Manchester Bioinformatics

THE UNIVERSITY of MANCHESTER

# RNA Structure, Evolution and Phylogenetics

**School of Biological Sciences**

Paul Higgs

David Hoyle

Cendrine Hudelot

Daniel Jameson

Stephen Morgan

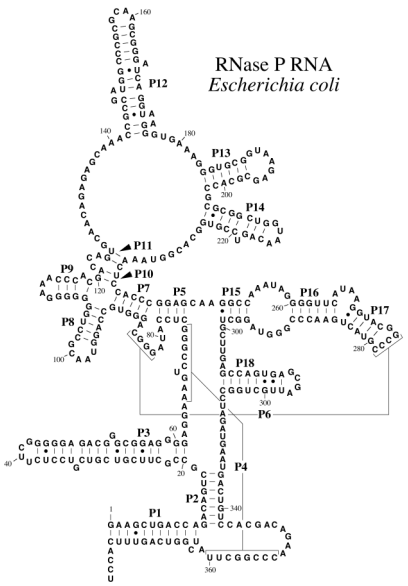
Nick Savill

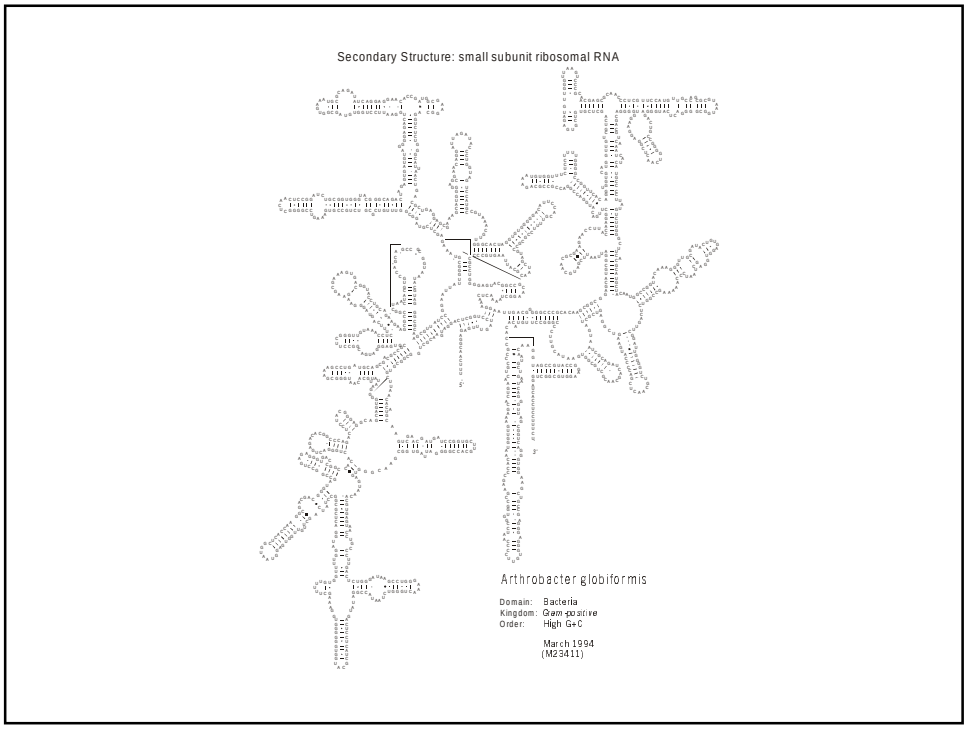
**Dept of Computer Science**

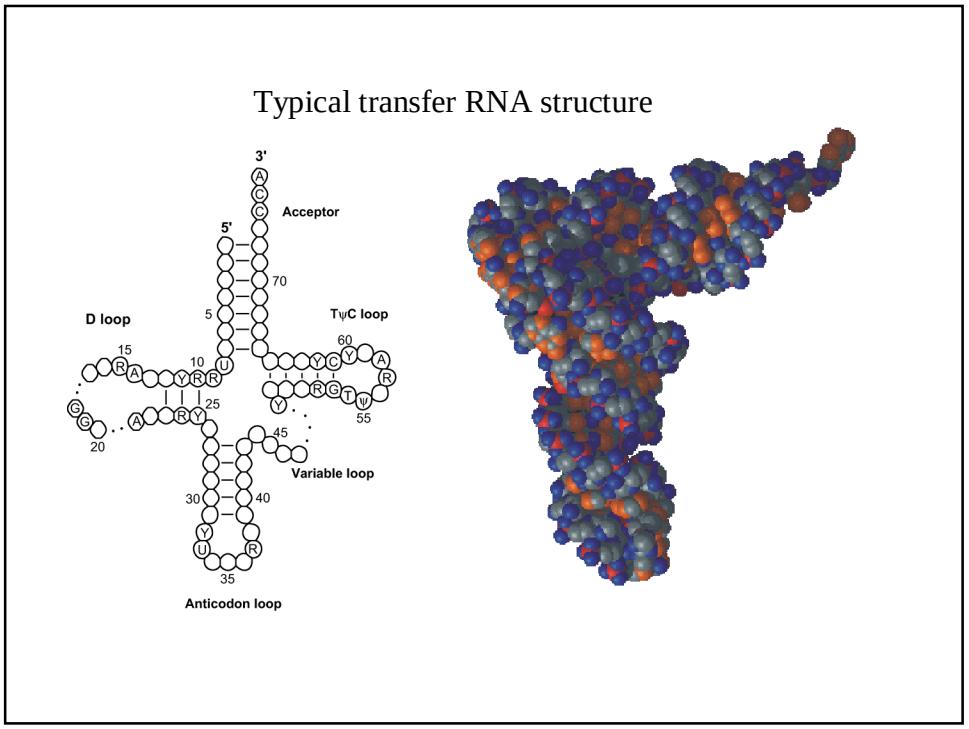
Magnus Rattray

Howsun Jow

**Supported by BBSRC**

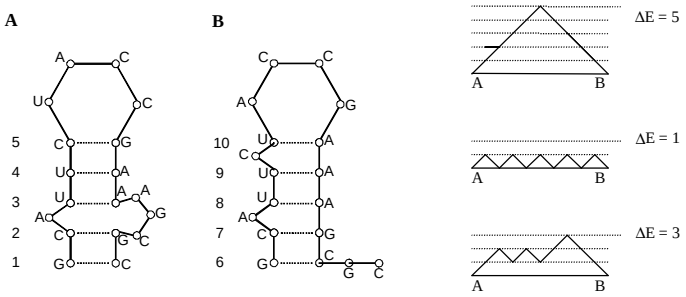



Secondary Structure: small subunit ribosomal RNA
Arthrobacter globiformis
