

從基因組序列的物理 性質看生命演化

教育部國家講座

中央大學松濤講座

2009年3月25日

李弘謙

Computational Biology Lab
Inst. Systems Biology & Bioinformatics
Dept. Physics & Inst. Biophysics
National Central University

如果把宇宙的生命縮成一年

- The Big Bang (13.7 Billion years ago) January 1st
- Origin of the Milky Way Galaxy (- 9.1 By) May 1st
- Origin of the Solar System (- 5.0 By) July 19th
- Formation of the Earth (- 4.6 By) August 1st
- Origin of life on Earth (- 3.8 By) August 20th
- Time of the oldest known fossils (bacteria and green blue algae) (-2.8 By) September 16th
- The first cells with nucleus appear (-2 By) November 9th
- Age of dinosaurs (-230 My) December 25th
- Appearance of the first man (-1My) 23:20 Dec 31st

“宇宙年”的最後一小時

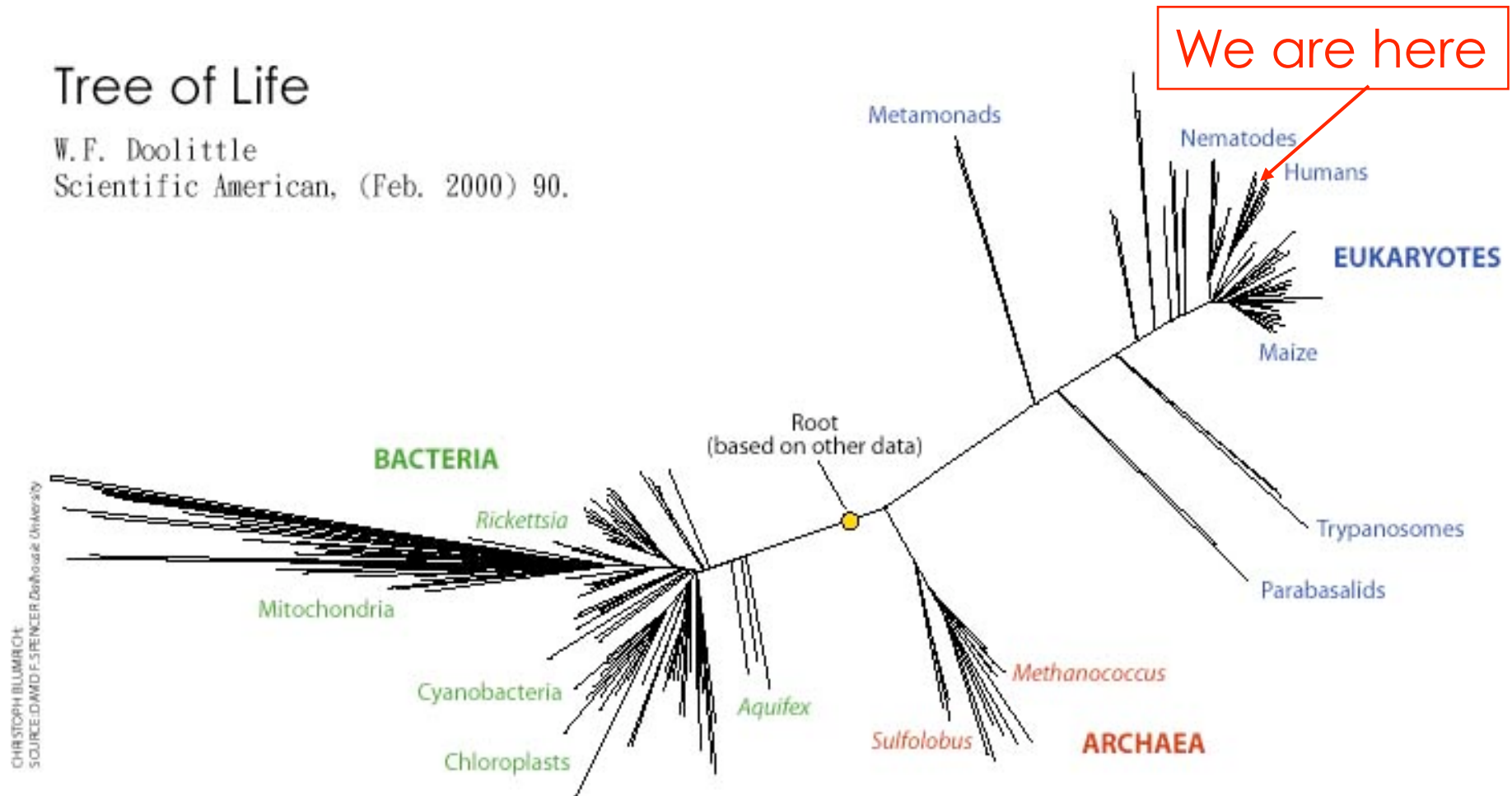
- Appearance of the first man 人類出現 23:20:00
 - Invention of agriculture 23:59:20
 - First dynasties in Sumer, Egypt, 夏朝 23:59:50
 - Roman Empire, birth of Christ 秦始皇 23:59:56
 - The Cruzades, Mayan civilization, 宋朝 23:59:58
 - Europe of the Renaissance, 鄭和下西洋 23:59:59
 - Worldwide expansion, space travel,
present time 00:00:01
- in the new year

生命的有極高的多樣性

Tree of Life

W.F. Doolittle

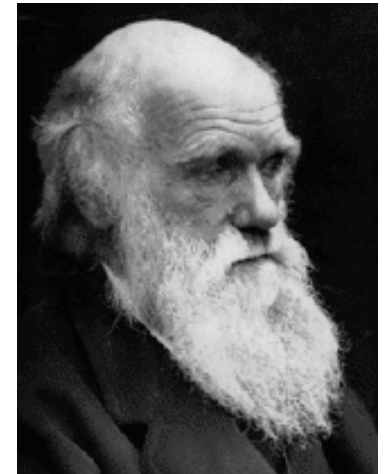
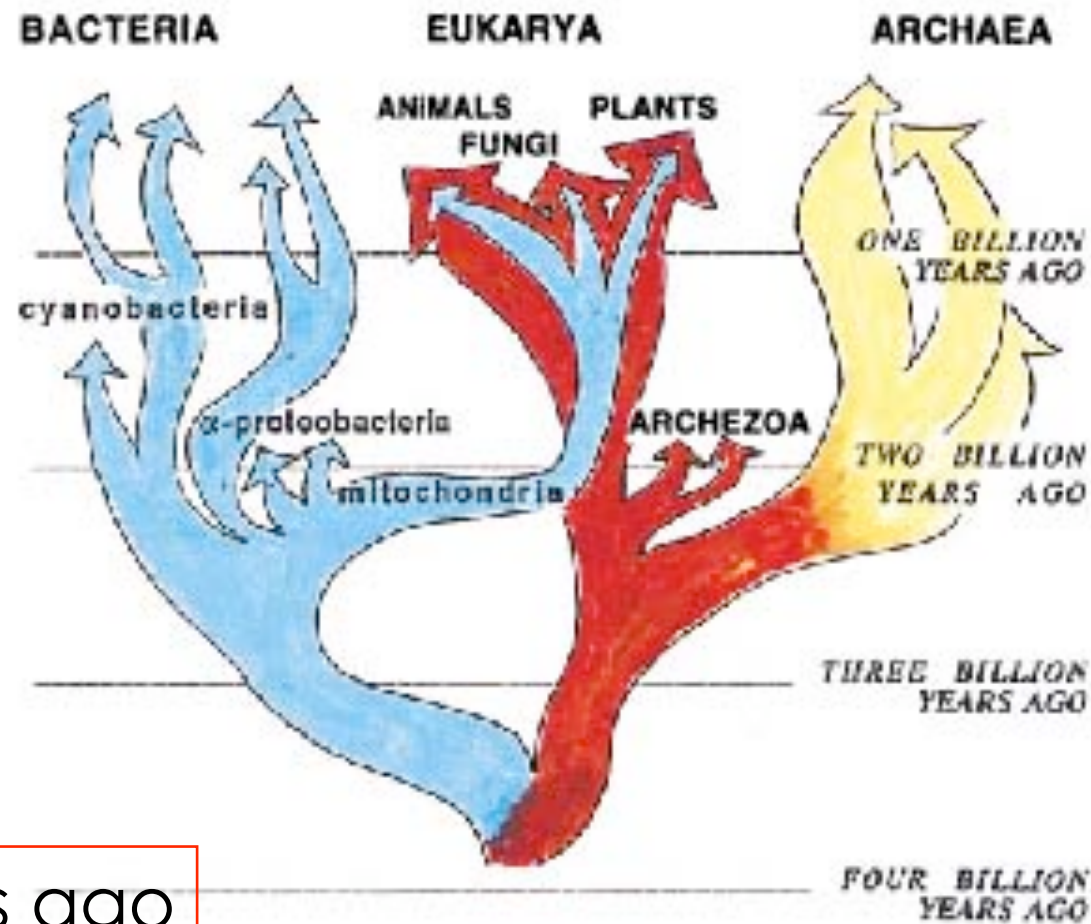
Scientific American, (Feb. 2000) 90.



