Biochem 3BP3

Evolutionary Biolgy

Week of Sept 27, 2021

Evolutionary Biology

Molecular evolution is a large sub-discipline

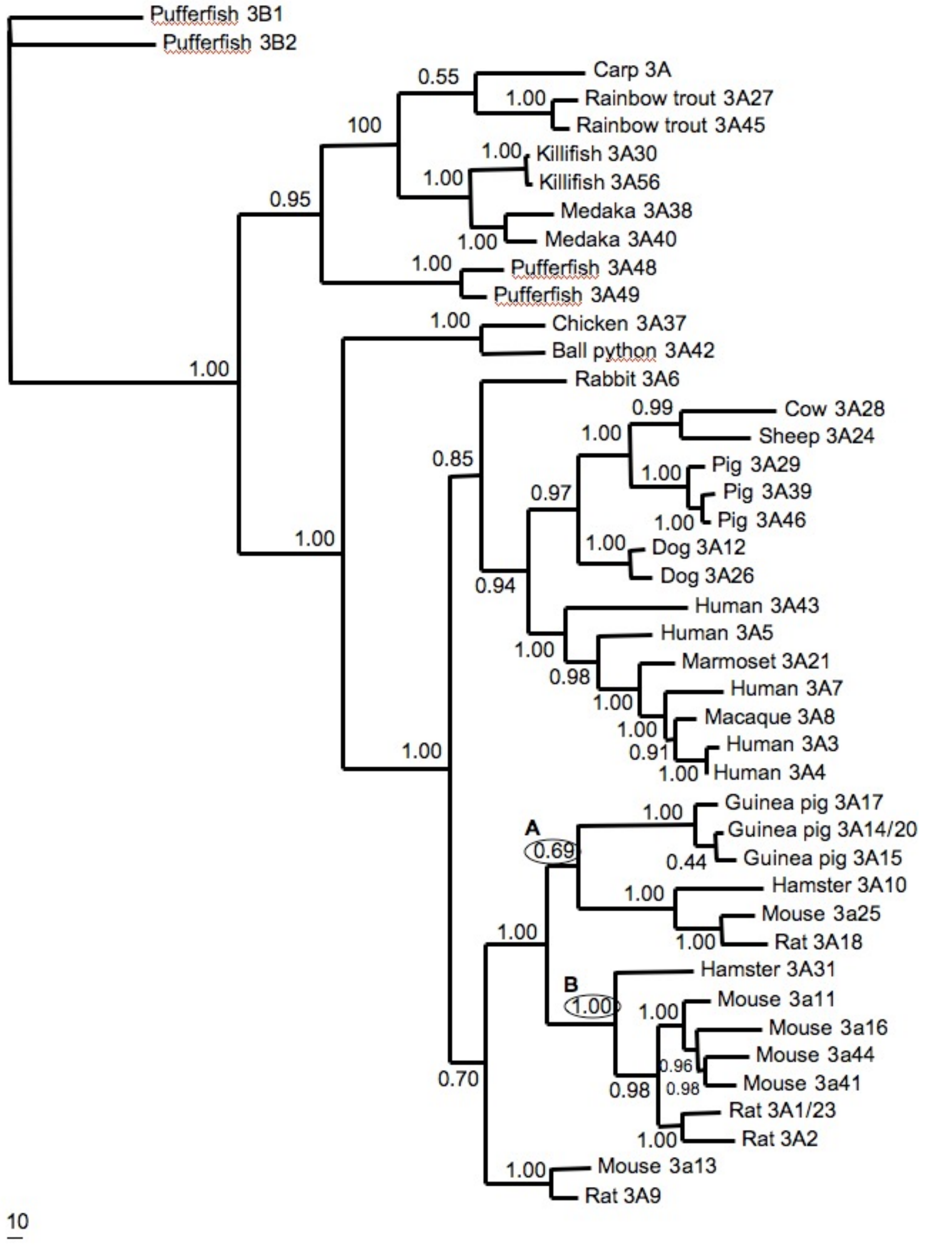
Bioinformatics plays a major role in evolutionary biology

Examples:

