Edge of Chaos and
The Order of Genomes

HC Lee
Institute Systems Biology & Bioinformatics
National Central University
December 10, 2008
Three questions I try to ask

- **WHAT** is the phenomenon?
  - What is strange/unusual/interesting?

- **HOW** did it happen?
  - (Physics)

- **WHY** did it happen?
  - (Biology)
Life is highly diverse and complex

Tree of Life
W.F. Doolittle

We are here
And it took a long time to get here

4 billion yrs ago

now
Evolution of life is recorded in genomes

• Genome is Book of Life
• A double helix - two strands of DNA
• DNA: String of four types of molecules – chemical letters - A, C, G, T
• Genome is a linear text written in four letters
• We believe all genomes have a common ancestor, or a small group of ancestors
Genome is an extremely complex

The 24 human genomes
A stretch of genome from the X chromosome of Homo sapien


The complete genome has 2,000,000 such pages

Genomes are BIG
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 represent 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 $L_m$.

\[
\tau_m = \binom{k}{m}2^k, \quad L_m^{\{\infty\}} = L\left(\binom{k}{m}p^mq^{k-m}\right),
\]
\[
\bar{f} = L/\tau, \quad \tau = 4^k. \quad \bar{f}_m^{\{\infty\}} = L_m^{\{\infty\}}/\tau_m
\]
An Order Index

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

• By design, ordered sequence has \( \phi = 1 \)

• Random sequence has \( \phi \sim L^{-1/2} \)
  • Infinitely long random sequence \( \phi = 0 \)
• Concatenating two dissimilar sequences reduces $\phi$

• E.g.
  • (semi-)ordered sequence
    AT…TATTATTAATATTTAGCCGCGCGCGCGC…GG
    and checker-board sequence
    …AGAGTGACAGTCTGTCTCACTG…
    both have $\phi = 1$
  • Concatenating the two yields $\phi = 0$
• An equivalent length for a \( \phi \)-valued sequence: 
\[
L_{eq}(\phi) \approx \phi^{-2}
\]
, the nominal length of a random sequence whose order index is
Degrade of ordered sequences by mutation and equivalent mutation rate

A critical point

Randomness is a fixed-point

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

- An equivalent mutation rate for a \( \phi \)-valued sequence:
  \[ \hat{\mu}_{eq}(\phi) \equiv \ln \phi^{-1/2} \]
  the nominal rate (per length) of a random sequence whose order index is \( \phi \)
• $\phi$ is essentially independent of length and composition

• Genomic $\phi$ congregates in a narrow range

$$\phi_g = 0.037 \pm 0.027$$
\(\phi\) 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 half as random as random sequences: Edge of Chaos

- 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

\[ \mu_{eq}(\phi)/\mu_c \approx 0.45 \pm 0.11 \]
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

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
Artificial sequences generated in “Random segmental duplication” model automatically driven to $\phi \sim \phi_g$ “fixed-point”

RSD model for genome growth:
- Randomly select a site
- Copy segment of some length
- Randomly select a site for insertion of copied segment

At left:
- 340 artificial sequences generated using single set of parameters.
- all have $\phi \sim \phi_g$
• $\phi \sim \phi_g$ 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”