生命演化了四十億年

Divergence of species
W.F. Doolittle, PNAS 94 (1997) 12751.

now



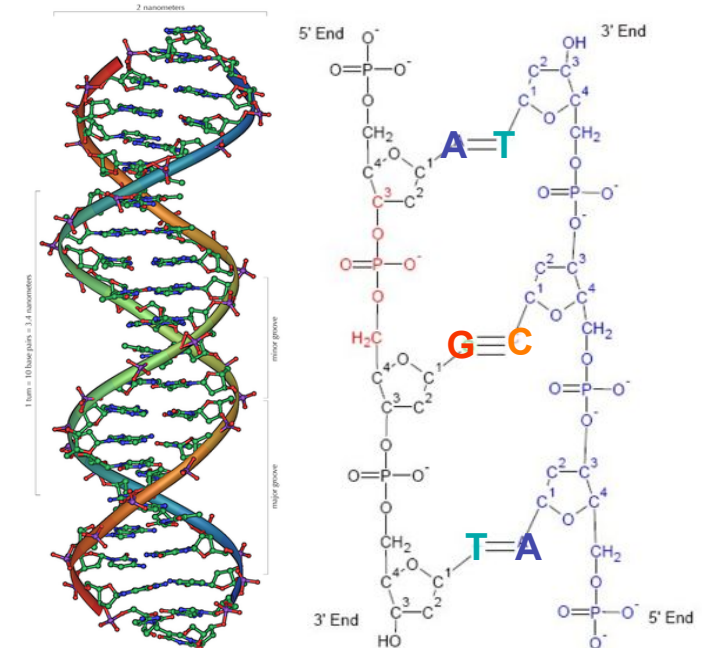
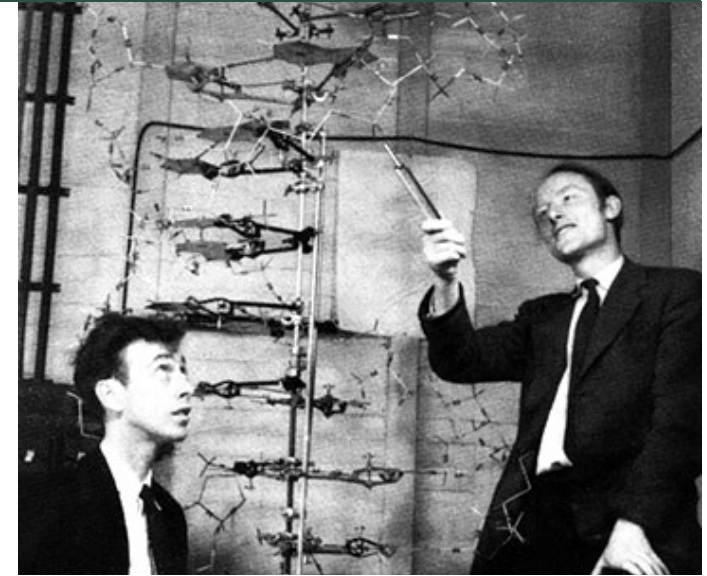
Charles Darwin
1809 -1882

物種起源(1859)

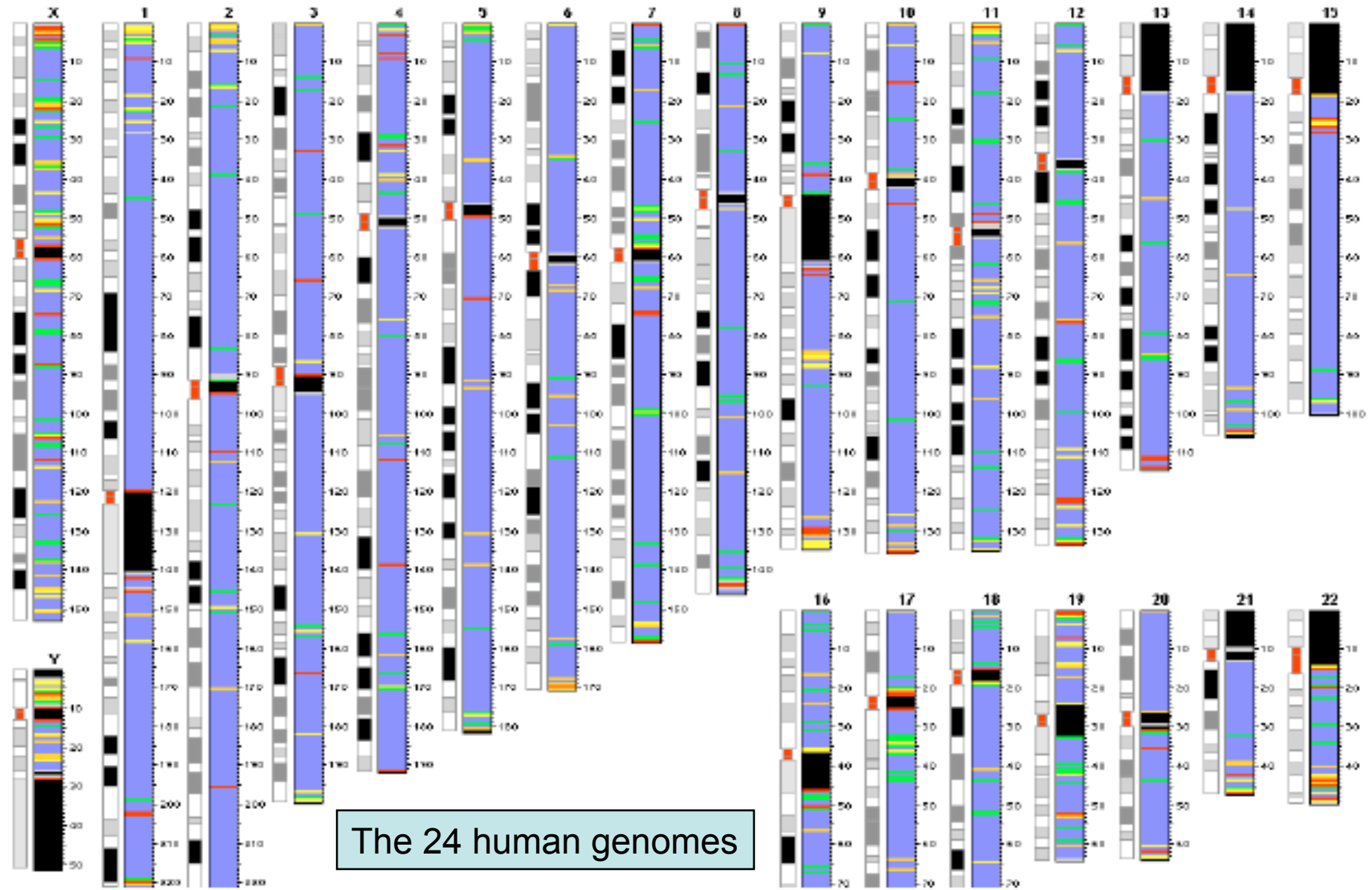
4 billion yrs ago

克力克與華生1954年發現DNA構造

- DNA 是核苷酸組成的雙螺旋
- 每一個生物的每一個細胞內都有一套完整的DNA—基因組 (基因圖譜)



基因組極為複雜



基因組 很大

A stretch of
genome from
the X chromo-
some of
Homo sapien

[http://
www.ncbi.nlm.nih.gov/
entrez/viewer.fcgi?val
=2276452&db
=Nucleotide
&dopt
=GenBank](http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?val=2276452&db=Nucleotide&dopt=GenBank)