- Evolution of drug resistance
- Gene duplication and functional diversification
- Host-pathogen co-evolution

Biochem 3BP3 is going to focus on phylogenetics – how genes are related as a guide to nomenclature and functional predictio

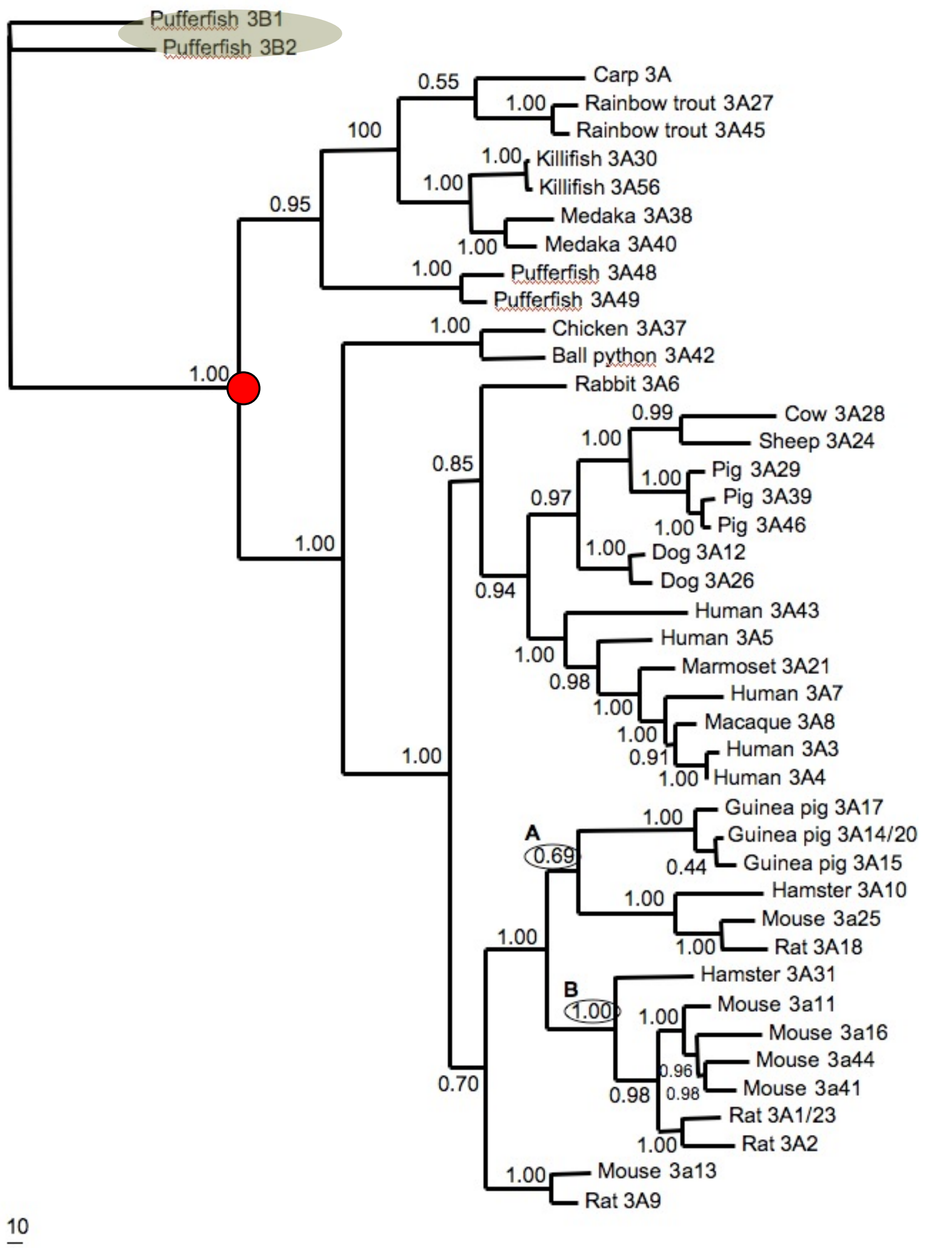
An Example: CYP3A Phylogeny



McArthur et al. 2003. J. Mol. Evol. 57: 200-211.

