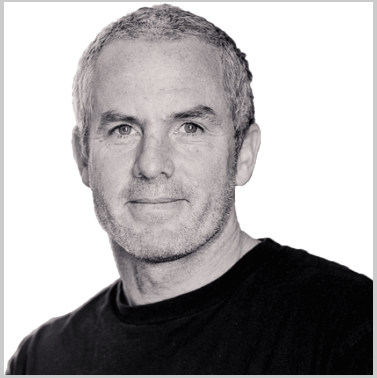
PartitionFinder

Paul B. Frandsen
Office of Research Information Services



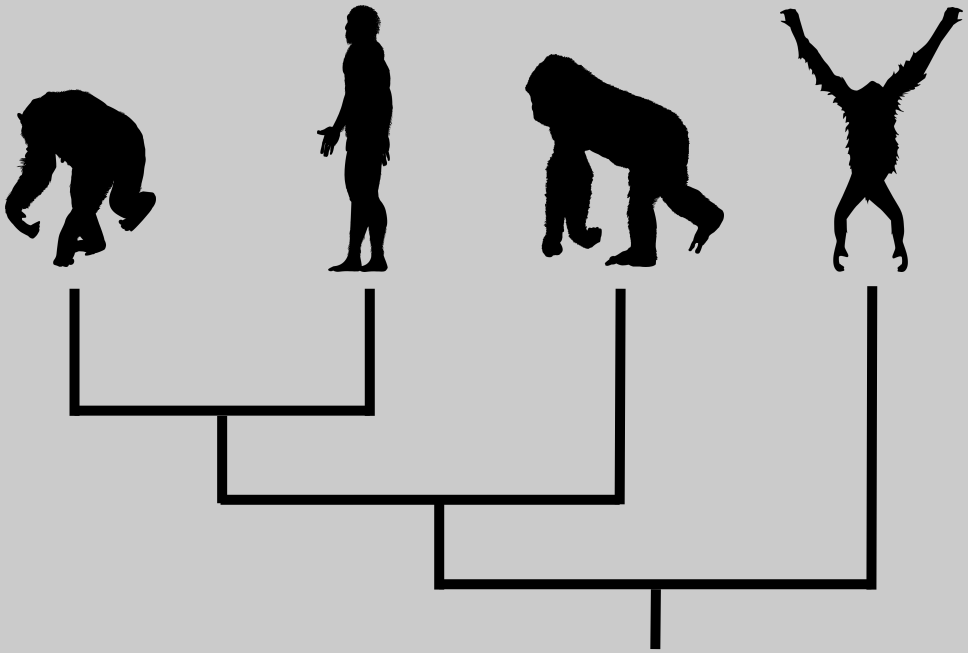
With Rob Lanfear and Brett Calcott


Google
SUMMER
OF
CODE
2013

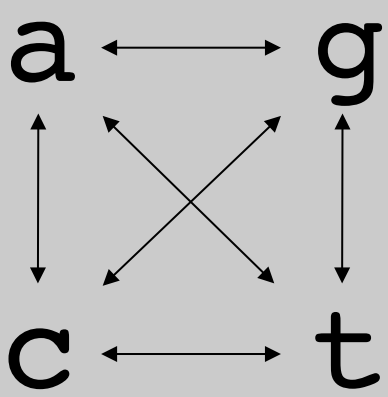

Google

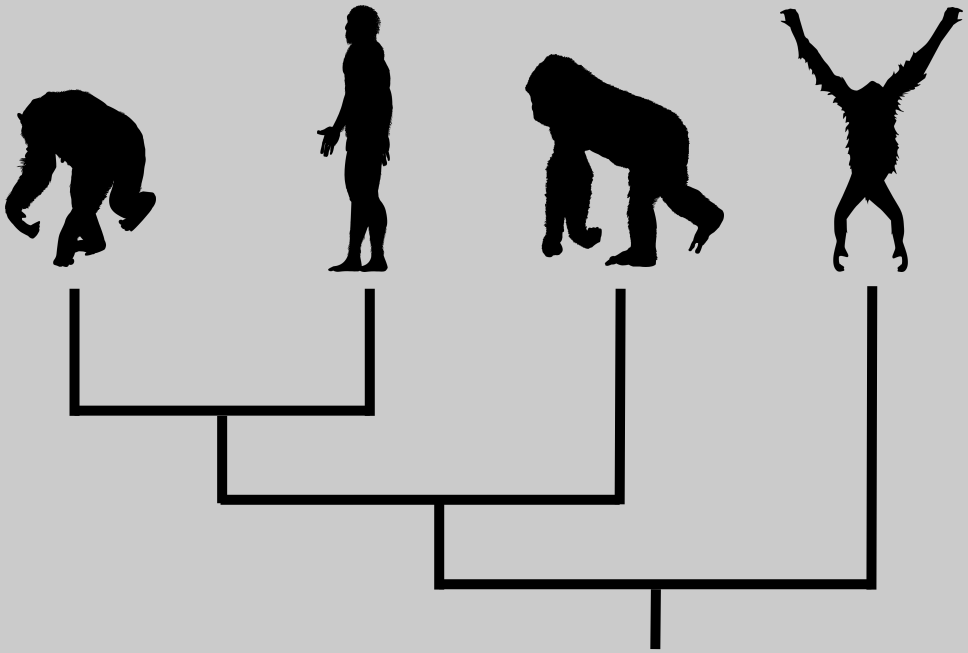
a c t g c c t g a c t g g c t g a c
g c t g a c g t c a t g g c t g a c
a g t g a c g t a g t g a g t g c g
a g c g c c t g c g c g a g c g c g

```
a c t g c c t g a c t g g c t g a c
g c t g a c g t c a t g g c t g a c
a g t g a c g t a g t g a g t g c g
a g c g c c t g c g c g a g c g c g
```



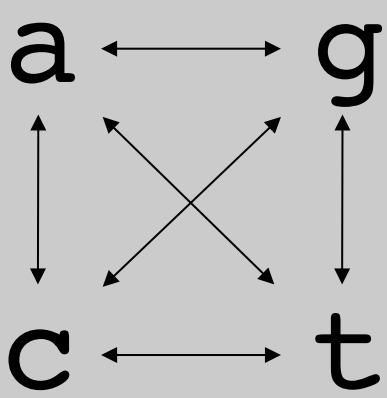
```
a c t g c c t g a c t g g c t g a c
g c t g a c g t c a t g g c t g a c
a g t g a c g t a g t g a g t g c g
a g c g c c t g c g c g a g c g c g
```

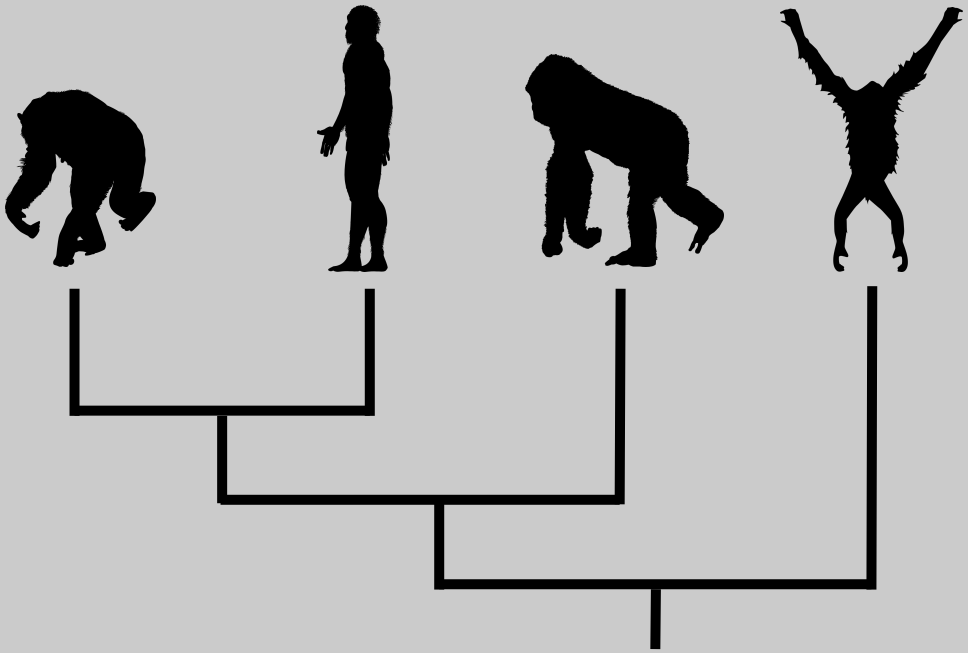




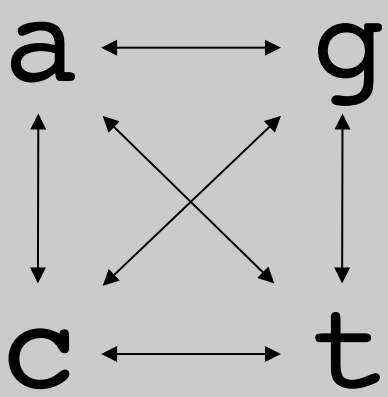
```
a c t g c c t g a c t g g c t g a c
g c t g a c g t c a t g g c t g a c
a g t g a c g t a g t g a g t g c g
a g c g c c t g c g c g a g c g c g
```

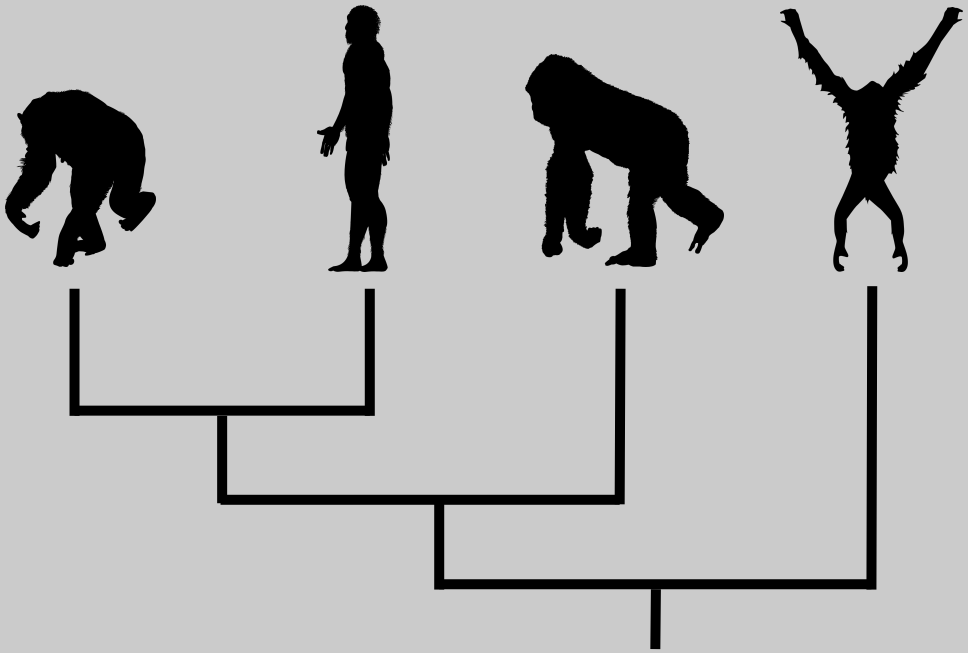
1 model





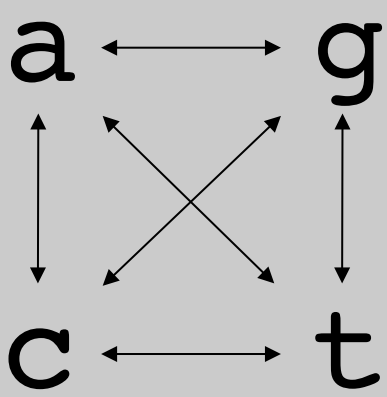
```
a c t g c c t g a c t g g c t g a c
g c t g a c g t c a t g g c t g a c
a g t g a c g t a g t g a g t g c g
a g c g c c t g c g c g a g c g c g
```