The complete
genome has
2,000,000 such
pages

```
1 tgctgagaaa acatcaagctg tgtttctcct tcccaaag acacttcgca gccctcttg
61 ggatccagcg cagcgcaagg taagccagat gcctctgctg ttgccctccc tgtgggctg
121 ctctctcac gccggcccc acctgggcca cctgtggcac ctgccaggag gctgagctgc
181 aaacccaat gaggggcagg tgctcccga gacctgttc ccacacgccc atcggtctgc
241 cccgggcttt gagttctccc aggcccctct gtgcaccctt ccctagcagg aacatgccgt
301 ctgccccctt gagctttgca aggtctcggg gataatagga aggtctttgc cttgcaggga
361 gaatgagta tccgtgctcc ctccgagggg gattctggag tccacagtaa ttgcagggt
421 gacactctgc cctgcaccgg gcgccccagc tctccccac ctccctctc catccctgtc
481 tccggctatt aagacggggc gctcaggggc ctgtaactgg ggaaggtata cccgccctgc
541 agaggtggac cctgtctgtt ttgattctg ttccatgtcc aaggcaggac atgaccctgt
601 tttggaatgc tgatttatgg attttccagg cactgtgcc ccagatacaa ttttcttga
661 cattaagaat acgtagagaa ctaaatgcat ttcttctta aaaaaaaaaa aaacaaaaa
721 aaaaaaaaaa aaacaaaaa actgtactta ataagatcca tgcctataag acaaggaac
781 acctctgtc atatatgtgg gacctcgggc agcgtgtgaa agtttacttg cagtttgcag
841 taaatgaca aagctaacac ctggcgtgga caatcttacc tagctatgct ctccaaaatg
901 tatttttct aatctgggca acaatgggtc catctcgggt cactgcaacc tccgcttccc
961 aggttcaagc gattctccgg cctcagcctc ccaagtagct gggaggacag gcacccgcca
1021 tgatgcccg ttaattttg tatttttagc agagatgggt ttcgccatg ttggccaggc
1081 tggctctgaa ctctgacct caggtgatcc gcctgccttg gcctcccaa gtgctgggat
1141 gacaggcgtg agccaccgcg cccagccagg aatctatgca ttgcctttg aatattagcc
1201 tccactgcc catcagcaaa aggcaaaaca gggtaccagc ctccgcccac ccctgaagaa
1261 taattgtgaa aaaatgtgga attagcaaca tgttggcagg attttgctg aggtataag
1321 ccacttcct catctgggtc tgagctttt tgtattcggg ctaccattc gttggttctg
1381 tagttcatgt ttcaaaaatg cagcctcaga gactgcaagc cgctgagta aatacaata
1441 gatttttaa gtgtatttat tttaaaaaa aaataaaatc acacataaga taaaacaaa
1501 cgaaactgac ttatacagt aaataaacg atgcctgggc acagtggctc acgcctgtca
```

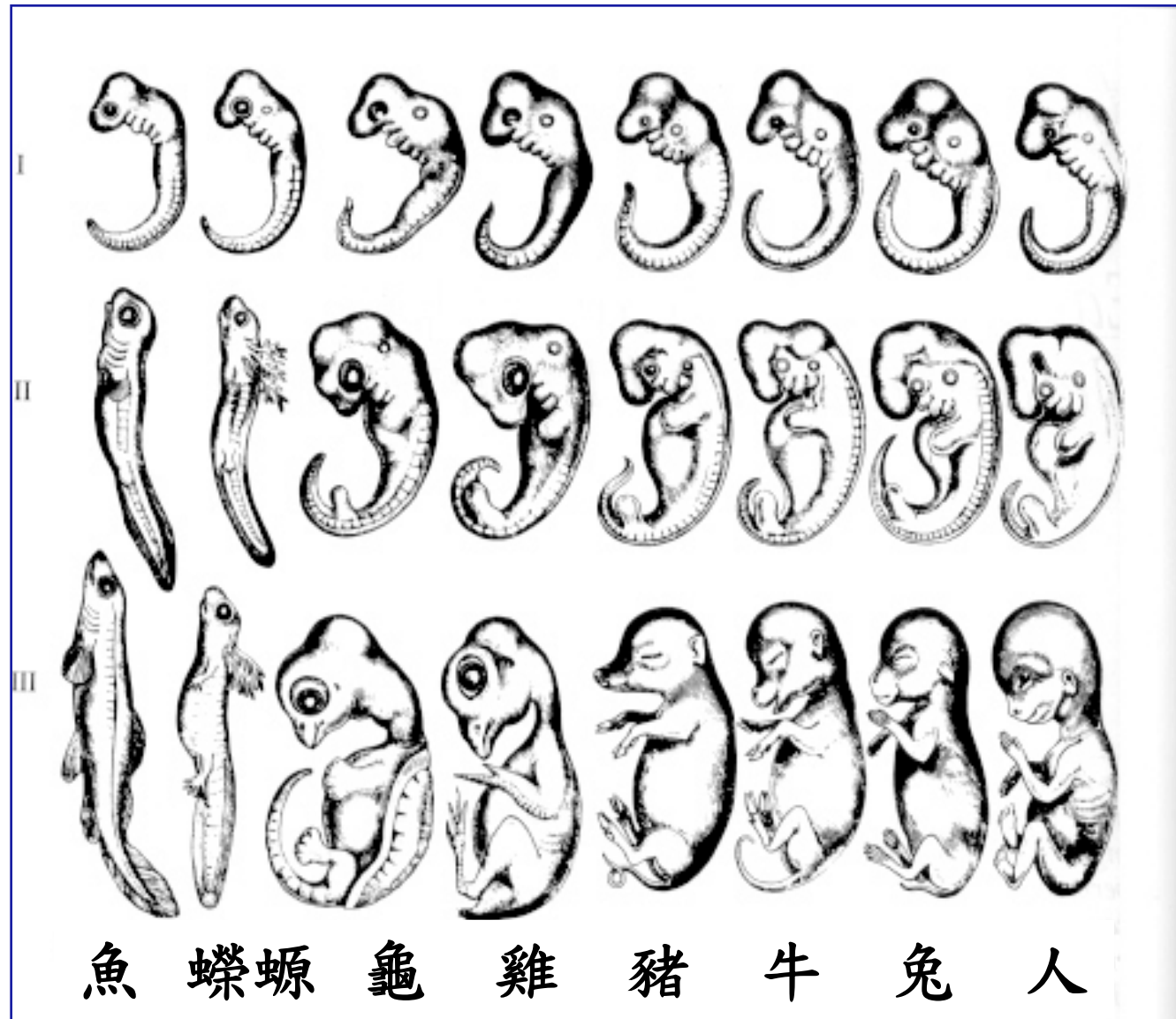
基因組是生命編碼的載體 像電腦中的硬碟

Hard Disc $\longleftrightarrow$ DNA (基因組)

Programs & documents $\longleftrightarrow$ Genes & other codes

Run program
copy document $\longleftrightarrow$ Transcribe
gene

從受精胚卵到胚胎，每個物種都把演化再走一遍



不同的物種，胚胎愈早期越相似

這暗示
基因組
裡載了很多演化的痕跡

我們的問題：基因組怎麼演化的？

- 物理與生物一般會問不一樣的問題
- 物理問原理、機制，生物問現象、功能
- 物理看共同性，生物看特一性

和宇宙起源與演化一樣， 生命的起源與演化也受制於物理定律



沒想到還有這麼多步要走吧？

如何從基因組的共同性找演化機制?

