

Lecture 7

Multiple sequence alignment

梁恭豪 Kung-Hao Liang

工研院生醫中心生物資訊部研究員
國立師範大學生物系兼任助理教授

Khliang@itri.org.tw

<http://flag.itri.org.tw/>

<http://reap.itri.org.tw/>

2003/4/10

Bioinformatics 名詞之創生

- 1987 年由當時任職於佛羅里達州立大學超級電腦中心之林華安博士 (Dr. Hwa A. Lim) 提出 Bioinformatics 一詞。
- 在定名之前，林博士曾以 bio-informatique, bioinformatique 稱之。
- 林博士對 bioinformatics 之定義
 - ◆ The study of biology-related information content and the associated information flow
- 第一屆國際 bioinformatics 會議於 1990 年 4 月在佛羅里達州立大學會議中心由林博士召開舉行，由 DOE, Florida technology and research authority, Thinking machines Corp., Digital Equipment Corp., Cray Research Inc. 贊助。

Thinking Machines

- 最多可集合 65536 CPUs 在一部電腦之中。
 - ◆ CM-200 : 8192-65536 CPUs
 - ◆ CM-2 : 16384-65536 CPUs with router chips
 - ◆ CM-5 : 16-16384 SPARC CPUs
- 利用 Hypercube 架構 , SIMD array type
- 在一個 2^n CPU 之系統中 , CPU 到 CPU 之間最多經過 n 次轉折。
- 此公司最後並未成功 , 猜測有可能是因為 PC Cluster 之興起而取代其潛在市場。

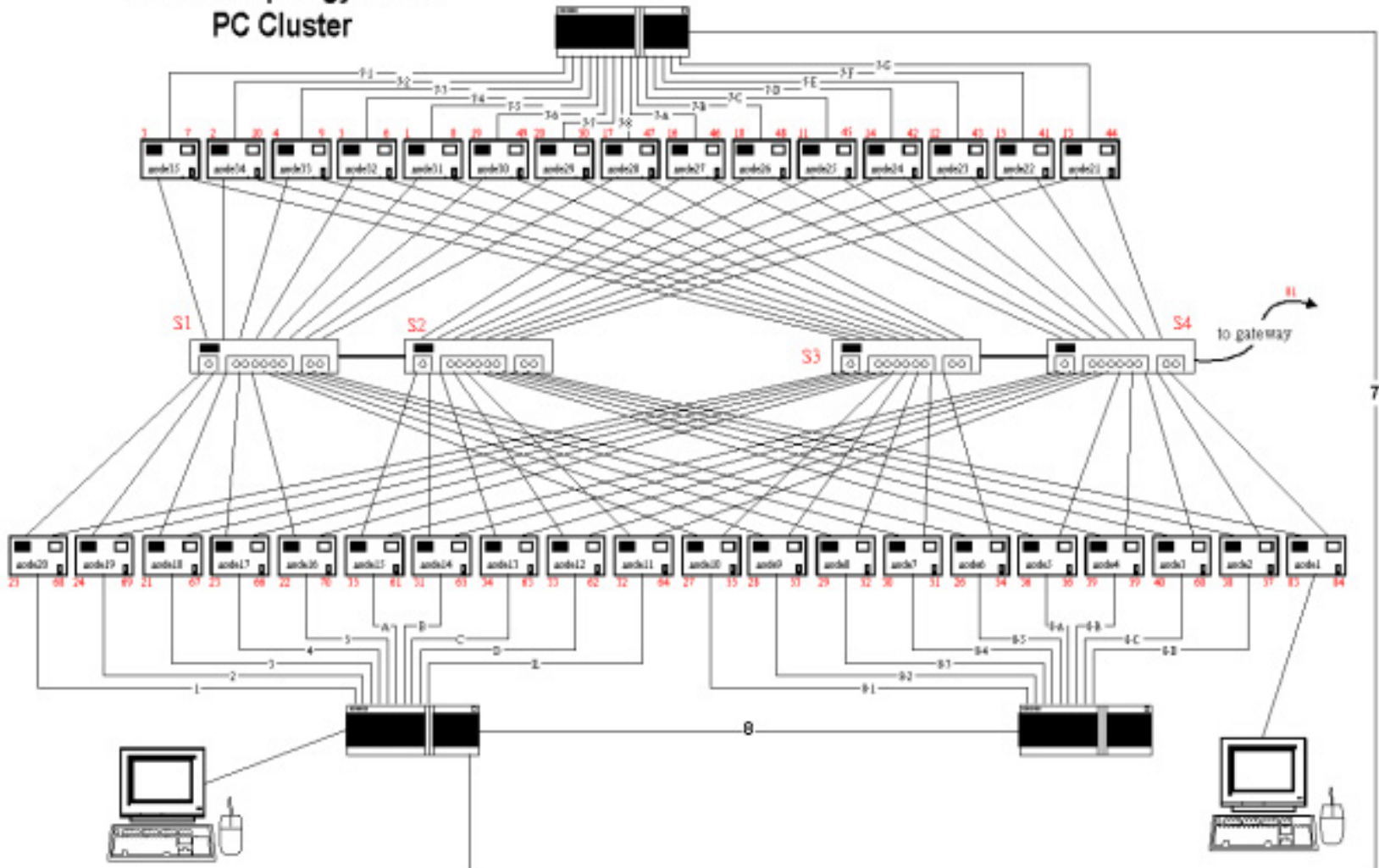
PC-Cluster and Servers in BMEC/ITRI



- PC-Cluster
 - ◆ PC-cluster 28 nodes
 - ◆ Dual Pentium 3, 1 GHz
 - ◆ SDRAM 1G
 - ◆ Hard Drive 36.4G*4 G
 - ◆ Red Hat 7.2
- Database Servers x 7
 - ◆ Dual Pentium 3, 1 GHz
 - ◆ SDRAM 2-4 G
 - ◆ Total Storage 2000 G

PC-Cluster in BMEC/I TRI

Network topology of the PC Cluster



本章學習重點

- 瞭解多序列比對之目的及應用
- 瞭解 Clustal W 運算法，步驟及相關背景知識
- 瞭解 guide tree, gene family tree, species tree 及 Phylogeny tree 間之關連
- 瞭解演化樹建構相關研究主題及文獻

排比分數 (Score)、期望值 (e-value) 與相似比例 (percent identity)

- **Score**
 - 基於 scoring matrix 對於各種排比方式之給分，並作為尋找最佳排比 (Optimum alignment) 過程中之參考。
- **E-value**
 - 判斷排比結果之 significance 之參考，並需考量資料庫之大小。數字越小，significance 越高。
- **Percent identity**
 - 完全相同之 base 個數 / 排比到的片段全長 * 100 %

相似度 (similarity) vs. 距離 (distance)

- **Similarity 與 distance 為反相關之概念**
- **Similarity**
 - **e.g. Percent identity**
- **Distance**
 - **Euclidean distance, Manhattan distance**
 - **Hamming distance (from telecom)**
 - **Edit distance : the minimum number of symbol insertions, deletions and substitutions that is required to transform a string into another**

多序列比對是瞭解新蛋白質 (novel protein) 功能之關鍵

- 通常未知功能之新蛋白質與已知蛋白質在整體序列的相似性上並不高 (i.e. percent identity less than 40 %), 因此不易由直接之序列比對找到相關之蛋白質家族。
- 解決的方法是將各個蛋白質家族中共通的部分 (通常為 conserved motif) 整理出來, 成為 pattern or profile。
- 再將新蛋白質對這些 pattern /profile 作比對, 由於這些重要的部位在序列間相似度很高, 即可增加成功機率。
- 多序列比對是用來建立 pattern / profile 的關鍵技術。

Pattern vs. Profile

- **Pattern**
 - ◆ e.g. Prosite pattern for PHI-Blast $\langle A-x-[ST](2)-x(0,1)-V$
- **Profile**
 - ◆ for protein families which lack strongly conserved features.
 - ◆ e.g. PROFILES (Gribskov et al PNAS 4355-4358 1987), PSSM (for PSI-BLAST) or HMM