Domain: Bacteria
Kingdom: Gram-positive
Order: High G+C
March 1994
(M23411)

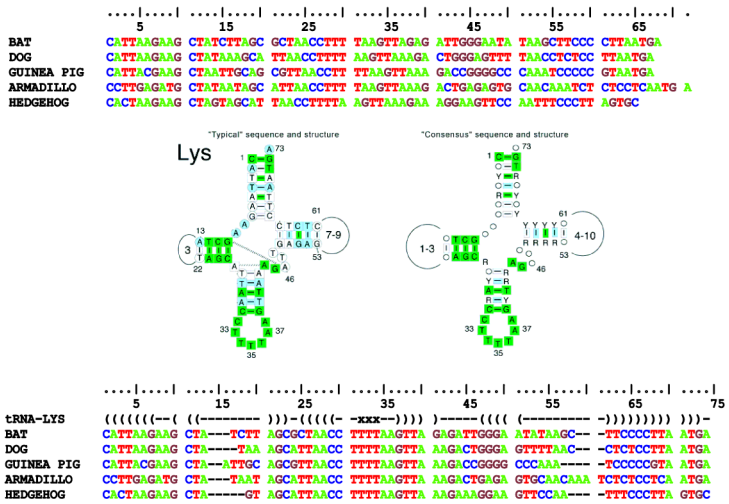

Typical transfer RNA structure
3'
Acceptor
5'
D loop
TψC loop
Variable loop
Anticodon loop

*Trailer No. 2* RNA as a Disordered System

Measure Barriers Between Groundstates




The relevant barrier height controlling kinetics at low temperature is the highest point on the lowest route



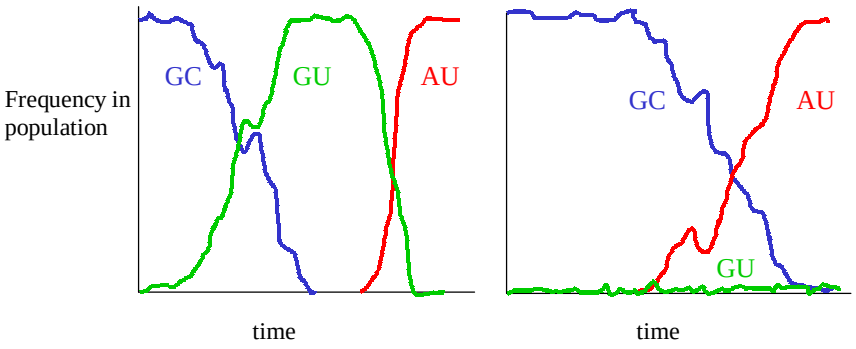
## Compensatory Substitutions

Two sides of the acceptor stem from a tRNA are shown.

