Some concepts to be discussed

• By examining at the textual property of genomes, we encounter/exploit the following concepts
  - Randomness and order
  - Second law of thermodynamics
  - (Shannon) Information and entropy
  - Distribution and its variance
  - Diversity and universality
  - Complexity and self-similarity
  - Neutral evolution and natural selection

• and arrive at a hypothesis and model for genome growth
We are here

Life is highly diverse and complex
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
A stretch of genome from the X chromosome of Homo sapien


The complete genome has 2,000,000 such pages
Evolution of Genomes and the Second Law of Thermodynamics

Genomes grew & evolved stochastically
  • modulated by natural selection
  • Bigger genomes carry more information than smaller ones

• The second law of thermodynamics:
  • the entropy of closed system can never decrease
  • a system that grows stochastically tends to acquire entropy
  • Increased randomness → more entropy

• Shannon information
  • Information decreases with increasing entropy

• How was genome able to simultaneously grow stochastically AND acquire information?
Characterization of Genomes

- **Primary characterization of genomes**
  - length in bp (base pair)
  - base composition $p = \frac{A+T}{A+T+C+G}$
  - word frequencies

- **Secondary characterization**
  - % coding region (microbials: ~85%; eukaryotes (2~50%)
  - number of genes (few hundred to 25K)

- **Tertiary characterization**
  - intron/exon (microbials, no; eukaryotes, yes)
  - other details
Complete Genomes are diverse

PF: *Plasmodium falciparum* (A eukaryotic Malaria causing parasite)
Consider \( \tau \) equally probable events occurring a total of \( L \) times.

Distribution of occurrence frequency characterized by

- mean frequency: \( f_{\text{ave}} = \frac{L}{\tau} \)
- SD (standard deviation) \( \Delta \); or
  - \( CV \) (coefficient of variation) = \( \frac{\Delta}{f_{\text{ave}}} \)
- Higher moments of distribution
Random events

- Random events given by Poisson distribution
  - $\Delta^2 = f_{\text{ave}}$, or, $(CV)^2 = 1/ f_{\text{ave}}$
  - That is, $(CV)^2 = \tau / L$

- For fixed $\tau$, $(CV)^2 \sim 1/L$
  - Large $L$ limit (thermodynamic limit): $L \sim \infty$, $CV \sim 0$

- For given $\tau$, if $CV$ is known, then
  - $L \sim \tau / (CV)^2$
Genome as text - Frequencies of k-mers

- Genome is a text of four letters - A,C,G,T
- Frequencies of k-mers characterize the whole genome
  - E.g. counting frequencies of 7-mers with a "sliding window"
  - Frequency set \( \{ f_i \mid i=1 \text{ to } 4^k \} \)

\[
N(\text{GT}T\text{A}C\text{C}C) = N(\text{GT}T\text{A}C\text{C}C) + 1
\]

...AACGTTACCCGCGTTATATG...
For genomes: events=word occurrence; type of events $\tau$=types of words = $4^k$; distr.=distr. of frequency of occurrence

Black/red: genome of *E. coli*
Green: matching random sequence

Huge difference between genome and random sequence
Given $\tau$ and CV, define effective length

$$L_{\text{eff}} = \frac{\tau}{(CV)^2}$$

- The $L_{\text{eff}}$ of complete genomes are far shorter than their actual lengths.
- For a given type of event (word counts) $L_{\text{eff}}$ is universal.
  - Actual length varies by factor $>1000$.
  - "Information" in genomes grows as $L$. 