- 基因組（應）有共同的祖先
- 基因組有共同的演化機制
- 基因組應有演化機制留下的共同腳印
- 這種腳印應在比較序列中不存在也不可能存在
 - 用隨機或其它人工序列作比較序列

我們發現的基因組的幾個共同性

基因組現象

- 自相似性
有效長度是常數 —
尺度不變
- 序指數是不大不小亂的
常數 — 亂中有序
- 高度的倒對稱

物理意義

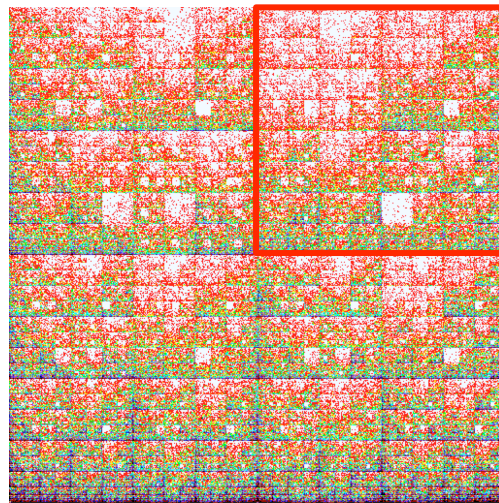
臨界態、臨界自組
複雜系統的特徵

高資訊量載體的特徵

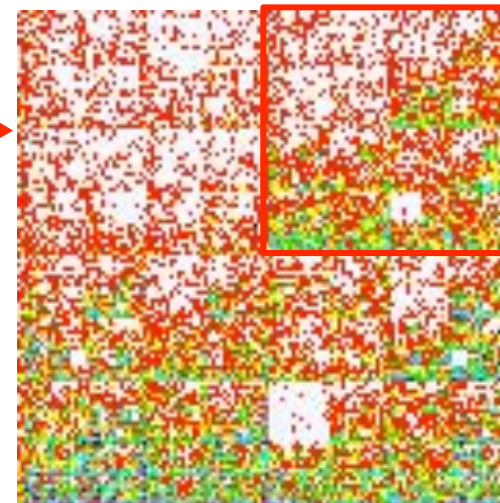
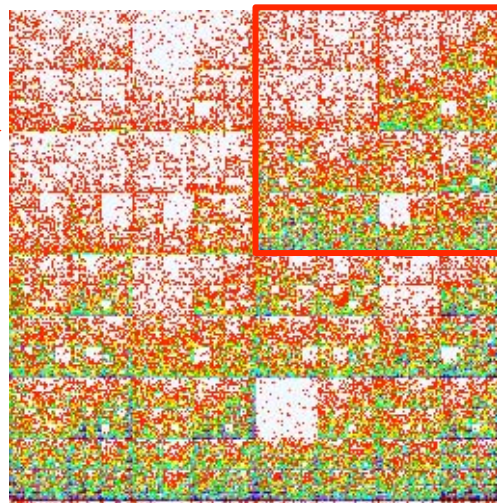
物質結構與動力學的
重要課題

- 楊振寧和李政道因
宇稱對稱工作得
1957年諾貝爾物理獎

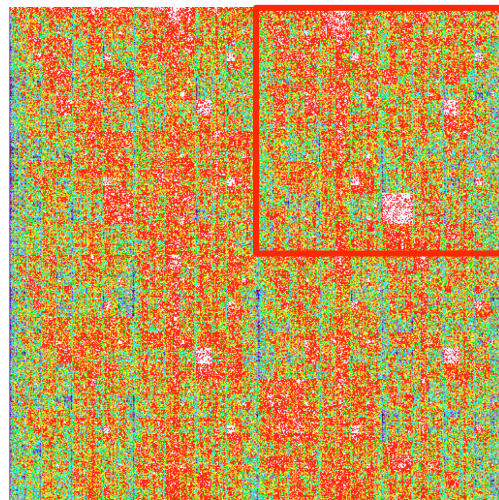
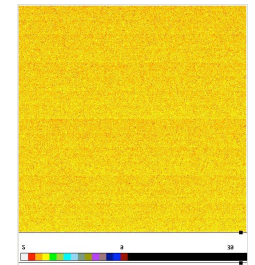
基因組的“畫像”有自相似性



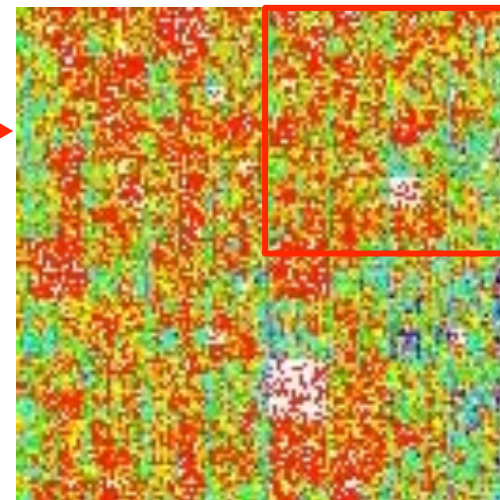
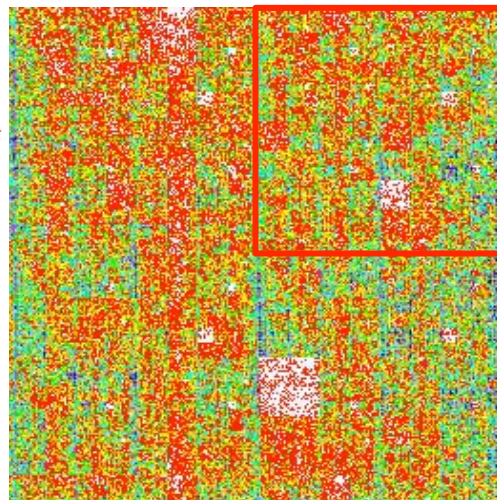
M. jannaschii