Due to structure conservation alignment is possible in widely different species.

| | 1234567 | 7654321 |
|---|---|---|
| | ((((((( | ))))))) |
| Bacillus subtilis | GGCUCGG | CCGAGCC |
| Escherichia coli | GCCCGGA | UCCGGGC |
| Saccharomyces cerevisiae | GCGGAUU | AAUUCGC |
| Drosophila melanogaster | GCCGAAA | UUUCGGC |
| Homo sapiens | GCCGAAA | UUUCGGC |

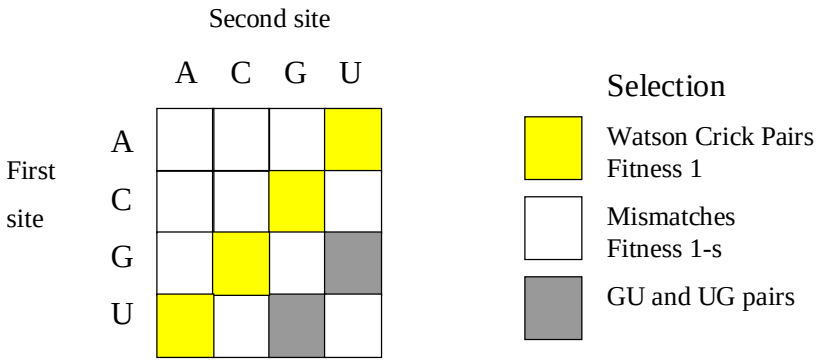
## Two possible mechanisms of compensatory substitutions




Two single substitutions

One double substitution

## Population genetics model

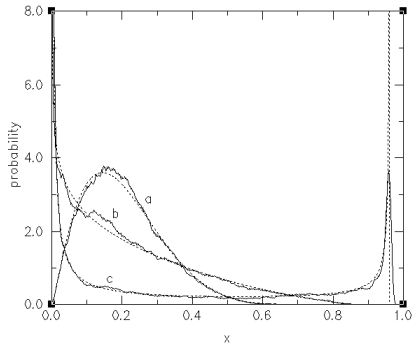



Population size N  Mutation rate u

Solve diffusion equation.

Higgs (1998) *Genetica*

## Probability F(x) of observing a fraction x of AU sequences




F(x) is controlled by a parameter $\alpha = 8Nu^2/s$

| | |
|---|---|
| $\alpha >> 1$ | all 4 WC pairs present simultaneously (curve a) |
| $\alpha << 1$ | population is monomorphic but switches between the 4 WC states (curve c) |

## Analysis of RNA sequence databases

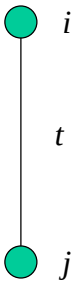
| | | tRNA mitoch. | tRNA general | tRNA archaea | Rnase P | SSU rRNA |
|---|---|---|---|---|---|---|
| G+C average | | 0.339 | 0.532 | 0.636 | 0.594 | 0.545 |
| G+C helical regions | | 0.448 | 0.681 | 0.829 | 0.730 | 0.674 |
| Frequencies | GC | 0.266 | 0.372 | 0.473 | 0.385 | 0.352 |
| | CG | 0.121 | 0.260 | 0.320 | 0.296 | 0.298 |
| | AU | 0.257 | 0.128 | 0.057 | 0.117 | 0.122 |
| | UA | 0.233 | 0.142 | 0.077 | 0.104 | 0.173 |
| | GU | 0.046 | 0.043 | 0.031 | 0.050 | 0.020 |
| | UG | 0.030 | 0.025 | 0.020 | 0.022 | 0.021 |
| | MM | 0.046 | 0.030 | 0.022 | 0.026 | 0.014 |
| Number of sequences | | 884 | 754 | 64 | 84 | 455 |
| Number of pairs | | 21 | 21 | 21 | 80 | 296 |

Selection for thermodynamically stable structures

Higgs (2000) *Quart. Rev. Biophysics*

## A Substitution Rate Model for RNA Stems for use with Molecular Phylogenetics

$P_{ij}(t)$ = probability of being in state $j$ at time $t$ given that ancestor was in state $i$ at time 0.




$$\frac{dP_{ij}}{dt} = \sum_k P_{ik} r_{kj}$$

States label pairs (AU, GC etc.) not single bases.

Model 7A is a General Reversible 7-state Model

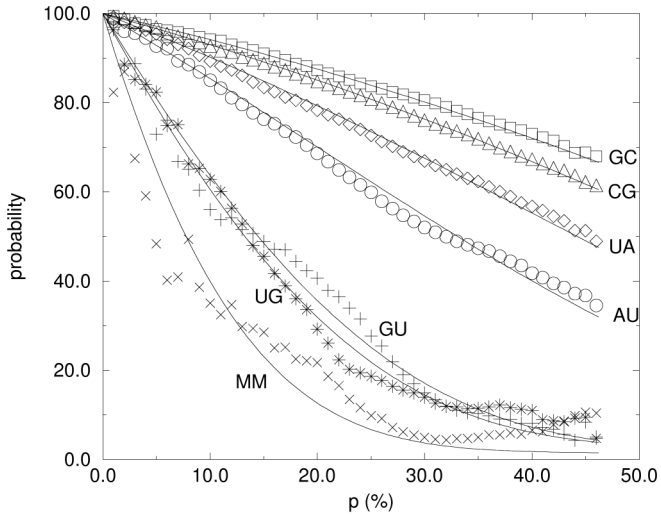
7 frequencies $\pi_i$ + 21 rate parameters $\alpha_{ij}$