Two big surprises from complete genomes.
Large CV, or small $L_{\text{eff}}$, implies more “information”

Compare $L_{\text{eff}}$ with true length $L$ for all complete genomes for 2-10 letter words

$$(CV_{\text{genome}})^2 = \tau/L_{\text{eff}}$$

$$(CV_{\text{random}})^2 = \tau/L$$

$$M_s = (CV_{\text{genome}})^2 / (CV_{\text{random}})^2 = L/L_{\text{eff}}$$

Note: technical details when $p$ not equal to 0.5
Shannon entropy

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

\[
H = - \sum_i f_i/L \log (f_i/L) = - \sum_f n_f f/L \log (f/L)
\]

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

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

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

\[
H_{\text{max}} = 2k \log 2
\]
Shannon information & coefficient of variation

- **Shannon information**: information decreases with increasing $H$. Define:

  $$ R = \log \tau - H $$

- Relation to **coefficient of variation** (for unimodal distribution)

  $$ R \equiv \ln\tau - H(\mathcal{F}) = L^{-1} \sum_i f_i \ln(f_i/\bar{f}) = (CV)^2/2 + ... $$

- Shannon information and coefficient of variation are equivalent measures

  (Note: technical detail re biased base composition important)
**R = log τ - 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 AT/CG = 50/50**

<table>
<thead>
<tr>
<th>$k$</th>
<th>$H$</th>
<th>$R$</th>
<th>$R_{ex}$</th>
<th>$H$</th>
<th>$R$</th>
<th>$R_{gen}/R_{ran}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3.9999</td>
<td>5.90 E-6</td>
<td>5.77 E-6</td>
<td>3.973</td>
<td>2.66 E-2</td>
<td>4500</td>
</tr>
<tr>
<td>3</td>
<td>5.9999</td>
<td>3.72 E-5</td>
<td>3.46 E-5</td>
<td>5.933</td>
<td>6.65 E-2</td>
<td>1922</td>
</tr>
<tr>
<td>4</td>
<td>7.9999</td>
<td>1.72 E-4</td>
<td>1.62 E-4</td>
<td>7.881</td>
<td>1.18 E-1</td>
<td>728</td>
</tr>
<tr>
<td>5</td>
<td>9.9993</td>
<td>7.26 E-4</td>
<td>7.53 E-4</td>
<td>9.821</td>
<td>1.79 E-1</td>
<td>246</td>
</tr>
<tr>
<td>6</td>
<td>11.999</td>
<td>2.94 E-3</td>
<td>2.90 E-3</td>
<td>11.75</td>
<td>2.74 E-1</td>
<td>94</td>
</tr>
<tr>
<td>7</td>
<td>13.988</td>
<td>1.18 E-3</td>
<td>1.17 E-3</td>
<td>13.66</td>
<td>3.35 E-1</td>
<td>29</td>
</tr>
<tr>
<td>8</td>
<td>15.955</td>
<td>4.78 E-2</td>
<td>4.71 E-2</td>
<td>15.53</td>
<td>4.69 E-1</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>17.798</td>
<td>2.02 E-1</td>
<td>1.88 E-1</td>
<td>17.26</td>
<td>7.33 E-1</td>
<td>3.0</td>
</tr>
<tr>
<td>10</td>
<td>19.xxx</td>
<td>x.xx E-1</td>
<td>5.24 E-1</td>
<td>19.xx</td>
<td>x.xx E-1</td>
<td>-</td>
</tr>
</tbody>
</table>
Results: color coded by organisms

Each point from one k-spectrum of one sequence; >2500 data points. Black crosses are microbial. Data shifted by factor $2^{10-k}$
Data from 14 Plasmodium chromosomes excluded; ~2400 data points. For each $k$, 268 data points form a narrow $M_\sigma \sim L$ “$k$ band”.
Each \( k \)-band defines a universal constant
\[
\frac{L}{M} = L_{\text{eff}} \sim \text{constant}
\]
(Effective root-sequence length)

- Defines a universal class
- Plasmodium has separate class:
  \[
  a = 0.146 \pm 0.012
  \]

Obeys
\[
\log L_r(k) = a \, k + B
\]

1989 pieces of data given be two parameters.
\[
a = 0.398 \pm 0.038 \\
B = 1.61 \pm 0.11
\]

Black: genome data; green: artificial
Self-similarity: "Sameness" at varying scales

We have seen: complete genome of length \( L \) has stats property of random sequence of length \( L_{\text{eff}} \ll L \)

- Question 1: what is the stats property of a \( L' > L_{\text{eff}} \) segment of the genome?
- Question 2: what happens when we concatenate two such segments?
Testing self-similarity

