)!} (1-s)^l \frac{(uL/s)^l}{l!} e^{-uL/s}$$

$$= \frac{1}{k!} \left(uL + (1-s)\frac{uL}{s} \right)^k e^{-uL/s}$$

$$= \frac{(uL/s)^k}{k!} e^{-uL/s} \qquad \text{Q.E.D.}$$

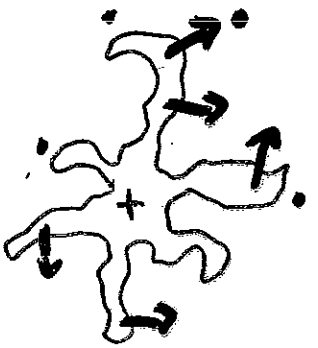
No error threshold:



Error threshold only exists if landscape stays flat as you go far from peak

Consequences of High Mutation Rates in viruses

1. Quasispecies structure :



Variability within population may help evade immune system

May help in developing drug resistance.

⇒ HIV treatment strategies

2. Rapid rate of sequence evolution — ie accumulation of mutations eg Flu.

— observable in laboratory.

3. Subject to accumulation of unfavourable mutations by chance
— Muller's Ratchet

- bottlenecks during transfer

- no meiosis but can get reassortment or recombination in certain cases.

SUMMARY

Quasispecies — a population of related sequences usually centred on a master sequence with high fitness

Parameters :

- L sequence length
- u error rate per nucleotide
- A_0/A_1 relative rate of replication of master sequence to the rest

Prob of exact replication $Q = (1-u)^L \approx e^{-uL}$

Master sequence can maintain itself if

$$QA_0 > A_1$$

$$\Rightarrow u < u_c \quad \text{where} \quad u_c = \frac{1}{L} \ln(A_0/A_1)$$

$\uparrow$

error threshold

Analogy with Physics

Error threshold = Phase transition

$u < u_c$ — Ordered low temp phase

$C_0 > 0$

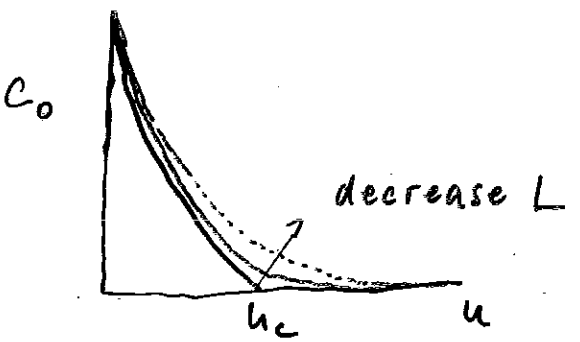
population is localized close to Master sequence

$u > u_c$ — high temp phase

$C_0 = 0$

population is delocalized over whole of sequence space

Finite Size effects



transition becomes rounded

Formal analogy made with Ising spin system (Lenthäuser)