

# Phylogenetic Analysis:

- ✓ *Phylogenetic reconstruction by Automatic Likelihood Model selector (PALM)*

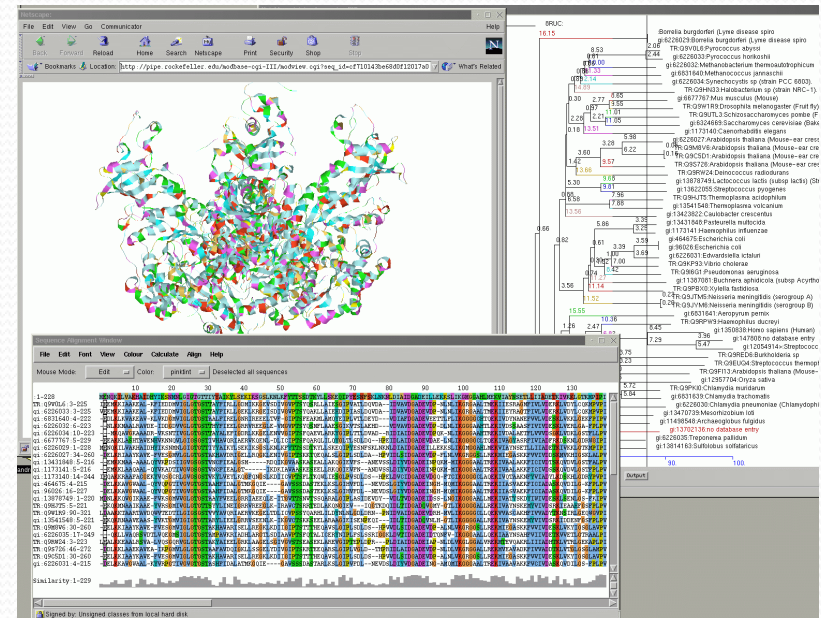
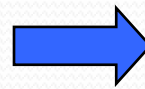
林仲彥 蘇聖堯



中央研究院資訊科學研究所

Oct 7, 2010

# Coding Characters and Defining Homology

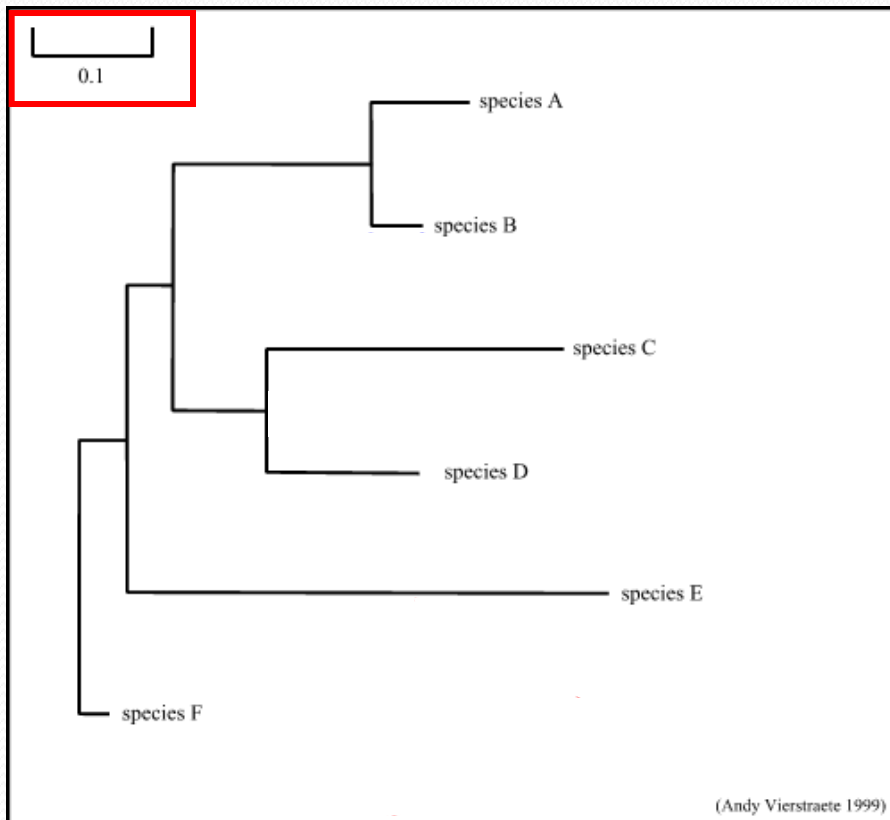


*Classical phylogenetic analysis  
by Morphology*

*Molecular phylogenetic analysis  
By Bio-Molecules*

# Phylogenetic Tree

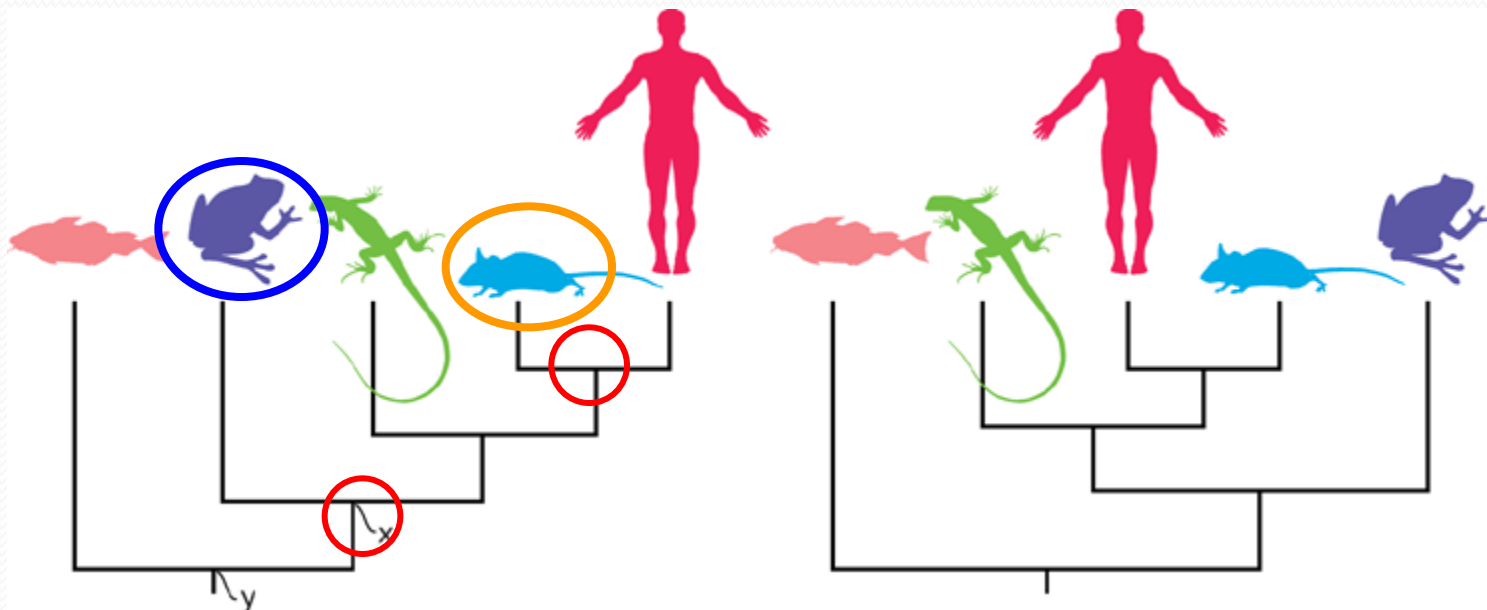
- The tree is composed of nodes connected by branches.