An Example: CYP3A Phylogeny

Outgroup choice is important

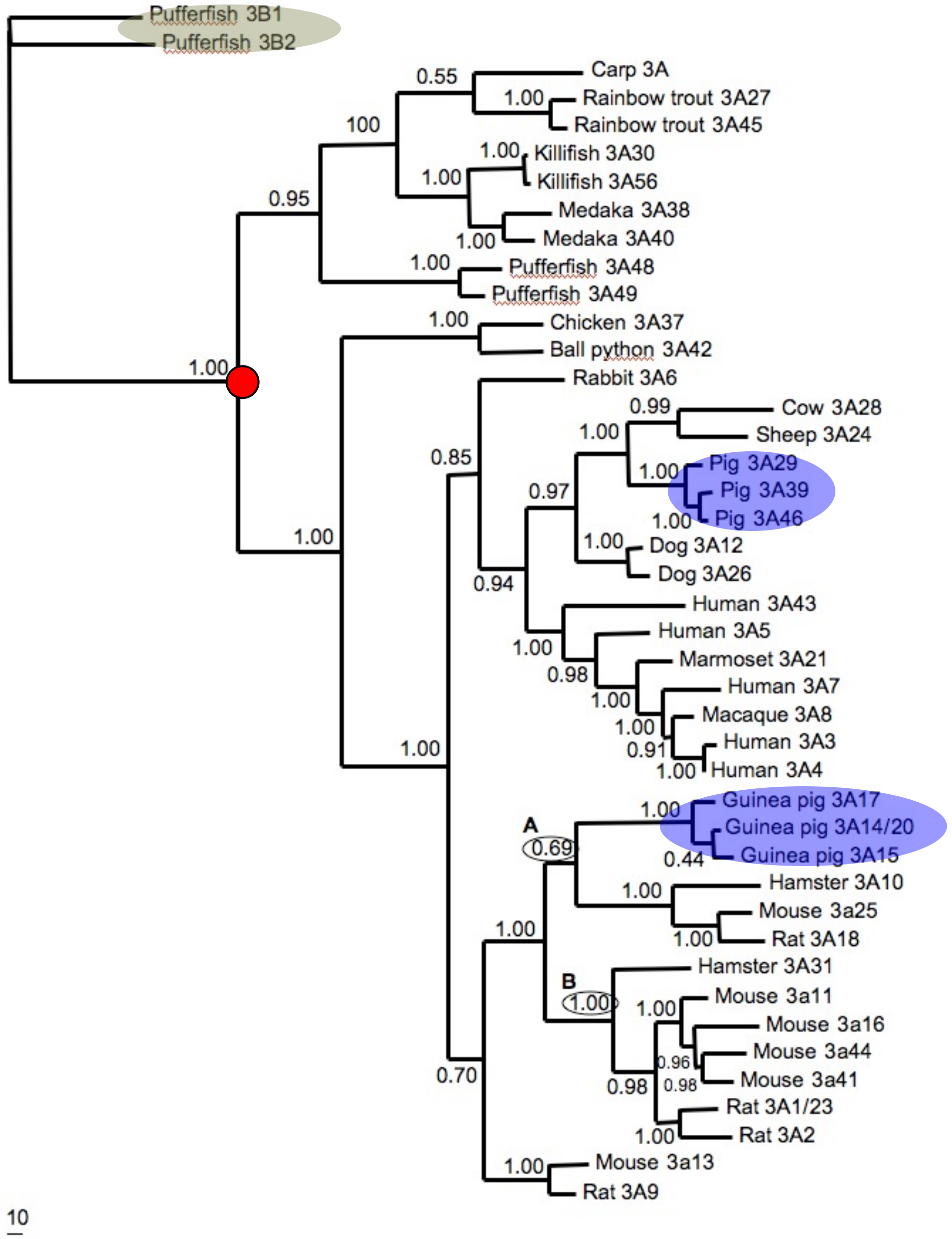


McArthur et al. 2003. J. Mol. Evol. 57: 200-211.