<table>
<thead>
<tr>
<th>Behavior of $(CV)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Random sequence</strong></td>
</tr>
<tr>
<td>Any segment $L \gg L' \gg L_{\text{eff}}$</td>
</tr>
<tr>
<td>Concatenation of two segments $L \gg L_1, L_2 \gg L_{\text{eff}}$</td>
</tr>
</tbody>
</table>
Two examples: H. sapien and E. coli: genomes are highly self-similar

Figure 1: RSW ($M_\sigma$) of $k$-spectra, $k=2$ to 10, of segments from the 246 Mb chromosome 1 of H. sapiens. Lengths of the segments are $1/2^n$ of full length, $n=1$ to 21, and for each length eight segments are randomly selected. Data for which segment length is less than $4^k$ are not included. Data for the same $k$ forms a $k$-band approximately linear in $L$ (red line), and each data point has been multiplied by factor of $2^{10-k}$ to delineate the $k$-bands for better viewing.
$L_u$ and $L_d$, $k=5$, all complete sequences

Figure 3: $L_u$ (the length above which all segments are similar to the genome; green bars) and $L_d$ (the length below which no segment is similar to the genome; red bars) for $k=5$ for all complete sequences in the main universality class. The blue (yellow) line is the position of $L_{\text{max}}$ ($L_{\text{min}}$).
Genomes are maximally self-similar

Table 3: Comparison of $4^k$ and mean values of $L_r(k)$ and $L_{sim}(k)$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$4^k$</th>
<th>$\langle L_r \rangle$</th>
<th>$\langle L_{sim} \rangle$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>16</td>
<td>310±200</td>
<td>690±570</td>
</tr>
<tr>
<td>3</td>
<td>64</td>
<td>680±350</td>
<td>1300±990</td>
</tr>
<tr>
<td>4</td>
<td>256</td>
<td>1690±760</td>
<td>2820±1700</td>
</tr>
<tr>
<td>5</td>
<td>1024</td>
<td>4450±1900</td>
<td>6690±3200</td>
</tr>
<tr>
<td>6</td>
<td>4096</td>
<td>12300±5200</td>
<td>16400±7200</td>
</tr>
<tr>
<td>7</td>
<td>16384</td>
<td>33600±15000</td>
<td>42700±18000</td>
</tr>
<tr>
<td>8</td>
<td>65536</td>
<td>89500±43000</td>
<td>109000±44000</td>
</tr>
</tbody>
</table>

- $L_{sim}$ is the average of prokaryotic $L_u$ & $L_d$ & eukaryotic $L_d$
- $L_{sim}$ barely $> L_r$ barely $> 4^k$,
- Hence genomes are almost maximally self-similar
Given a sequence of events. Consider the distribution of intervals between adjacent events

- Random events: distribution is exponential \([d_{ave} = \text{average interval}]\)
  
  \[N(d) \sim N_0 \exp(-d/d_{ave})\]

- Conversely, if distribution is exponential, then infer events occurred randomly (or vice versa)
Words occurred in genome randomly

- Have already seen genomes are highly non-random

- Yet, distributions of words intervals in genomes are universally exponential to a high degree of accuracy
Interval distribution is exponential in random sequence as expected. But not so in genome!
Summary of genome data

• **Universality class** – for fixed word length $k$, $L_{\text{eff}}$ is (approximately) the same for all genomes
  \[ \log L_{\text{eff}}(k) = ak + B \]
  $a$, $B$ are universal constants

• **Maximally self-similar**

• $k$-mer intervals have exponential distribution

• **What is the cause of these properties?**
• If we take random sequence of length \( L_0 \) and replicate it \( n \) times, then total sequence length (L) is \( nL_0 \) but \( L_{\text{eff}} \) of sequence remains \( L_0 \).

• Smaller \( L_{\text{eff}} \) implies higher degree of ORDER.

• Larger \( L_{\text{eff}} \) implies higher degree of RANDOMNESS.