- **node** : a node represents a taxonomic unit.
  - Internal nodes
  - External nodes
- **branch (edge)**: defines the relationship between the taxa.
- **branch length** : often represents the number of changes that have occurred in that branch.
- **root** : is the common ancestor of all taxa.
- **distance scale** : scale which represents the number of differences between sequences (e.g. 0.1 means 10 % differences between two sequences)

# Trees Only Represent The Order Of Branching

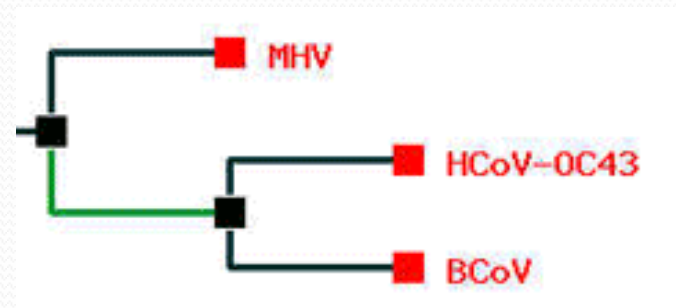
- Same topology in a different style
  - Both trees have identical topologies, with some of the internal nodes rotated.



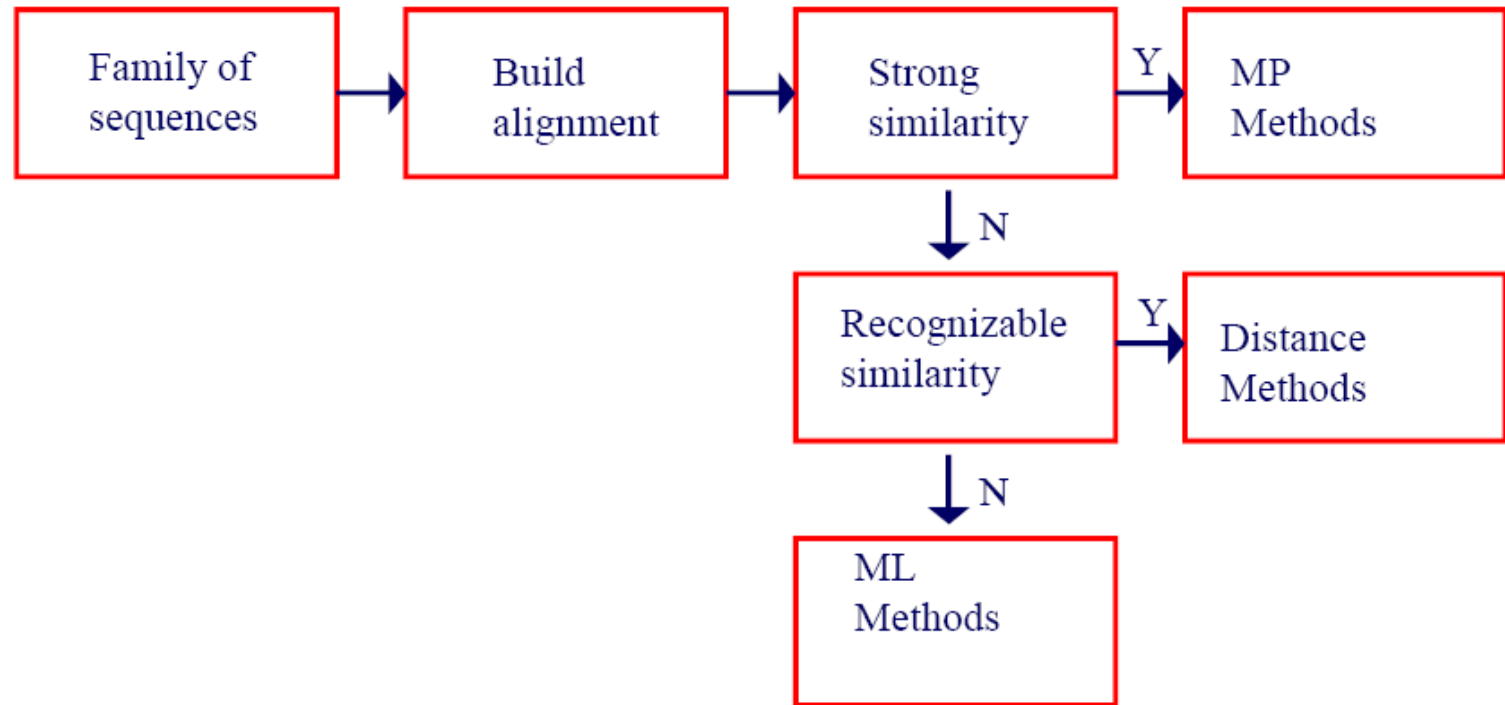
(David A. Baum et al., *Science* 11 November 2005:Vol. 310. no. 5750, pp. 979 – 980)

# *The Ways to Construct the tree*

- Distance-matrix methods
  - Neighbor-joining
  - Fitch-Margoliash method
  - Using outgroups
- Maximum parsimony
  - Branch and bound
  - Sankoff-Morel-Cedergren algorithm
  - MALIGN and POY
- Maximum likelihood
- Bayesian inference



# Flowchart of Analysis



(Mount, *Bioinformatics*)

# *Distance Method, MP and ML*

- Which method should we choose?
- The main disadvantage of distance-matrix methods is their inability to efficiently use information about local high-variation regions that appear across multiple subtrees.
- ML is broadly similar to the maximum-parsimony (MP) method, but **maximum likelihood allows additional statistical flexibility** by permitting varying rates of evolution across both lineages and sites.
- ML, a better choice?

