Sequence Space
All possible sequences of a given length composed of a given alphabet
- binary sequences 0 and 1
- RNA ACGU
- DNA ACGT
- proteins 20 amino acids

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

<table>
<thead>
<tr>
<th>L</th>
<th>$\Omega$ (K=2)</th>
<th>$\Omega$ (K=4)</th>
<th>$\Omega$ (K=20)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>binary</td>
<td>RNA</td>
<td>proteins</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>64</td>
<td>160000</td>
</tr>
<tr>
<td>12</td>
<td>4096</td>
<td>$16.7 \times 10^6$</td>
<td>$4 \times 10^{15}$</td>
</tr>
<tr>
<td>100</td>
<td>$1.27 \times 10^{30}$</td>
<td>$1.61 \times 10^{60}$</td>
<td>$1.27 \times 10^{130}$</td>
</tr>
</tbody>
</table>

cf total human population $5 \times 10^9$
Avogadro's number $6 \times 10^{23}$

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

<table>
<thead>
<tr>
<th>d</th>
<th>$\omega(d)$ (K=2)</th>
<th>$\omega(d)$ (K=4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>12</td>
<td>36</td>
</tr>
<tr>
<td>2</td>
<td>66</td>
<td>594</td>
</tr>
<tr>
<td>3</td>
<td>220</td>
<td>5940</td>
</tr>
<tr>
<td>4</td>
<td>495</td>
<td>40095</td>
</tr>
<tr>
<td>5</td>
<td>792</td>
<td>192456</td>
</tr>
<tr>
<td>6</td>
<td>924</td>
<td>673596</td>
</tr>
<tr>
<td>7</td>
<td>792</td>
<td>1732104</td>
</tr>
<tr>
<td>8</td>
<td>495</td>
<td>3247695</td>
</tr>
<tr>
<td>9</td>
<td>220</td>
<td>4330260</td>
</tr>
<tr>
<td>10</td>
<td>66</td>
<td>3897234</td>
</tr>
<tr>
<td>11</td>
<td>12</td>
<td>2125764</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>531441</td>
</tr>
</tbody>
</table>

Mean Hamming distance between two random sequences is

$$<d> = (K-1) \frac{L}{K}$$
**Qβ system**


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 $L$:

$a_i$ = conc of sequence $i$, $i = 1 \ldots 2^L$.

$A_i$ = replication rate specific for sequence $i$.

$D_i$ = death / breakdown rate specific for sequence $i$.

Step 1

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

Excess production of seq $i$ is $E_i = A_i - D_i$.

$\Rightarrow$ what happens?

Step 2

Add dilution term in order to keep total concentration constant.

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

$dilution \ rate$

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

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

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

Mean excess production

$\Rightarrow$ what happens?
Need to account for replication error

\[ q = \text{fidelity per base} \]

\[ u = 1 - q = \text{error rate per base} \]

\[ Q_{ij} = \text{probability that sequence } i \text{ is produced by replication of sequence } j \]

\[ Q_{ij} = (1-u)^{d_{ij}} \cdot u^{d_{ij}} \]

where \( d_{ij} = \text{Hamming dist} \)

Prob of correct replication of whole sequence is

\[ Q_{ii} = (1-u)^L \approx e^{-uL} \text{ if } u \ll 1 \]

\( uL \) is average number of mistakes per replication.

\[ x_i = (Q_{ii} A_i - D_i) x_i + \sum_{j \neq i} Q_{ij} A_j x_j - E x_i \]

Note \( \sum_i x_i = 0 \) still

what happens?
Look at special case.
Assume $D_i$ is independent of $i$.
Simplifies to

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

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

Can a good sequence beat many bad ones?

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

$$\begin{array}{c}
0 \\
\leadsto 0 \\
\leadsto 0 \\
0 \\
0
\end{array}$$

$$\begin{array}{c}
0 \\
\leadsto 0 \\
\leadsto 0 \\
0 \\
0
\end{array}$$

$$\begin{array}{c}
0 \\
\leadsto 0 \\
\leadsto 0 \\
0 \\
0
\end{array}$$

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

At equil. $\dot{x}_0 = 0$

$$\tilde{A} = x_0 A_0 + (1-x_0) A_1$$

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

$$x_0 = 0 \quad \Rightarrow \quad x_0 = \frac{e^{-uL} A_0 - A_1}{A_0 - A_1}$$
$\alpha_0 \to 0$ when $u \to u_c$

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

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

A higher fitness sequence can survive more errors. Longer sequences require fewer errors.

Issue of self-replicating molecules at the origin of life.