- 2 constraints = 26 free parameters

| | | 1 *AU* | 2 *GU* | 3 *GC* | 4 *UA* | 5 *UG* | 6 *CG* | 7 *MM* |
|---|---|---|---|---|---|---|---|---|
| 1 | *AU* | * | $\pi_2\alpha_{12}$ | $\pi_3\alpha_{13}$ | $\pi_4\alpha_{14}$ | $\pi_5\alpha_{15}$ | $\pi_6\alpha_{16}$ | $\pi_7\alpha_{17}$ |
| 2 | *GU* | $\pi_1\alpha_{12}$ | * | $\pi_3\alpha_{23}$ | $\pi_4\alpha_{24}$ | $\pi_5\alpha_{25}$ | $\pi_6\alpha_{26}$ | $\pi_7\alpha_{27}$ |
| 3 | *GC* | $\pi_1\alpha_{13}$ | $\pi_2\alpha_{23}$ | * | $\pi_4\alpha_{34}$ | $\pi_5\alpha_{35}$ | $\pi_6\alpha_{36}$ | $\pi_7\alpha_{37}$ |
| 4 | *UA* | $\pi_1\alpha_{14}$ | $\pi_2\alpha_{24}$ | $\pi_3\alpha_{34}$ | * | $\pi_5\alpha_{45}$ | $\pi_6\alpha_{46}$ | $\pi_7\alpha_{47}$ |
| 5 | *UG* | $\pi_1\alpha_{15}$ | $\pi_2\alpha_{25}$ | $\pi_3\alpha_{35}$ | $\pi_4\alpha_{45}$ | * | $\pi_6\alpha_{56}$ | $\pi_7\alpha_{57}$ |
| 6 | *CG* | $\pi_1\alpha_{16}$ | $\pi_2\alpha_{26}$ | $\pi_3\alpha_{36}$ | $\pi_4\alpha_{46}$ | $\pi_5\alpha_{56}$ | * | $\pi_7\alpha_{67}$ |
| 7 | *MM* | $\pi_1\alpha_{17}$ | $\pi_2\alpha_{27}$ | $\pi_3\alpha_{37}$ | $\pi_4\alpha_{47}$ | $\pi_5\alpha_{57}$ | $\pi_6\alpha_{67}$ | * |

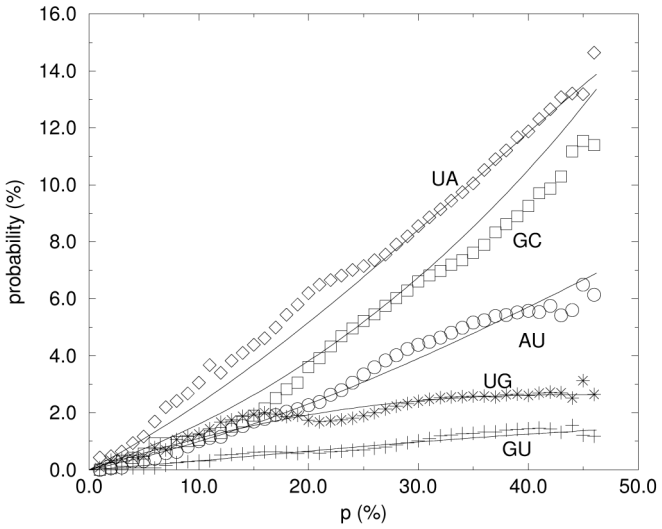
Probability of remaining in same state $P_{ii}$

SSU rRNA sequences from Eubacteria

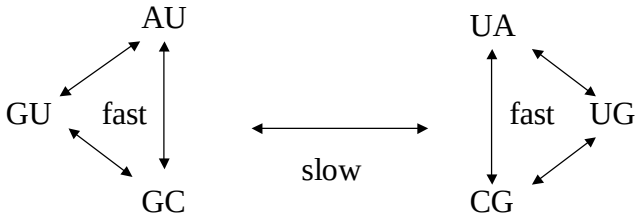


Probability $P_{ij}$ of changes from CG to other pairs

SSU rRNA from Eubacteria




What is going on?




Selection against GU and UG is weaker than against mismatches.

Double transitions are faster than double transversions.

Double transitions are faster than single transitions to GU and UG states. This is explained by the theory of compensatory substitutions.

## Analysis of RNA Substitution Rates

| | | tRNA mitoch. | tRNA general | tRNA archaea | Rnase P | SSU rRNA |
|---|---|---|---|---|---|---|
| Mutabilities | GC | 0.67 | 0.49 | 0.45 | 0.65 | 0.55 |
| | CG | 0.84 | 0.83 | 0.89 | 0.60 | 0.66 |
| | AU | 0.86 | 1.46 | 4.01 | 1.46 | 1.40 |
| | UA | 0.77 | 1.24 | 1.78 | 1.09 | 0.93 |
| | GU | 2.44 | 1.96 | 1.85 | 1.72 | 3.92 |
| | UG | 3.32 | 5.01 | 3.00 | 2.84 | 4.36 |
| | MM | 2.32 | 0.99 | 0.86 | 5.24 | 7.84 |
| Double transitions / Double transversions | | 4.7 | 1.7 | 2.3 | 3.1 | 2.1 |
| Double transitions / Transitions to GU or UG | | 1.6 | 2.0 | 8.9 | 3.6 | 2.8 |