# *Maximum Likelihood*

- Conditional probability of the data (Aligned sequences) given a hypothesis (a model of substitution with a set of parameter  $\theta$ , and the tree  $\tau$ , including topology and branch lengths)

$$L(\tau, \theta) = \text{Prob}(\text{Data} | \tau, \theta)$$

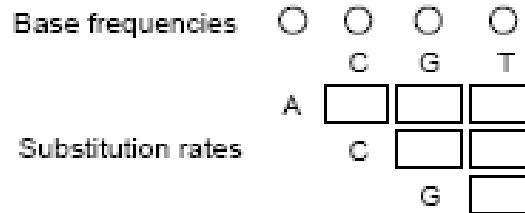
Or

$\text{Prob}(\text{Aligned Sequences} | \text{tree, model of evolution})$



# Illustration of DNA substitution Model

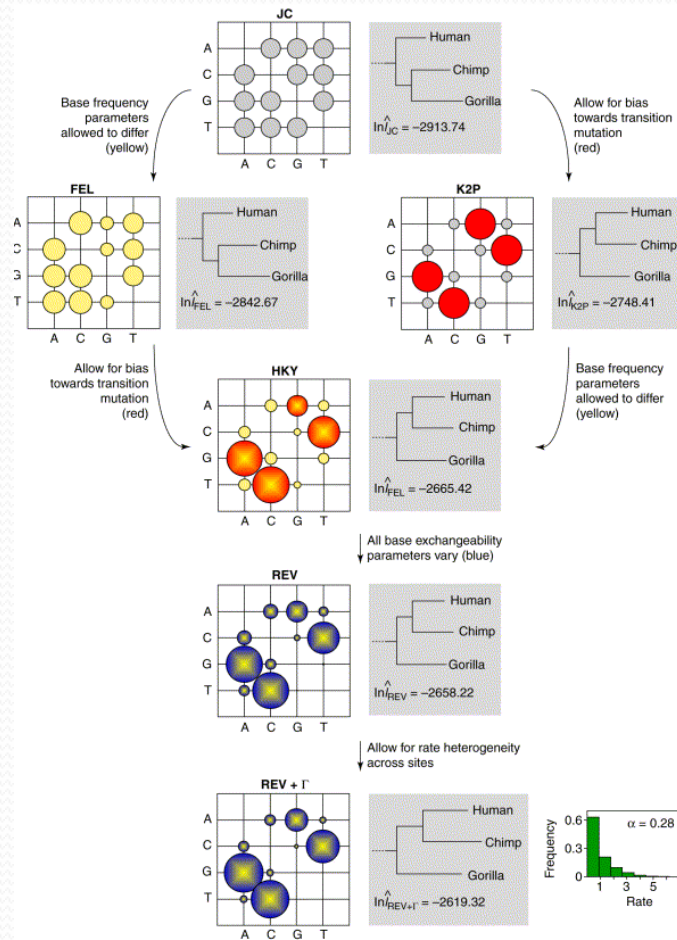
$$Q = \begin{pmatrix} -(x_1 + x_2 + x_3) & x_1 & x_2 & x_3 \\ \frac{\pi_1 x_1}{\pi_2} & -\left(\frac{\pi_1 x_1}{\pi_2} + x_4 + x_5\right) & x_4 & x_5 \\ \frac{\pi_1 x_2}{\pi_3} & \frac{\pi_2 x_4}{\pi_3} & -\left(\frac{\pi_1 x_2}{\pi_3} + \frac{\pi_2 x_4}{\pi_3} + x_6\right) & x_6 \\ \frac{\pi_1 x_3}{\pi_4} & \frac{\pi_2 x_5}{\pi_4} & \frac{\pi_3 x_6}{\pi_4} & -\left(\frac{\pi_1 x_3}{\pi_4} + \frac{\pi_2 x_5}{\pi_4} + \frac{\pi_3 x_6}{\pi_4}\right) \end{pmatrix}$$



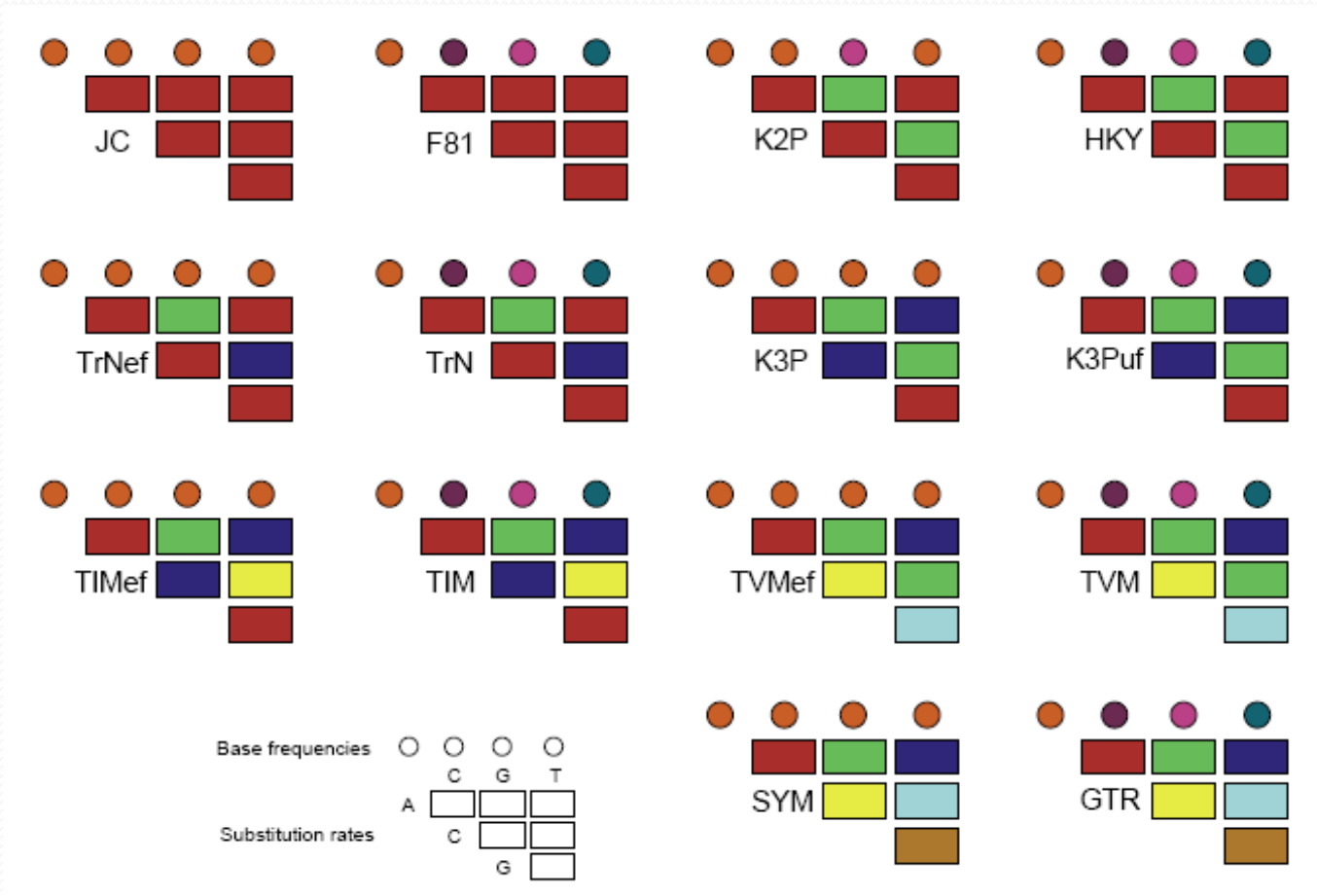
GTR (for four characters, as is often the case in phylogenetics) requires 6 substitution rate parameters ( $x_1 \sim x_6$ ), as well as 4 equilibrium base frequency parameters.