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

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


<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome Size</th>
<th>Mutation Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteriophage M13</td>
<td>$6.4 \times 10^3$</td>
<td>$7.2 \times 10^{-7}$</td>
</tr>
<tr>
<td>&quot; T2</td>
<td>$1.6 \times 10^5$</td>
<td>$2.7 \times 10^{-8}$</td>
</tr>
<tr>
<td>E. coli</td>
<td>$4.7 \times 10^6$</td>
<td>$4.1 \times 10^{-10}$</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>$1.4 \times 10^7$</td>
<td>$2.8 \times 10^{-10}$</td>
</tr>
<tr>
<td>N. crassa</td>
<td>$4.2 \times 10^7$</td>
<td>$4.5 \times 10^{-11}$</td>
</tr>
</tbody>
</table>

mean number per genome ~ 0.03

cf. 0.1 - 1.0 for RNA viruses.

implication of error threshold for origin of life
Length 50

\[ 1 - q = u \]

\[ \frac{A_0}{A_1} = 10 \Rightarrow u_c = 0.046 \]
Discrete time equations - Biologist's way

Population of asexual organisms with discrete generations

Fitness \( w_i \): average number of offspring of an individual with gene sequence \( i \)

\[ q_{ij} = \text{prob of mutation from sequence } j \text{ to } i \]

\[ = u^d (1-u)^{k-d} \]

\( x_i \): 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 \( O \)

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 $C_k = \sum x_i$ for all sequences at dist $k$

$$C_k(t+1) = \frac{1}{\bar{W}} \sum_{l=0}^{L} M_{kl} W_{l} C_{l}(t)$$

The probability that a sequence in class $l$ mutates to a sequence in class $k$ is

$$\bar{W} = \sum_{l=0}^{L} W_{l} C_{l}$$

$m$ is the number of back mutations

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

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

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

\[
M_{kl} \approx \binom{L-l}{k-l} u^{k-l} (1-u)^{l-k+l}
\]

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

So, at equil:

\[
C_k = \frac{1}{W} \sum_{l \leq k} \frac{(uL)^{k-l}}{(k-l)!} e^{-uL} \quad w_c C_0 \times
\]

\[
k=0 : \quad C_0 = \frac{1}{W} w_c e^{-uL} C_0
\]

\[
\Rightarrow C_0 = 0 \quad \text{or} \quad \bar{w} = w_c e^{-uL}
\]

In master sequence landscape \( \bar{w} = w_0 C_0 + w_1 (1-C_0) \):

\[
C_0 (w_0 - w_1) + w_1 = w_0 e^{-uL}
\]

\[
C_0 = \frac{w_0 e^{-uL} - w_1}{w_0 - w_1}
\]

As before!
\[ \sum_{k=1}^{\infty} c_k = \frac{1}{W} \sum_{k=1}^{\infty} e^{-uL} \left( uL w_0 c_0 + w_1 c_1 \right) \]

\[ (w_0 - w_1) c_1 = uL w_0 c_0 \]

\[ c_1 = \frac{uL w_0 c_0}{w_1 - w_0} \]

Example 2 - Multiplicative landscape

\[ w_k = (1-s)^k \]

Each mutation reduces fitness by the same factor

\[ c_k = \frac{1}{W} \sum_{l=1}^{k} (\frac{uL}{k-l})^{k-l} \frac{e^{-uL}}{(k-l)!} (1-s)^l c_l \]

Solution is:

\[ c_l = (\frac{uL/s}{l})^{l} \frac{e^{-uL/s}}{l!} \]

---

Poisson dist
Proof this:

we know that \( \bar{w} = w_0 e^{-ul} \)

\[ \text{independent of landscape shape!} \]

Substitute soln.

\[ C_k = \sum_{k \leq l} \frac{(ul)^{k-l}}{(k-l)!} \left( 1-s \right)^l \left( \frac{ul}{s} \right)^k \frac{e^{-ul/s}}{l!} \]

\[ = \frac{1}{k!} \left( ul + (1-s)ul \right)^k e^{-ul/s} \]

\[ = \frac{(ul/s)^k e^{-ul/s}}{k!} \quad \text{Q.E.D.} \]

No error threshold:

Error threshold 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 - i.e., accumulation of mutations
   - Observable in laboratory
   - Examples: Flu

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
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 $Q A_0 > A_1$

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

$\uparrow$ error threshold
Analogy with Physics

Error threshold = Phase transition

\[ n < n_c \quad \text{ordered low temp phase} \]
\[ C_0 > 0 \]
population is localized close to master sequence

\[ n > n_c \quad \text{high temp phase} \]
\[ C_0 = 0 \]
population is delocalized over whole of sequence space

Finite Size effects

\[ C_0 \]
\[ \text{decrease } L \]
transition becomes rounded

Formal analogy made with Ising spin system (Leutkäuser)

\[ P_r \]
\[ \text{time direction} \]