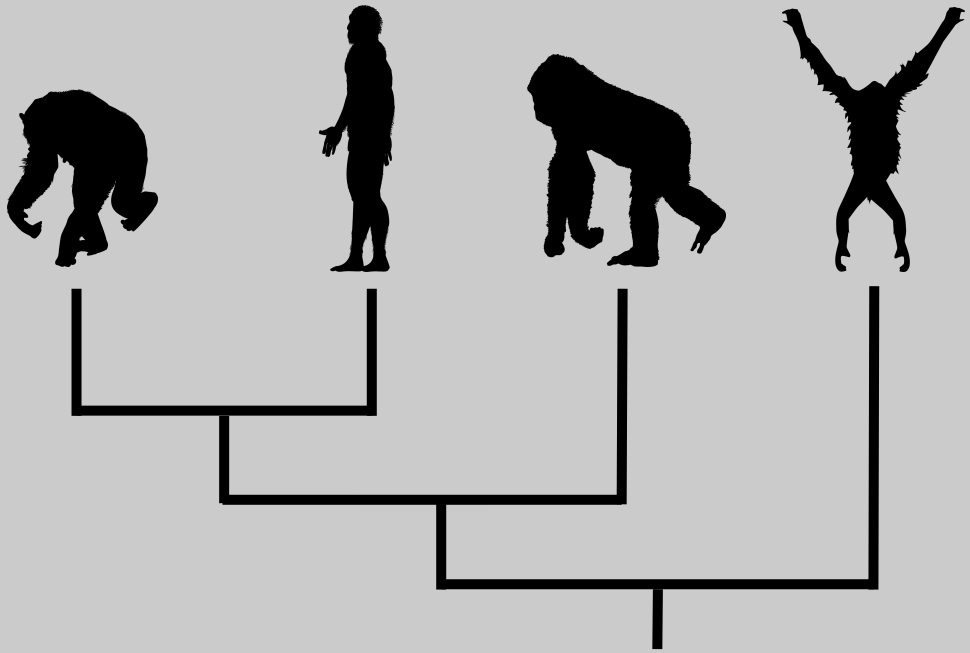


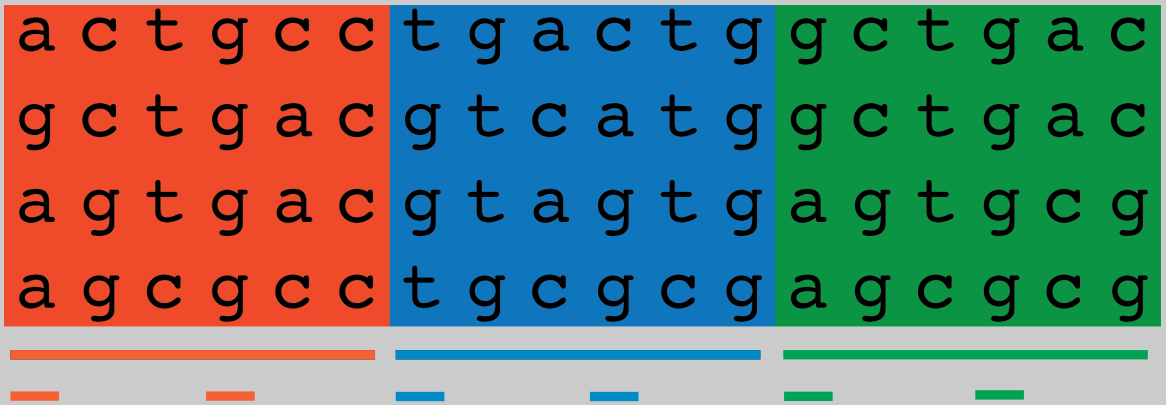


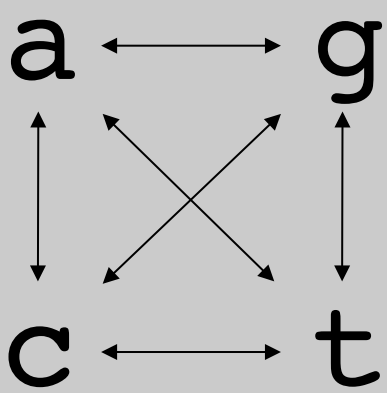
a c t g c c	t g a c t g	g c t g a c
g c t g a c	g t c a t g	g c t g a c
a g t g a c	g t a g t g	a g t g c g
a g c g c c	t g c g c g	a g c g c g

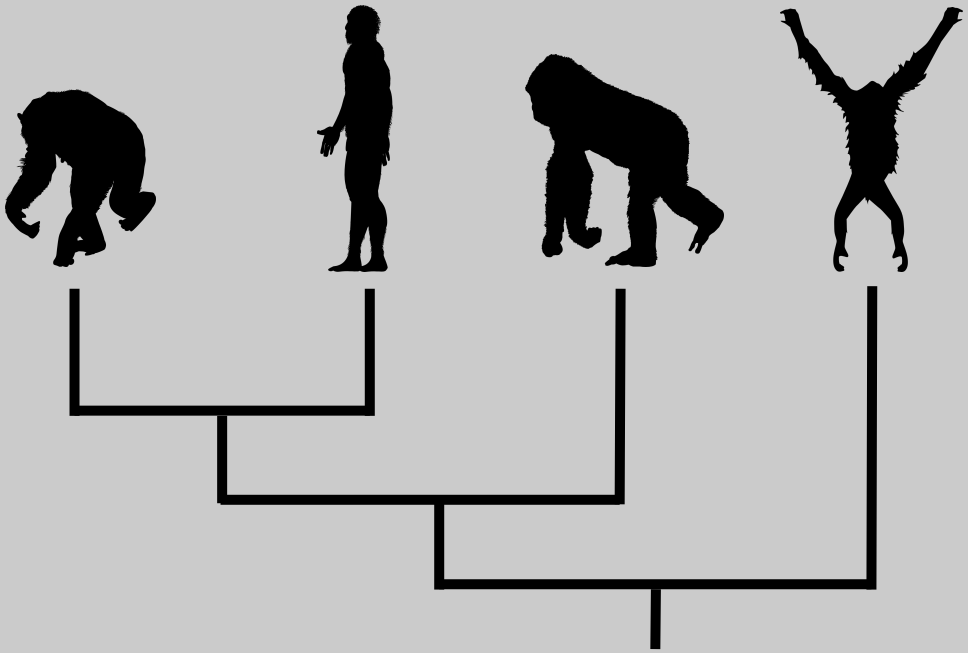
3 models

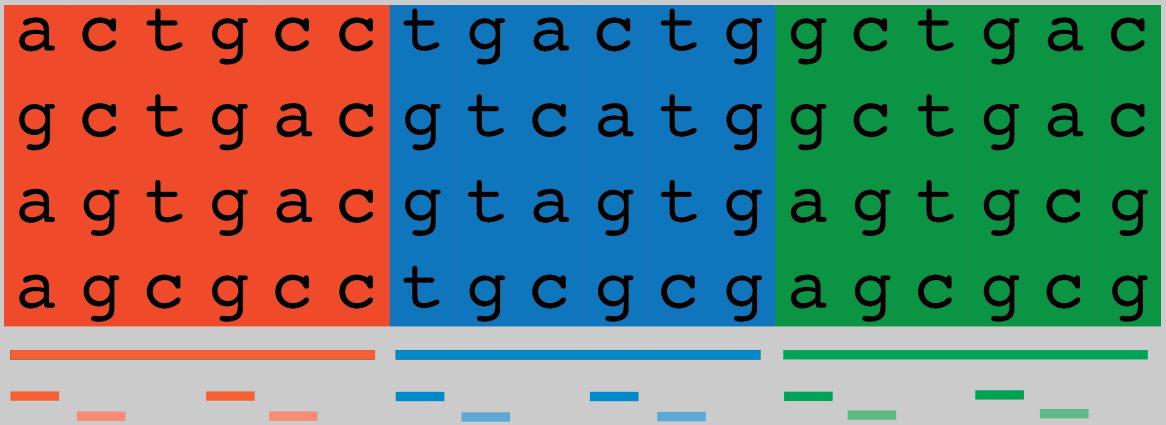


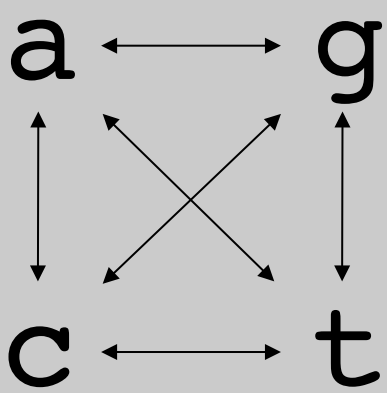


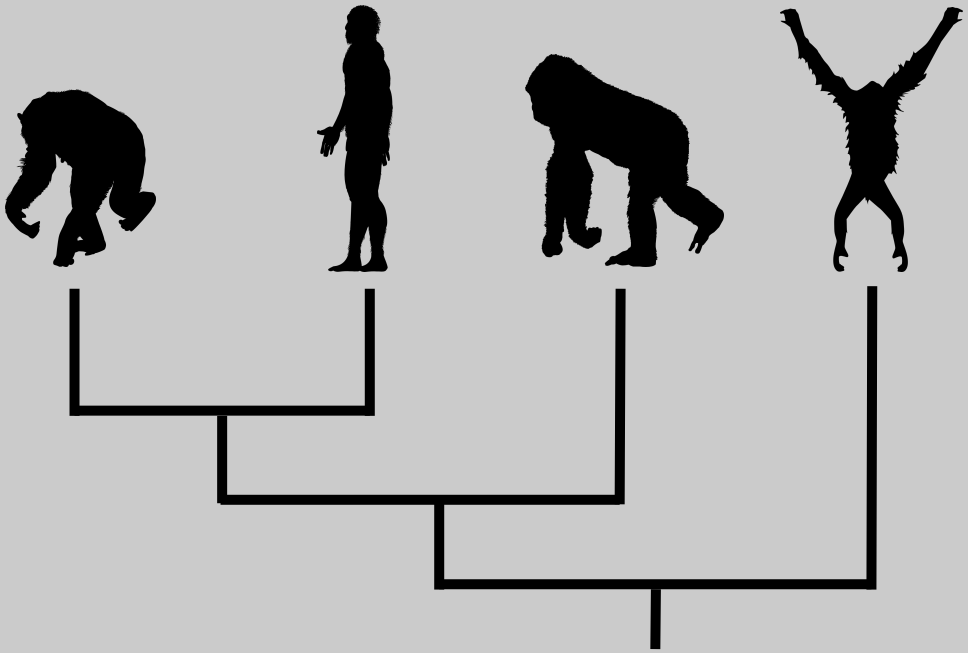

a c t g c c t g a c t g g c t g a c
g c t g a c g t c a t g g c t g a c
a g t g a c g t a g t g a g t g c g
a g c g c c t g c g c g a g c g c g

