Lecture Two

Genome: Information Carrier at the “Edge of Chaos”

HC Lee
National Central University
Life at the edge of chaos

- Edge of chaos
  - Computational system
  - Cellular automata
  - Transition to criticality

- Life at the Edge of chaos
  - Life involves complex computation
  - Technical apparatus for description still missing

- Genome as Life
  - Chaos as a state of near randomness
  - Textual complexity of a genome represent computational ability
  - Dynamics of genome evolution
Consider genome with fractional AT-content $p$ (then fractional GC-content $q=1-p$)
  - When $p>0.5$, there will be more AT-rich words than GC-rich words
Partition $k$-letters words ($k$-mer) into sets, called $m$-sets $S_m$, $m=0,1,...,k$; elements in $S_m$ are $k$-mers with $m$ ATs, $\bigcup_m S_m = S$
Total number of kinds of $k$-mers is $\tau=4^k$, of kinds of $k$-mers in $m$-set is $\tau_m$. Let $u$ be a $k$-mer,

\[ L_m = \sum_{u \in S_m} f_u \quad \sum_{u \in S_m} f_u = L - k + 1 \quad \sum_{u \in S_m} f_u = \sum_m L_m = L \]

\[ \tau_m = \binom{k}{m} 2^k, \quad L_m^{\infty} = L \binom{k}{m} p^m q^{k-m}, \]

\[ \bar{f} = L/\tau, \quad \tau = 4^k, \quad \bar{f}_m^{\infty} = L_m^{\infty} / \tau_m \]
Shannon entropy (briefly)

- **Shannon entropy** for a system frequency set \( \{f_i| \Sigma_i f_i = L\} \) or a spectrum \( \{n_f\} \) is

  \[
  H = - \Sigma_i f_i/L \log (f_i/L) = - \Sigma_f n_f \cdot f/L \log (f/L)
  \]

- Suppose there are \( \tau \) types of events: \( \Sigma_i = \tau \). Then \( H \) has **maximum value** when every \( f_i \) is equal to \( N/\tau \):

  \[
  H_{max} = \log \tau
  \]

- For a genomic \( k \)-frequency set: \( \tau = 4^k \), \( L = \) genome length.

  \[
  H_{max} = 2k \log 2
  \]
Let $D$ be coefficient of invariance (SD/mean) of distribution of frequency of occurrence of $k$-mers:

$$D = (CV)^2 = \left(\frac{\sigma}{\bar{f}}\right)^2 = \frac{1}{\tau \bar{f}^2} \sum_{u \in G_k} (f_u - \bar{f})^2$$

- Random sequence: $D \sim L^{-1/2}$
- Define $l_{\text{eff}}$ for a genome to be the random sequence length having the genomic $D$
- Genomes have (almost) universal $l_{\text{eff}}$, about 150-600 bases long (for 2-mers)
Shannon information & relative spectral width

- **Shannon information**: information is decrease in $H$: define
  \[ R = \log \tau - H \]

- Relation to **relative spectral width** (for unimodal distribution)
  \[ R = \sigma^2/2 + O(\sigma^3) \]

- Shannon information and relative spectral width ("fluctuation part" from Lecture 1) are equivalent measures

Shannon called $R/H_{\text{max}}$ redundancy; Gatlin (1972) called $R$ divergence

\[ R = \log \tau - H \] is a good definition

Table 1: Shannon entropy \( H \) and information \( R \) in units of \( \log 2 \) in the \( k \)-spectra of the genome sequence of \( P. \ aerophilum \) and of the random sequence obtained by randomizing the genome. \( R_{ex} \) is the expected information in a random sequence. Sequences have \( \text{AT/CG= 50/50} \)