隨機序
列畫像

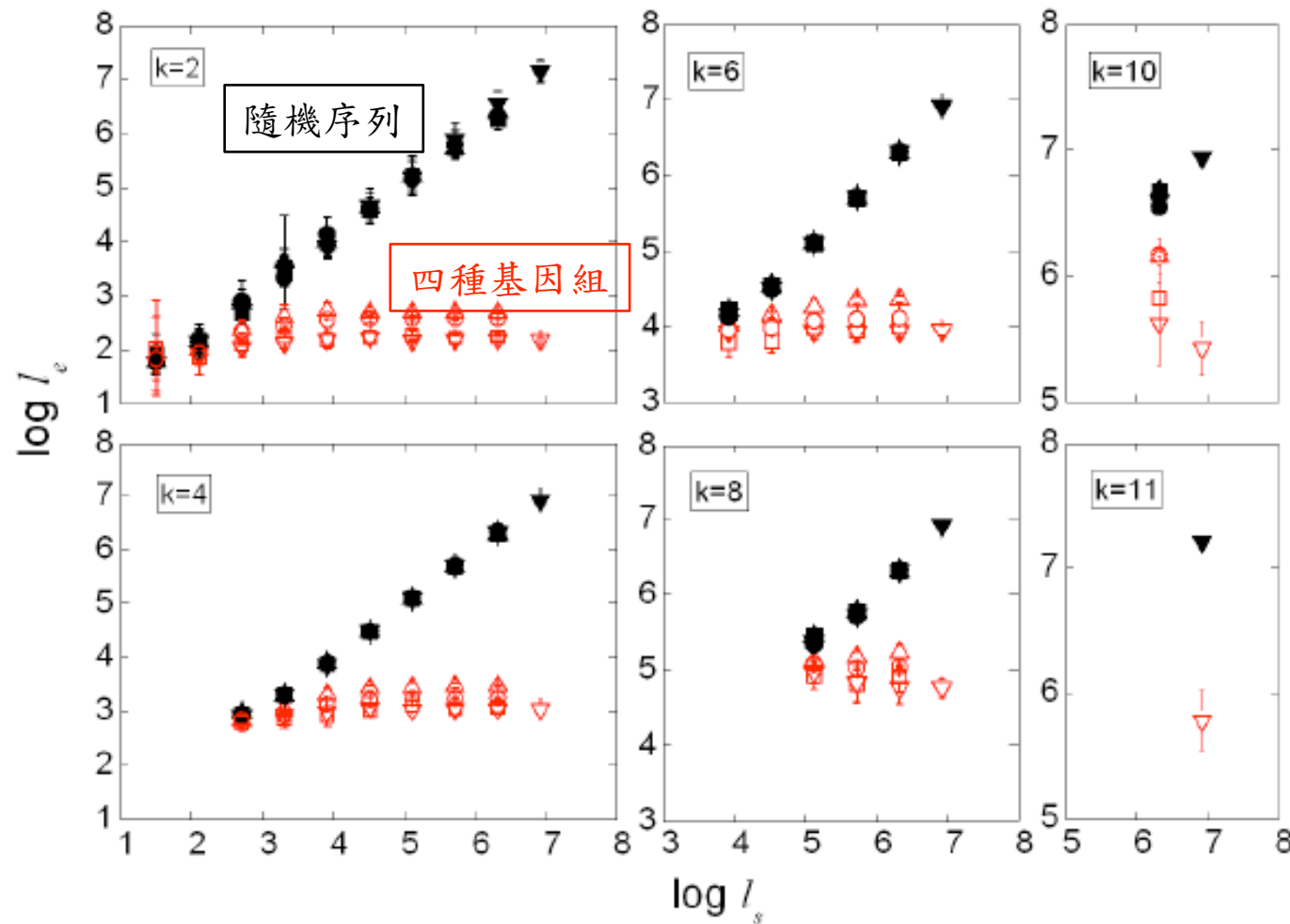


A. fulgidus



隨機序列的“有效長度”與真實長度相同

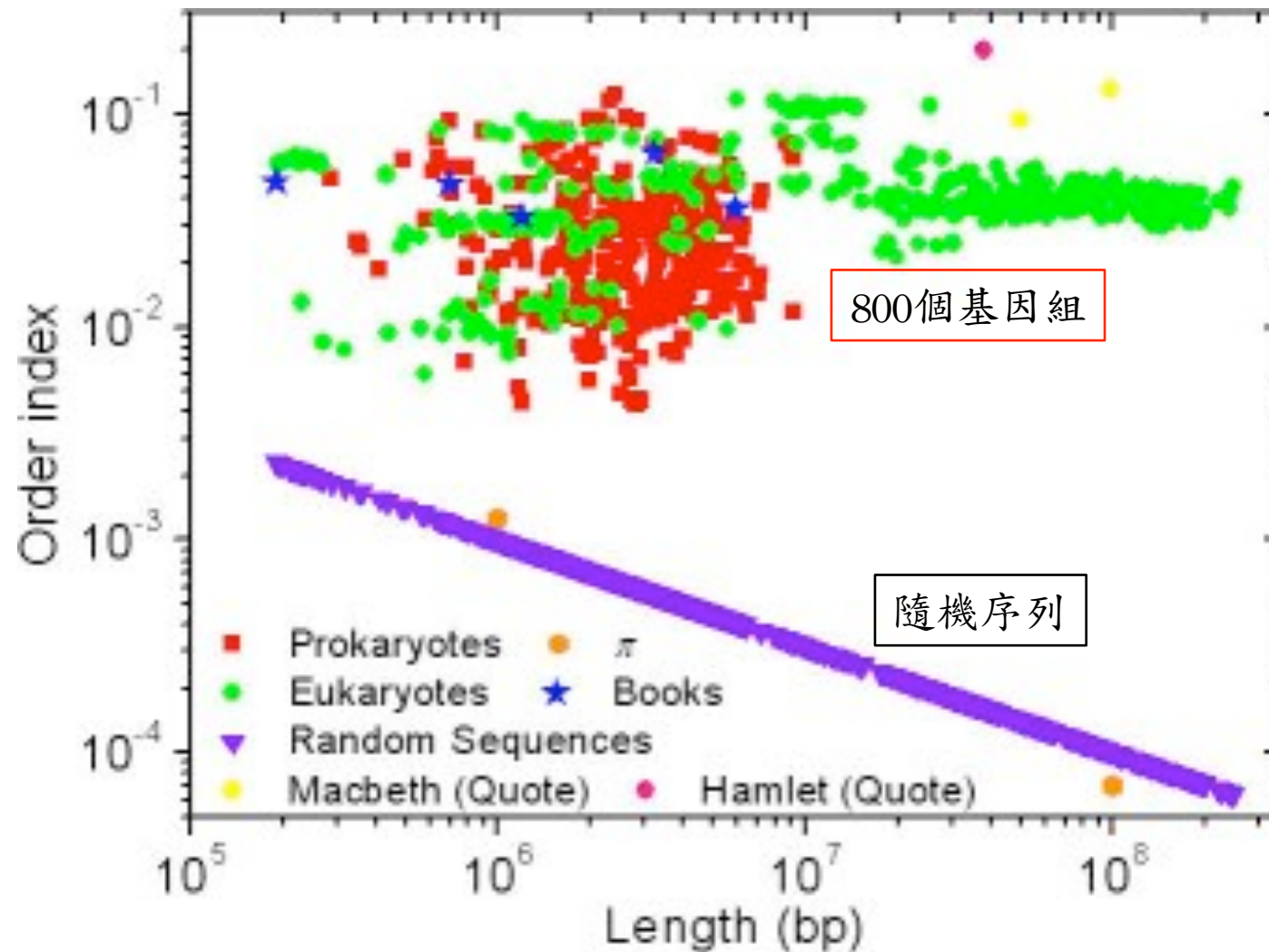
基因組的有效長度是普適常數



有效長度是一種亂度的度量。亂度愈大長度愈長，反之愈短。

四種物種: 大腸桿菌、線蟲、芥末藍、人類 (其它物種亦同)

隨機序列的“序指數”隨長度遞減
基因組的序指數是（近似）普適常數

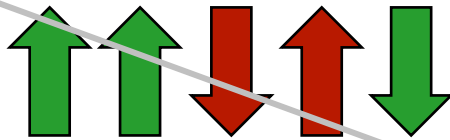


基因組中的倒字串對稱



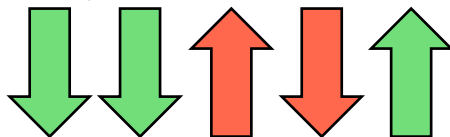
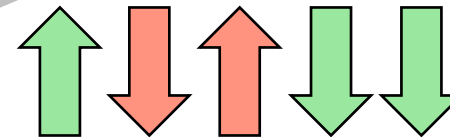
反像

物 (五字串)



直角鏡

倒像 (倒 = 反 + 對偶)



