Lecture Three
Symmetries in the Genome

HC Lee
National Central University
Symmetry in Nature

• Symmetry has been considered as an aspect of beauty in mathematics, physics, chemistry, evolution, human appearance, and psychology.

• The cause of symmetry is usually subtle*, and the pursuit of it often leads to a deep understanding of the possessor of the symmetry.

• *“Subtle is the Lord, but malicious He is not.”
  - Albert Einstein
Symmetry in Genomes

• Symmetry: key to structure and dynamical principles

• Chargaff’s parity rule (1951)
  – In DNA, A=T and C=G
  – Watson and Crick’s double helix (1953)

• Chargaff’s second parity rule (1968)
  – A~T and C~G in SINGLE strand of DNA

Question: Is CPR2 part of a general phenomenon? If so, what is it? What is the source of the symmetry? What can it tell us?

Reverse, complement, & Inverse symmetries

- Conjugation (example) of AAGTC
  - Reverse: CTGAA
  - Complement: TTCAG
  - Inverse (reverse-complement): GACTT

- Symmetry - measure of balance of word frequencies of conjugate pairs
Inverse symmetry and the (Yang-Lee) parity

If we represent the four nucleotides by two pairs of spins, red (C/G) and green (A/T) (for example, \text{TCGAA} is TCGAA), then inverse symmetry is exactly the same as (the Yang-Lee) parity in physics.
If for every word (k-mer) there is a $\rho$-symmetry conjugate of that word somewhere else in a sequences, then we say the sequence has perfect $\rho$-symmetry.

We found genomes have neither reverse nor complement symmetry but have excellent inverse symmetry.
Symmetry index $\chi$

\[
\chi^2 = \frac{1}{2N} \sum_{(u, u^+)} \left( \frac{f_u - f_{u^+}}{\sigma_{mu}} \right)^2
\]

where $\rho = r, c, or i$.

$u$: a k-letter word; $u^+$, its $\rho$-conjugate

$f_u$: occurrence frequency of $u$ in a sequence

$\rho$: all $\rho$-conjugate pairs

$\sigma_{mu}$: the standard deviation of complete set (of words) to which $u$ and $u^+$ belong

Index determined by relative, not absolute, frequency difference

$\chi_\rho \sim 1$, no $\rho$-symmetry; $\chi_\rho = 0$, perfect $\rho$-symmetry
$\chi$ vs. $S^1 = 1 - \left( \frac{\sum_{u \in S} |f_u - f_u^\dagger|}{\sum_{u \in S} f_u + f_u^\dagger} \right)$

$L^1$-distance ($S^1$) is not a good measure of symmetry.
χ vs. $S^1 = 1 - \left( \sum_{u \in S} |f_u - f_u^+| \right) / \left( \sum_{u \in S} f_u + f_u^+ \right)$