**-AGTCA
CATTTT
GACTCT
CTGTCC**

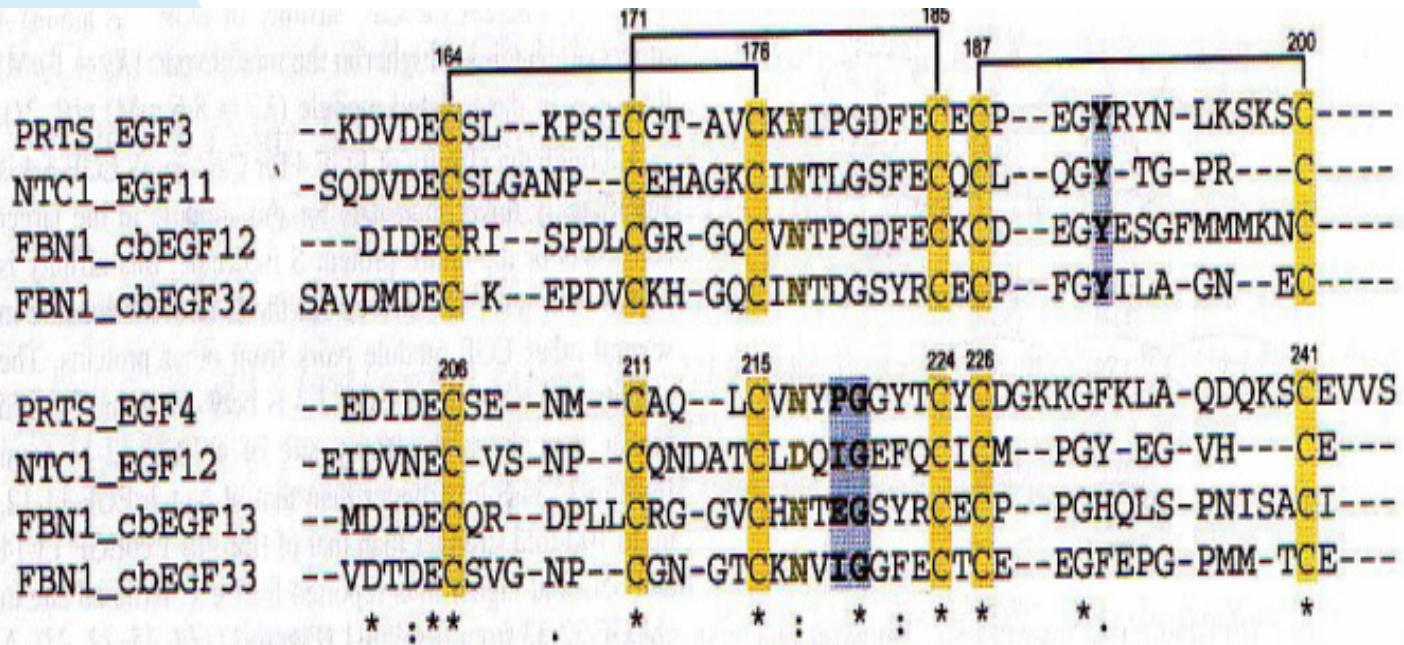
	L1	L2	L3	L4	L5	L6
A		0.75				0.25
T		0.25	0.25	1.00	0.25	0.50
G	0.25		0.50			
C	0.50		0.25		0.75	0.25

多序列比對之應用

- 尋找具有重要生物功能的功能區
- 辨認蛋白質中之 module/domain , 輔助 structure prediction
- 建立蛋白質家族之 pattern / profile
- 分子演化分析的前奏
 - ◆ Guide tree
 - ◆ Gene Family tree
 - ◆ Speciation/Phylogeny tree
- 設計 PCR primer

多序列比對 (Multiple Sequence Alignment)

- Clustal W, GCG-PileUp
- Demo!



Where can I find sequences for Multiple Sequence Alignment ?

- Use a query sequence to extract homologous sequences from GenBank. These sequences can then be compiled into a **multiple sequence FASTA format** file.
- multiple sequences can also be obtained from Sean Eddy's paper in the web
 - ◆ <http://www.genetics.wustl.edu/eddy/publications/tigs-9808/worm2.fa>

>F18C5.8 CE02657 (ST. LOUIS)

```
MENLNPACASEDVKNALTSPIIMMLSHGFILMIIVVSFITTALAVQTLWYKNVFPFCTKNLLLSAIVNGIFHQSTVAEIRLKTVYHLIRYSNAPCSILFQSSDCFYDNFLYYQTALFSSFYCVSLFLDRLFSLNPRSFYNHQTLGFI  
VFLILQIICPIAIQFWTFHDSYTSYVPMCNYPASSVSGTKFYFINDSRIIIMGTIFMCSFLYIHNKSREKRMI  
FNVTNTDSRYKSYENFLATRAVCIIFSQITCLGITSFVPSIFNQFRQSSISPDWFHLILAFMAGATYSNFFLPLIV  
IYETQLIIAHRHKLKIKLKSQKEEFSDFASLDFVWEREANKKKKTQLVQ
```

>F28C12.3 CE09750 7TM RECEPTOR PROTEIN (CAMBRIDGE)

```
MNKTAEELLDSRKCASEGLTNALTSFMMKMNFSFIVTIRQIYRSIVYAAEPCKLLYLSSDCVFELHFYYLTN  
YFSTYSVFSLTFDRLISHYKSKFYHINQYFIAISLLVFQLFLTLLSFYIAFYGVPLAGYIPMCTYHPALSVYYST  
INNVRTGVMVSCIVTIFVYYLSVNLEKRIQKTSYSPGERYSAYENVTTSSKVCILIVLQFSCIMISSFGVNLLM  
MSEKSMSEKVFHTIVPFLPGVTYANLCLPLVIFIQTTLTIRTRKMRIAVM TSMYGDVKDHMNRLKKSWE
```

Clustal W

- Most commonly used global MSA tool.
- Progressive approach.
- Could be used and downloaded from several sites, e.g. [EMBL-EBI](#) , [PIR](#) (Georgetown university) and [BCM Search Launcher](#) (Baylor college of medicine)

Clustal W

ClustalW - Microsoft Internet Explorer

EMBL-EBI
European Bioinformatics Institute

SEQUENCE ANALYSIS

ClustalW Submission Form

Clustal W is a general purpose multiple sequence alignment program for DNA or proteins. It produces biologically meaningful multiple sequence alignments of divergent sequences. It calculates the best match for the selected sequences, and lines them up so that the identities, similarities and differences can be seen. Evolutionary relationships can be seen via viewing Cladograms or Phylograms.

[Download Software](#)

ALIGNMENT TITLE	ALIGNMENT	OUTPUT FORMAT	OUTPUT ORDER	COLOR ALIGNMENT
<input type="text" value="NONE"/>	<input type="text" value="full"/>	<input type="text" value="aln whnumbers"/>	<input type="text" value="signed"/>	<input type="text" value="no"/>

KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="percent"/>	<input type="text" value="def"/>	<input type="text" value="def"/>

MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="def"/>	<input type="text" value="def"/>

TREE GRAPH		PHYLOGENETIC TREE		
TYPE	DISTANCES	TREE TYPE	CORRECT DIST.	IGNORE GAPS
<input type="text" value="cladogram"/>	<input type="text" value="hida"/>	<input type="text" value="none"/>	<input type="text" value="off"/>	<input type="text" value="off"/>

Taskbar: 開始, 動..., a_c..., OU..., lect..., Micr..., Clus..., Pho..., 下午 04:04

- <http://www.ebi.ac.uk/clustalw/index.html>

Try the MSA yourself!

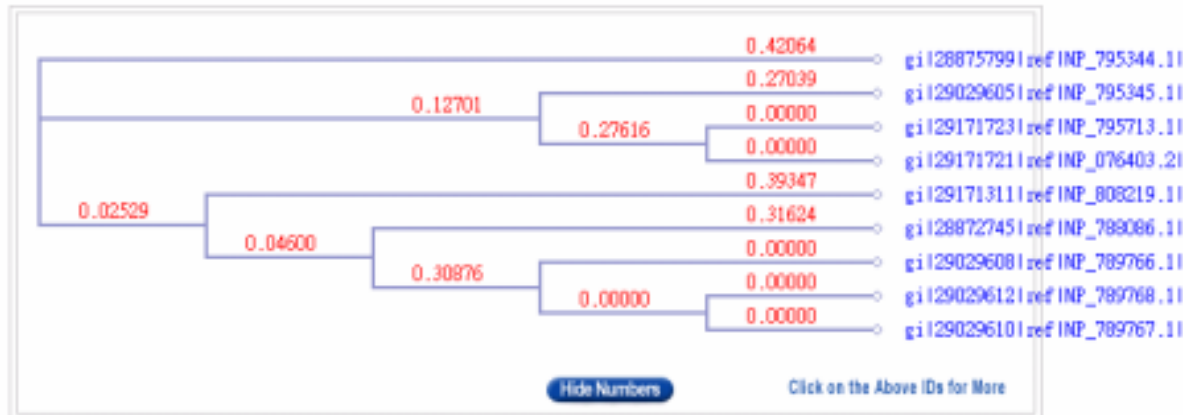
- Goto NCBI-Entrez and then search 'Protein' for 'rhodopsin human'
- Click on the top 10 sequences
- Display 'Fasta'
- copy the sequences and put them in a multiple FASTA sequence file.
- Go to EMBL-EBI-ClustalW site
 - ◆ Upload the multiple sequence file and 'Run'
 - ◆ scroll down the webpage to see the results and the guide tree
 - ◆ use JalView to manipulate the result
- Try also the PIR site to see the better guide tree
 - ◆ <http://pir.georgetown.edu/pirwww/search/multiIn.html>



PIR Multiple Alignment

[Site Map](#)[Site Search](#)Text Search Protein Databases: [About PIR](#)[Databases](#)[Search & Retrieval](#)[Download](#)[Support](#)

TREE VIEW:



Branch lengths are indicated by numbers.