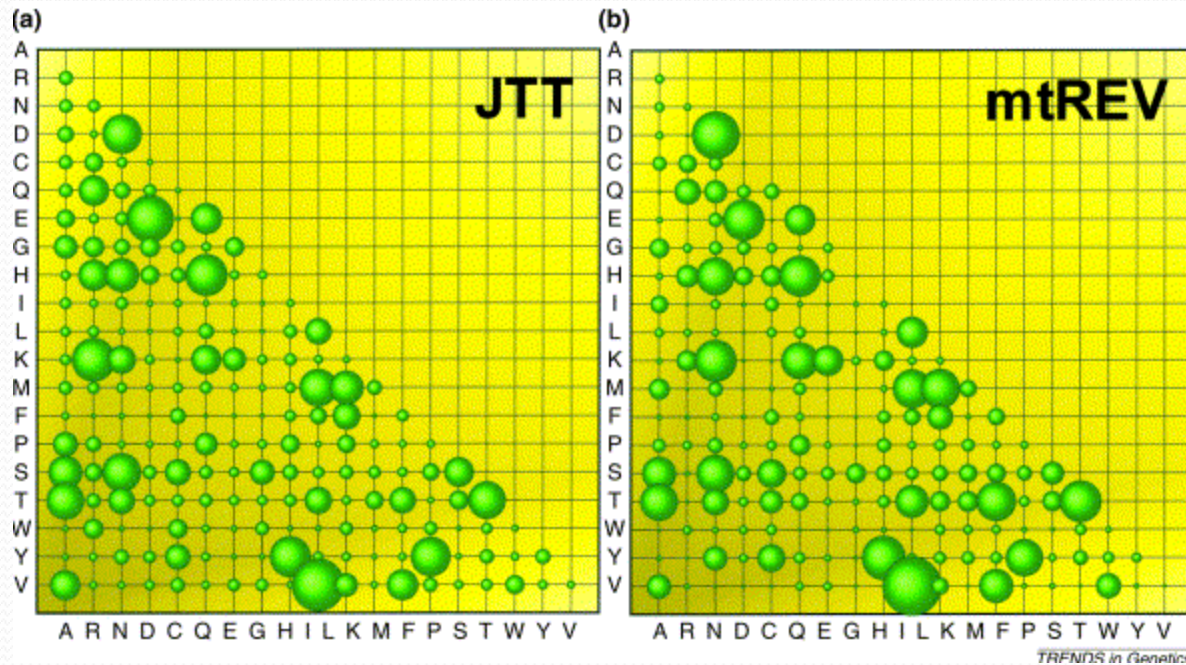
# Relationships Among Some Standard Models Of Nucleotide Evolution



# Illustration of Models for DNA



# Models of Amino Acid Replacement

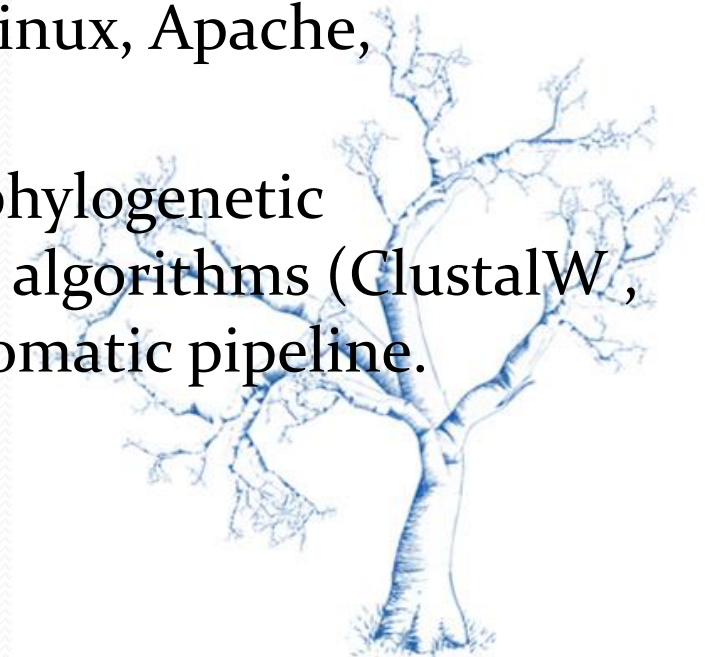


# *Background for PALM*

- Model fitting in phylogenetics has been suggested for many years, yet **many authors still arbitrarily choose their models**, often using the default models implemented in standard computer programs for phylogenetic estimation.
- Here, we want to show the way that a best-fit model can be readily identified. Consequently, given the relevance of models, model fitting should be routine in any phylogenetic analysis that uses models of evolution.

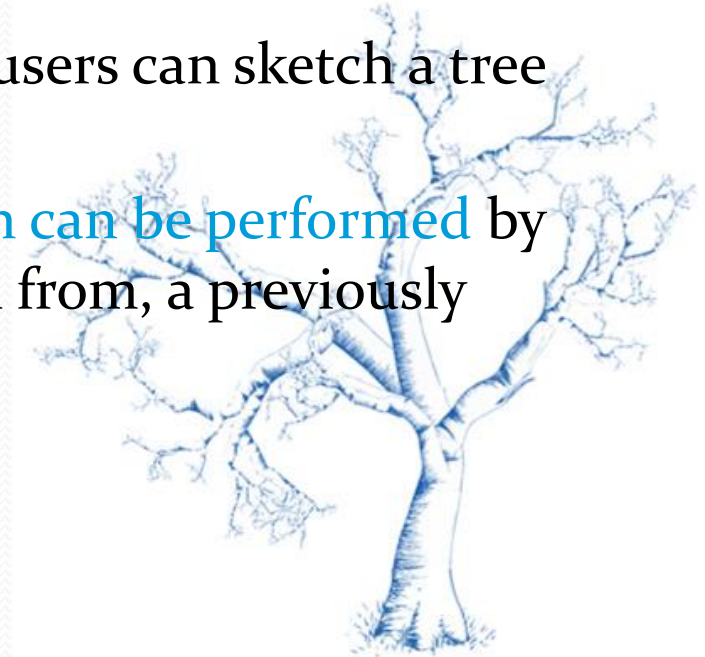
# Motivation I

- Provide a **seamless way** to conduct the **complex phylogenetic analysis** for Biologists
- An integrated and user-optimized framework for biomolecular phylogenetic analysis
- PALM uses an open-source LAPP (Linux, Apache, PostgreSQL, PHP) structure and
- PALM infers genetic distances and phylogenetic relationships using well-established algorithms (ClustalW, PhyML, ProtTest, Modeltest) in automatic pipeline.



# Motivation II

- Model can be selected by following methods including hierarchical likelihood ratio tests (hLRTs), Akaike information criterion (AIC), and Bayesian information criterion (BIC)
- PALM can help user to construct the tree with bootstrap based on best substitution model chosen by maximum likelihood.
- Through a user-friendly web interface, users can sketch a tree effortlessly in multiple steps
- Furthermore, iterative tree construction can be performed by adding sequences to, or removing them from, a previously submitted job



# Component Programs of PALM

- PhyML 3.0
- ModelTest 3.7
- ProtTest 2.0
- ClustalW 2.0.8
- Seqret (EMBOSS)



The screenshot displays the PALM web interface. At the top, there is a logo featuring a palm tree and the text "PALM" in a stylized font, with the subtitle "Phylogenetic reconstruction by Automatic Likelihood Model selector". Below the logo are four navigation icons: Home, Demo, Help, and Contact.