<table>
<thead>
<tr>
<th>( k )</th>
<th>Random sequence</th>
<th>Genome sequence</th>
<th>( R_{gen}/R_{ran} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( H )</td>
<td>( R )</td>
<td>( R_{ex} )</td>
</tr>
<tr>
<td>2</td>
<td>3.9999</td>
<td>5.90 E-6</td>
<td>5.77 E-6</td>
</tr>
<tr>
<td>3</td>
<td>5.9999</td>
<td>3.72 E-5</td>
<td>3.46 E-5</td>
</tr>
<tr>
<td>4</td>
<td>7.9999</td>
<td>1.72 E-4</td>
<td>1.62 E-4</td>
</tr>
<tr>
<td>5</td>
<td>9.9993</td>
<td>7.26 E-4</td>
<td>7.53 E-4</td>
</tr>
<tr>
<td>6</td>
<td>11.999</td>
<td>2.94 E-3</td>
<td>2.90 E-3</td>
</tr>
<tr>
<td>7</td>
<td>13.988</td>
<td>1.18 E-3</td>
<td>1.17 E-3</td>
</tr>
<tr>
<td>8</td>
<td>15.955</td>
<td>4.78 E-2</td>
<td>4.71 E-2</td>
</tr>
<tr>
<td>9</td>
<td>17.798</td>
<td>2.02 E-1</td>
<td>1.88 E-1</td>
</tr>
<tr>
<td>10</td>
<td>19.xxx</td>
<td>x.xx E-1</td>
<td>5.24 E-1</td>
</tr>
</tbody>
</table>
An Order Index $\phi$

$L_m$ of random sequence of infinite length

$L_m$: frequency sum of $k$-mers in $m$-set

$\phi \equiv \frac{1}{(2 - 2(p^k + q^k))} \sum_m \frac{1}{L} \left| L_m - L_m^{\{\infty}\} \right|$

Total sequence length = $\Sigma_m L_m$

$\phi$ of (semi-)ordered sequence

Note: $\phi$ is a measure of differential in averages
An (semi-)ordered sequence
AT...TATTATATTAATATTTAGCCGGGCGGGCGC...GG
or a checker-board sequence
...AGAGTGACAGTCTGTCTTCACTG...
have $\phi \sim 1$

A random sequence has
$\phi \sim L^{-1/2} \sim 0$ for large $L$
\[ \phi \text{ scales as } L^{-1/2} \text{ for random sequences} \]

Depends only weakly on \( k \) or AT-content (\( p \)).
Averaged over \( k \) and \( p \):

\[ \phi^{\{\text{ran}\}}(k; p) = c_\phi L^{-\gamma_\phi}, \quad \gamma_\phi = 0.500 \pm 0.005, \quad c_\phi = 1.0 \pm 0.2 \]
An equivalent length $L_\phi$ for order index

Use the relation

$$\phi \{\text{ran}\} \approx L^{-1/2}$$

to define an (order-index) Equivalent Length for a $\phi$-valued sequence:

$$L_\phi(\phi) = \phi^{-2},$$

the nominal length of a (non-random) sequence whose order index is $\phi$.

(Note. Unlike the CV-defined and $k$-dependent $L_e$, $L_\phi$ is essentially independent of $k$.)
\( \phi \) decreases exponentially with increasing number of point mutations

For an ordered sequence, \( \phi \) drops exponentially from \( \phi=1 \) as a function of \( N_\mu \) until it reaches a critical point when the sequence has become random.

\( N_\mu : \# \) of random mutations

\[
\phi = \begin{cases} 
\exp \left(-2N_\mu/L\right), & N_\mu \lesssim N_{\mu c}; \\
\phi_c \approx L^{-1/2}, & N_\mu > N_{\mu c} 
\end{cases}
\]
ϕ decreases exponentially with increasing number of point mutations

If sequence already has ϕ₀ < 1, then as a function of Nμ

ϕ = ϕ₀exp(-2Nμ/L),

until ϕ reaches its critical value ϕ_c \sim L^{-1/2}.