(For best printout, use IE or use Netscape 6.0 or higher browser.)

MULTIPLE ALIGNMENT:

```
gi|29171723|ref|NP_795713.1|  ----MNTT-----VMQGFNRS-ERCPR-----  
gi|29171721|ref|NP_076403.2|  ----MNTT-----VMQGFNRS-ERCPR-----
```

◆ <http://pir.georgetown.edu/pirwww/search/multiIn.html>

Clustal W outputs

- Multiple sequences alignment result
 - ◆ some system (e.g. DiAlign) will give the alignment result in FASTA format.
- A guide tree in the Newick tree format
 - ◆ <http://evolution.genetics.washington.edu/phylip/newicktree.html>

Trees and Graph Theory

- **Three basic data structures**
 - ◆ lists, trees and graphs
- **Trees**
 - ◆ Topology (branch order)
 - ◆ Branch (edge) length (weight)
 - ◆ rooted and un-rooted trees
 - ◆ leaf nodes and internal nodes

Progressive MSA, Clustal W and Trees

- J. D. Thompson, D. G. Higgins and T. J. Gibson, **“CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice,”** *Nucleic Acid Research*, vol.22, no.22, 4673-4680, 1994. (EBI)
- <http://bimas.dcrn.nih.gov/clustalw/clustalw.html>

Time and Space complexity

- Time and space are important resources for a computational system
- Time and space complexity can be described using the big O notation
 - ◆ The time and space consumption is a function of input data size n .
 - ◆ Thus, the big O notation is generally presented as follows
 - ☞ Time complexity = $O(n)$ -- linear
 - ☞ Space complexity = $O(n^2)$ -- quadratic

Computational Rationale of Progressive approach for global MSA

- 若利用 dynamic programming 進行長度各為 n_1 與 n_2 之雙序列全域比對，則
 - ◆ time complexity = $O(n_1 * n_2)$
- 當長度一致均為 n 時，time complexity = $O(n^2)$
- 若用 dynamic programming 於 i 條長度為 n_j 之多序列全域最佳化 (Optimization) 比對 ($i \ll n_j$)，則
 - ◆ time complexity 為 $O(n_1 * n_2 * n_3 * \dots * n_j)$
- 當長度一致均為 n 時，time complexity = $O(n^i)$
- 若將此序列兩兩比對，則可大幅減少計算量為
 - ◆ $C(l, 2) n^2 = O(n^2)$
- Progressive approach，如 Clustal W 為 $O(n^2)$
- Question：是否可以避開 local optimum 達到 global optimum??

Biological Rationale of Progressive approach for global MSA

- **Da-Fei Feng and R. F. Doolittle, “Progressive sequence alignment as a prerequisite to correct phylogenetic trees,” *J Mol. Evol* 25:351-360, 1987.**
- **Sequences are aligned progressively, beginning with the most similar pair and continuing with the addition of the next most similar sequence**
- **Once a gap, always a gap**
- **Assumes that divergent evolution is binary in nature**
- **Putting more trust in the comparison of recently diverged sequences than in those evolved in distant past**
- **Parsimony**

Algorithm of Clustal W Part 1

- Progressive approach (漸進式之方法)
- 1. Pairwise alignment using
 - ◆ 1.1 the **full** mode, which utilizes dynamic programming with the affine gap penalties, i.e. opening (GOP) and extending (GEP).
 - ◆ 1.2 the **fast** mode. Parameters such as ktup, window size etc must be set. (not described in paper)
- 2. Calculate distance of all pairs of sequences.
 - ☞ Distance = 1 – percent identity
- 3. Distance matrix construction using the pairwise alignment of all pairs of sequences

Distance Matrix of a set of globins

Hbb_human 1						
Hbb_horse 2	0.17					
Hba_human 3	0.59	0.60				
Hba_horse 4	0.59	0.59	0.13			
Myg_Phyca 5	0.77	0.77	0.75	0.75		
Glb5_Petma 6	0.81	0.82	0.73	0.74	0.80	
Lg2_Luplu 7	0.87	0.86	0.86	0.88	0.93	0.90
	1	2	3	4	5	6

Algorithm of Clustal W Part 2

- **4. Guide tree construction**
 - ◆ **4.1 UPGMA – unweighted pair-group method with arithmetic mean (old version)**
 - ◆ **4.2 Neighbor-joining method + mid point method (new version)**
 - ☞ **Neighbor-joining method >> Unrooted guide tree**
 - ☞ **Mid point method >> Rooted guide tree**

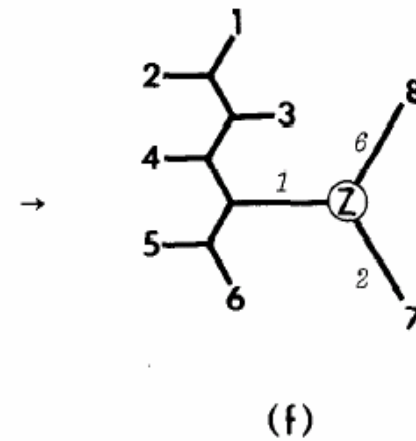
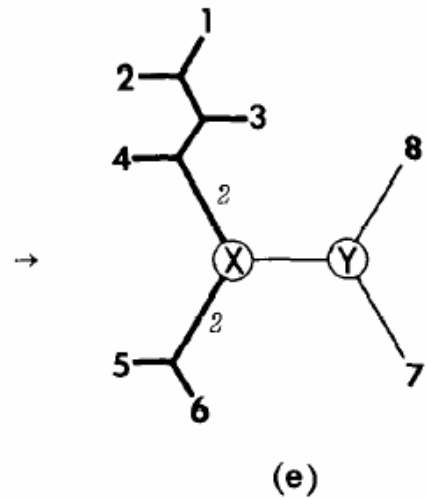
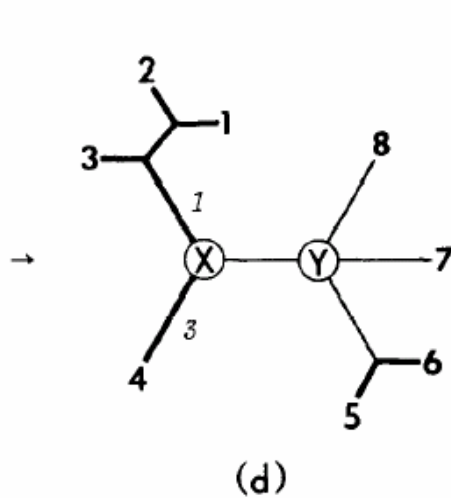
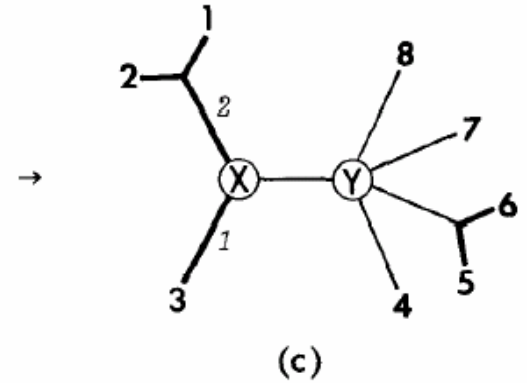
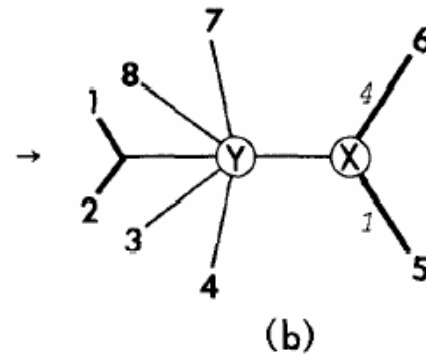
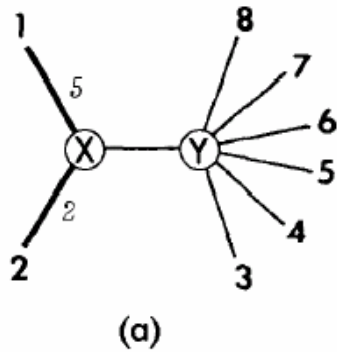
Guide Tree : UPGMA