• Small \( L_{\text{eff}} \) of genomes suggests many DUPLICATIONS.
A Universal Model for Genome Growth

A model: at a **universal initial length**, genomes grew (and diverged) by **maximally stochastic segmental duplication**

1. **Universal initial length** - Common ancestor(?), universal $L_{eff}$.
2. **Segmental duplication** – $L$-independent CV
3. **Maximum stochasticity** – self-similarity, random word interval

**Self copying** – strategy for retaining and multiple usage of hard-to-come-by coded sequences (i.e. genes)
\[ \chi^2 = \left\langle \left[ \frac{(L_r)_{\text{model}} - (L_r)_{\text{gen}}}{\Delta (L_r)_{\text{gen}}} \right]^2 \right\rangle \]

Model parameter search: favors very small \( L_0 \)
Model has three universal parameters - generates universal $L_{\text{eff}}$.

Effective root-sequence length

Red & blue symbols are from (same) model sequences.
Model sequences are maximally self-similar - $L_{sim}$ agrees with data

Note: Model predates data

But model has smaller spread

Model is too smooth
• Estimate rates for human
  \[ r_S \sim 2 \text{ /site/By}, \quad r_D \sim 3.4/\text{Mb/My} \]
• Human genome grew 15-20% last 50 My
• References
    • Estimated silent site substitute rates for plants and animals range from 1 to 16 (/site/By) (Li97)
    • Humans: \( r_S \sim 2 \) (Lynch00) or 1 (Liu03) /site/By.
    • Animal gene duplication rate \( \sim 0.01 \) (0.002 to 0.02) per gene per My (Lynch00)
    • Human (coding region \( \sim 3\% \) of genome) translates to 3.9/Mb/My.
    • Human retrotransposition event rate \( \sim 2.8/\text{Mb/My} \) (Liu03)
• Average rates from model if $T = 4 \text{ By}$
  \[ <r_s> \sim 0.25/\text{site}/\text{By}, \quad <r_D> \sim 0.50/\text{Mb}/\text{My} \]

• About 7~8 time smaller than recent sequence divergence estimates

• Arguments
  - Can estimate substitution and duplication rate if assign total growth time
  - Human genome still growing last 50 My
  - Hence assume total growth time for human genome $T \sim 4 \text{ By}$
Empirical rates $r_{S,D}$ for last $\Delta T \sim 50$ My are terminal rates.

Model rates $<r_{S,D}>$ averaged over whole growth history, hence

$$<r_{S,D}> < r_{S,D}$$

Given constant duplication rate $r_D$ per length per unit time and constant average duplicated segment length $\lambda$, then genome grew exponentially. Fit to data gives

$$L(t) \sim 1 \ (Mb) \ exp(t/0.5 \ (By))$$
Remarks on \[ L(t) \sim 1 \text{ (Mb)} \exp(t/0.5 \text{ (By)}) \]

• Our model can be reconciled with alignment based data on evolution
• HS genome grew by \(~12\%\) last 50My
  – Liu et al. grew by \(~15-19\%\) last 50My
• Does not imply \( L=1 \text{ Mb at } t=0 \)
• Does imply at \( t \sim 500\text{My}, L \sim 1 \text{ Mb} \)
Summary on genome data and growth model

- Genomes form a *universality class* defined by:
  - universal effective lengths
  - maximally self-similarity
  - Random correlation between words

- Genome-like sequences are generated by simple growth model characterized by:
  - Three universal parameters
  - maximally stochastic segmental duplication
  - Very early onset of duplication process

- For HS genome, model consistent with evolution rates extracted by sequence divergence methods
More phenomena explained by model

- Preponderance of **homologous genes** in all genomes
- Numerous **pseudogenes** in eukaryotic genomes (85% microbial genome is coded)
- Genome is full of **non-coding repeats**
- Large-scale genome “**rearrangements**”
- Rapid **rate of evolution** - random self-copying is an extremely efficient way for information accumulation; growth by random self-copying is likely the result of natural selection
- Many more ...
Are genes “spandrels”? 