<table>
<thead>
<tr>
<th>k</th>
<th>Chr.</th>
<th>$S^1_c$</th>
<th>$S^1_{c,R}$</th>
<th>$\chi_c$</th>
<th>$\chi_{c,R}$</th>
<th>$S_i^1$</th>
<th>$S_{i,R}^1$</th>
<th>$\chi_i$</th>
<th>$\chi_{i,R}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>E. coli</td>
<td>0.9280</td>
<td>0.9992</td>
<td>0.9563</td>
<td>1.1359</td>
<td>0.9974</td>
<td>0.9991</td>
<td>0.0345</td>
<td>1.1925</td>
</tr>
<tr>
<td></td>
<td>HS1</td>
<td>0.8866</td>
<td>0.9998</td>
<td>1.1006</td>
<td>1.1927</td>
<td>0.9992</td>
<td>0.9996</td>
<td>0.0093</td>
<td>1.4425</td>
</tr>
<tr>
<td>3</td>
<td>E. coli</td>
<td>0.8863</td>
<td>0.9983</td>
<td>0.9607</td>
<td>1.0558</td>
<td>0.9965</td>
<td>0.9982</td>
<td>0.0255</td>
<td>1.0602</td>
</tr>
<tr>
<td></td>
<td>HS1</td>
<td>0.8509</td>
<td>0.9996</td>
<td>0.8821</td>
<td>1.1241</td>
<td>0.9992</td>
<td>0.9996</td>
<td>0.0061</td>
<td>1.1587</td>
</tr>
<tr>
<td>4</td>
<td>E. coli</td>
<td>0.8323</td>
<td>0.9960</td>
<td>0.9465</td>
<td>1.0086</td>
<td>0.9943</td>
<td>0.9963</td>
<td>0.0307</td>
<td>0.9497</td>
</tr>
<tr>
<td></td>
<td>HS1</td>
<td>0.8001</td>
<td>0.9993</td>
<td>0.8703</td>
<td>1.0733</td>
<td>0.9989</td>
<td>0.9993</td>
<td>0.0065</td>
<td>1.1097</td>
</tr>
<tr>
<td>5</td>
<td>E. coli</td>
<td>0.7765</td>
<td>0.9918</td>
<td>0.9320</td>
<td>1.0177</td>
<td>0.9905</td>
<td>0.9921</td>
<td>0.0399</td>
<td>0.9706</td>
</tr>
<tr>
<td></td>
<td>HS1</td>
<td>0.7590</td>
<td>0.9988</td>
<td>0.8792</td>
<td>1.0511</td>
<td>0.9984</td>
<td>0.9988</td>
<td>0.0066</td>
<td>1.0207</td>
</tr>
<tr>
<td>6</td>
<td>E. coli</td>
<td>0.7328</td>
<td>0.9839</td>
<td>0.9188</td>
<td>1.0110</td>
<td>0.9824</td>
<td>0.9846</td>
<td>0.0611</td>
<td>0.9671</td>
</tr>
<tr>
<td></td>
<td>HS1</td>
<td>0.7159</td>
<td>0.9976</td>
<td>0.8888</td>
<td>1.0191</td>
<td>0.9973</td>
<td>0.9976</td>
<td>0.0091</td>
<td>1.0082</td>
</tr>
</tbody>
</table>

$L^1$-distance ($S^1$) is **not** a good measure of symmetry
$f_u$ vs. $f_{u+}$ plots are good visuals, but can be misleading.
Reverse and complement symmetry absent in all genomes on all scales

\( \chi_l : \chi \text{ of segment of length } l \)

segment length
Global and local symmetry not the same

Structure in segmental $\chi_{\text{inverse}}$

$r_{\chi} = \chi_{\text{background}} (\chi_{\text{global}})^{-1}$
No reverse/complement symmetry but strong global inverse symmetry

Global $\chi_{\text{inverse}}$ scales with chromosome length
Classification by $\chi_{\text{inverse}}$

- 50% eubacteria types A & B
- Archaea all types C & D
- Eukaryotes: multi-cells type D; all types B & C are unicells
The $\chi$-plot

- $\chi(a+a) \sim \chi(c+c) \sim 1$
  - no local IS
- $\chi(a+b) \sim 1$
  - no mutual IS if both segments from same half
- $\chi(a+c) \sim 0.2$
  - high mutual IS if segments from different half

Inference:
- No local IS anywhere
- Fore and aft of chromosome have high mutual IS
Four types of chromosomes have characteristic $\chi$-plots

Type A: No local IS, bisected by ter/ori sites into two mutual high IS halves

Type B: Like type A, but higher local IS background

Type C: Hybrid of Band D

Type D: Homogeneous high local & mutual IS
IS type correlated with taxonomy

Phylum Firmicutes: all are type A
P-values for distributions of $\Delta_u = |f_u - f_u^*| / \sigma_m$
Inverse segmental duplication (ISD) generates IS

Absence of similar mechanism for generating reverse/complement symmetries may explain their absence.
Types of segmental duplications

Segmental duplication

Inverse segmental duplication

Whole-genome duplication

Inverse whole-genome duplication
Proximal and distal ISDs generate different effects

**Distal** ISD enhance global symmetry but not local symmetry

**Proximal** ISD enhance local symmetry and global symmetry
Inferences from $\chi$-plots on chromosome evolution

- Whole-chromosome ISD (WISD)
- Few $dist$-SDs (but possibly many $prox$-SDs)
- Very few $prox$-ISDs

- WISD
- Few $dist$-SDs
- Low to medium level of $proximal$ SDs

- WISD
- Some $dist$-SDs (or chrom. Re-arrangements)
- Various level of $prox$-ISDs

- With or w/o WISD
- Unconstrained SDs
- Saturating amount of $prox$-ISDs
Mosaic of prokaryotic $\chi$-plots invites "$\chi$-archeology"
**k-mer skews**