The main content area is titled "Input Sequences" and contains several form fields:

- Input type:** Radio buttons for "Sequence in FASTA format" and "Aligned sequence in PHYLIP format".
- Sequence type:** Radio buttons for "DNA" and "Protein".
- Sequences\*:** A large text input area for pasting sequences, with a "Clear Input" button below it. There is also a file selection button labeled "浏览..." and a checkbox for "example file".
- Number of bootstrap data sets:** A dropdown menu set to "100" and a checkbox for "Print bootstrap information".
- Job Note:** A text input field.
- Enter your email\*:** A text input field.

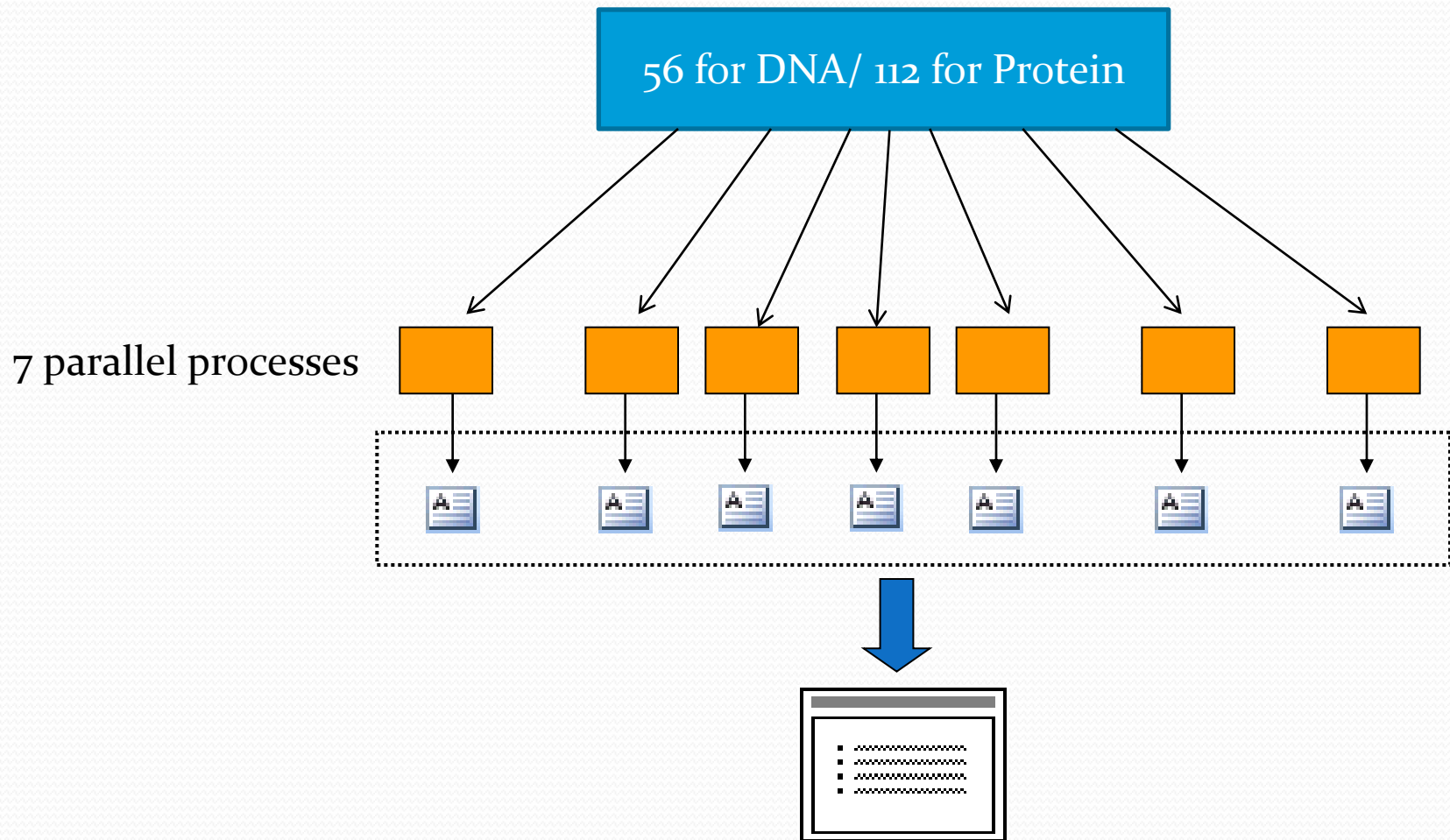
At the bottom, there is a section for "Advanced Option" with a dropdown menu for "Number of substitution rate categories" set to "4".



# *Models Used in PALM*

- For DNA (56 models)
  - JC69, K80, F81, HKY, TrN, TrNef, K3P, K3Puf, TIM, TIMef, TVM, TVMef, SYM, GTR
  - +I, +G
- For Protein (112 models), **Time consuming**
  - LG, DCMut, JTT, MtREV, MtMam, MtArt, Dayhoff, WAG, RtREV, CpREV, Blosum62, VT, HIVb, HIVw
  - +I, +G, +F

# *Distribution Computing by PalmMonitor for Each Substitution Model*



# Decreasing Time by PALMmonitor

- According the algorithm used in PALM, models will take a lot of time to calculate the value of maximum likelihood.

• JTT, MtREV	1h:22:21
• MtMam, MtArt	<u>1h:51:25</u>
• Dayhoff ,WAG	1h:33:51
• RtREV, CpREV	1h:33:50
• Blosum62, VT	1h:15:58
• HIVb, HIVw	1h:35:02
• LG, DC, Mut	1h:38:36



Over 10 Hours

- All Models by PALMmonitor

1h:11:11

Source: 99 sequences with 247 residues for each

# Parallel Computing on Bootstrapping

DNA : [DNA\\_Big\\_24.phy](#) (24 sequences, average 5000 bps substitution model: HKY85- Default)

Bootstrap	100	1000
Runtime	11h:47m:5s	~120 h
Runtime (5 cores/ 8 cores)	2h:53m:2s	17h:52m:15s



1/6 ↓

Protein : [Pseq](#) (20 sequences, average 820 a.a., substitution model: LG - Default)

