Inverse Symmetry in Genomes and Whole-Genome Inverse Duplication

HC Lee
Graduate Institute of Systems Biology and Bioinformatics and Department of Physics
National Central University

May 7-9 2008, Kunming Yunan, China
Outline

- Symmetry in genome
- Reverse, complement, inverse
- Symmetry index $\chi$
- Global and local symmetry
- The $\chi$-plots
- Segmental duplication (SD)
- Direct, inverse, proximal, distal, whole-genome SD
- Genome growth and evolution
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
Symmetry index $\chi$

$$\chi_\rho^2 = \frac{1}{2N_\rho} \sum_{(u, u^\dagger) \in \mathcal{P}_\rho} \left( \frac{f_u - f_{u^\dagger}}{\sigma_{mu}} \right)^2, \quad \rho = r, c, or i$$

- $u$: a $k$-letter word; $u^+$, its $\rho$-conjugate
- $f_u$: occurrence frequency of $u$ in a sequence
- $\mathcal{P}_\rho$: all $\rho$-conjugate pairs
- $\sigma_{mu}$ is the standard deviation of set (of words) to which $u$ and $u^+$ belong

$\chi_\rho \sim 1$, no $\rho$-symmetry; $\chi_\rho = 0$, perfect $\rho$-symmetry

Note: $\chi$-index is much better than all known distance measures
Reverse and complement symmetry absent in all genomes on all scales

\( \chi_l : \chi \) of segment of length \( l \)

segment length
Strong global inverse symmetry in all chromosomes

Global $\chi_{\text{inverse}}$ scales with chromosome length
Global and local symmetry not the same

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

$\chi_{\text{global}} = \chi_{\text{background}}(\chi_{\text{global}})^{-1}$

Global and local symmetry not the same
• 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

$\chi(a+a)$
$\chi(a+b)$
$\chi(a+c)$

$5'$
$5'$
$3'$
$3'$

Type A. C. acetobutylicum

$\chi(c+c)$

Segment site (Mb)

$100$ kb segment

ter (terminal of replication)
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 and mutual IS
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"
First order, intuitive interpretation of $\chi_{\text{inv}}$

- Let $v$ be the fraction of chromosome length generated by ISD. To lowest order in mean-field theory

  $$\chi_{\text{inv}} \sim 1 - 2v$$

- Averaged over 786 complete chromosomes

  $$\chi_{\text{inv,global}} \sim 0.073 +/- 0.066, \text{ or } v \sim 0.46$$

That is, most chromosomes have close to saturated amount of ISD generated segments.
Some other results

• Inverse-symmetry breakpoints are close to origin/terminal sites of replication
  – $\chi$-scanning is powerful tool for locating ori/ter sites

• Model of genome growth based on SD and ISD can explain patterns of $\chi$-plots and scale-dependence of $\chi$

• Base- and $k$-mer skewss are natural products of random drift.
  – Reverse and complement skews ALWAYS large
  – Rise of local/global inverse-symmetry causes fall of local/global inverse skews
• People
  – Sing-Guan Kong, Physics, NCU
  – Hong-Da Chen, Physics, NCU
  – Wen-Lang Fan, Physics, NCU

• Financial support from
  – National Science Council
  – Ministry of Education
  – National Center for Theoretical Sciences
  – National Central University
Thanks for your attention