Let μ = Nμ/L be mutation density (mutation per site)

- Equivalent mutation density for a ϕ-valued sequence:

\[ \bar{\mu}_{eq}(\phi) \equiv \ln \phi^{-1/2} \]

, the nominal mutation density needed to bring an ordered sequence to a state of ϕ
The critical mutation density

Given

\[ \phi = \begin{cases} \exp(-2N_\mu/L), & N_\mu \lesssim N_{\mu c}; \\ \phi_c \approx L^{-1/2}, & N_\mu > N_{\mu c} \end{cases} \]

The critical mutation density that will bring an ordered sequence to a state of randomness is given by

\[ \phi_c = \exp(-2N_\mu/L) = \exp(-2 \mu_c) = L^{-1/2} \]

That is:

\[ \mu_c = (1/4) \ln L \]

(For \( L = 10^6 \) to \( 10^8 \),
\[ \mu_c = 3.4 \) to 4.6/site)
Equivalent mutation density is a path-independent state quantity

• Want to test whether \( \phi \) (or equivalently, \( \mu_{eq} \)) is akin to a potential energy, or a quantity that defines a state, but not how that state is arrived at.

• The 4.6 Mb \( E. \colli \) DID NOT arrive at its present state by random point mutation from an ordered sequence.
• It is measured to have \( \phi = 0.049 \), or \( \mu_{eq} = 1.5 \)/site.
• The critical \( \mu_{eq} \) for a 4.6 Mb sequence is \( \mu_c = \frac{1}{4} \ln (4.6 \times 10^6) = 3.8 \)/site.
• If we assume is \( \phi \) a path-independent state quantity, then we predict it will take \( (\mu_c - \mu_{eq}) \times L = 1.1 \times 10^7 \) additional mutations to randomize \( E. \colli \)
• The actual number is found to be \( 1.1(+/-)0.1 \times 10^7 \)
Considered as a dynamical system driven by mutations, the state of randomness is a fixed point in $\phi$ space.

[Fixed point. Consider a function $f(x)$ mapping a point $x$ in a space to another point $x'$ in the same space. Then $x_*$ is a fixed point of $f$ if 

$$f(x_*) = x_*.$$

Here the action causing the mapping is mutation, and the space is the $\phi$-space. Mutation takes the sequence from one $\phi$ to another $\phi$. When the sequence is random, mutation maps $\phi_c$ back to itself.]
For ~800 complete genomes extant in GenBank, $\phi$ is essentially length- and base-composition-independent.

- Genomic $\phi$ congregates in a narrow range

$$\phi_g \equiv 0.037 \pm 0.027$$
Coding (genic) and non-coding parts have similar $\phi$

- Dynamics of genome evolution leading $\phi$ to $\phi_g$ is not under strong (genic) selection pressure
- Predominant characteristics is neutral
• $\mu_{\text{eq}} \sim 1.8 \text{ b}^{-1}$ implies a genome is as random as an ordered sequence becomes after each site has on average been mutated 1.8 times.
• Genome is at the Edge of Chaos
Genomes form a universality class in $\phi$

Vast majority of genomes have

$$\ln \phi_g = -3.49 \pm 0.65$$

or

$$\phi_g = 0.031^{+0.028}_{-0.015}$$

P-value of genomes belonging to the universality class defined by $\phi_g$
37 (out of ~800) chromosomes belong to universality class in $\phi$ with $P<0.05$.