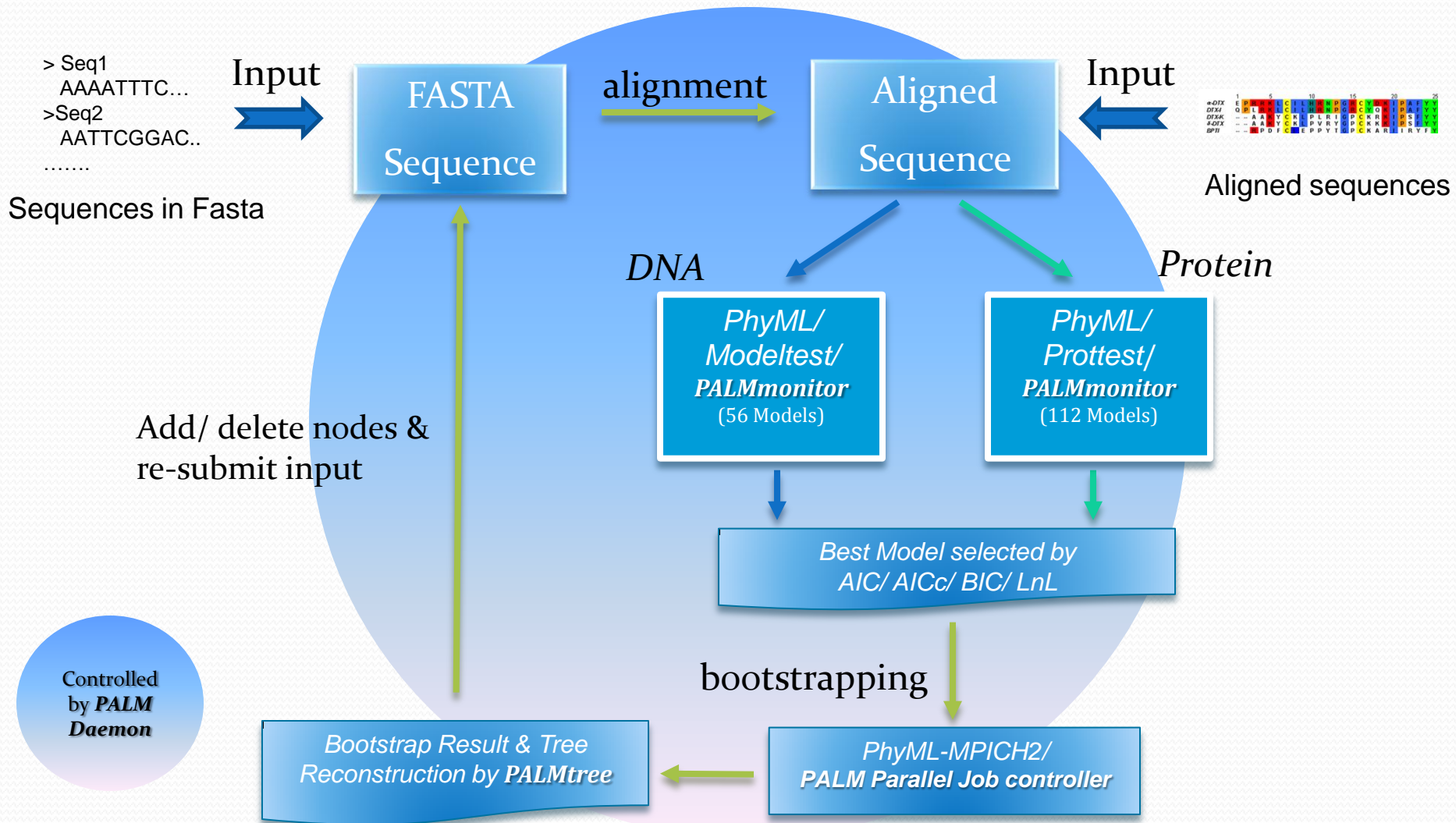
Bootstrap	100	1000
Runtime	17h:31m:33s	~175 h
Runtime (5 cores/ 8 cores)	5h:15m:19s	36h:55m:10s

1/6 ↓

# *Input and Output of PALM*

- Input format (Protein and DNA)
  - Fasta format
  - Phylip format: Aligned Sequences
  - User tree (if submitted and valid)
- Output
  - Tree topology by php and GD library
  - Tree file in Newick format
  - Aligned Sequence in phylip format
  - Best model selected by PALM

# Flowchart of PALM



# Result of PALM



PALM Result

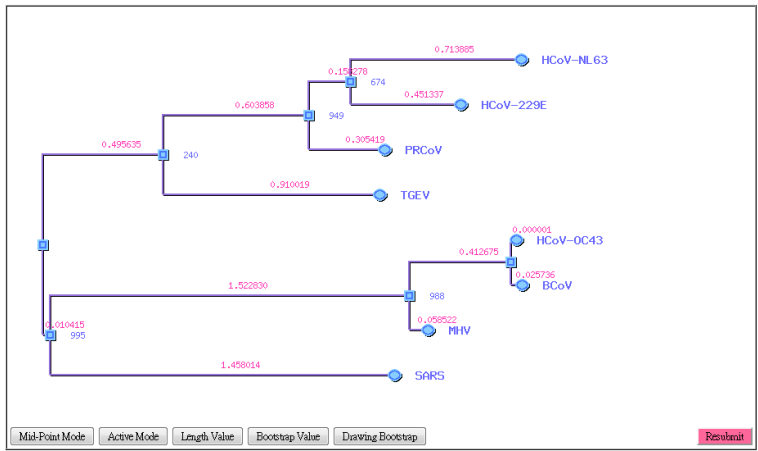
Job ID	20080821060606361	Number of Substitution Rate Category	4
Job Note	test for speed in protein	Model Selection Criterion	LnL
Sequence Type	Protein	Optimization of Tree Topology	Yes
Number of Bootstrap	1000	Optimization of Branch Length	Yes
Starting Tree	BIONJ		

Best Model Selected	JTT+I+G+F
Model Selection Criterion	LnL
AIC	2336.50
-lnL	1134.25

Model	deltaAIC	AIC	-lnL*	AICw
JTT+I+G+F	2.00	2336.50	-1134.25	0.12
JTT+G+F	0.00	2334.50	-1134.25	0.33
WAG+I+G+F	2.78	2337.28	-1134.64	0.08
WAG+G+F	0.78	2335.28	-1134.64	0.23
WAG+I+F	5.17	2339.67	-1136.83	0.03
WAG+F	3.50	2338.00	-1137.00	0.06

Original File	20080821060606361
Phylip File	20080821060606361.phy
Phylogenetic Tree (Newick)	tree20080821060606361.txt
ProtTest Information	ProtTest_20080821060606361.txt
Bootstrap Tree	20080821060606361_phyml_boot_trees.txt
Bootstrap Statistic data	20080821060606361_phyml_boot_stats.txt

The job is computed approximately in 47 minute(s).



Mid-Point Mode Active Mode Length Value Bootstrap Value Drawing Bootstrap

Result

# Demonstration of PALM



**Input Sequence**

Input Type  Sequence in FASTA format  
 Aligned sequence in PHYLIP format

Sequence Type  DNA  Protein

Example File

Sequences

Clear Input

Number of Bootstrap Data Sets

Job Note

Enter Your Email\*

\*: email is optional. Users also can receive the notification immediately via mail when jobs are done.