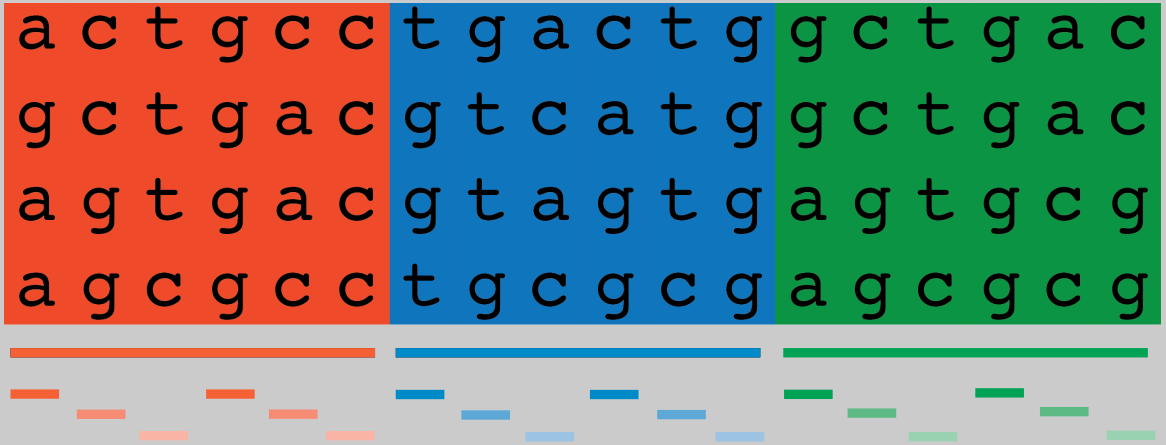

a
g
c
t

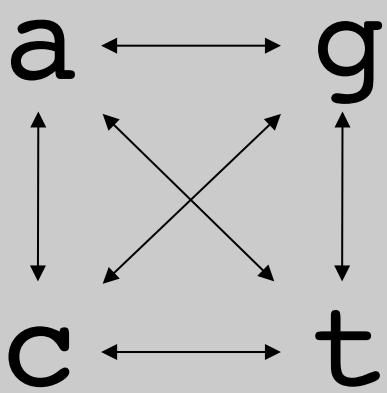


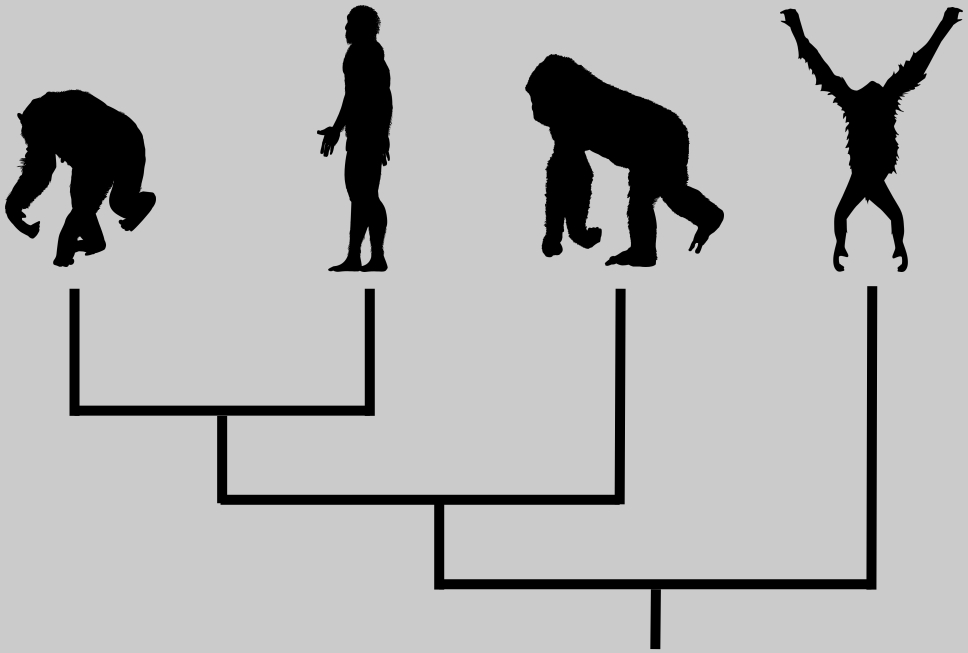

a c t g c c t g a c t g g c t g a c
g c t g a c g t c a t g g c t g a c
a g t g a c g t a g t g a g t g c g
a g c g c c t g c g c g a g c g c g

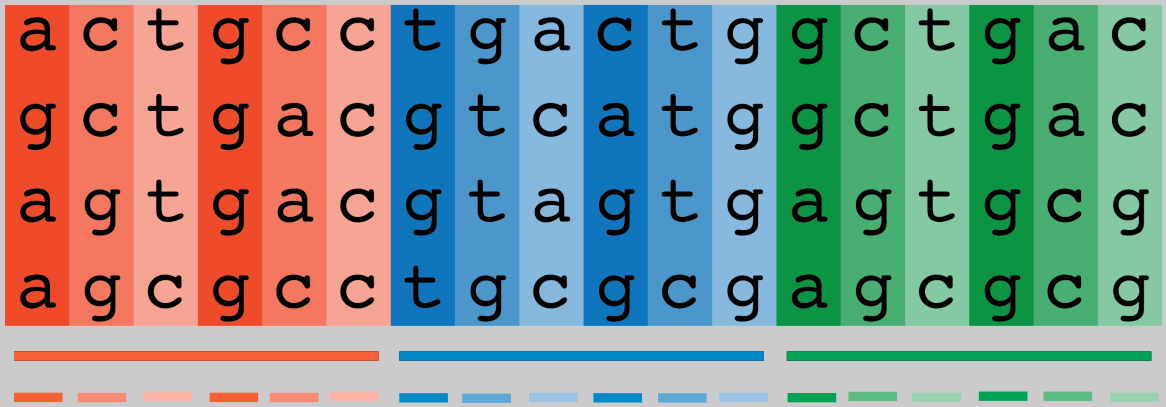

a
g
c
t



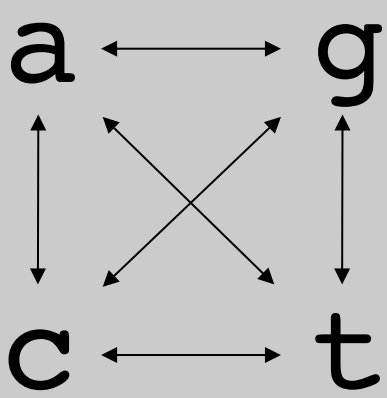

a c t g c c t g a c t g g c t g a c
g c t g a c g t c a t g g c t g a c
a g t g a c g t a g t g a g t g c g
a g c g c c t g c g c g a g c g c g

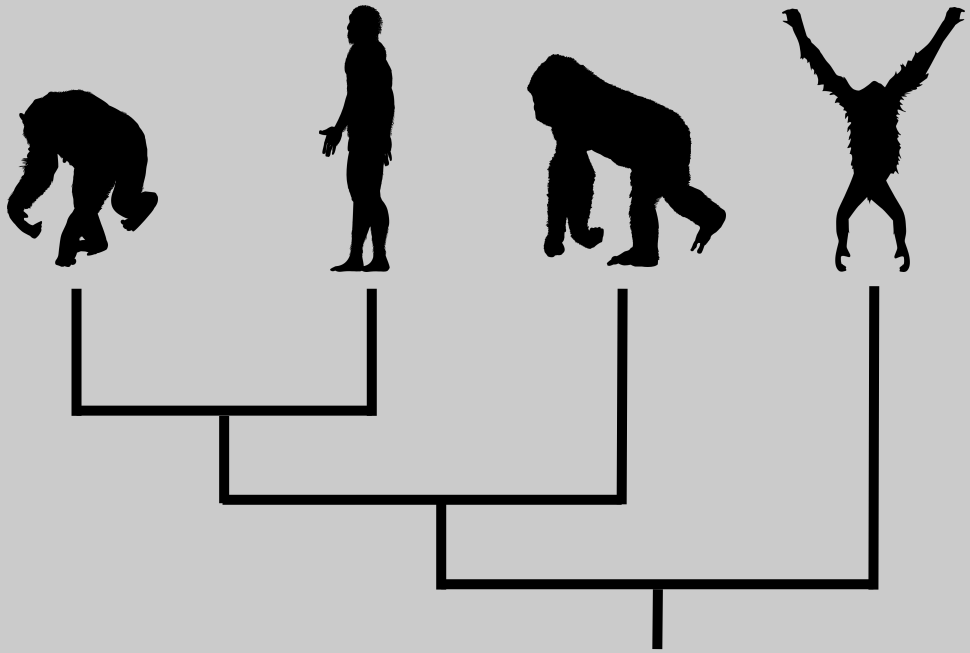

a
g
c
t




a c t g c c t g a c t g g c t g a c
g c t g a c g t c a t g g c t g a c
a g t g a c g t a g t g a g t g c g
a g c g c c t g c g c g a g c g c g

9 models


a
g
c
t



```
actgcctgactggctgac
gctgacgtcatggctgac
agtgacgtagtgagtgcg
agcgcctgcgcgagcgcg
```

1

```
actgcctgactggctgac
gctgacgtcatggctgac
agtgacgtagtgagtgcg
agcgcctgcgcgagcgcg
```

3

```
actgcctgactggctgac
gctgacgtcatggctgac
agtgacgtagtgagtgcg
agcgcctgcgcgagcgcg
```

9

a	c	t	g	c	c	t	g	a	c	t	g	g	c	t	g	a	c
g	c	t	g	a	c	g	t	c	a	t	g	g	c	t	g	a	c
a	g	t	g	a	c	g	t	a	g	t	g	a	g	t	g	c	g
a	g	c	g	c	c	t	g	c	g	c	g	a	g	c	g	c	g

3

a	c	t	g	c	c	t	g	a	c	t	g	g	c	t	g	a	c
g	c	t	g	a	c	g	t	c	a	t	g	g	c	t	g	a	c
a	g	t	g	a	c	g	t	a	g	t	g	a	g	t	g	c	g
a	g	c	g	c	c	t	g	c	g	c	g	a	g	c	g	c	g