Thermodynamic properties influence Evolutionary properties

## Model Selection

Savill, Hoyle & Higgs (2001) *Genetics*

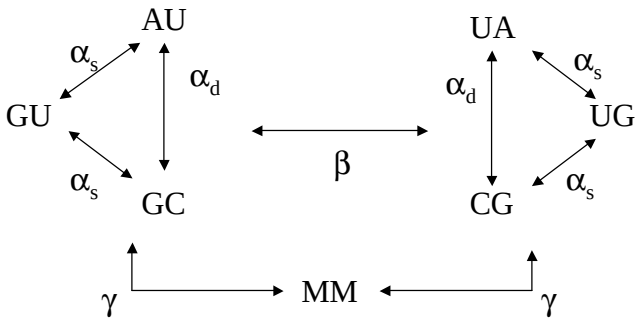
How complex must a model be to explain the data?

How many parameters are justified?

Many other rate models have been proposed:

- Schöniger & von Haeseler
- Muse
- Rzhetsky
- Tillier & Collins

Use likelihood-based statistical tests to distinguish between models.




## Model 7D

Also called OTRNA by Tillier & Collins (1998)

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| | | *AU* | *GU* | *GC* | *UA* | *UG* | *CG* | *MM* |
| 1 | *AU* | * | $\pi_2\alpha_s$ | $\pi_3\alpha_d$ | $\pi_4\beta$ | $\pi_5\beta$ | $\pi_6\beta$ | $\pi_7\gamma$ |
| 2 | *GU* | $\pi_1\alpha_s$ | * | $\pi_3\alpha_s$ | $\pi_4\beta$ | $\pi_5\beta$ | $\pi_6\beta$ | $\pi_7\gamma$ |
| 3 | *GC* | $\pi_1\alpha_d$ | $\pi_2\alpha_s$ | * | $\pi_4\beta$ | $\pi_5\beta$ | $\pi_6\beta$ | $\pi_7\gamma$ |
| 4 | *UA* | $\pi_1\beta$ | $\pi_2\beta$ | $\pi_3\beta$ | * | $\pi_5\alpha_s$ | $\pi_6\alpha_d$ | $\pi_7\gamma$ |
| 5 | *UG* | $\pi_1\beta$ | $\pi_2\beta$ | $\pi_3\beta$ | $\pi_4\alpha_s$ | * | $\pi_6\alpha_s$ | $\pi_7\gamma$ |
| 6 | *CG* | $\pi_1\beta$ | $\pi_2\beta$ | $\pi_3\beta$ | $\pi_4\alpha_d$ | $\pi_5\alpha_s$ | * | $\pi_7\gamma$ |
| 7 | *MM* | $\pi_1\gamma$ | $\pi_2\gamma$ | $\pi_3\gamma$ | $\pi_4\gamma$ | $\pi_5\gamma$ | $\pi_6\gamma$ | * |

## Model 16D and related models

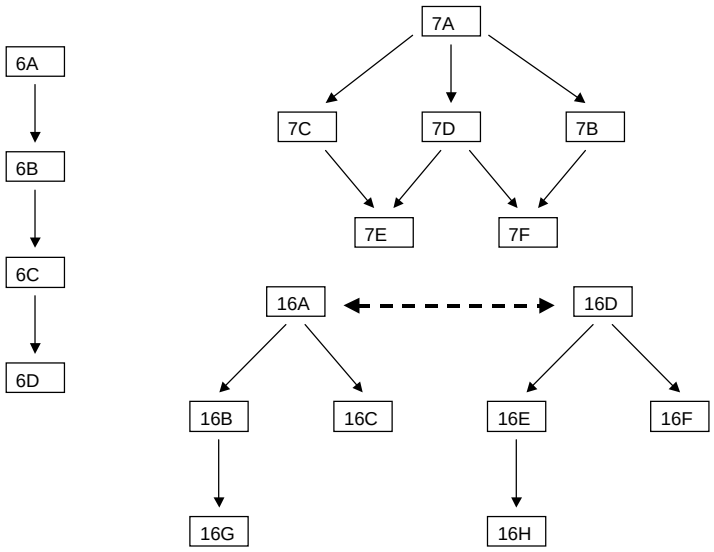
Proposed by Muse (1995)

Considers all 16 states.

Only single substitutions allowed. Distinguishes between paired and mismatch states.