**Advanced Option**

Number of Substitution Rate Categories

Starting Tree (Newick Format)  Build BioNJ tree  User tree

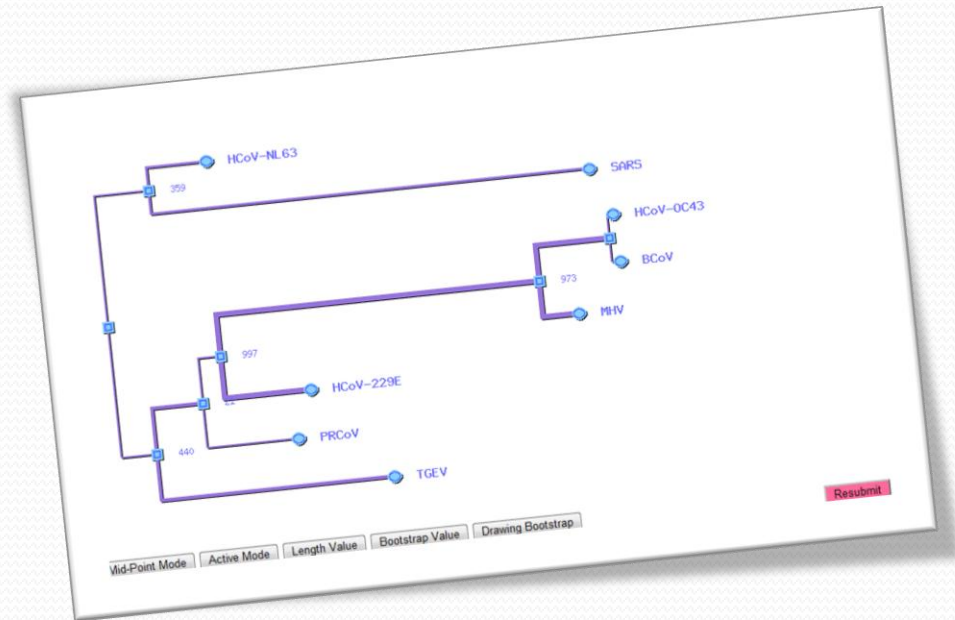
Model Selection Criterion

Optimize Tree Topology and Branch Lengths  Yes  No

Submit Reset

Current Status in Queue: 1 job(s) remaining in the queue.

© 2008 Systems Biology & Network Biology Lab,



Access : <http://palm.iis.sinica.edu.tw>



# Demo Flash of PALM



**Demo** ( Please click the following vedio clip)

**English** 中文

**1. Create a Job**



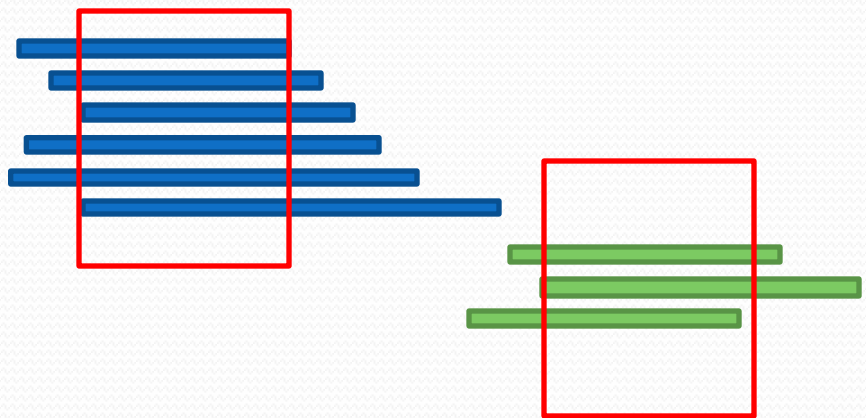
Users can paste their specific and interesting sequences in the below area, or select the Example File with related options.