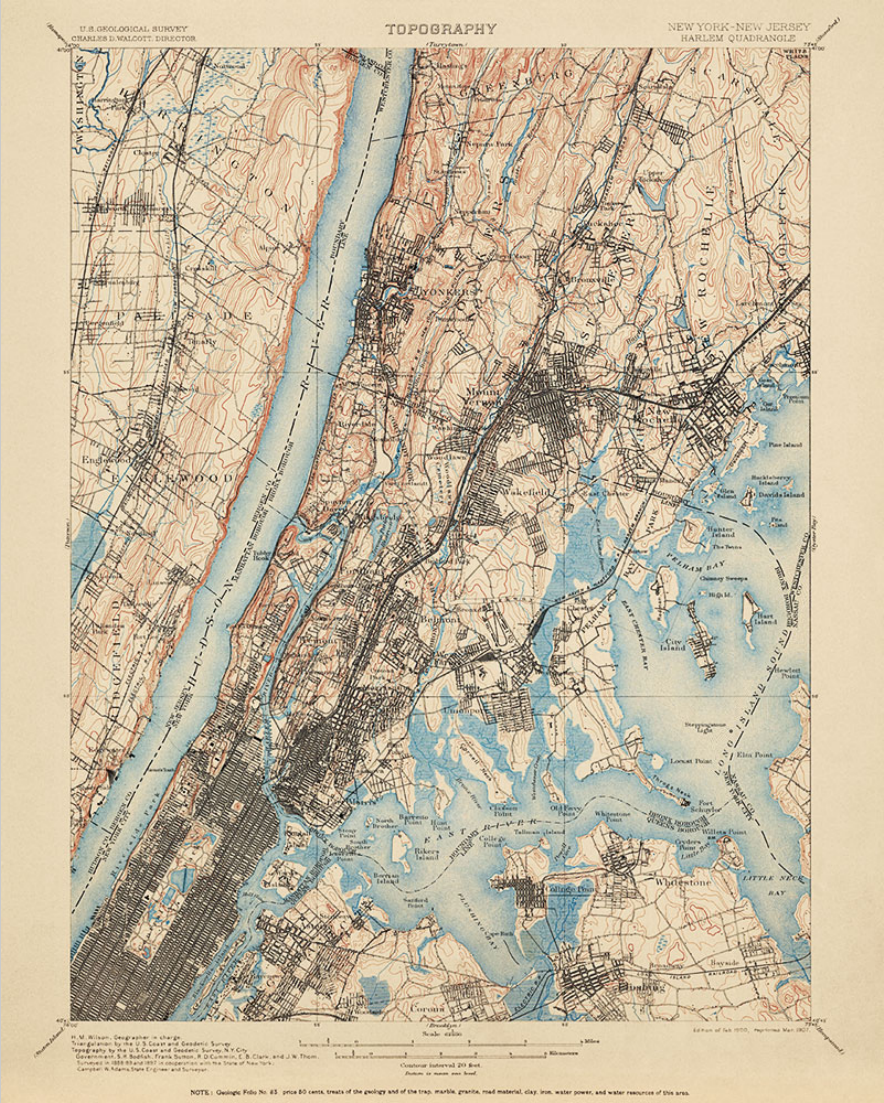
9

a	c	t	g	c	c	t	g	a	c	t	g	g	c	t	g	a	c
g	c	t	g	a	c	g	t	c	a	t	g	g	c	t	g	a	c
a	g	t	g	a	c	g	t	a	g	t	g	a	g	t	g	c	g
a	g	c	g	c	c	t	g	c	g	c	g	a	g	c	g	c	g

6

a	c	t	g	c	c	t	g	a	c	t	g	g	c	t	g	a	c
g	c	t	g	a	c	g	t	c	a	t	g	g	c	t	g	a	c
a	g	t	g	a	c	g	t	a	g	t	g	a	g	t	g	c	g
a	g	c	g	c	c	t	g	c	g	c	g	a	g	c	g	c	g

2


U.S. GEOLOGICAL SURVEY
CHARLES D. WALCOTT, DIRECTOR
TOPOGRAPHY
NEW YORK-NEW JERSEY
HARLEM QUADRANGLE
NOTE: Geologic Folio No. 83, price 50 cents, treats of the geology and of the trap, marble, granite, road material, clay, iron, water power, and water resources of this area.

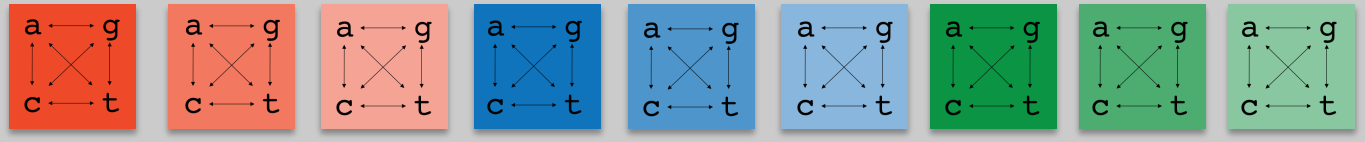

Metropolitan Transportation Authority
MTA New York City Subway
Long Island Sound
THE BRONX
QUEENS
MANHATTAN
BROOKLYN
Hudson River
Key
Part-time line extension
Local service only
All trains stop (local and express service)
Free subway transfer
Free out-of-system subway transfer (excluding single ride ticket)
Normal service
Additional express service
Station Name
Terminal
Accessible station
© 2007 Metropolitan Transportation Authority
Design: Michael Hertz Associates, NYC
4/07

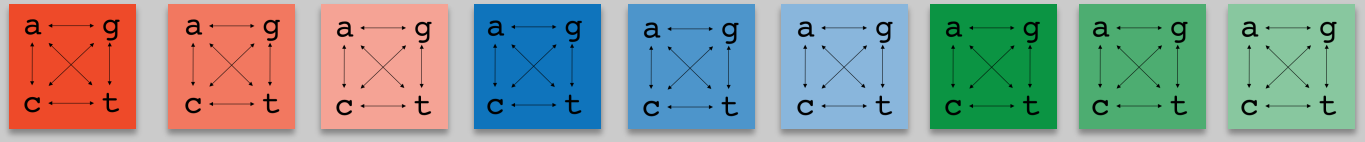
Information theoretic metrics

- AIC, AICc, BIC
- Balancing act between over-parameterization and under paramaterization