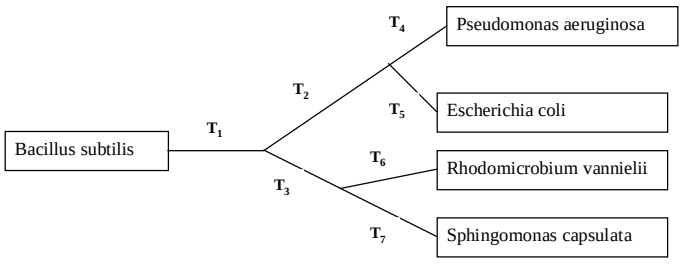
| | *AU* | *GU* | *GC* | *UA* | *UG* | *CG* | *AA* | *AC* | *AG* | *CA* | *CC* | *CU* | *GA* | *GG* | *UC* | *UU* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *AU* | * | $\alpha\pi_G\phi/\lambda$ | 0 | 0 | 0 | 0 | $\beta\pi_A/\lambda$ | $\alpha\pi_C/\lambda$ | $\beta\pi_G/\lambda$ | 0 | 0 | $\beta\pi_C/\lambda$ | 0 | 0 | 0 | $\beta\pi_U/\lambda$ |
| *GU* | $\alpha\pi\lambda/\varphi$ | * | $\alpha\pi_C\lambda/\phi$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | $\beta\pi_C/\phi$ | $\beta\pi_A/\phi$ | $\beta\pi_G/\phi$ | 0 | $\beta\pi_U/\phi$ |
| *GC* | 0 | $\alpha\pi_U\phi/\lambda$ | * | 0 | 0 | 0 | 0 | $\alpha\pi_A/\lambda$ | 0 | 0 | $\beta\pi_C/\lambda$ | 0 | $\beta\pi_A/\lambda$ | $\beta\pi_G/\lambda$ | $\beta\pi_U/\lambda$ | 0 |
| *UA* | 0 | 0 | 0 | * | $\alpha\pi_G\phi/\lambda$ | 0 | $\beta\pi_A/\lambda$ | 0 | 0 | $\alpha\pi_C/\lambda$ | 0 | 0 | $\beta\pi_G/\lambda$ | 0 | $\beta\pi_C/\lambda$ | $\beta\pi_U/\lambda$ |
| *UG* | 0 | 0 | 0 | $\alpha\pi_A\lambda/\varphi$ | * | $\alpha\pi_C\lambda/\phi$ | 0 | 0 | $\beta\pi_A/\phi$ | 0 | 0 | 0 | 0 | $\beta\pi_G/\phi$ | $\beta\pi_C/\phi$ | $\beta\pi_U/\phi$ |
| *CG* | 0 | 0 | 0 | 0 | $\alpha\pi_U\phi/\lambda$ | * | 0 | 0 | $\beta\pi_A/\lambda$ | $\alpha\pi_A/\lambda$ | $\beta\pi_C/\lambda$ | $\beta\pi_U/\lambda$ | 0 | $\beta\pi_G/\lambda$ | 0 | 0 |
| *AA* | $\beta\pi_U\lambda$ | 0 | 0 | $\beta\pi_U\lambda$ | 0 | 0 | * | $\beta\pi_C$ | $\alpha\pi_G$ | $\beta\pi_C$ | 0 | 0 | $\alpha\pi_G$ | 0 | 0 | 0 |
| *AC* | $\alpha\pi_U\lambda$ | 0 | $\alpha\pi_G\lambda$ | 0 | 0 | 0 | $\beta\pi_A$ | * | $\beta\pi_G$ | 0 | $\beta\pi_C$ | 0 | 0 | 0 | $\beta\pi_C$ | 0 |
| *AG* | $\beta\pi_U\lambda$ | 0 | 0 | 0 | $\beta\pi_U\phi$ | $\beta\pi_C\lambda$ | $\alpha\pi_A$ | $\beta\pi_C$ | * | 0 | 0 | 0 | 0 | $\alpha\pi_G$ | 0 | 0 |
| *CA* | 0 | 0 | 0 | $\alpha\pi_U\lambda$ | 0 | $\alpha\pi_G\lambda$ | $\beta\pi_A$ | 0 | 0 | * | $\beta\pi_C$ | $\beta\pi_U$ | $\beta\pi_G$ | 0 | 0 | 0 |
| *CC* | 0 | 0 | $\beta\pi_G\lambda$ | 0 | 0 | $\beta\pi_G\lambda$ | 0 | $\beta\pi_A$ | 0 | $\beta\pi_A$ | * | $\alpha\pi_U$ | 0 | 0 | $\alpha\pi_U$ | 0 |
| *CU* | $\beta\pi\lambda$ | $\beta\pi_G\phi$ | 0 | 0 | 0 | $\beta\pi_G\lambda$ | 0 | 0 | 0 | $\beta\pi_A$ | $\alpha\pi_C$ | * | 0 | 0 | 0 | $\alpha\pi_U$ |
| *GA* | 0 | $\beta\pi_U\phi$ | $\beta\pi_C\lambda$ | $\beta\pi_U\lambda$ | 0 | 0 | $\alpha\pi_A$ | 0 | 0 | $\beta\pi_C$ | 0 | 0 | * | $\alpha\pi_G$ | 0 | 0 |
| *GG* | 0 | $\beta\pi_U\phi$ | $\beta\pi_C\lambda$ | 0 | $\beta\pi_U\phi$ | $\beta\pi_C\lambda$ | 0 | 0 | $\alpha\pi_A$ | 0 | 0 | 0 | $\alpha\pi_A$ | * | 0 | 0 |
| *UC* | 0 | 0 | $\beta\pi_G\lambda$ | $\beta\pi\lambda\lambda$ | $\beta\pi_G\phi$ | 0 | 0 | $\beta\pi_A$ | 0 | 0 | $\alpha\pi_C$ | 0 | 0 | 0 | * | $\alpha\pi_U$ |
| *UU* | $\beta\pi_G\lambda$ | $\beta\pi_G\lambda$ | 0 | $\beta\pi_G\lambda$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | $\alpha\pi_G$ | 0 | 0 | $\alpha\pi_G$ | * |