**TABLE II.** Chromosomes belonging to the universal set defined by Eq. (8) with $P<0.05$.

<table>
<thead>
<tr>
<th>Name</th>
<th>Accession no.</th>
<th>$\bar{\phi}$</th>
<th>$P$ value</th>
<th>Name</th>
<th>Accession no.</th>
<th>$\bar{\phi}$</th>
<th>$P$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. aureus</em></td>
<td>9 strains</td>
<td>~4.4(−3)</td>
<td>~3.0(−3)</td>
<td><em>A. marginale</em></td>
<td>4842</td>
<td>4.45(−3)</td>
<td>3.15(−3)</td>
</tr>
<tr>
<td><em>S. epidermidis</em></td>
<td>4461</td>
<td>4.87(−3)</td>
<td>4.89(−3)</td>
<td><em>C. felis</em></td>
<td>7899</td>
<td>5.20(−3)</td>
<td>6.66(−3)</td>
</tr>
<tr>
<td><em>L. johnsonii</em></td>
<td>5362</td>
<td>5.58(−3)</td>
<td>9.18(−3)</td>
<td><em>S. hemolyticus</em></td>
<td>7168</td>
<td>5.79(−3)</td>
<td>1.08(−2)</td>
</tr>
<tr>
<td><em>S. epidermidis</em></td>
<td>2976</td>
<td>6.49(−3)</td>
<td>1.77(−2)</td>
<td><em>M. mobile 163 K</em></td>
<td>6908</td>
<td>6.80(−3)</td>
<td>2.14(−2)</td>
</tr>
<tr>
<td><em>T. denitrificans</em></td>
<td>7404</td>
<td>7.12(−3)</td>
<td>2.58(−2)</td>
<td><em>L. acidophilus</em></td>
<td>6814</td>
<td>7.34(−3)</td>
<td>2.90(−2)</td>
</tr>
<tr>
<td><em>G. sulfurreducens</em></td>
<td>2939</td>
<td>7.40(−3)</td>
<td>2.99(−2)</td>
<td><em>F. tularensis</em></td>
<td>7880</td>
<td>7.50(−3)</td>
<td>3.15(−2)</td>
</tr>
<tr>
<td><em>W. succinogenes</em></td>
<td>5090</td>
<td>7.51(−3)</td>
<td>3.17(−2)</td>
<td><em>C. hydrogenoformans</em></td>
<td>7503</td>
<td>1.23(−1)</td>
<td>3.20(−2)</td>
</tr>
<tr>
<td><em>M. hungatei</em></td>
<td>7796</td>
<td>7.75(−3)</td>
<td>3.57(−2)</td>
<td><em>F. tularensis</em></td>
<td>6570</td>
<td>7.90(−3)</td>
<td>3.84(−2)</td>
</tr>
<tr>
<td><em>C. caviae</em></td>
<td>3361</td>
<td>7.94(−3)</td>
<td>3.91(−2)</td>
<td><em>M. succiniciproducens</em></td>
<td>6300</td>
<td>1.15(−1)</td>
<td>4.04(−2)</td>
</tr>
<tr>
<td><em>C. abortus</em></td>
<td>4552</td>
<td>8.06(−3)</td>
<td>4.14(−2)</td>
<td><em>X. fastidiosa 9a5c</em></td>
<td>2488</td>
<td>8.12(−3)</td>
<td>4.25(−2)</td>
</tr>
<tr>
<td><em>P. marinus</em></td>
<td>7335</td>
<td>8.19(−3)</td>
<td>4.37(−2)</td>
<td><em>S. tokodaii</em></td>
<td>3106</td>
<td>8.47(−3)</td>
<td>4.96(−2)</td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>Chr V</td>
<td>6.00(−3)</td>
<td>1.26(−2)</td>
<td><em>S. cerevisiae</em></td>
<td>Chr XV, III</td>
<td>~7.7(−3)</td>
<td>~3.5(−2)</td>
</tr>
<tr>
<td><em>S. cerevisiae</em></td>
<td>Chr VI</td>
<td>8.43(−3)</td>
<td>4.87(−2)</td>
<td><em>A. mellifera</em></td>
<td>8 chrs.</td>
<td>~1.1(−1)</td>
<td>~4.8(−2)</td>
</tr>
</tbody>
</table>

---

$^a$4842 indicates the accession no. NC_004842.

$^b$The value 4.4(−3) means $4.4 \times 10^{-3}$.

$^c$The nine strains, in order of increasing $P$ value, are 3923, 2953, 7793, 7795, 2952, 7622, 2951, 2758, and 2745.

$^d$The eight chromosomes, in order of increasing $P$ value, are XV, X, XII, II, IV, V, I, and XI.
Artificial sequence with genomic $L_e$ generated in RSD model has genomic $\phi$.