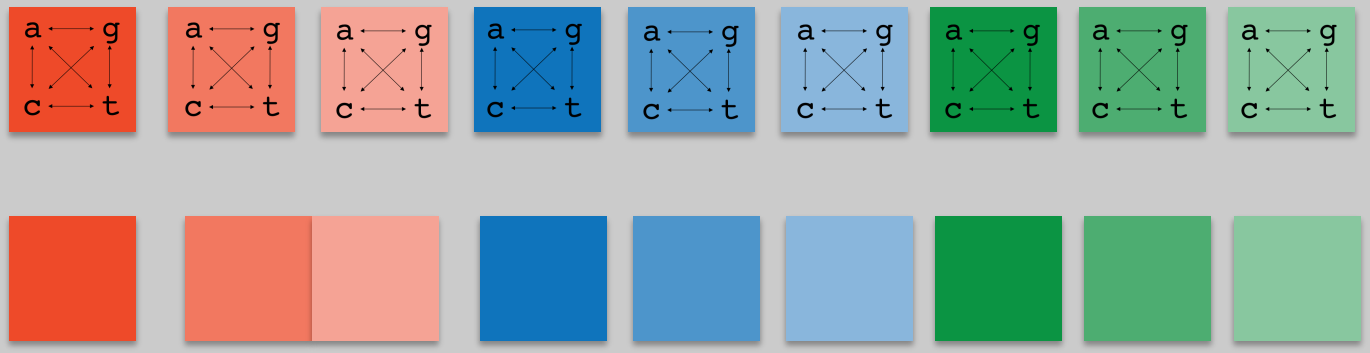

a
g
c
t


a
g
c
t
a
g
c
t

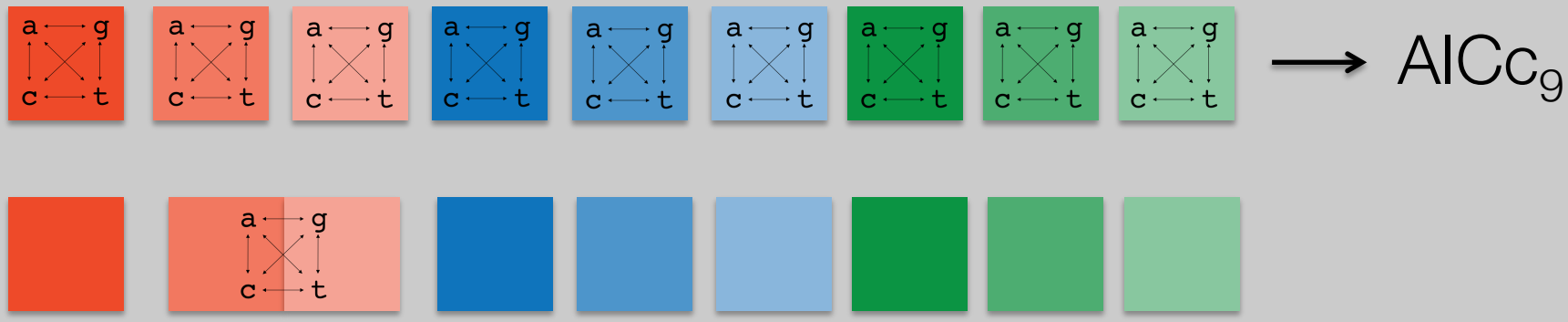

a g
c t
a g
c t
a g
c t
a g
c t
a g
c t
a g
c t
a g
c t
a g
c t
a g
c t

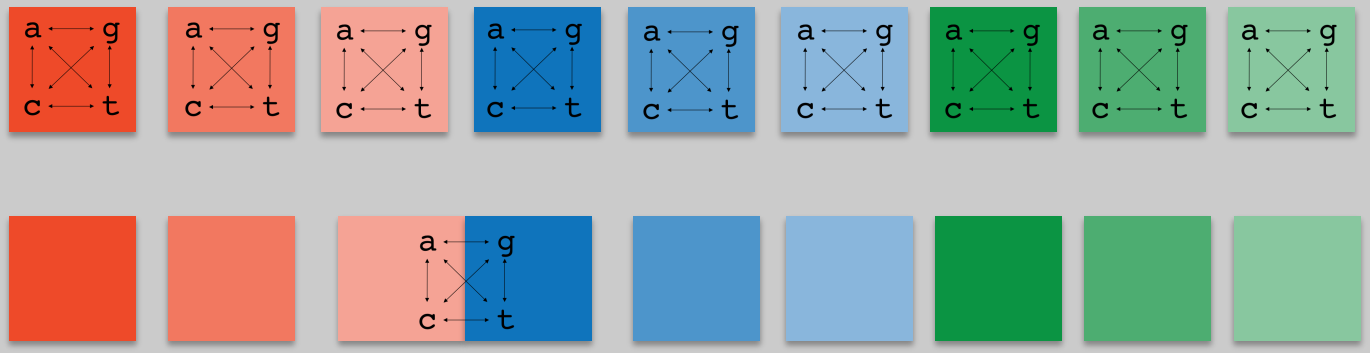


$\longrightarrow$ $AlCc_9$

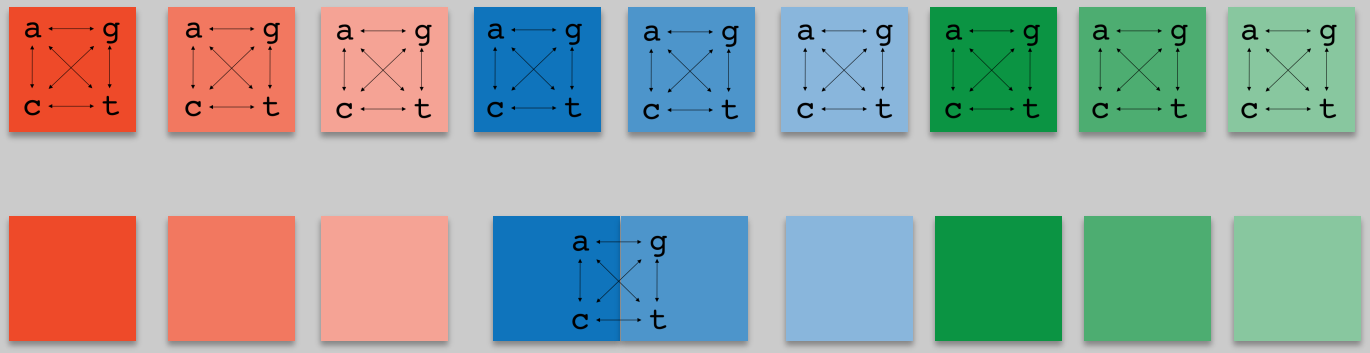


$\longrightarrow$ $AlCc_9$


a g
c t
AlCc9

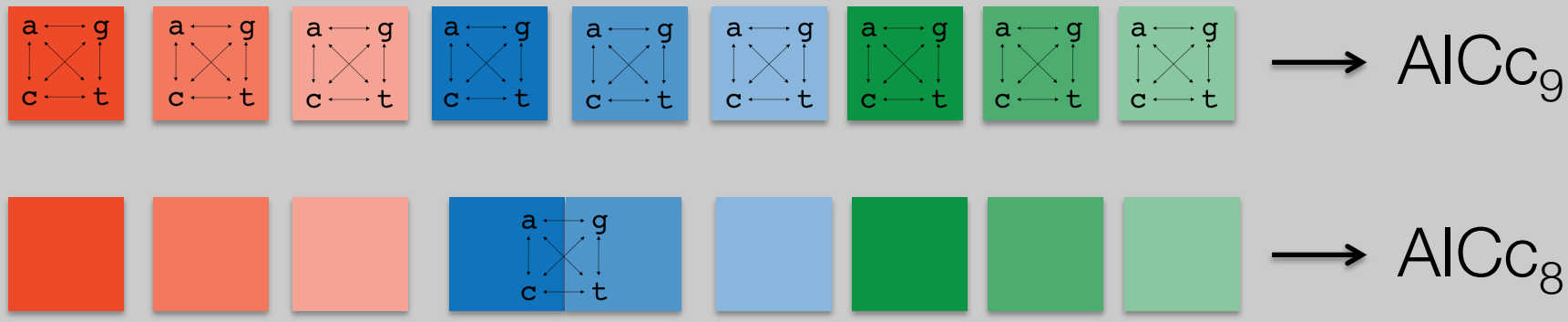


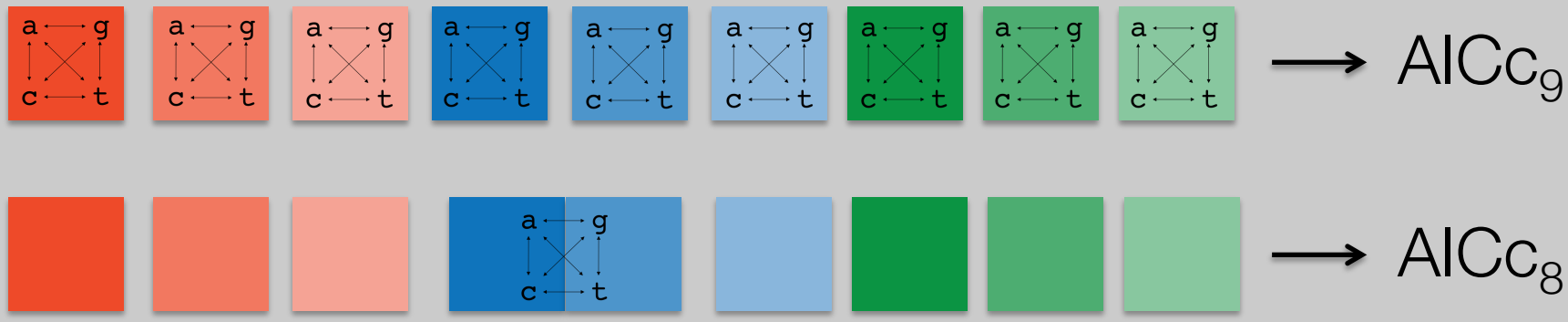
$\longrightarrow$ $AlCc_9$



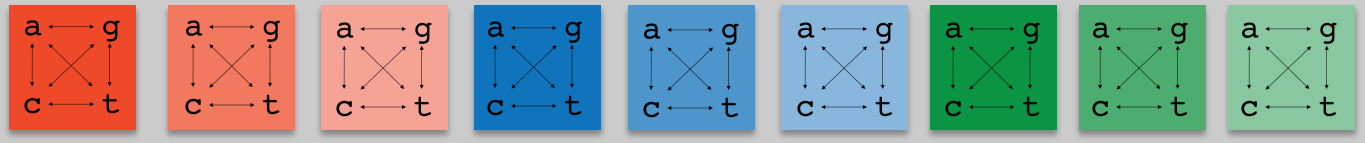
$\longrightarrow$ $AlCc_9$

$$C_2^9 = 36$$

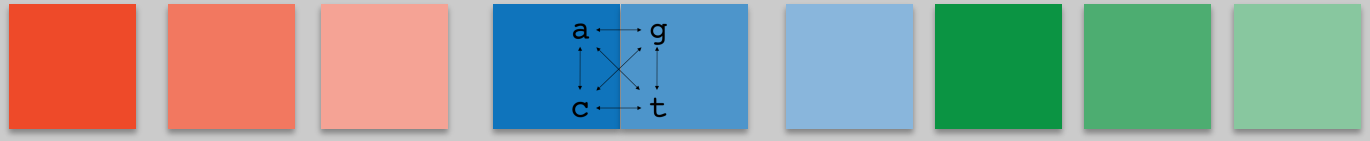

a g
c t
AlCc9
AlCc8



If: $AICc_8 < AICc_9$

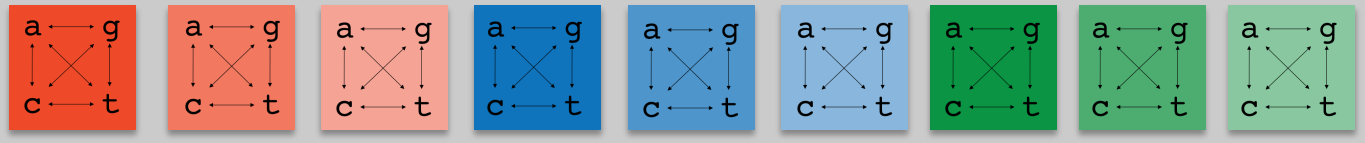


$\longrightarrow$ $AICc_9$

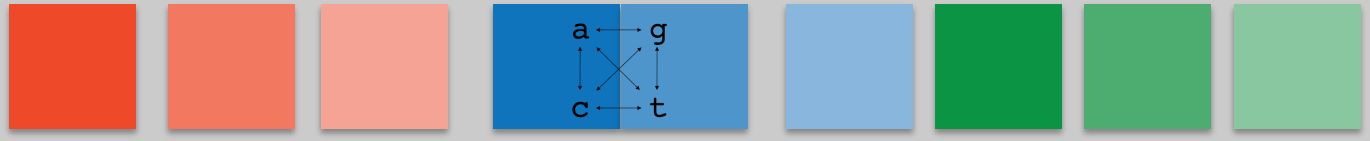


$\longrightarrow$ $AICc_8$

If: $AICc_8 < AICc_9$



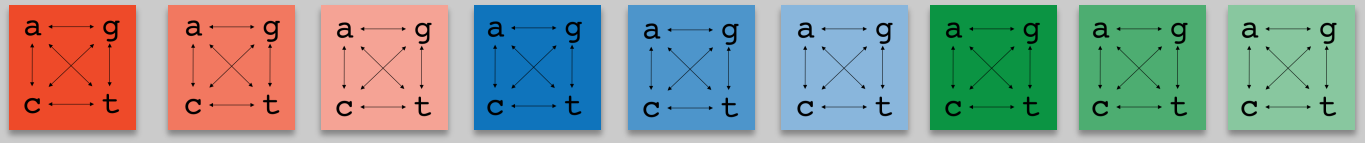
$\longrightarrow$ $AICc_9$



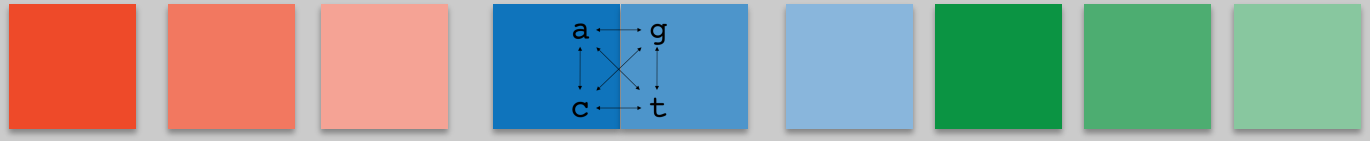
$\longrightarrow$ $AICc_8$

If: $AICc_8 < AICc_9$

$\longrightarrow$ $AICc_7$



$\longrightarrow$ $AICc_9$

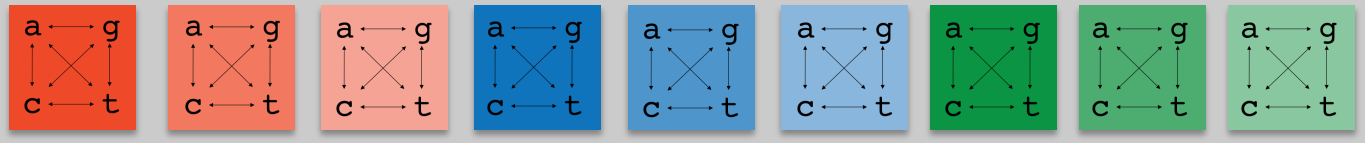


$\longrightarrow$ $AICc_8$

