Genomes – at the edge of chaos with maximum information capacity

Biophysics Camp for Students
Graduate Institute of Biophysics
National Central University
2007 November 27-28

HC Lee
Computational Biology Lab
Inst. Systems Biology & Bioinformatics
Dept. Physics & Inst. Biophysics
National Central University
Life is highly diverse and complex.
And it took a long time to get here

Divergence of species

BACTERIA
- cyanobacteria
- α-proteobacteria

EUKARYA
- ANIMALS
- FUNGI
- PLANTS
- mitochondria

ARCHAEA
- ARCHEZOA

ONE BILLION YEARS AGO

TWO BILLION YEARS AGO

THREE BILLION YEARS AGO

FOUR BILLION YEARS AGO

now

4 billion yrs ago
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 randomness
  - Textual complexity of a genome represents computational ability
  - Dynamics of genome evolution
Terminology & Notations

• 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, elements are $k$-mers with $m$ ATs; we have $m=0,1,\ldots,k$

• Total number of kinds of $k$-mers is $\tau = 4^k$, of $k$-mers is $L$ (sequence length), of kinds of $k$-mers in $m$-set is $\tau_m$, of $k$-mers in $m$-set is $L_m$.

$$\tau_m = \binom{k}{m} 2^k, \quad L^\{\infty\} = L \left( \binom{k}{m} p^m q^{k-m} \right),$$

$$\bar{f} = L / \tau, \quad \tau = 4^k, \quad \bar{f}_m^\{\infty\} = L_m^\{\infty\} / \tau_m$$
An Order Index

\[ \phi \equiv \frac{1}{(2 - 2(p^k + q^k))} \sum_{m} \frac{1}{L} \left| L_m - L_m^{\infty} \right| \]

- An (semi-)ordered sequence
  \[ AT\ldots TATTAATTTA\text{GCCG} \text{GGCGGC} \text{GCGC} \ldots \text{GG} \]
  or a checker-board sequence
  \[ \ldots \text{AGAGTGACAGTC} \text{TGTC} \text{TCACTG} \ldots \]
  have \( \phi = 1 \)

- A random sequence has \( \phi \sim L^{-1/2} \)
Order index for random sequences and equivalent length

- An equivalent length for a f-valued sequence: 
  \[ L_{eq}(\phi) \equiv \phi^{-2} \], the nominal length of a random sequence whose order index is
Degrade of ordered sequences by mutation and equivalent mutation rate

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

Critical point

• An equivalent mutation rate for a \( \phi \)-valued sequence:

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

, the nominal rate (per length) of a random sequence whose order index is \( \phi \)
For 786 complete genomes extant in GenBank, $\phi$ is essentially length- and base-composition-independent. Total length $2 \times 10^{10}$ bases.

- Genomic $\phi$ congregates in a narrow range

$$\phi_g \equiv 0.037 \pm 0.027$$
\( \phi \) is essentially the same in coding (genic) and non-coding parts and, in genes, the same in mRNA and non-mRNA parts.

- Dynamics of genome evolution leading \( \phi \) to \( \phi_g \) is not under strong (genic) selection pressure.
- Predominant characteristics is neutral.
Genomes are half as random as random sequences

- $\mu_{eq} \sim 1.8 \text{ b}^{-1}$ 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

$\frac{\mu_{eq}(\phi)}{\mu_c} \sim 0.45 \pm 0.11$
Genomes resides in a small distinct space characterized by $\phi \sim \phi_g$ in the space of sequences.

Inference: Genomes are driven to a fixed-point in the space of sequences by the dynamics of a robust evolutionary process.
The $\phi \sim \phi_g$ “fixed-point” is shared by literature classics.

Six literature 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.
• $\phi \sim \phi_g$ are high information capacity states

• The observed shortness of $L_{eq}$ suggests that the neutral process is dominated by (non-deleterious) segmental duplications

• No difference in coding and non-coding part suggest process is random/neutral. A random sequence has no information
  - Random: low free-energy, easy access

• Random process can only built infrastructure, not information; actual information is acquired in mostly point mutation event
  - 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.
Thank you

Lee Lab: Google “HC Lee”