Families of Nested Models

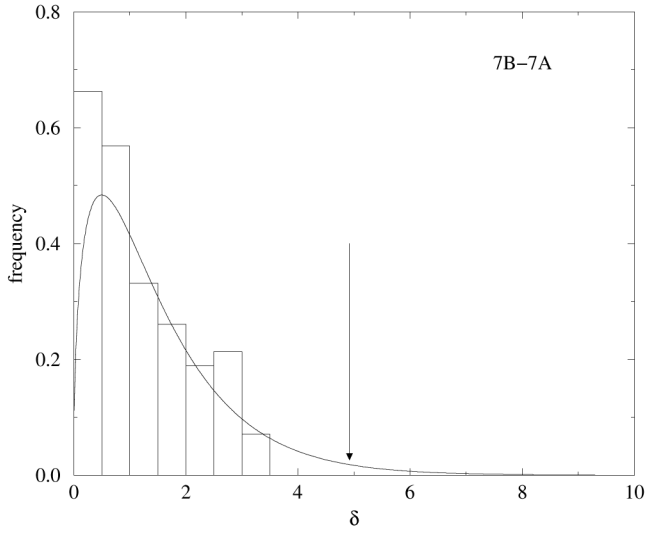



Test models using “known” phylogeny of five bacteria.

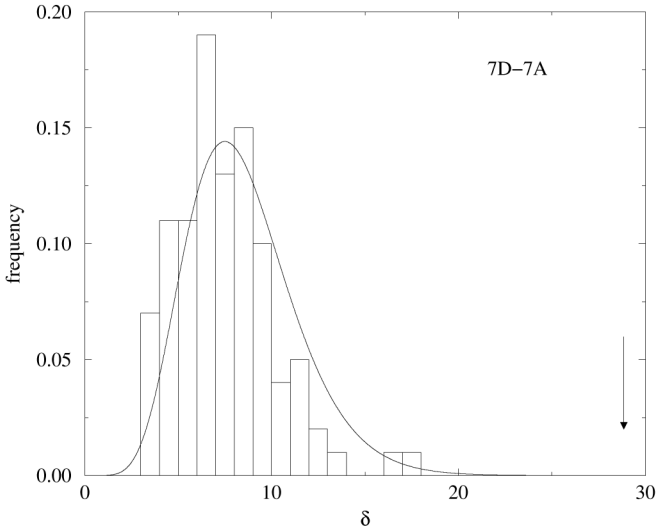
Small sub-unit ribosomal RNA sequences.



Likelihood Ratio Test shows that rates of double substitutions are non-zero.

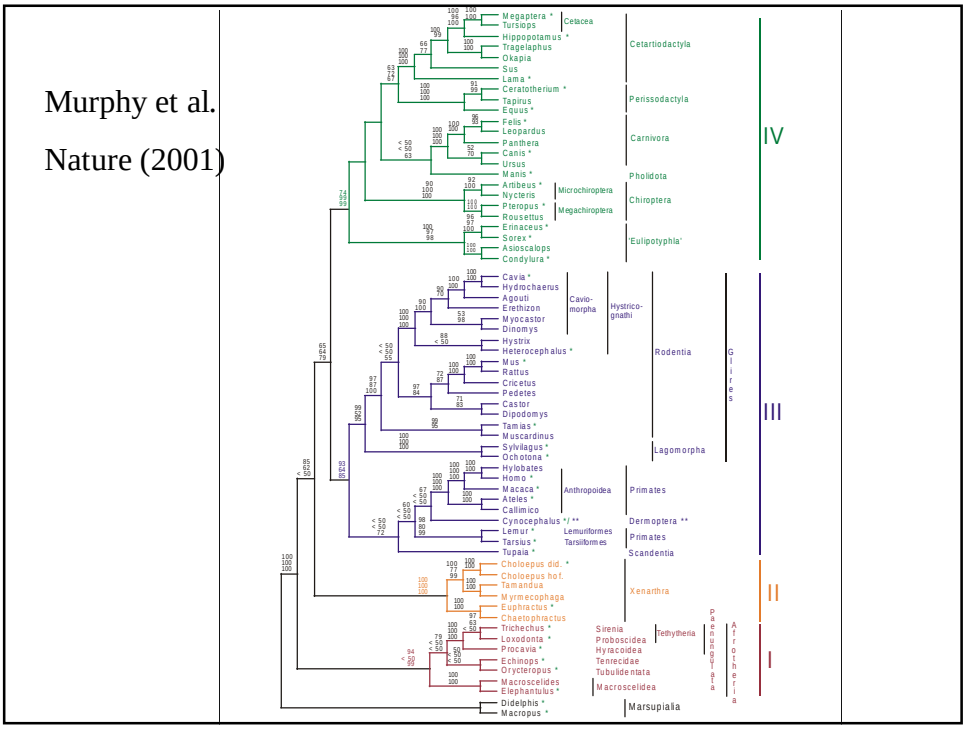



Likelihood Ratio Test shows General Reversible Model 7A is preferable to 7D.