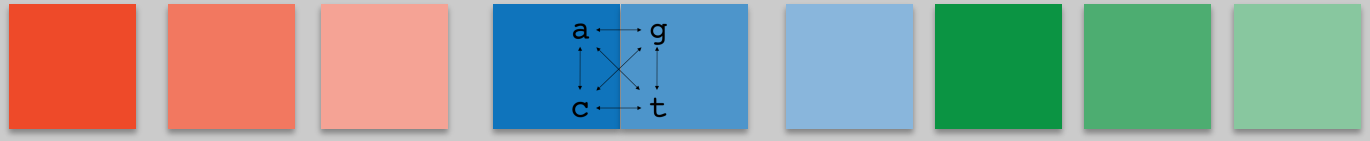
If: $AICc_8 < AICc_9$

$\longrightarrow$ $AICc_7$

If: $AICc_7 < AICc_8$



$\longrightarrow$ $AICc_9$

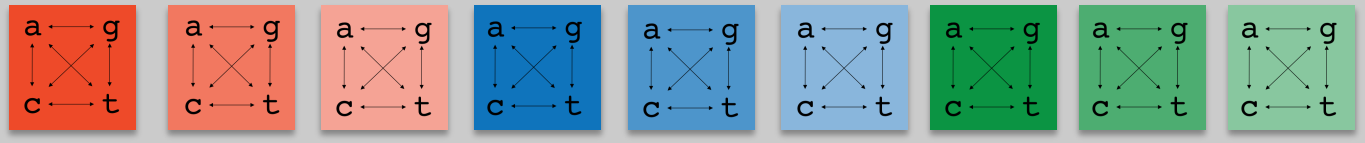


$\longrightarrow$ $AICc_8$

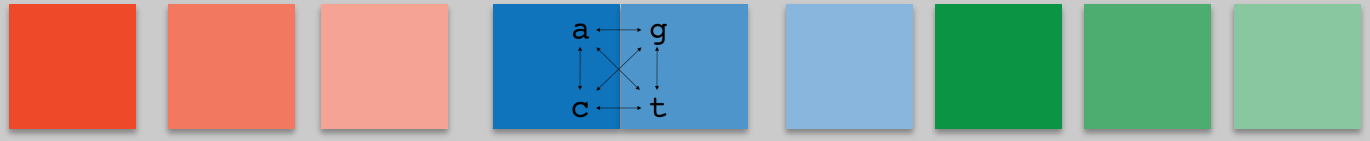
If: $AICc_8 < AICc_9$

$\longrightarrow$ $AICc_7$

If: $AICc_7 < AICc_8$



$\longrightarrow$ $AICc_9$

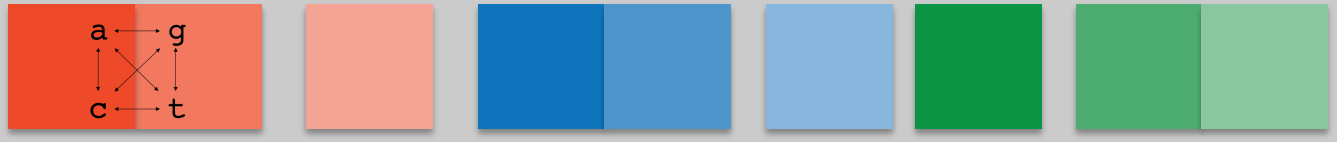


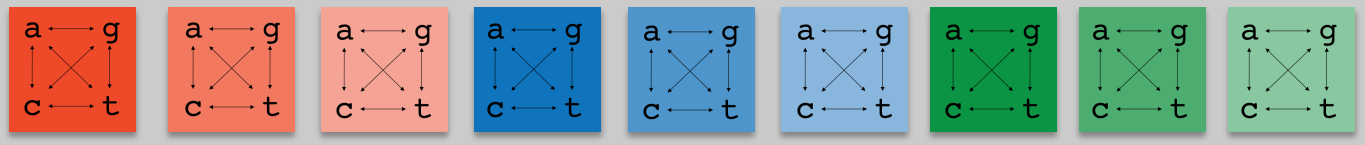
$\longrightarrow$ $AICc_8$

If: $AICc_8 < AICc_9$

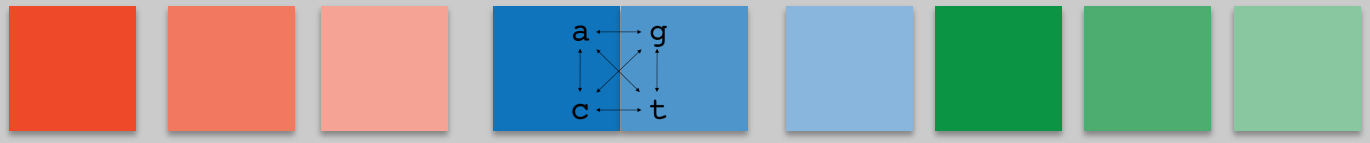
$\longrightarrow$ $AICc_7$

If: $AICc_7 < AICc_8$





$\longrightarrow$ $AICc_9$

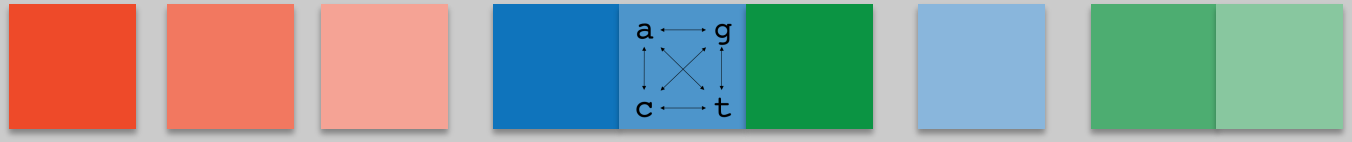


$\longrightarrow$ $AICc_8$

If: $AICc_8 < AICc_9$

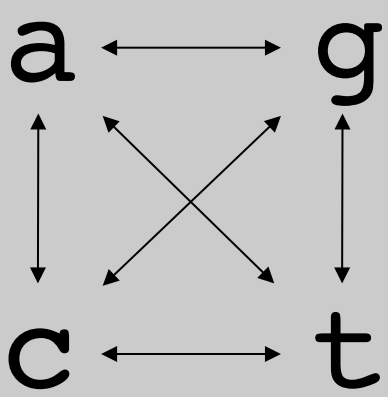
$\longrightarrow$ $AICc_7$

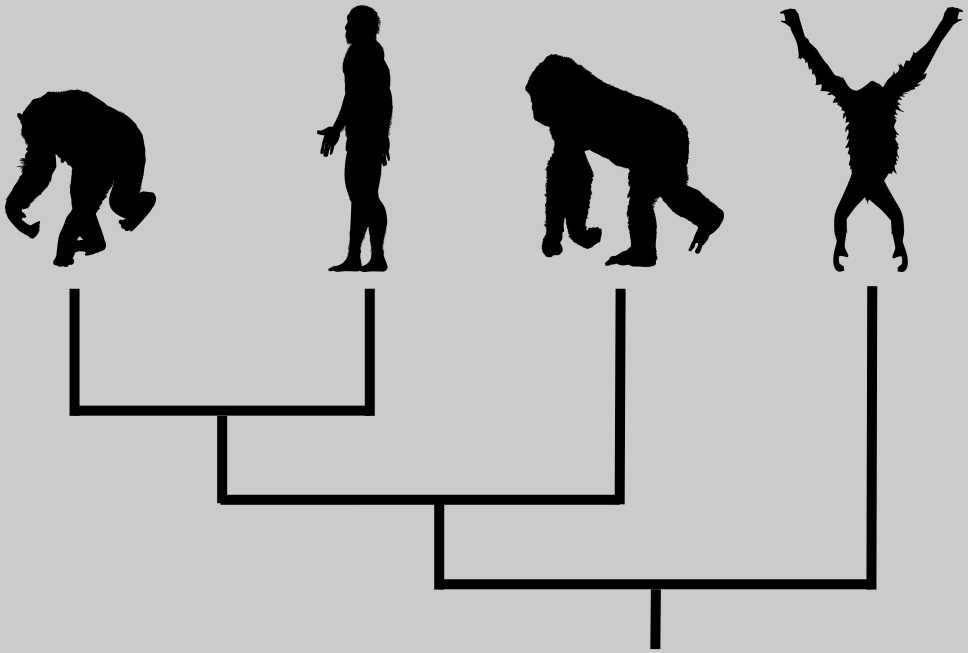
If: $AICc_7 < AICc_8$



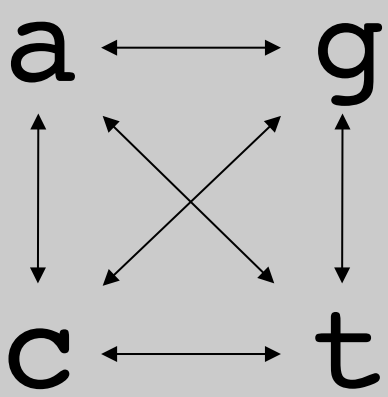
$\longrightarrow$ $AICc_6$

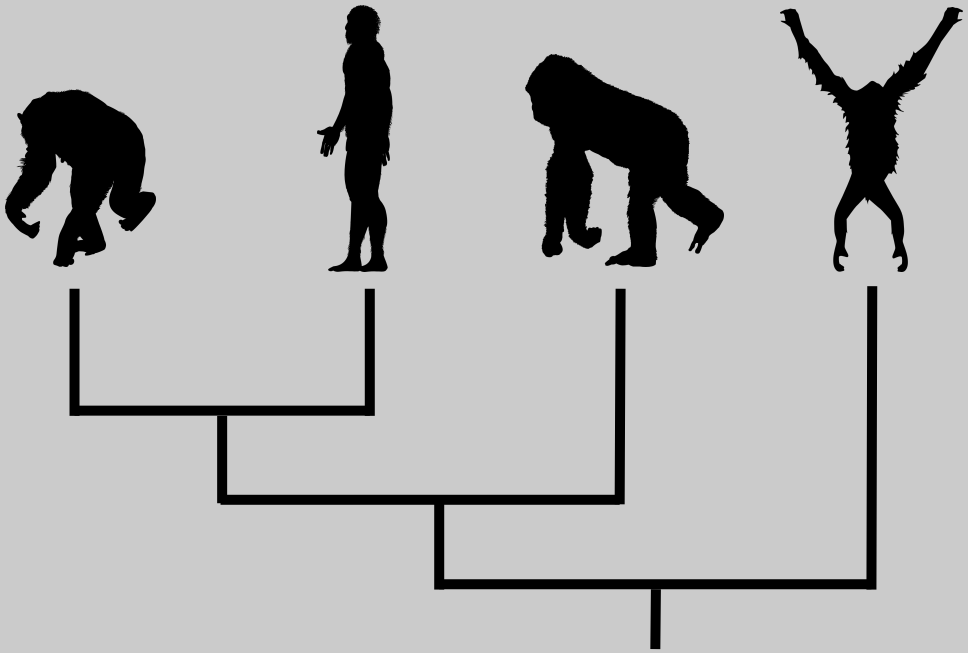
```
a c t g c c t g a c t g g c t g a c
g c t g a c g t c a t g g c t g a c
a g t g a c g t a g t g a g t g c g
a g c g c c t g c g c g a g c g c g
```