- **Unweighted pair-group method with arithmetic mean**
 - ◆ 由 Distance Matrix 中，尋找 distance 最小之兩條序列 (or OTUs)，將之群聚 (clustering) 起來，同時平均分派 branch length。
 - ◆ OTU : operational taxonomic unit
 - ◆ 調整群聚之後之 distances
 - ◆ 重複以上程序，直到所有序列均被納入 tree 中。
- **Problems**
 - ◆ UPGMA 每次只考慮一對 nodes，而不是整個 tree
 - ◆ 對於一對 nodes，分離點到兩端點距離相同之假設常常與實情不符，導致錯誤的 topology
- <http://www.icp.ucl.ac.be/~opperd/private/upgma.html>

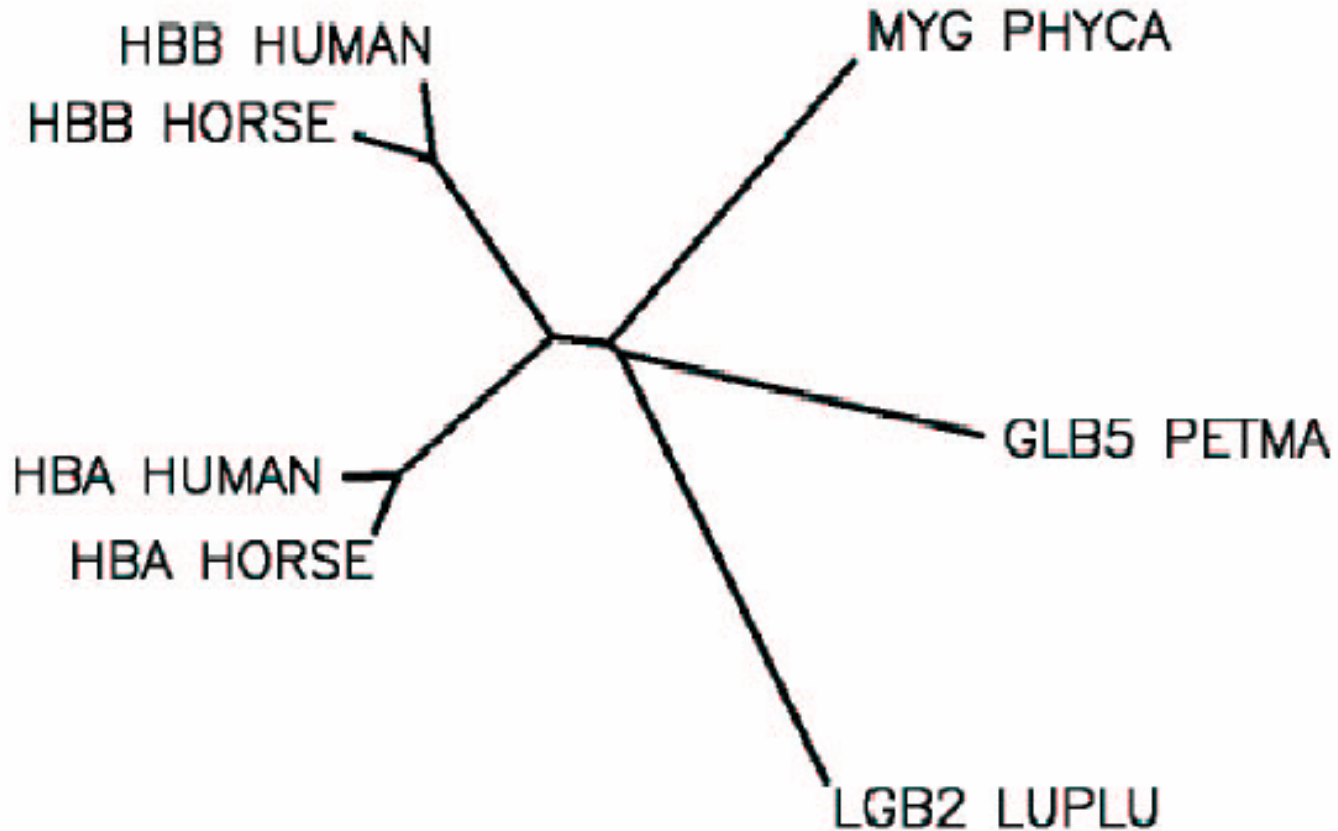
Guide Tree : Neighbor Joining

- **N.Saitou and M Nei, “The neighbor-joining method: a new method for reconstructing phylogenetic trees,” *Mol Biol Evol* 1987 Jul 4(4): 406-25**
 - ◆ Try to minimize the total length of the tree
 - ◆ N sequences ($N \geq 3$)
 - ◆ Starts with a star-like tree, centered at ‘X’
 - ◆ Obtain ‘Y’ which results in the smallest sum as the new center.
 - ◆ Iteratively obtain new ‘Y’ until number of branch of ‘Y’ is 3.
 - ◆ Set ‘Y’ as ‘Z’ and the topology of tree is fixed.
 - ◆ Using the least square estimation to set the branch lengths

Neighbor Joining (Saitou and Nei 1987)



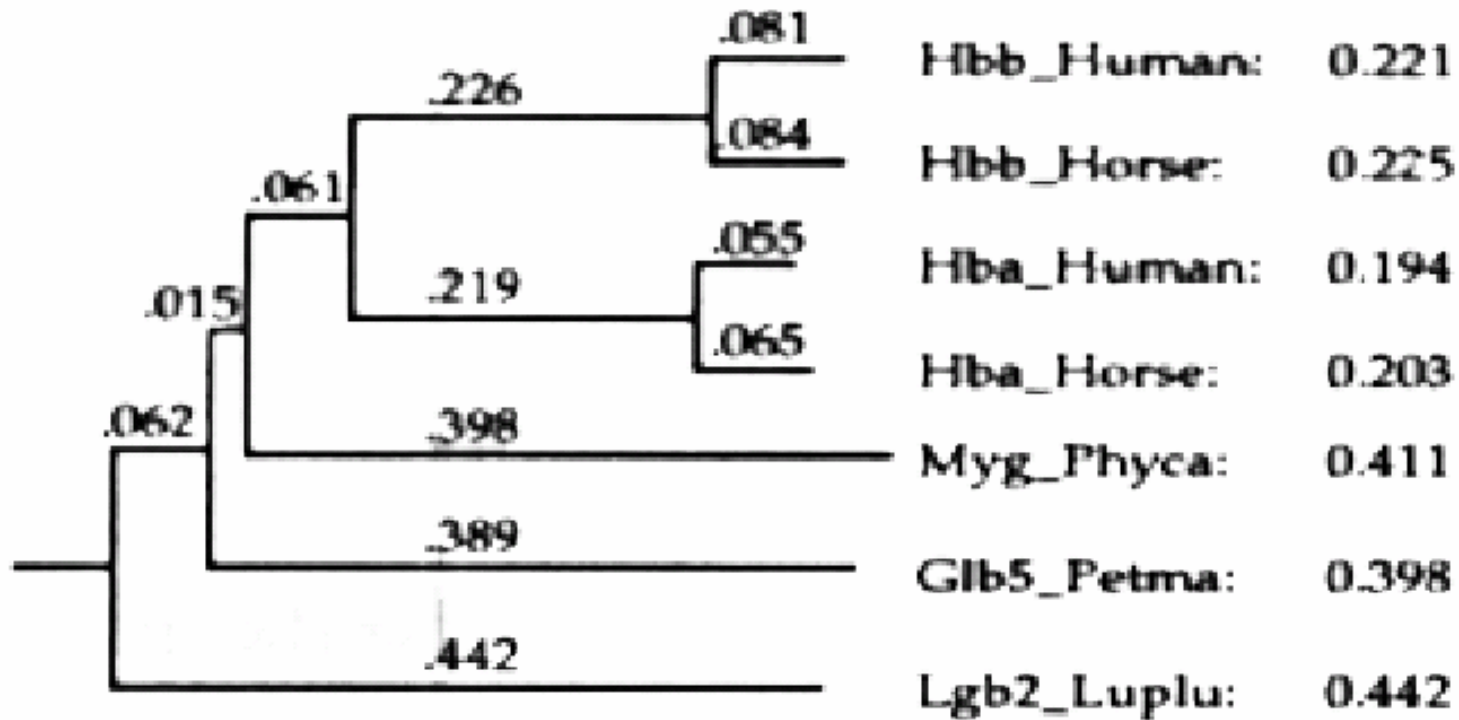
An un-rooted tree of a set of globins



Guide Tree : mid-point method

- **J.D. Tompson, D.G.Higgins and T.J.Gibson “Improved sensitivity of profile searches through the use of sequence weights and gap extension,” CABIOS, vol.19, no.1, 19-29, 1994.**
 - ◆ **Determine the branch length using the topology of the un-rooted tree, the distance matrix and the linear algebra.**
 - ◆ **For every internal node in the un-rooted tree, bisect the tree into left tree and right tree. The total distance of the sequences to this internal node is then calculated.**
 - ◆ **The real node of the tree can be inferred from above information.**

A rooted tree with branch lengths



- Branch lengths determines weights for each sequence. The most divergent sequence receives the highest weights.