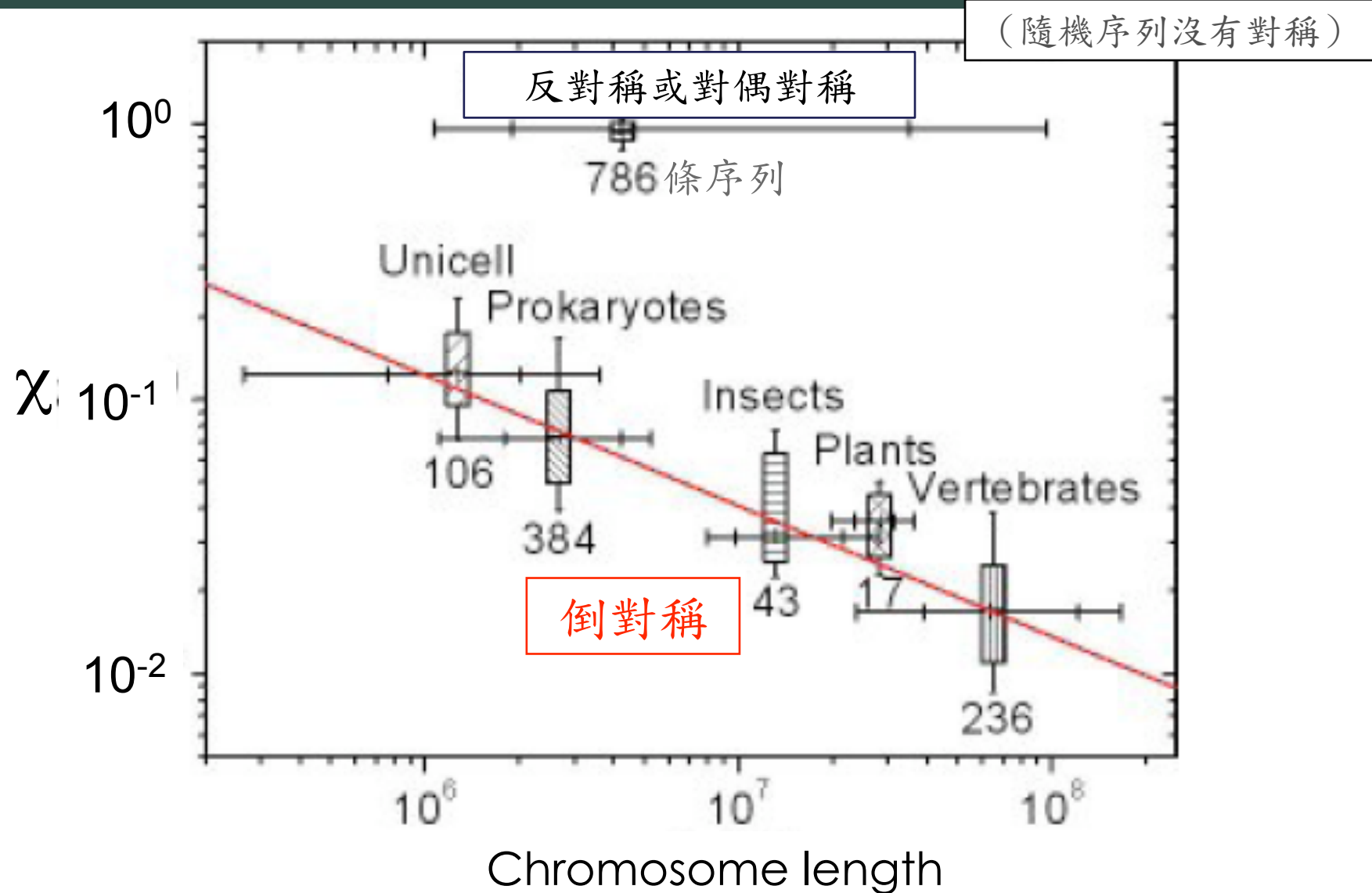
對偶像

在一個序列裡，如果相對於每一個字串
都能找到它的倒字串，
則稱該**序列**有完整的倒對稱。

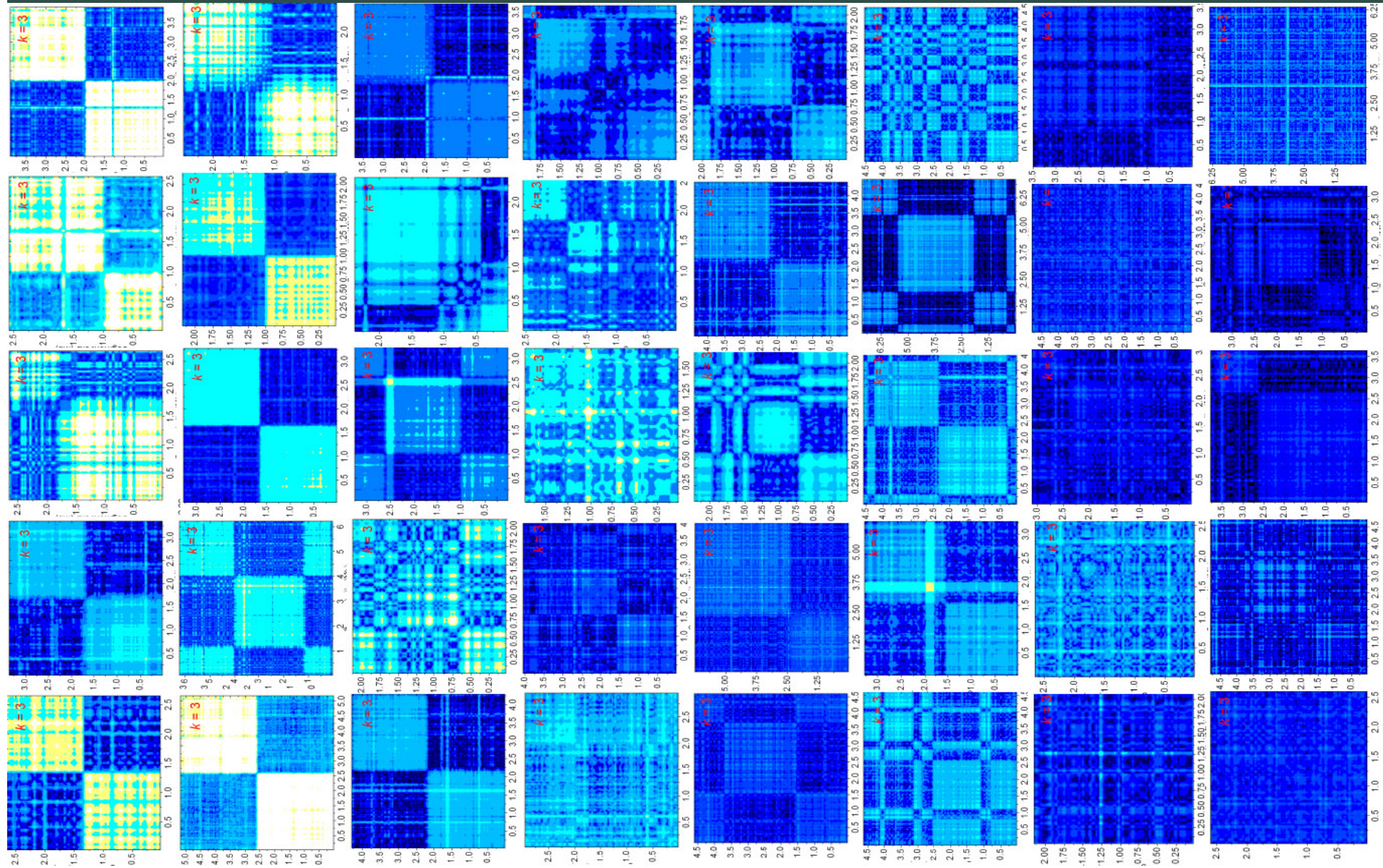
反對稱及對偶對稱亦同。

我們發現基因組沒有反對稱或對偶對稱
但是有高度的倒對稱。

基因組沒有反對稱或對偶對稱 ($\chi \sim 1$)
但是有很強的倒對稱 ($\chi \ll 1$)



基因組有各式各樣的倒對稱形式

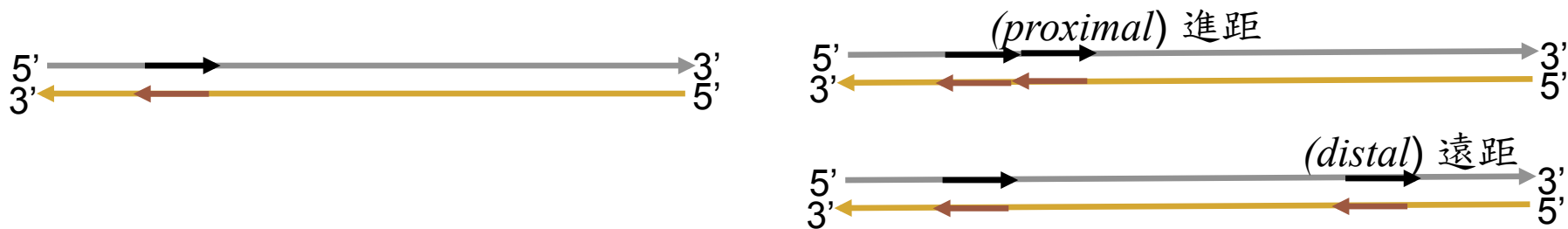


基因組的共同性是 隨機片斷複製的腳印

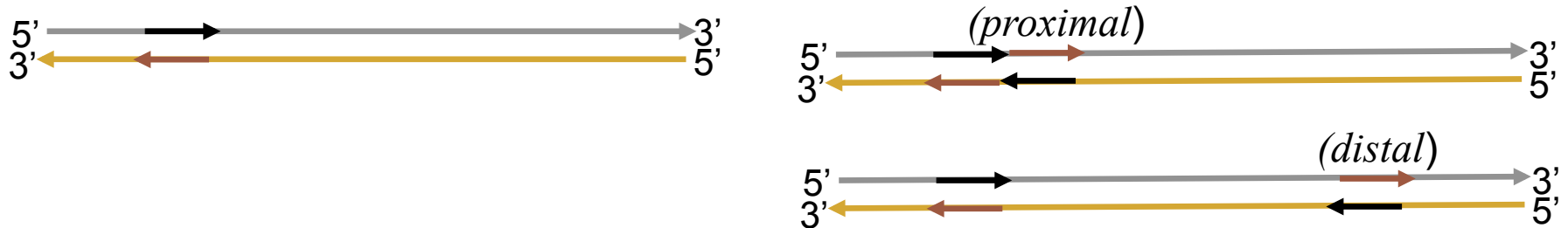
- 隨機片斷複製中的“片斷複製”讓基因組維持高度的序
 - 隨機片斷複製中的“隨機”也讓基因組有高度的亂度、或複雜度
-
- 我們以隨機片斷複製為基本機制建立很簡單的基因生長模型、製造模擬的基因組