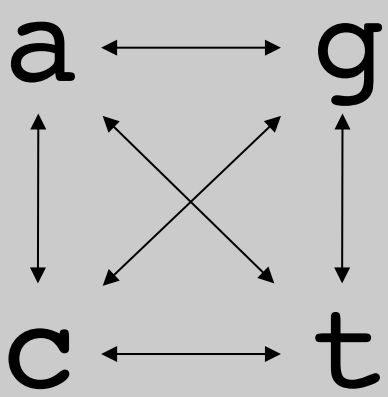


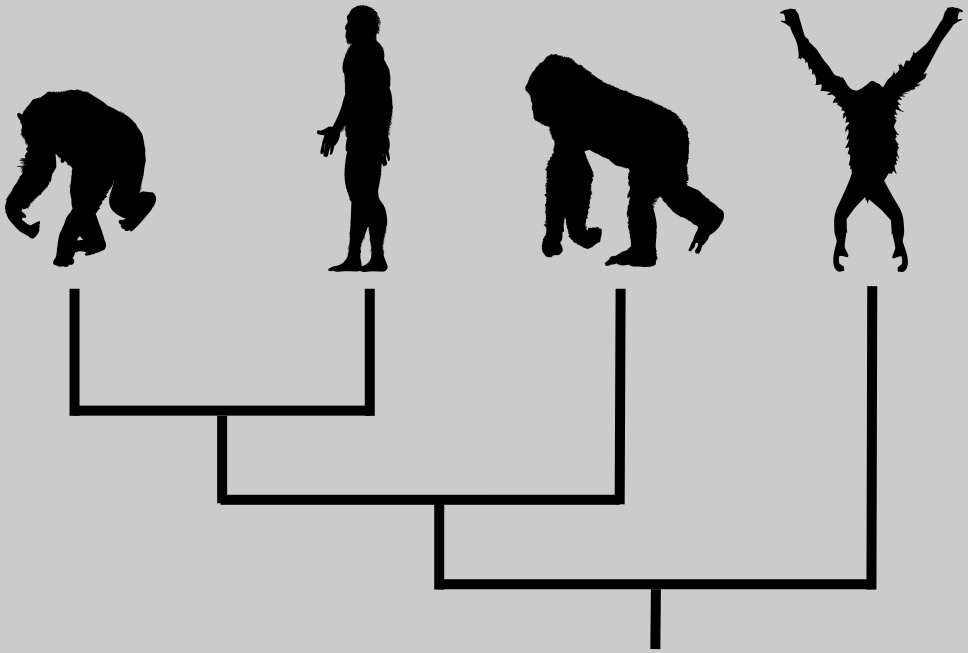
```
a c t g c c t g a c t g g c t g a c
g c t g a c g t c a t g g c t g a c
a g t g a c g t a g t g a g t g c g
a g c g c c t g c g c g a g c g c g
```



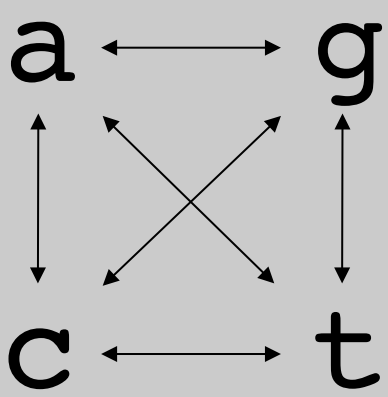


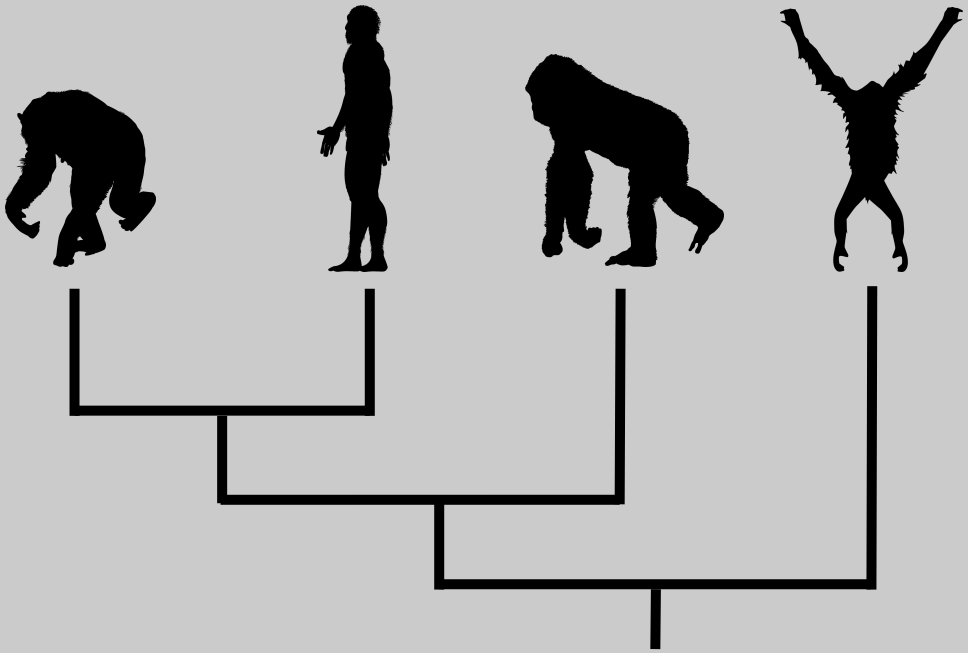
```
a c t g c c t g a c t g g c t g a c
g c t g a c g t c a t g g c t g a c
a g t g a c g t a g t g a g t g c g
a g c g c c t g c g c g a g c g c g
```

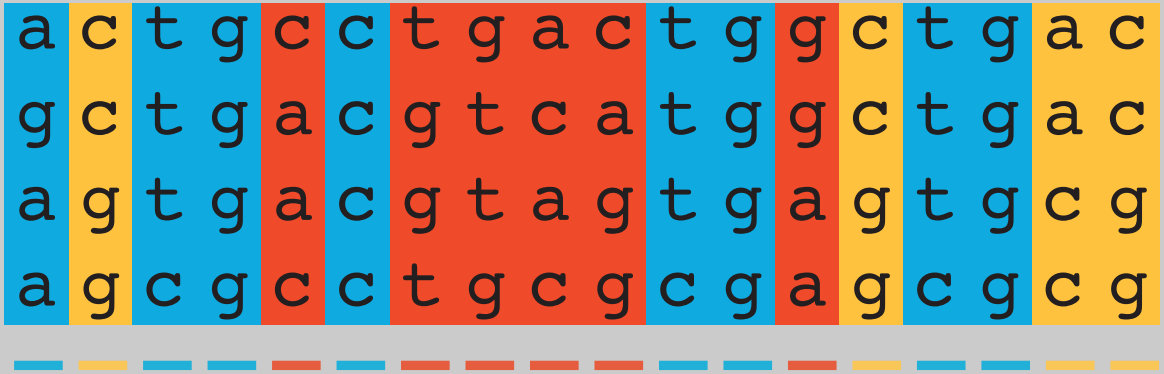


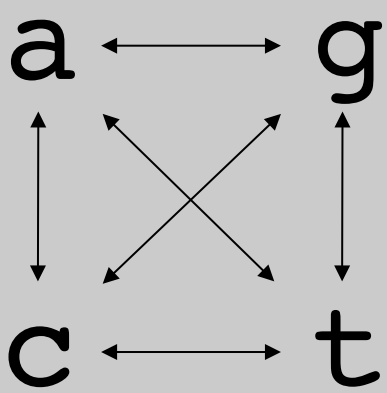


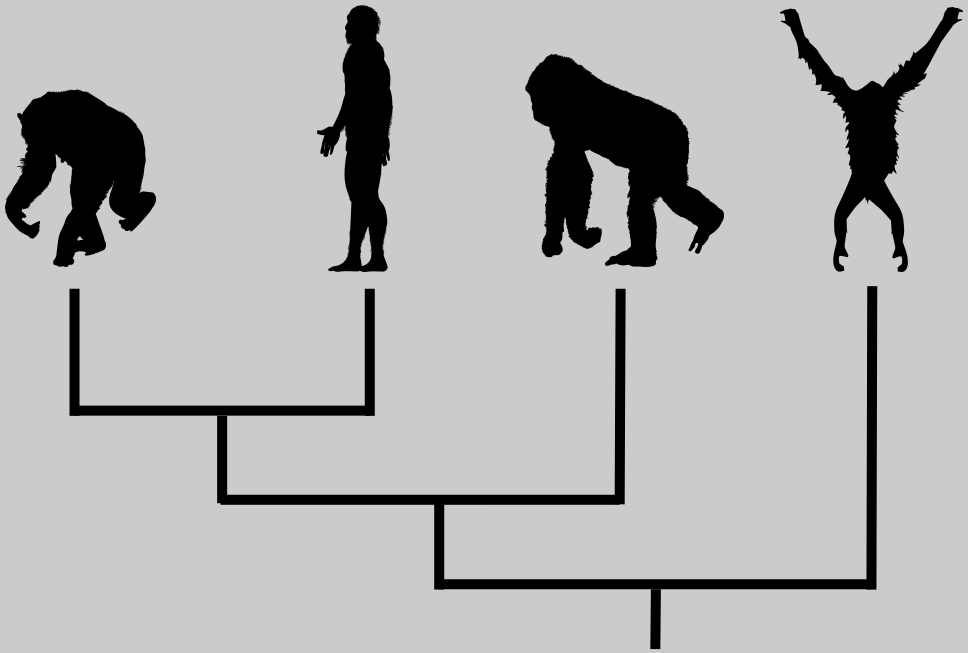
```
a c t g c c t g a c t g g c t g a c
g c t g a c g t c a t g g c t g a c
a g t g a c g t a g t g a g t g c g
a g c g c c t g c g c g a g c g c g
```






a c t g c c t g a c t g g c t g a c
g c t g a c g t c a t g g c t g a c
a g t g a c g t a g t g a g t g c g
a g c g c c t g c g c g a g c g c g


a
g
c
t



Algorithms available in PartitionFinder 2

Predefined data blocks:

user

greedy

hcluster

rcluster

rclusterf

Without predefined data blocks:

kmeans

PartitionFinder: Combined Selection of Partitioning Schemes and Substitution Models for Phylogenetic Analyses