Scoring Strategy for Alignment

- 在利用 dynamic programming 計算排比之例子中，我們採用一個簡單的給分策略
 - Match：正分
 - Mismatch：負分
 - Gap 負分
- 在 Gap 的給分中，可以用較細緻的 affine gap
 - Gap open penalty (GOP)
 - Gap extension penalty (GEP)

Algorithm of Clustal W Part 3

- **5. Progressive alignment**
 - ◆ **Align the sequences following the branching order of the guide tree, from the tips of the tree toward the root**
 - ◆ **sequence is weighted during alignment**
- **6. Heuristic on gaps**
 - ◆ **short stretches of hydrophilic residues usually indicate loop or random coil regions and are unlikely to have gaps**
 - ◆ **GOP and GEP is determined on lots of heuristic factors, e.g. where there existing gaps in the neighborhood.**

Clustal W Method Summary

- Assumes sequence similarity provides the phylogeny information, thus the alignment can follow it.
- Does not guarantee global optimum.
 - ◆ This problem can be alleviated using iterative or stochastic sampling procedure.
- The alignment obtained from the initial stage can effect the final stage of the alignment
- Not reliable for highly divergent sequences
- DBClustal = Blastp database search + ClustalW
 - ◆ **J. D. Tompson et al.**, “DBClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches,” **Nucleic acids research**, vol.28, no 15 pp.2919-2926, 2000.
 - ◆ <http://igbmc.u-strasbg.fr:8080/ballast.html>

Can you follow me ???

- 以上對於 UPGMA, neighbor joining, mid-point method 等等介紹屬於演算方法中較深入之部分，不易理解。
- 若同學學習的目的是運用 MSA 來分析 protein family, 作 phylogenetic study 等等生醫研究，則只要掌握該算法的精神即可。重點是要懂得使用適當的工具，採用適當的參數，以提高研究結果的正確性。
- 若同學目的是改進 MSA 算法，以便達成更強大的功能，則必須深入理解。
- 對於有志於朝生醫資訊演算法發展的同學，這些算法都是基本功夫，瞭解這些算法是很好的練習。

Other MSA related approaches

■ DIALIGN

◆ Global alignment using locally Matched Block

- ☞ <http://bioweb.pasteur.fr/seqanal/interfaces/dialign2-simple.html>
- ☞ <http://www.genomatix.de/cgi-bin/dialign/dialign.pl>
- ☞ <http://www.gsf.de/biodv/dialign.html>

■ MultiAlign

◆ Global MSA with hierarchical clustering

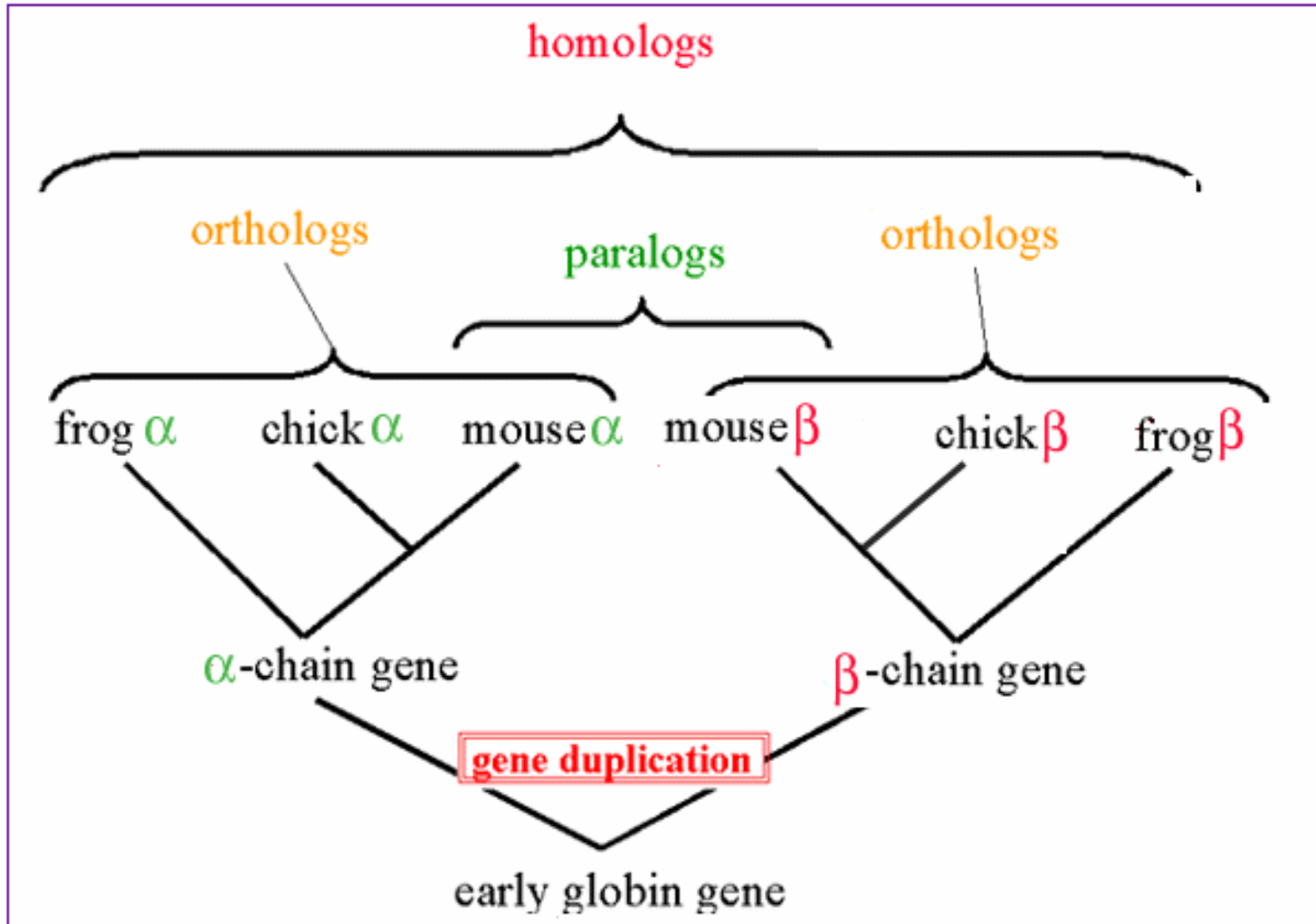
- ☞ <http://prodes.toulouse.inra.fr/multalin/multalin.html>

■ MEME

◆ Motif finding for a group of sequences , 適合序列間有組成相同但次序不同的 domain 時 (e.g.. circular permutation)。