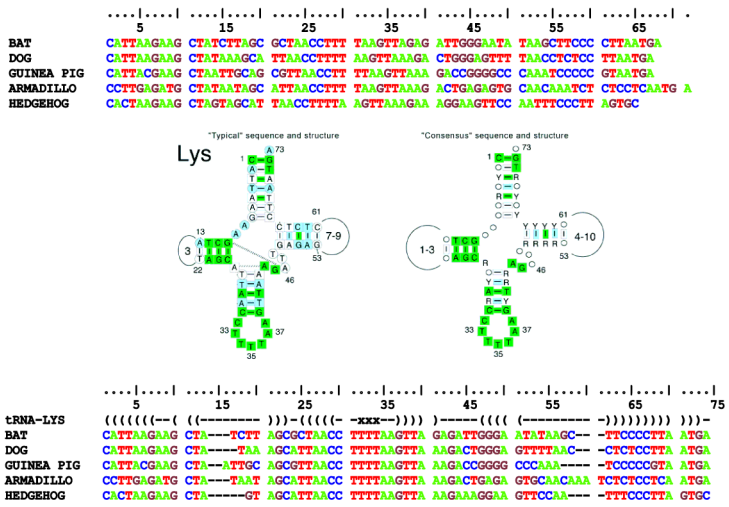


Murphy et al.

Nature (2001)




Afrotheria / Laurasiatheria




## RNA Phylogeny Package

Soon to be released …

Uses a range of evolutionary models designed specifically for RNA paired regions (as well as standard single site models).

Secondary structure is specified in the alignment as bracket notation.

Exhaustive Maximum Likelihood search for small numbers of species or small numbers of clusters.

Statistical tests to distibuish between proposed trees: Kishino-Hasegawa (important to consider correlation between sites)

Markov Chain Monte Carlo methods for tree searching.

Sub-trees in the Laurasiatheria Group

Carnivora

(((Harbour Seal, Gray Seal), Dog), Cat)

Perissodactyla

((Donkey, Horse),(Indian Rhino, White Rhino))

Cetartiodactyla

((((Blue Whale, Sperm Whale), Hippo),(Cow, Sheep)), Pig)

Chiroptera

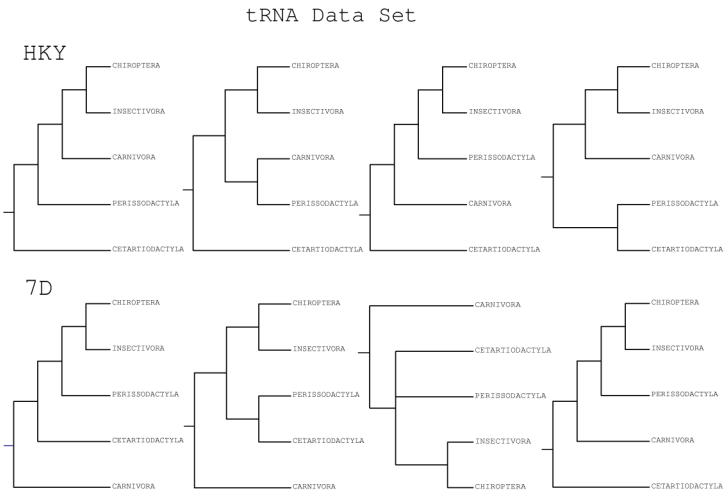
((Long-tailed Bat, Fruit-eating Bat),(Flying Fox, Flying Fox))

Insectivora

(Hedgehog, Mole)

Outgroup: Rabbit

## Maximum Likelihood Cluster Arrangements



## Preliminary Results with MCMC

43 mammals with completely sequenced mitochondrial genomes

SSU + LSU + 14 tRNAs on the same strand

Study paired regions using the 7D model

- Well-defined groups: 100% support
  - e.g. Primates, Marsupials, Rodents, Carnivores.
- Four large-scale clades well-defined
  - Afrotheria (100%), Laurasiatheria (97%), Primates+Glires (86%), Xenarthra (100%)
- Insectivores+Bats (87%)
- A few wandering species: e.g. Pig

Manchester Bioinformatics

THE UNIVERSITY of MANCHESTER

Thoughts for the day…

Thermodynamic parameters affect evolutionary parameters.

Rate models for phylogenies need to be tailored to the sequences studied.