Robert Lanfear,*[,1] Brett Calcott,[1,2] Simon Y. W. Ho,[3] and Stephane Guindon[4]

greedy

METHODOLOGY ARTICLE Open Access

Selecting optimal partitioning schemes for phylogenomic datasets

Robert Lanfear[1,2*†], Brett Calcott[3†], David Kainer[1], Christoph Mayer[4] and Alexandros Stamatakis[5,6]

rcluster, hcluster

METHODOLOGY ARTICLE Open Access

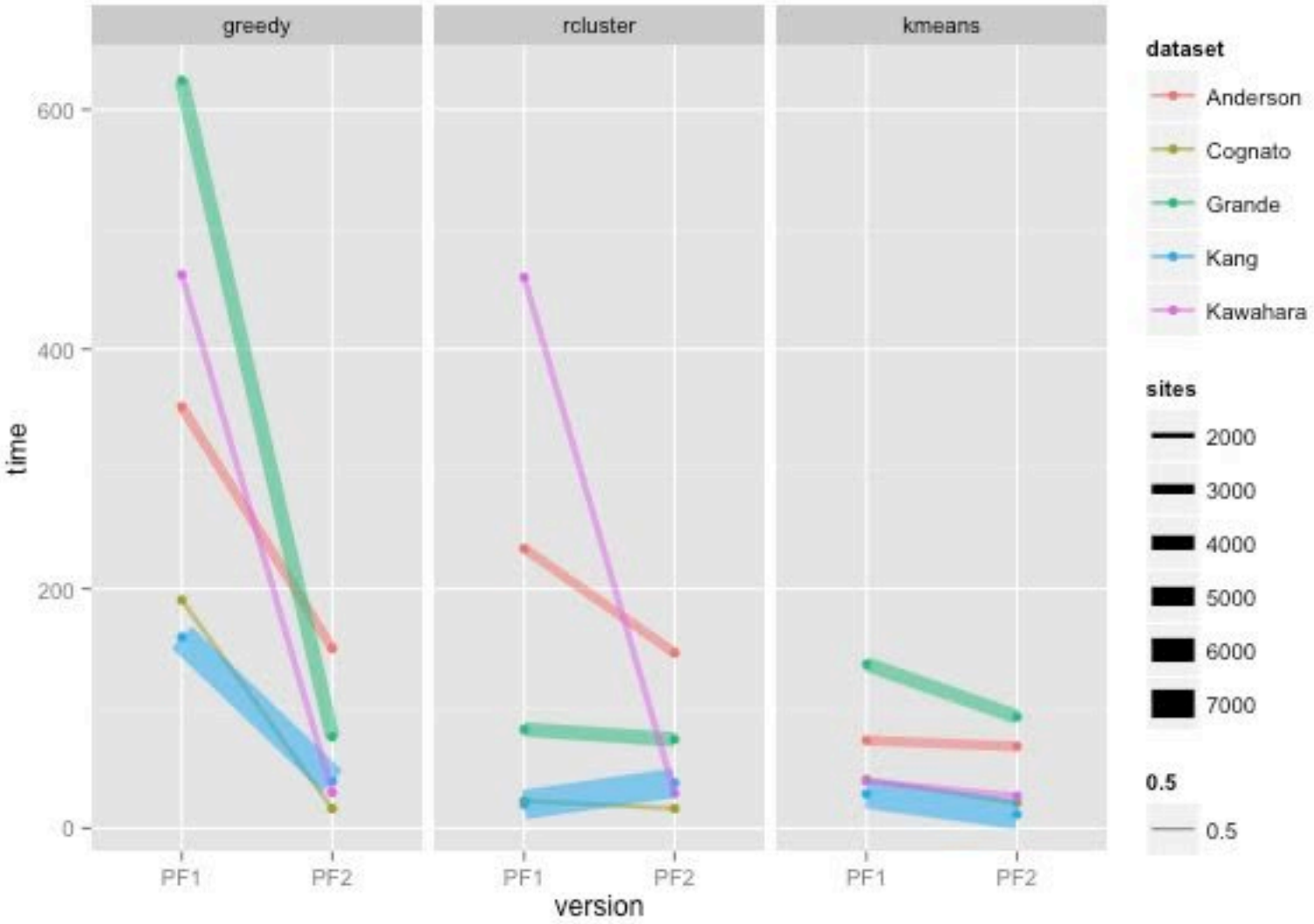
Automatic selection of partitioning schemes for phylogenetic analyses using iterative *k*-means clustering of site rates

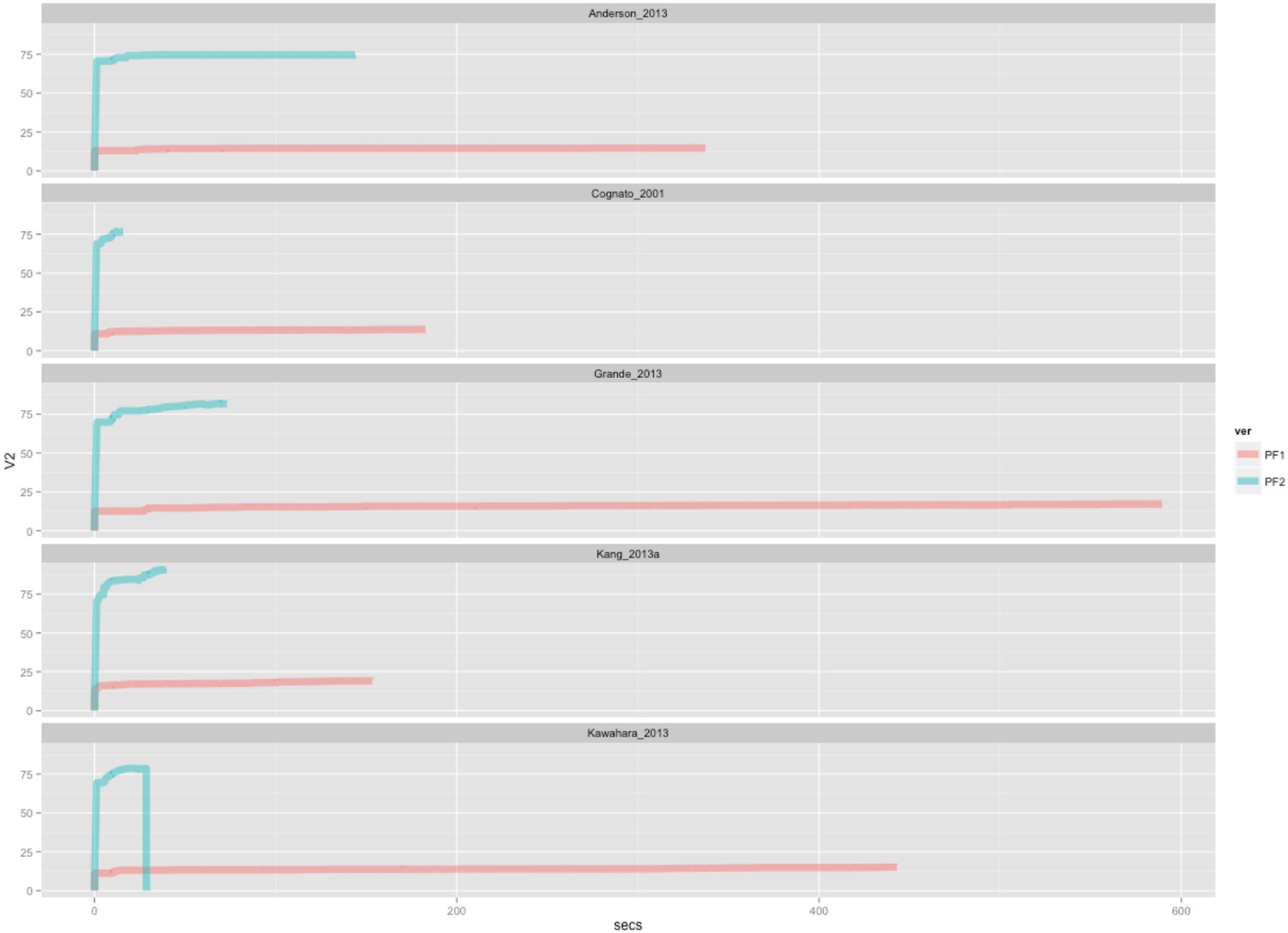
Paul B Frandsen[1,2*], Brett Calcott[3], Christoph Mayer[4] and Robert Lanfear[5,6,7]

kmeans

PartitionFinder2

- Out now on GitHub!
- Installed on Hydra
- New features:
  - Uses NumPy for faster computation
  - kmeans and rclusterf algorithms are included
- And soon…
  - Morphology! (with help from April Wright of Iowa State University)


greedy
rcluster
kmeans
time
600
400
200
0
PF1
PF2
version
dataset
Anderson
Cognato
Grande
Kang
Kawahara
sites
2000
3000
4000
5000
6000
7000
0.5
0.5


Anderson_2013
Cognato_2001
Grande_2013
Kang_2013a
Kawahara_2013
V2
secs
0
25
50
75
0
200
400
600
ver
PF1
PF2

PartitionFinder support

- Excellent manual included with GitHub repo (https://github.com/brettc/partitionfinder)
- Google group: https://groups.google.com/forum/#!forum/partitionfinder
- Questions specific to Hydra installation: contact me

Tutorial

- ssh into Hydra
- Navigate to your /pool directory
- Copy /pool/genomics/frandsenp/examples/nucleotide to your /pool directory
- Change your directory into the nucleotide directory
- Inside should be a partition_finder.cfg file and an alignment called test.phy

The PartitionFinder config file

partition_finder.cfg

```
## ALIGNMENT FILE ##
alignment = test.phy;

## BRANCHLENGTHS: linked | unlinked ##
branchlengths = linked;

## MODELS OF EVOLUTION: all | allx | mrbayes | beast | gamma | gammai <list> ##
models = all;

# MODEL SELECCTION: AIC | AICc | BIC #
model_selection = aicc;

## DATA BLOCKS: see manual for how to define ##
[data_blocks]
Gene1_pos1 = 1-789\3;
Gene1_pos2 = 2-789\3;
Gene1_pos3 = 3-789\3;
Gene2_pos1 = 790-1449\3;
Gene2_pos2 = 791-1449\3;
Gene2_pos3 = 792-1449\3;
Gene3_pos1 = 1450-2208\3;
Gene3_pos2 = 1451-2208\3;
Gene3_pos3 = 1452-2208\3;

## SCHEMES, search: all | user | greedy | rcluster | hcluster | kmeans ##
[schemes]
search = greedy;

#user schemes go here if search=user. See manual for how to define.#
```

Line 1, Column 1 | Spaces: 4 | Plain Text

Tutorial

- Open the Hydra QSubGen web app (https://hydra-3.si.edu/tools/QSubGen)
- Select the short queue
- Select multi-thread and 4 CPUs
- Select “sh” for your shell
- Load the bioinformatics/partitionfinder/2.0pr13 module

Tutorial

- For PartitionFinder on Hydra you must call the PartitionFinder.py executable directly. Note: this contradicts the instructions in the user manual
- Enter the command: PartitionFinder.py nucleotide --raxml -p $NSLOTS

Tutorial

- Enter the command: PartitionFinder.py nucleotide --raxml -p $NSLOTS
- --raxml specifies that you will use raxml for your likelihood calculations
- -p specifies the number of CPU threads to use. Note: this is very important to specify on Hydra because if you do not specify it, the default in PartitionFinder is to use all CPUs.

Tutorial

- Choose an informative job name
- Select -cwd
- Either download and upload your job file to Hydra or copy and paste it into your favorite text editor on Hydra. I save my job file as “pf_nuc.job”.