- ☞ <http://bioweb.pasteur.fr/seqanal/motif/meme/>

Homologs



基因體長段複製 (Duplication) 相關文獻

- **Color vision gene**

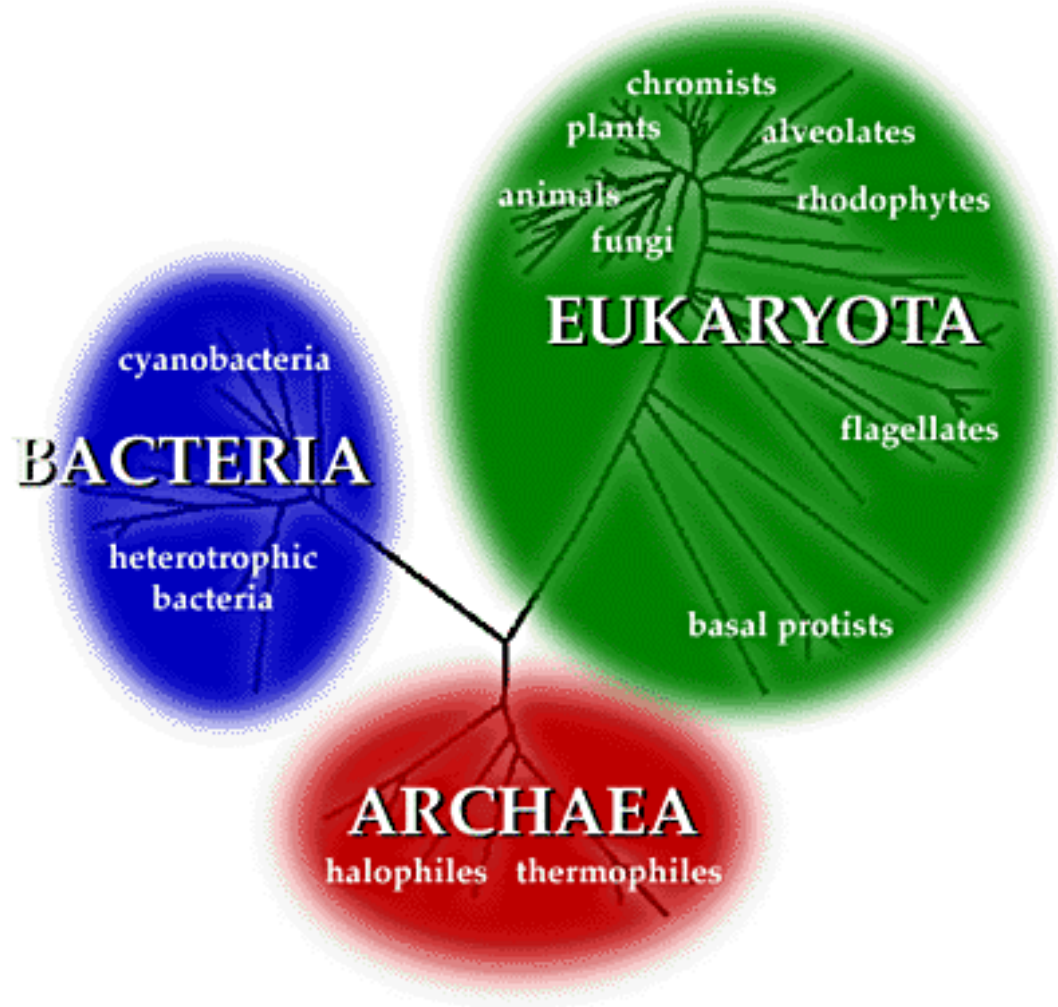
- ☞ Shyue, S.-K., D. Hewett-Emmett, H. G. Sperling, D.M. Hunt, J.K. Bowmaker, J.D. Mollon and Wen-Hsiung Li (1995) Adaptive evolution of colour vision genes in higher primates. *Science* 269:1265-1267

- **Yeast genome:**

- ◆ **446 duplicated genes in found in 55 duplicated regions. These genes make up 13% of proteins**

- ☞ K.H. Wolfe and D.C. Shields, “Molecular evidence for an ancient duplication of the entire yeast genome,” *Nature* 387, 708-713, 1997

演化樹建構 (Phylogeny)



Algorithm and software that synchronize gene family trees with Phylogenetic trees

- **Kevin Chen, Dannie Durand and Martin Farach-Colton** “**Notung: A program for dating gene duplications and optimizing gene family trees,**” **Journal of Computational Biology,** vol.7, numbers 3/4 pp.429-447, 2000.

Phylogenetic trees vs. gene family trees

- A guide tree is more likely to be a gene family tree. It doesn't guarantee to be a true phylogenetic tree.
- Observations : within a gene family tree, there are
 - ◆ speciation internal node and duplication internal node

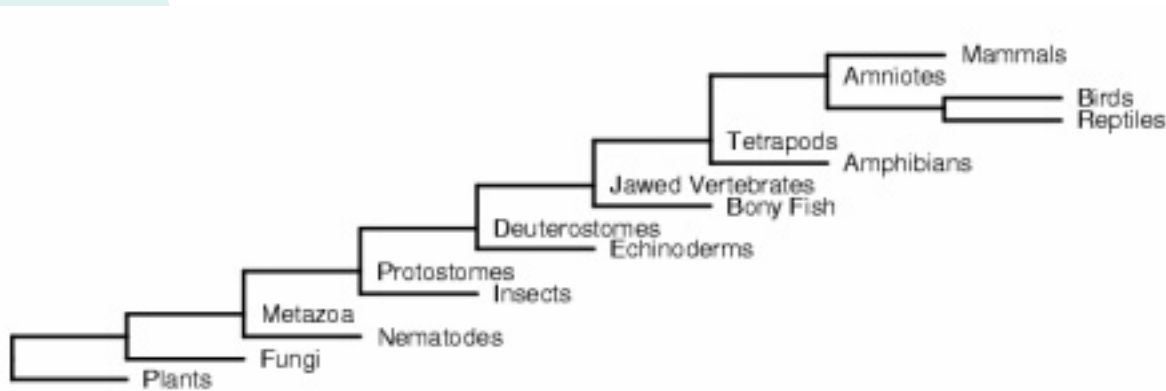
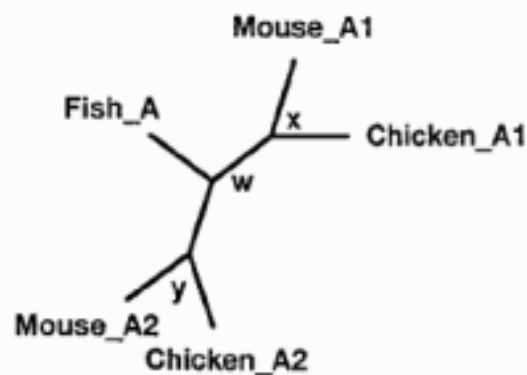
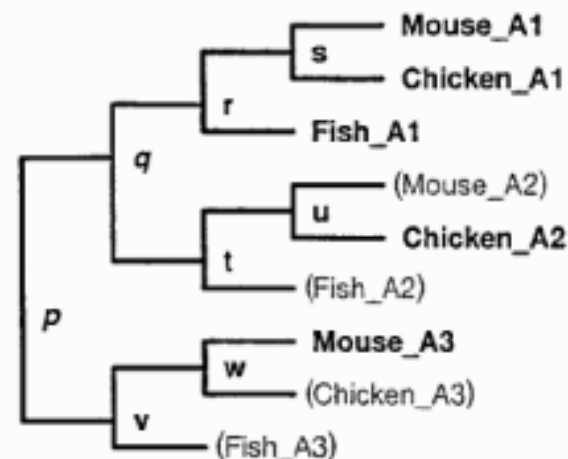


Figure 2: A species tree showing major speciation events in the eukaryote lineage. This tree was derived from the University of Arizona Tree of Life project [13] and the NCBI Taxonomy database [16]

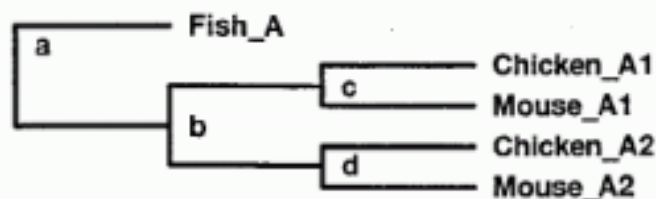
- See the University of Arizona, Tree of Life Project at
 - ◆ <http://tolweb.org/tree/phylogeny.html>



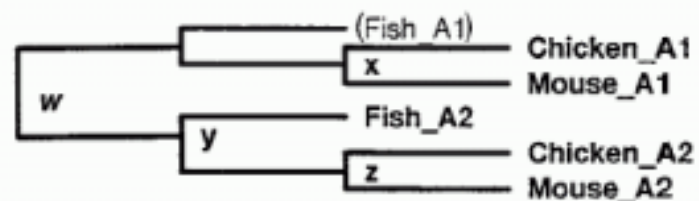
(a)



(d)



(b)



(c)

FIG. 3. Gene family trees for the hypothetical gene family, A, with five known gene sequences (two in mouse, two in chicken and one in fish). (a) An unrooted GFT for A. (b) – (d) Three alternate rootings of the GFT in (a). Duplication nodes are shown in italics and missing genes are shown in parentheses.