片斷複製種類

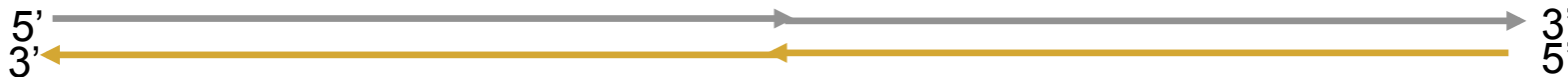
(正) 片斷複製



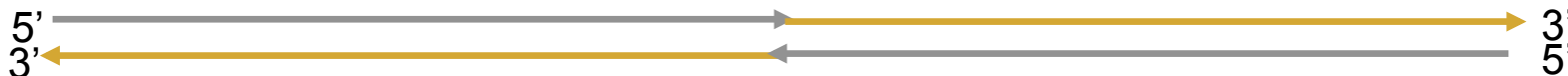
倒片斷複製



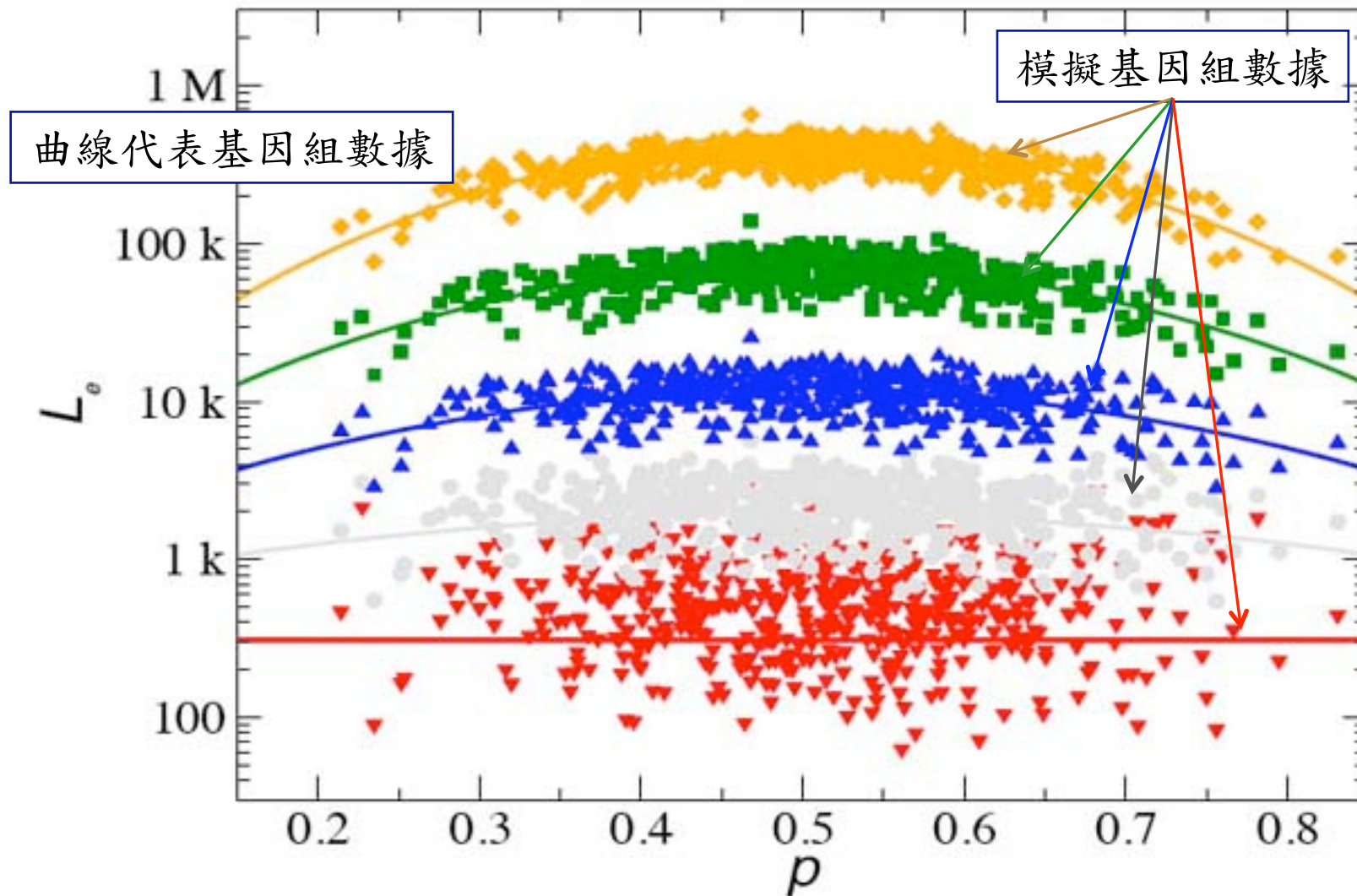
全基因組複製



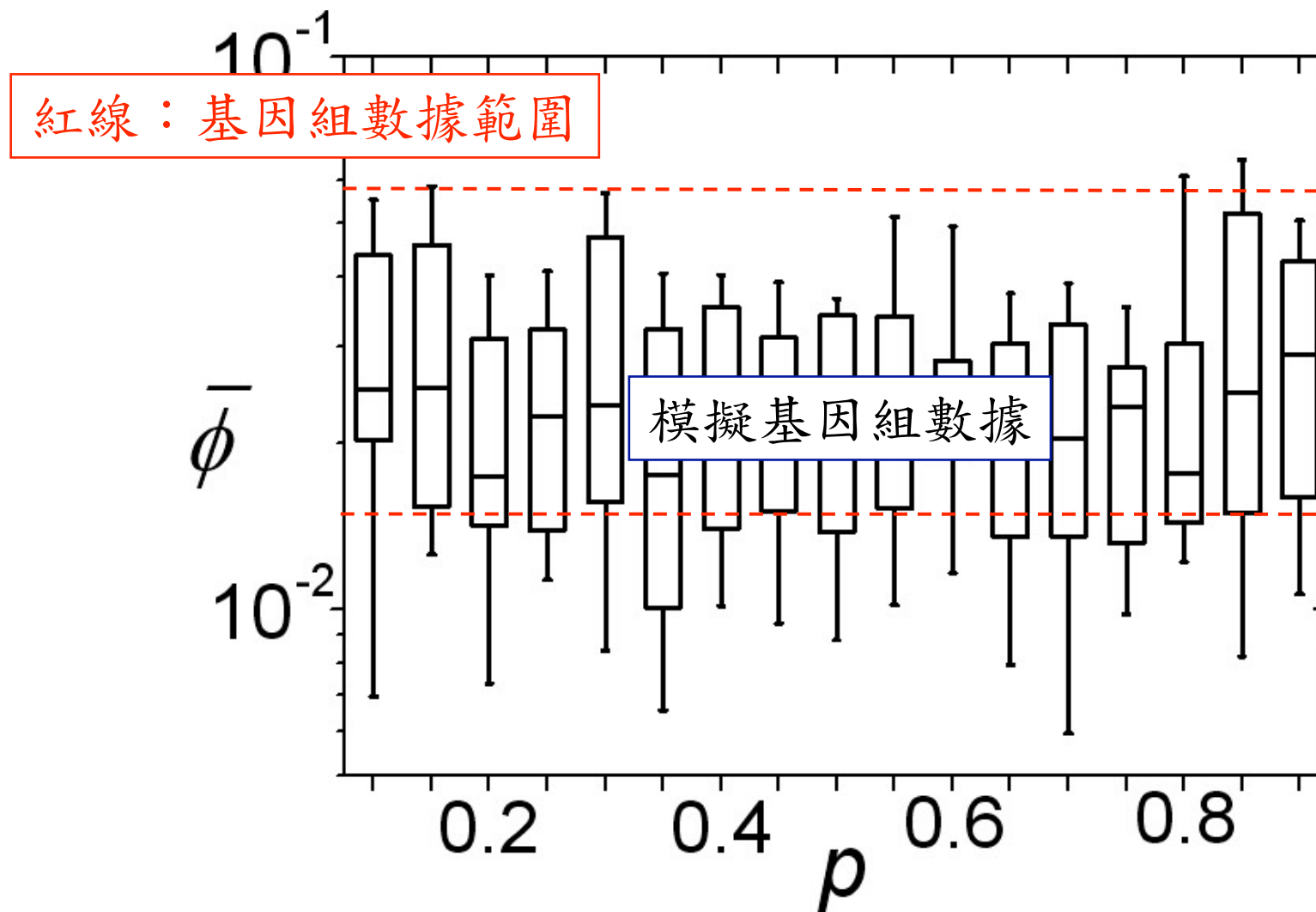
全基因組倒複製



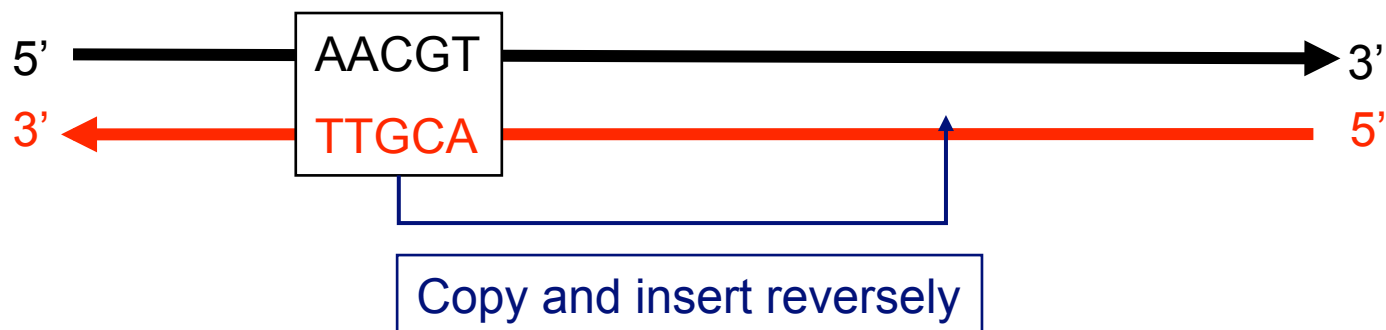
模擬基因組與真基因組的 有效長度相似



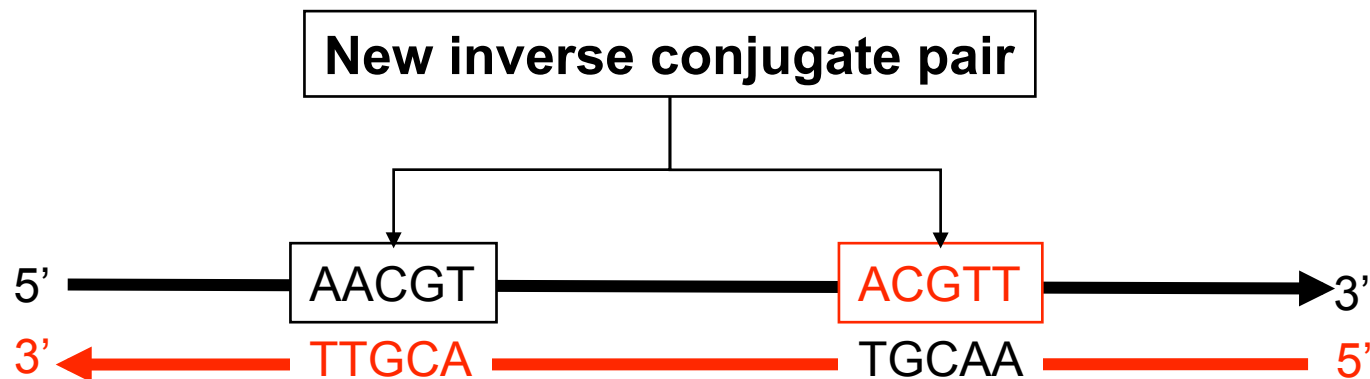
同一組模擬基因組也有與真基因組相近的序指數



倒片斷複製促進倒對稱



After inverse duplication:



No simple mechanism for generating r and c symmetries