Red lines, range of genomic $\phi$

$\phi$ of model generated sequence

Equivalent length

$10^{-1}$ $10^{-2}$ $10^{-3}$ $10^{-4}$ $10^{-5}$ $10^{-6}$ $10^{-7}$ $10^{-8}$ $10^{-9}$ $10^{-10}$ $10^{-11}$

$\rho$

$0.2$ $0.4$ $0.6$ $0.8$ $1.0$ $1.2$ $1.4$ $1.6$ $1.8$ $2.0$ $2.2$ $2.4$ $2.6$ $2.8$ $3.0$ $3.2$ $3.4$ $3.6$ $3.8$ $4.0$ $4.2$ $4.4$ $4.6$ $4.8$ $5.0$ $5.2$ $5.4$ $5.6$ $5.8$ $6.0$ $6.2$ $6.4$ $6.6$ $6.8$ $7.0$ $7.2$ $7.4$ $7.6$ $7.8$ $8.0$ $8.2$ $8.4$ $8.6$ $8.8$ $9.0$ $9.2$ $9.4$ $9.6$ $9.8$ $10.0$
An empirical function of information capacitance

Define an “information capacity” function $I(z)$ such that: the variable $z$ is a scaling function of $\phi$ with $z|_{\phi=0}=0$ and $z|_{\phi=1}=1$, and $I$ has two minima at $I(0)=I(1)=0$ and a maximum at $z|_{\phi=\phi_g}=0.5$. The simplest solution is

$$I(z) = -z \ln z - (1-z)\ln(1-z); \quad z = \phi^{0.21}.$$  \hfill (10)

Say “information capacitance”, not “information”, because $\phi_g$ is universal but sequence length is not; not “information density”, because not every sequence with $\phi_g$ has information.
Genomes reside in a small distinct region characterized by $\phi \sim \phi_g$ in the space of sequences.
Recall: Sequences are driven by random mutations to a state-of-randomness fixed point. Similarly, genomes are driven by the dynamics of a robust evolutionary process (yet to be identified) to a fixed-point $\phi_g$. 
The two types of fixed points are driven by different dynamics

- **Random sequence fixed point**
  - $\phi_c \sim L^{-1/2}$, depends on sequence length
  - Driven by random point mutation

- **Genome fixed point**
  - $\phi_c = 0.016-0.059$ is universal (and independent on length)
  - Driven by “robust evolution process”
  - Our guess: random (+plus tandem) segmental duplication + plus point mutation
The $\phi \sim \phi_g$ “fixed point” shared by literature classics

The six classics:
- *The Bible*, King James Version;
- *Sonnets*, William Shakespeare;
- *Oliver Twist*, Charles Dickens;
- *Remembrance of Things Past*, Marcel Proust;
- *Ulysses*, James Joyce;
- *A Moveable Feast*, Earnest Hemingway
• Making pseudogenomes of classics
  – (adjlsy) to A; (chiopq) to C; (efgnvxc) to G; (bkmrtuw) to T
  – All six classics have $p_A \sim p_C \sim p_G \sim p_T = 0.250 +/- 0.007$, or $p \sim 0.50 +/- 0.02$.

• The rants (repeated 1M times)
  – “Though this be madness, yet there is method in’t” (Hamlet)
  – “All the perfumes in Arabia will not sweeten this hand” (MacBeth)

• $\pi$: equivalent length close to true length
  – Highly complex sequence, yet low information content.
Characteristics of Information carriers:
- Has “maximum” $I_c$ independent of length
- Are quasi-random – has $\frac{1}{2}$ mut'ns needed for randomization
- Has $\phi$ equivalent to random sequence 250 – 10000 b

Order index
Equivalent length
Equivalent Mutation density
Conjectural Inferences

- $\phi \sim \phi_g$ are high information capacitance states
- The observed shortness of $L_\phi$ suggests that the neutral process is dominated by (fixed, hence non-deleterious) segmental duplications
- No difference in coding and non-coding part suggest process is random/neutral
  - Random: low free-energy, easy access
- Random process can only built infrastructure, not information; actual information is acquired in mostly fitness driven point mutation events
  - Selective: difficult to access
A two-step genome growth

- Genome growth by a two-step process:
  - One neutral, robust, infrastructure-building and universal
  - The other selective, fine-tuning, information-gathering and diverse
  - Example: paradigm of accidental gene duplication followed by mutation driven subfunctionalization

- The twin-processes acted in a ratchet-like, complementary manner, driving the genome, in successive stages, to a state of maximum information capacity, and helping it to acquire, at each stage, near-maximum information content.
End of Lecture Two