From universality

\[ |f_u - f_{ut}| \approx |f'_u - f'_{ut}| \approx \chi_{i,bg} \sqrt{2\sigma_{m_u}}, \]

\[ \sqrt{2\sigma_m} \approx \sqrt{2\bar{\sigma}} \approx 0.20L(3.2)^{-k}. \]
BLASTing genes on (+ vs +), (- vs -), and (+ vs -) strands

(+ vs +) Genes in 5’ end

(- vs -) Genes in 5’ end

Type A, NC_003030

(+ vs -) Genes in 5’ end on (+) strand highly similar to genes in 5’ end on (-) strand

Type D, NC_000911

(+ vs +), (- vs -) & (+ vs -) all show genes uniformly distributed over entire genome.

Let $\langle \Delta f \rangle = \sqrt{2\sigma}$ be the average frequency difference of an unrelated pair (of $k$-mers). This means a typical pair of $k$-mers have respective frequencies $f_- = \bar{f} - \langle \Delta f \rangle / 2$ and $f_+ = \bar{f} + \langle \Delta f \rangle / 2$, so that on average the fraction of unrelated $k$-mers that are "paired-up" is

$$2f_- / \bar{f} = 2v_0 = (\bar{f} - \langle \Delta f \rangle / 2) / \bar{f} = 1 - \sqrt{2\sigma} / 2\bar{f}.$$ 

Similarly, if $\langle \Delta f \rangle_\rho$ is the average frequency difference in a $\rho$-conjugate pair, then the fraction of $k$-mers paired with their respective $\rho$-conjugate partners is

$$2v_\rho = (\bar{f} - \langle \Delta f \rangle_\rho / 2) / \bar{f} = 1 - \langle \Delta f \rangle_\rho / 2\bar{f}.$$  \hspace{1cm} (3)

The mean field value of $\chi_\rho$ is, by definition

$$\langle \chi_\rho \rangle = \langle \Delta f \rangle_\rho / \sqrt{2\sigma} = (\langle \Delta f \rangle_\rho / 2\bar{f}) \left( \sqrt{2\sigma} / 2\bar{f} \right)^{-1} = (1 - 2v_\rho) / (1 - 2v_0).$$
In mean-field approximation, fraction of a sequence that causes it to manifest inverse-inverse symmetry is

\[ 2 \delta v_{\text{inv}} \sim (1-2v_0)(1-<\chi_{\text{inv}}>) \]

(1-2v_0) is the fraction of unrelated k-mers that are paired-up. With \(<\chi_{\text{inv}}> \sim 0.073+/- 0.066\)

\[ 2\delta v_{\text{inv}} \sim 0.25+/- 0.15 \]

That is, about one quarter of chromosomes is composed of ISD generated segments.
Genome growth model based on random segmental duplication

Band fits universality class of equivalent length

Fraction of inverse duplication

\[ \sim 2 \delta v_{\text{inv}} \sim 0.34\pm 0.05 \]

Minimal genome growth model


Summary I

• All chromosome have strong global IS
• Local and global reverse symmetry (RS) and complement symmetry (CS) are absent in genomes
• Chargaff’s second parity rule is a reduction of global IS, not of CS
• (Large scale) IS is likely generate by ISD
• There is no comparable mechanism that can generate RS or CS
• IS of chromosomes have types
  – Type A: No local IS; whole genome inverse duplication (WGID)
  – Type D: Only local IS; saturating amount of prox-ISDs

• Skews in single base AND k-mers conjugates can be quantitatively accounted for in terms of IS

• Origin & terminal sites play prominent roles in IS
• Model of genome growth based on SD and ISD can explain genomic patterns IS
• IS-type classification has clear correlation with taxonomy, and suggests an ancient origin of WGID. Attempt to determining time of occurrence under way.

• Not sure why global IS is so strong. Does promote equal utilization of both strands, but A and D types represent different strategies
End of Lecture Three
Related papers


• **People**
  - Sing-Guan Kong, Physics, Systems Biology, NCU
  - Hong-Da Chen, Physics, Systems Biology, NCU
  - Wen-Lang Fan, Physics, NCU; Genome Center, Sinica
  - Zi-Ting Hsu, Systems Biology, NCU
  - Prof. B. Zheng, Inst. Mod. Phys., Zhejiang U.

• **Financial support from**
  - National Science Council
  - Ministry of Education
  - National Center for Theoretical Sciences
  - National Central University