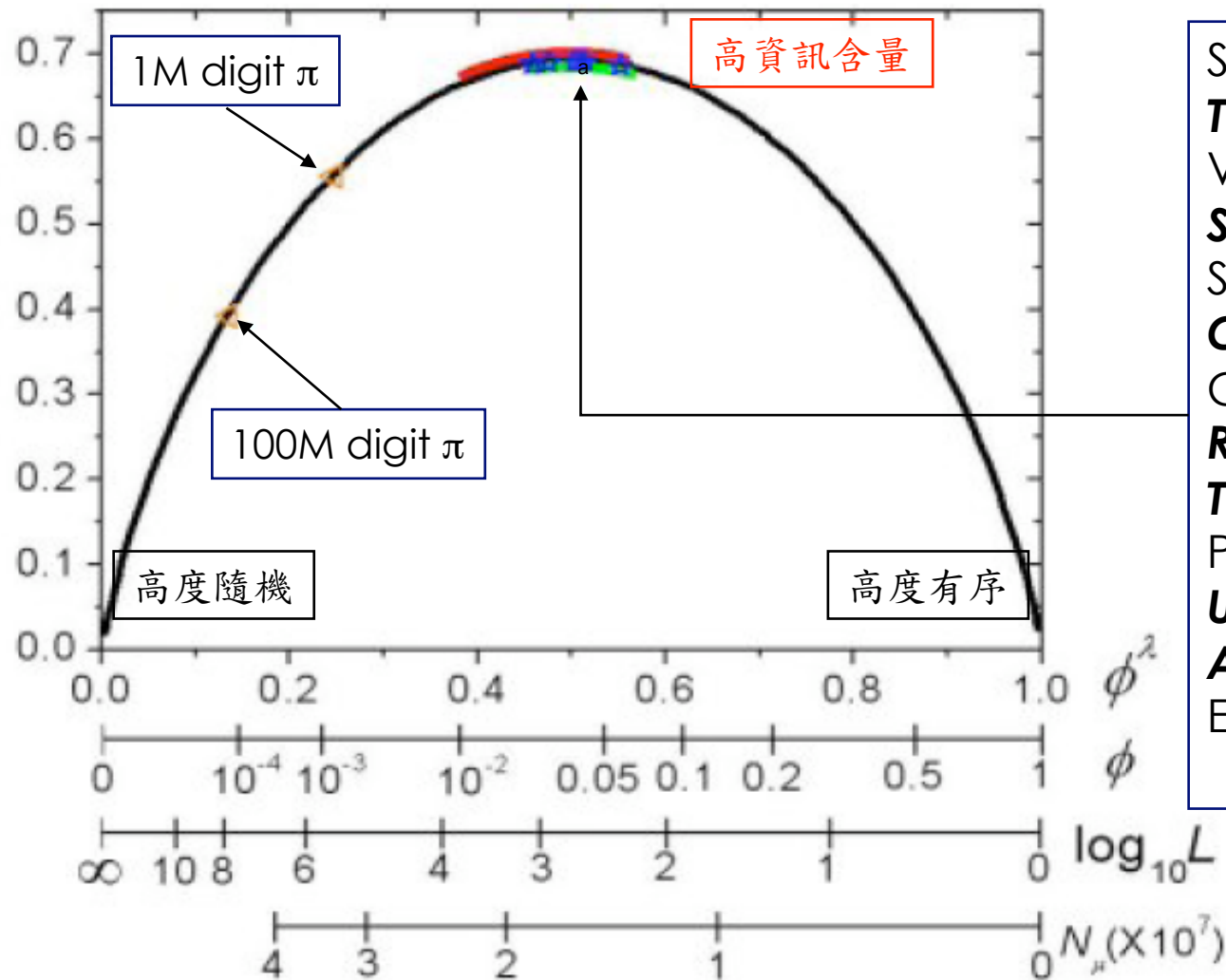
小結

- 基因組有若干不尋常共的共同性
- 我們認為這些共同性來自於隨機片斷複製是基因組的主要生長機制
- 我們的模擬計算支持這種看法
- 沒有其它機制可以給更好或更簡單的解釋

演化為甚麼要選擇隨機片斷複製 作為基因組的主要生長機制？

- 因為隨機片斷複製是基因組：
 - 最有效最快速累積難以獲得的高價值資訊的方法
(自我抄襲最容易累積資訊)
 - 最有效最快速使基因組成為高資訊量載體的方法

基因組與經典文學著作的高資訊含量有定量的相似



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

結論

基因組的主要生長機制是
隨機片斷複製

或

基因組是一個盲的自我複製者

(註: 這個看法與許多今天沒有提到的生物現象吻合)



感謝替我作苦工的小朋友們！

謝謝大家