- **Spandrels**
  - In _architecture_. The roughly triangular space between an arch, a wall and the ceiling
  - In _evolution_. Major category of important evolutionary features that were originally side effects and did not arise as adaptations (Gould and Lewontin 1979)

- Duplications to a genome are what the construction of arches, walls and ceilings are to a cathedral
- Codons are the spandrels and genes are décorations in the spandrels
## Cross-disciplinary Similarities

<table>
<thead>
<tr>
<th>Control</th>
<th>Economics</th>
<th>Biological</th>
<th>Genomes</th>
<th>Games</th>
</tr>
</thead>
<tbody>
<tr>
<td>theory</td>
<td></td>
<td>cells</td>
<td></td>
<td></td>
</tr>
<tr>
<td>process</td>
<td>activities</td>
<td>phenotypic</td>
<td>coding</td>
<td>board</td>
</tr>
<tr>
<td>variables</td>
<td></td>
<td>features</td>
<td>regions</td>
<td>config</td>
</tr>
</tbody>
</table>
Characteristics of a *cas* (Holland)

A complex adaptive system, *cas*, is an evolving, perpetually novel set of interacting agents where

- There is no universal competitor or global optimum.
  There is no BEST genome (organism)

- There is great diversity, as in a tropical forest, with many niches occupied by different kinds of agents.
  There is great diversity in genomes

- Innovation is a regular feature – equilibrium is rare and temporary
  Genomes evolve continuously

- Anticipations change the course of the system.
  ??
• Genome is the system, genes and other codes are the agents
• Duplicated segments are the building blocks, site replacements (mutations) are the innovations
• Metaphor
Questions to be answered

• How to reconcile with or quantify effect of natural selection
• Can model be refined to differentiate coding and non-coding regions?
• Can model be extended to describe the rise of genes and gene families, regulatory sequences, …?
• Is model consistent with phylogeny?
• Can model say anything about the origin of life? (RNA world?)
‘Replicators’

• Early in the RNA world, RNA genes are thought to have been able to directly copy themselves (albeit imperfectly).

• Such RNAs are described as ‘replicators’.

A. Poole, The RNA world & LUCA
The Eigen Limit: a paradox of prebiotic evolution

- The amount of information that can be maintained in a genome is limited by the accuracy (fidelity) of replication.
Avoiding Catch-22

- Imagine the best replicator is shorter than max. length dictated by error threshold
- A longer mutant with higher copying fidelity can emerge, which allows new max. length
- Longer sequences are sequentially possible, and stepwise increase in replication fidelity occurs.
More fundamental problems

• Cause for variation in base composition – why is base composition different from organism to organism but (almost uniform in a genome?
• How to reconcile universal growth model with apparent genome specific substitution rate?
• HS genome is still growing (our luck) but microbial genomes must gained its current size 2-3 billion years ago. How do microbial genomes maintain “universal” stats property under effect of constant mutation but with no growth?
• Can our model be COMPLETELY WRONG?
Some recent papers on segmental duplication and evolution

  Recent Segmental Duplications in the Human Genome
  Jeffrey A. Bailey, et al. (Eichler group)
  http://www.sciencemag.org/cgi/content/full/297/5583/1003

- Genome Res. 2003 March 1; 13(3): 358-368.
  Analysis of Primate Genomic Variation Reveals a Repeat-Driven Expansion of the Human Genome. Ge Liu, et al (Eichler group)

  Gene Duplication and Evolution. Michael Lynch
  http://www.sciencemag.org/cgi/content/full/297/5583/945

- Science, Vol 290, Issue 5494, 1151-1155, 10 November 2000
  The Evolutionary Fate and Consequences of Duplicate Genes
  Michael Lynch and John S. Conery
  http://www.sciencemag.org/cgi/content/full/290/5494/1151
‘Accessible’ reading on LUCA and the RNA world
(Anthony Poole)

• http://www.actionbioscience.org/newfrontiers/

• Ridley (2000) The search for LUCA. Natural History November pp. 82-85.


Our papers are found at Google: HC Lee

Thank you!