Three functions of Notung

- **Identify duplication events for unambiguous gene family tree by comparing it with a species tree.**
 - **Find a best root for an un-rooted gene family trees which is consistent with phylogeny**
 - **Optimize rooted gene family trees which contain edges of weak confidence**
- [Kevin Chen, Dannie Durand and Martin Farach-Colton](#) “Notung: A program for dating gene duplications and optimizing gene family trees,” [Journal of Computational Biology](#), vol.7, numbers 3/4 pp.429-447, 2000.

A gene family tree, produced by neighbor joining heuristics, may be inconsistent with the phylogeny

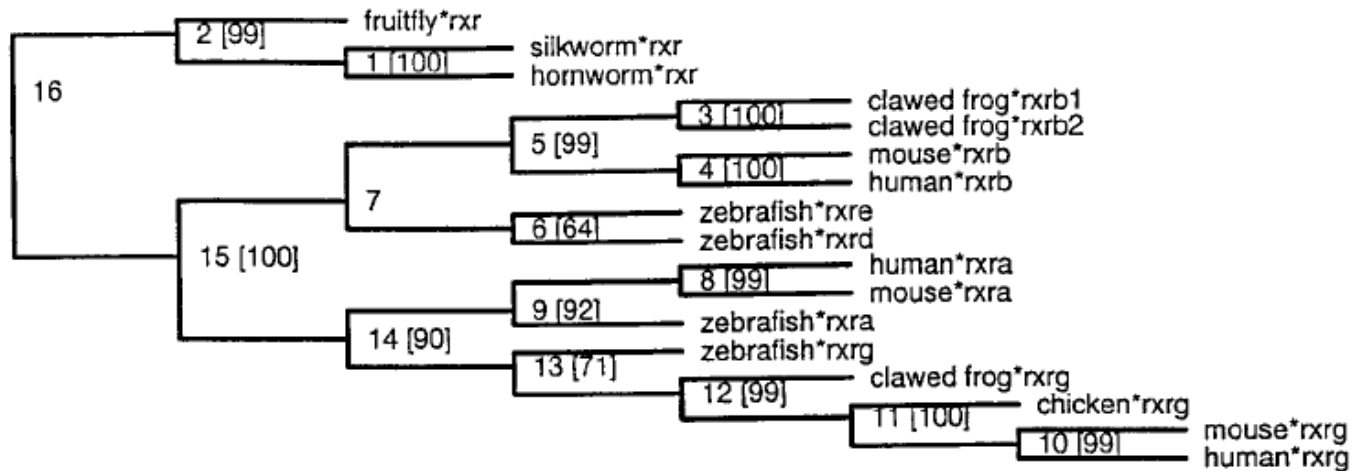


FIG. 1. A rooted Neighbor Joining tree for the RXR family reproduced from Hughes (1998). Interior nodes are labeled numerically. Labels in square brackets represent the percentage of bootstrap samples supporting that branch leading from the label to the root. Values $\leq 50\%$ are not shown.

- **A.L. Hughes, “phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosome 6,9, and 1,” MBE, 15, 854-870, 1998**

1. Duplication event identification

- Goal: identify species ($M(v)$) and duplication events of nodes (v) in a gene family tree
- bottom up approach, $M(v) = \text{least common ancestor} (M(l(v)), M(r(v)))$
- a node is a duplication node iff $M(v) = M(l(v))$ or $M(v) = M(r(v))$ or both
- duplication event occurs at
 - ◆ Upper bound $U(v) = M(av)$ $av = \text{ancestor of } v$
 - ◆ Lower bound $L(v) = M(v)$
- [Kevin Chen, Dannie Durand and Martin Farach-Colton](#) “Notung: A program for dating gene duplications and optimizing gene family trees,” [Journal of Computational Biology](#), vol.7, numbers 3/4 pp.429-447, 2000.

Result of duplication event identification for RXR family

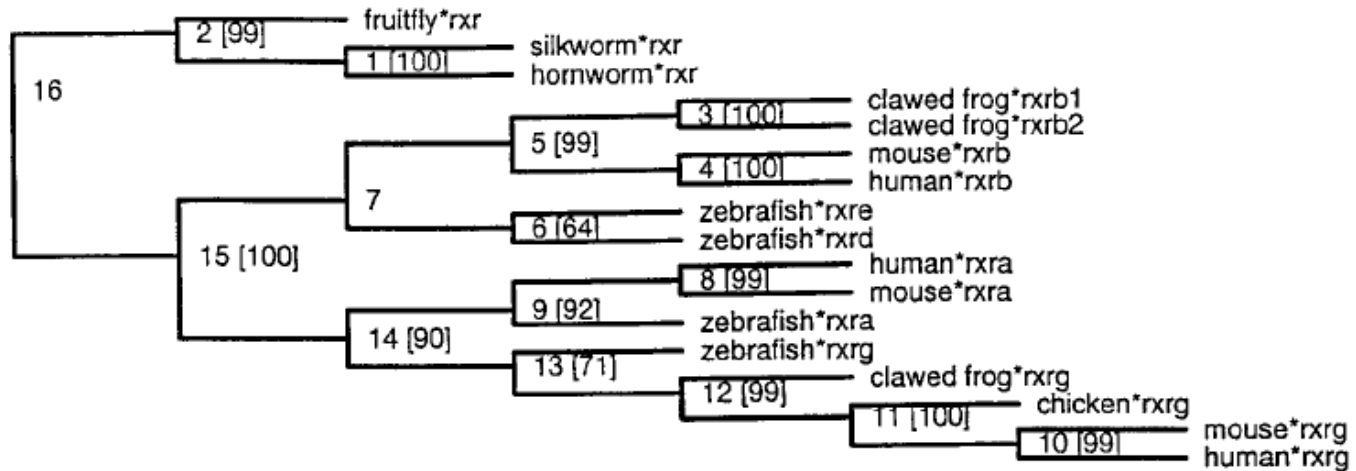


FIG. 1. A rooted Neighbor Joining tree for the RXR family reproduced from Hughes (1998). Interior nodes are labeled numerically. Labels in square brackets represent the percentage of bootstrap samples supporting that branch leading from the label to the root. Values $\leq 50\%$ are not shown.

Duplication at 15	Lower bound: jaw	Upper bound: pro
Duplication at 14	Lower bound: jaw	Upper bound: pro
Duplication at 6	Lower bound: zebrafish	Upper bound: jaw
Duplication at 3	Lower bound: clawed frog	Upper bound: tet

2. Find best root for un-rooted gene family tree

$$C_{dl}(T_G) = c_\lambda \cdot \lambda + c_\delta \cdot \delta,$$

- **lamda** : number of gene loss
- **delta** : number of duplication nodes
- **For every node in the un-rooted tree, calculate C_{dl} . Report the node which gives the smallest C_{dl} as root.**
- **Kevin Chen, Dannie Durand and Martin Farach-Colton** “Notung: A program for dating gene duplications and optimizing gene family trees,” **Journal of Computational Biology**, vol.7, numbers 3/4 pp.429-447, 2000.