<http://palm.iis.sinica.edu.tw/demo.html>

# Some Suggestions

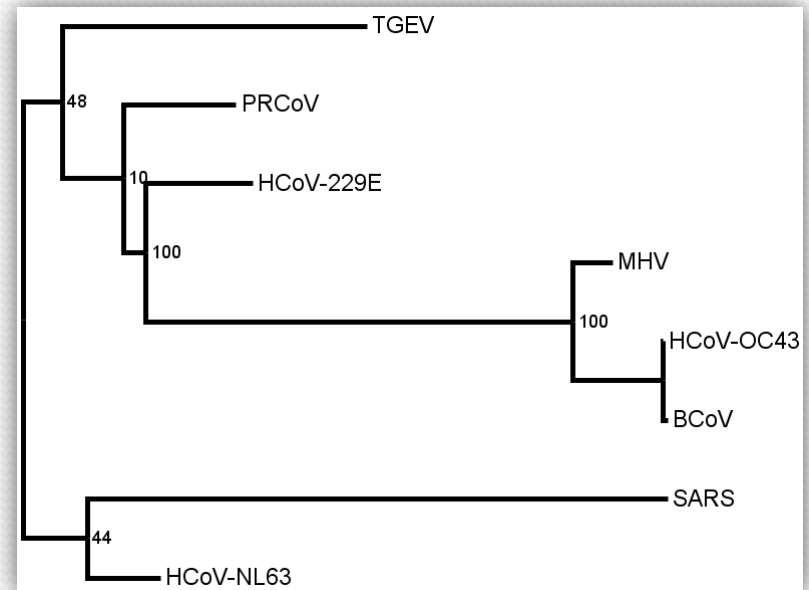
- Please be careful to choose the sequences
- Only well aligned sequences lead to meaningful phylogenetic result.
- RNA editing may introduce bias during analysis. Avoid those regions that may have such conditions.

```
Q56740_BOVIN -----MRDDRYSRNSVFKLIDLDQKCFVADNFKKQIIMSLSKAVVLMGKSTMRRAIKHLEHW--FALE 76
RLAO_HUMAN -----MRDDRYSRNSVFKLIDLDQKCFVADNFKKQIIMSLSKAVVLMGKSTMRRAIKHLEHW--FALE 76
RLAO_MOUSE -----MRDDRYSRNSVFKLIDLDQKCFVADNFKKQIIMSLSKAVVLMGKSTMRRAIKHLEHW--FALE 76
RLAO_RAT -----MRDDRYSRNSVFKLIDLDQKCFVADNFKKQIIMSLSKAVVLMGKSTMRRAIKHLEHW--FALE 76
RLAO_CICK -----MRDDRYSRNSVFKLIDLDQKCFVADNFKKQIIMSLSKAVVLMGKSTMRRAIKHLEHW--FALE 76
RLAO_RABY -----MRDDRYSRNSVFKLIDLDQKCFVADNFKKQIIMSLSKAVVLMGKSTMRRAIKHLEHW--FALE 76
O72DC3_HHAB -----MRDDRYSRNSVFKLIDLDQKCFVADNFKKQIIMSLSKAVVLMGKSTMRRAIKHLEHW--FALE 76
RLAO_ICTFD -----MRDDRYSRNSVFKLIDLDQKCFVADNFKKQIIMSLSKAVVLMGKSTMRRAIKHLEHW--FALE 76
RLAO_DROME -----MYTSEKNSRNSVFKLIDLDQKCFVADNFKKQIIMSLSKAVVLMGKSTMRRAIKHLEHW--KALE 76
RLAO_DICED -----MGRSLSKRKLIFERKRLITTYQKIVARDFVRSLSKIKRSKSLGAVLMGKSTMRRAIKHLEHW--DEED 75
Q54LP0_DICED -----MGRSLSKRKLIFERKRLITTYQKIVARDFVRSLSKIKRSKSLGAVLMGKSTMRRAIKHLEHW--DEED 75
RLAO_PLAFS -----MGLALSKRKLIFERKRLITTYQKIVARDFVRSLSKIKRSKSLGAVLMGKSTMRRAIKHLEHW--DEED 76
RLAO_SHLAC -----MGLAVTTKSKKLVDFEAEDEKTKTKTILKAEIETFAKLEIKKIKKRDVRSVVEDEMTALDHW--DQK 79
RLAO_SULTO -----MGLAVTTKSKKLVDFEAEDEKTKTKTILKAEIETFAKLEIKKIKKRDVRSVVEDEMTALDHW--DQK 80
RLAO_SULSO -----MGRALALQKRVASLSEKLEKELKNSNTLIGMLIETFAKLEIKKIKKRDVRSVVEDEMTALDHW--DQK 80
RLAO_AYRBP -----MREIVVQKTRKLEKDFKRLMREDEKLEKRVVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--DDH 86
RLAO_PYRAB -----MREKICLVYRKYVLAQVYKELVYELGKAVVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--DDH 85
RLAO_METAC -----MREKIHTEKPKQKDEKEMKELQKVKVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--ETD 78
RLAO_METMA -----MREKIHTEKPKQKDEKEMKELQKVKVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--ETD 78
RLAO_AECFU -----MAYVLEK-----DQVYVAVVEKELHSEKRVAVVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--DQV 75
RLAO_METFA -----MREKIHTEKPKQKDEKEMKELQKVKVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--ETD 88
RLAO_METTH -----MREKIHTEKPKQKDEKEMKELQKVKVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--ETD 74
RLAO_METTL -----MREKIHTEKPKQKDEKEMKELQKVKVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--ETD 82
RLAO_METVA -----MREKIHTEKPKQKDEKEMKELQKVKVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--ETD 82
RLAO_METJA -----MREKIHTEKPKQKDEKEMKELQKVKVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--ETD 81
RLAO_PYRAB -----MREKIHTEKPKQKDEKEMKELQKVKVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--ETD 77
RLAO_PYRHO -----MREKIHTEKPKQKDEKEMKELQKVKVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--ETD 77
RLAO_PYRFO -----MREKIHTEKPKQKDEKEMKELQKVKVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--ETD 77
RLAO_PYRKO -----MREKIHTEKPKQKDEKEMKELQKVKVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--ETD 76
RLAO_HALMA -----MREKIHTEKPKQKDEKEMKELQKVKVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--ETD 79
RLAO_HALVO -----MREKIHTEKPKQKDEKEMKELQKVKVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--ETD 79
RLAO_HALSA -----MREKIHTEKPKQKDEKEMKELQKVKVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--ETD 79
RLAO_TYRAC -----MREKIHTEKPKQKDEKEMKELQKVKVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--ETD 72
RLAO_TYRVO -----MREKIHTEKPKQKDEKEMKELQKVKVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--ETD 72
RLAO_PICTO -----MREKIHTEKPKQKDEKEMKELQKVKVVEADITLTPYKAVKSLKAVVLMGKSTMRRAIKHLEHW--ETD 72
truel 1 ..... 10 ..... 20 ..... 30 ..... 40 ..... 50 ..... 60 ..... 70 ..... 80 ..... 90
```



# Bootstrap (BS) Analysis

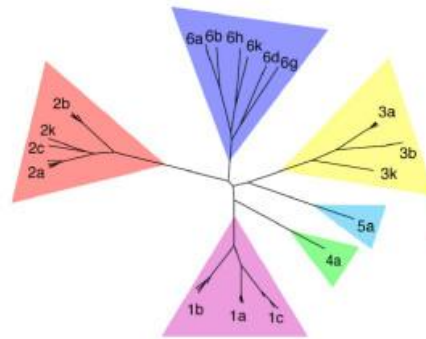
- Bootstrap analysis is the most often used method for statistical evaluation of phylogenies.
- In general:
  - **BS >95%: Often close to 100% confidence in that branch**
  - **BS >75%: Often close to 95% confidence in that branch**
  - BS <75% : Maybe a correct clade due to the original bias cannot be corrected by the re-sampling process.



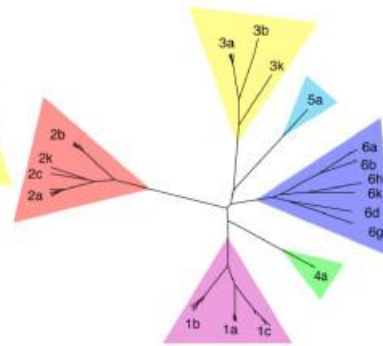
# Input Sequences Make the Tree Different

HIV

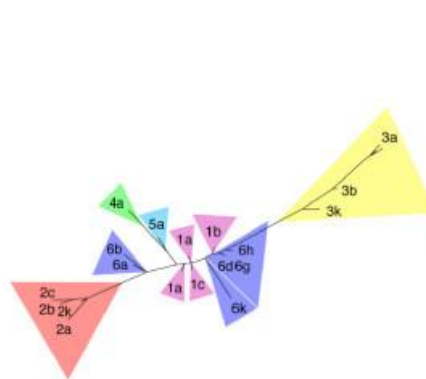
(a) Complete Genome



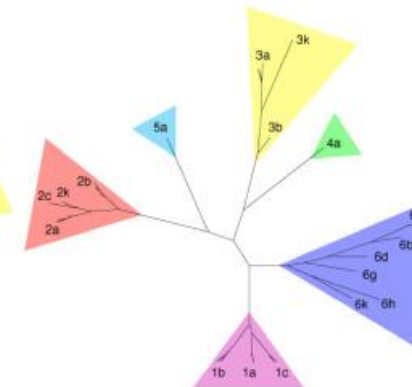
(b) Polyprotein



(c) 5' UTR



(d) Okamoto region of NS5B



Hraber *et al.* *Virology Journal* 2006  
3:103 doi:10.1186/1743-422X-3-103



# *Future Plans for PALM*

- Integrate more substitution models into PALM
- Improve and optimize the performance of whole pipeline
- MrBayes will be implemented into this system for Bayesian inference.
- Parallel computing for large scale, ie. 16S RNA tree (near 2000-4000 sequences) reconstruction in metagenomics for revealing microbial community



# Acknowledgement



國家衛生研究院  
National Health Research Institutes

*Chieh-Hwa Lin*  
*Shu-Jun Hsu*  
*Chia-Ling Chen*  
*Fan-kai Lin*  
*Li-Wei Lai*  
*Chao A. Hsiung*



中央研究院  
資訊科學研究所  
Institute of Information Science  
Academia Sinica

*Sheng-Yao Su*  
*Tengi Huang*  
*Pao-Han Kuo*  
*Chun-Zen Lo*