Nearest Neighbor Interchange

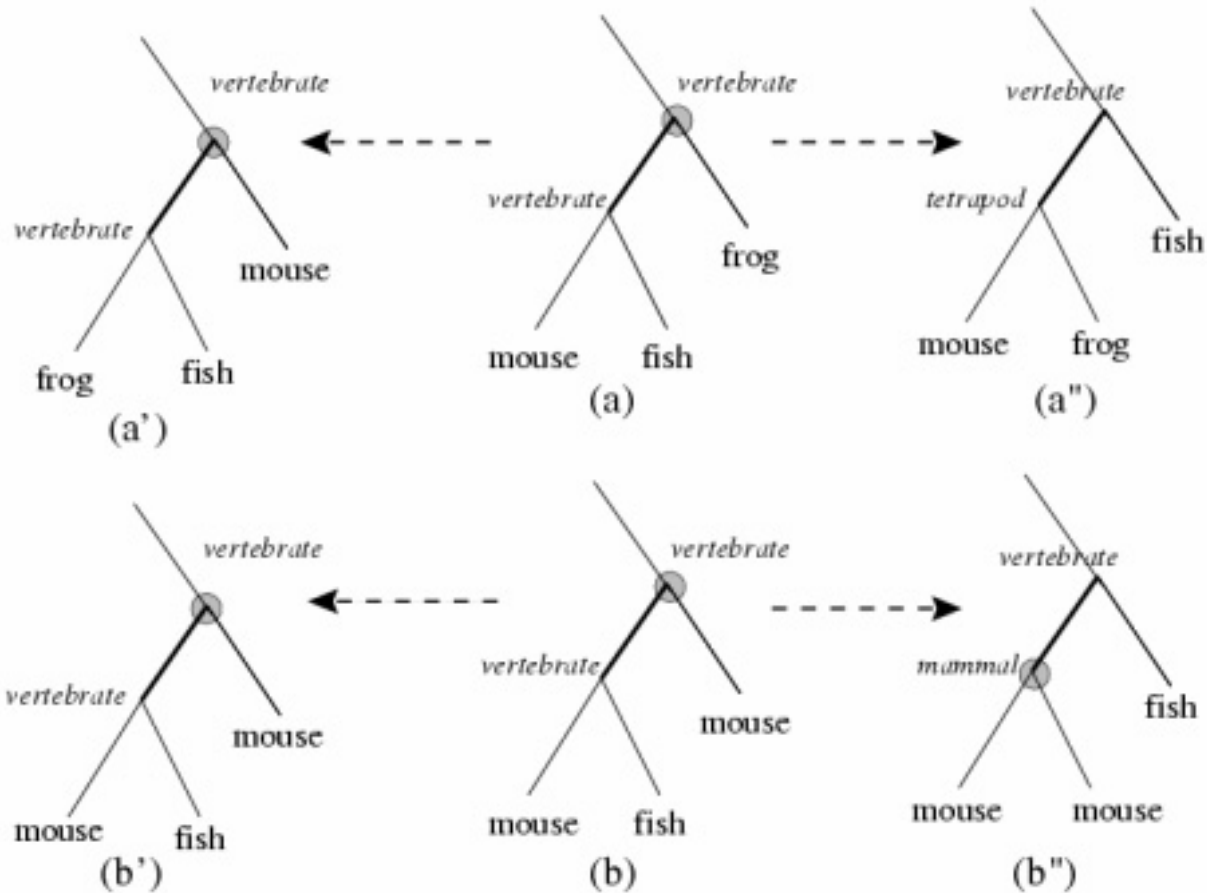


Figure 6: Two tree fragments, each with the three possible Nearest Neighbor Interchanges around the edge shown in bold. Duplication nodes are shown as grey circles.

3. Optimize rooted GFTs

- Nearest neighbor interchange (NNI)
- Bottom-up greedy heuristics
 - ◆ For every weak edge, apply NNI and see which configuration can optimize C_{dl}
- Kevin Chen, Dannie Durand and Martin Farach-Colton “Notung: A program for dating gene duplications and optimizing gene family trees,” *Journal of Computational Biology*, vol.7, numbers 3/4 pp.429-447, 2000.

Algorithm A:

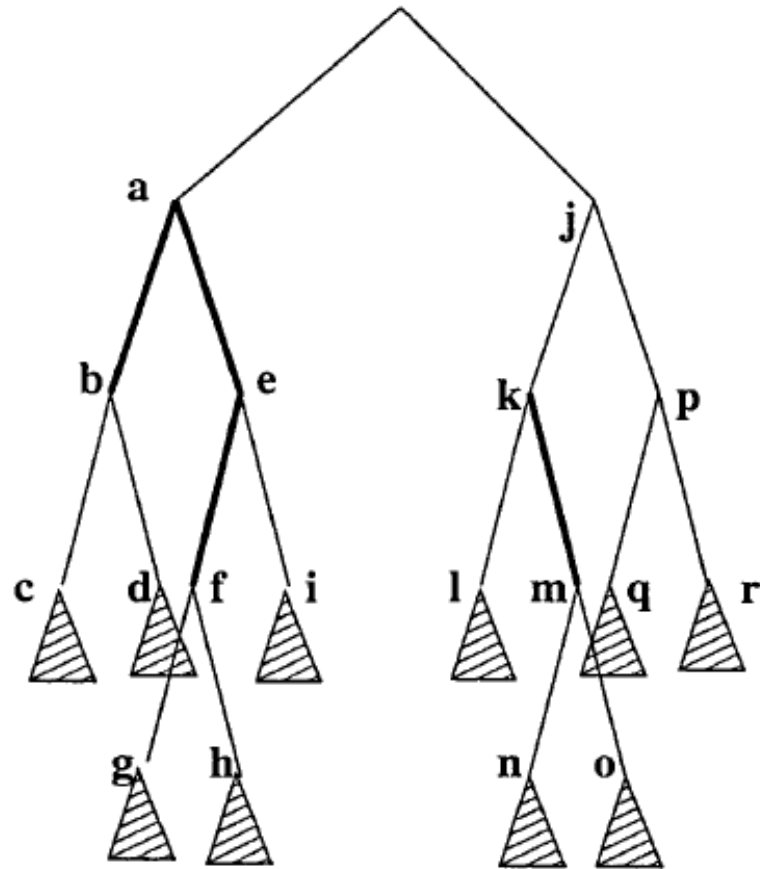
Compute $M(x)$ for every node in T_G .

For each T_i ,

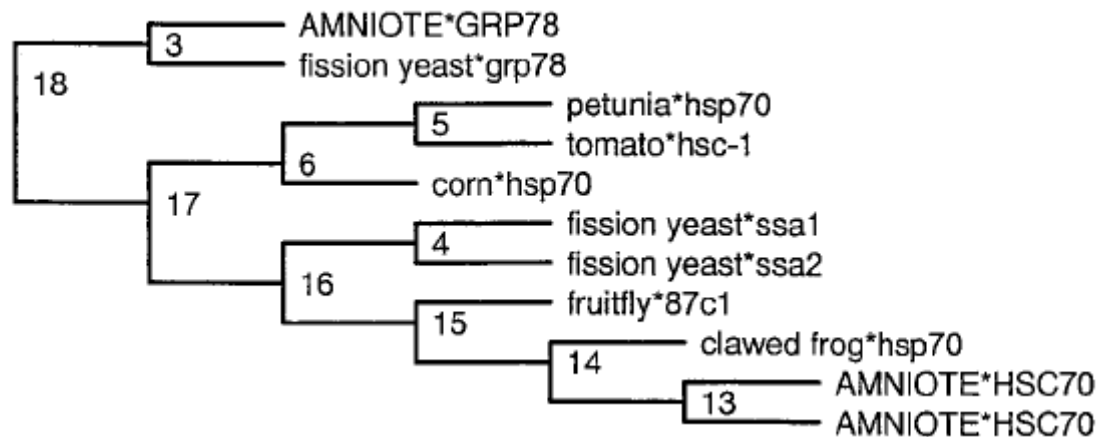
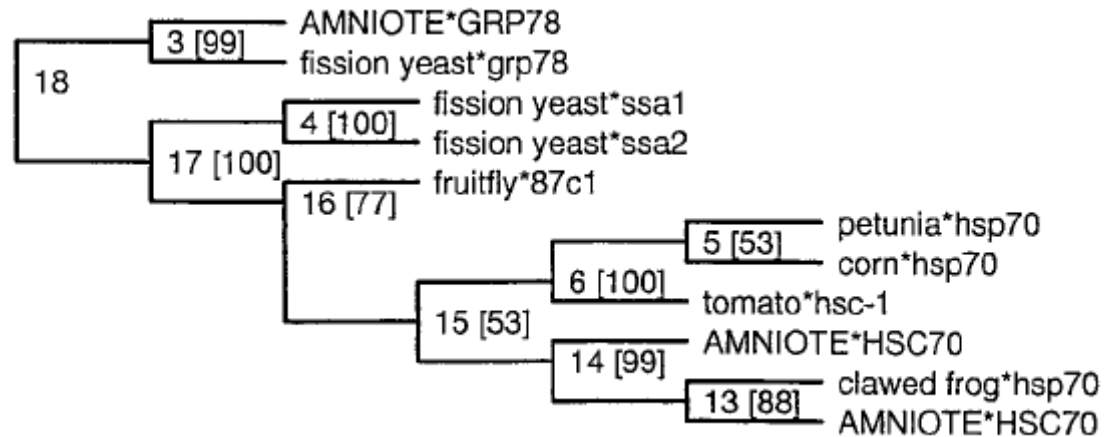
 Compute the optimal rearrangement tree, T_i^* , by exhaustive enumeration.

 Replace T_i with T_i^* in T_G .

Recompute $M(x)$ for every node in T_G .



The HSP70 (Hughes 1998) tree before and after NNI rearrangements. The Phylogeny is changed from (fungi, (insects, (plants, vertebrate))) to (plants, (fungi, (insects, vertebrates)))



延伸閱讀

- Phylogeny Programs
 - ◆ <http://evolution.genetics.washington.edu/phylip/software.html>
- Wen-Hsiung Li, Molecular Evolution, Sinauer Associates, Publishers, 1997
 - ◆ (作者為中研院李文雄院士)