An Example: CYP3A Phylogeny

Outgroup choice is important

Trees can reflect gene duplication



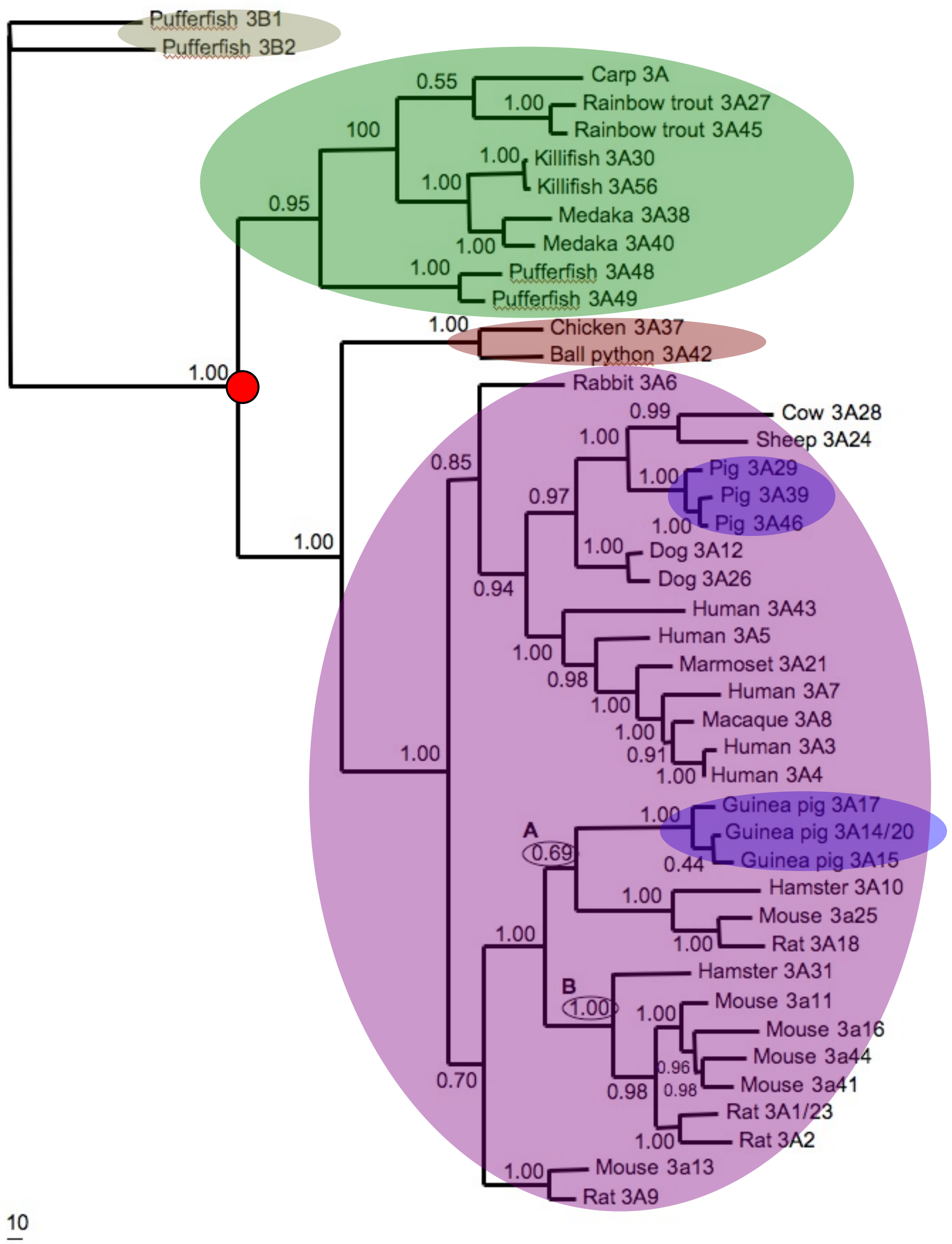
McArthur et al. 2003. J. Mol. Evol. 57: 200-211.

An Example: CYP3A Phylogeny

Outgroup choice is important

Trees can reflect gene duplication

Trees can reflect speciation



McArthur et al. 2003. J. Mol. Evol. 57: 200-211.

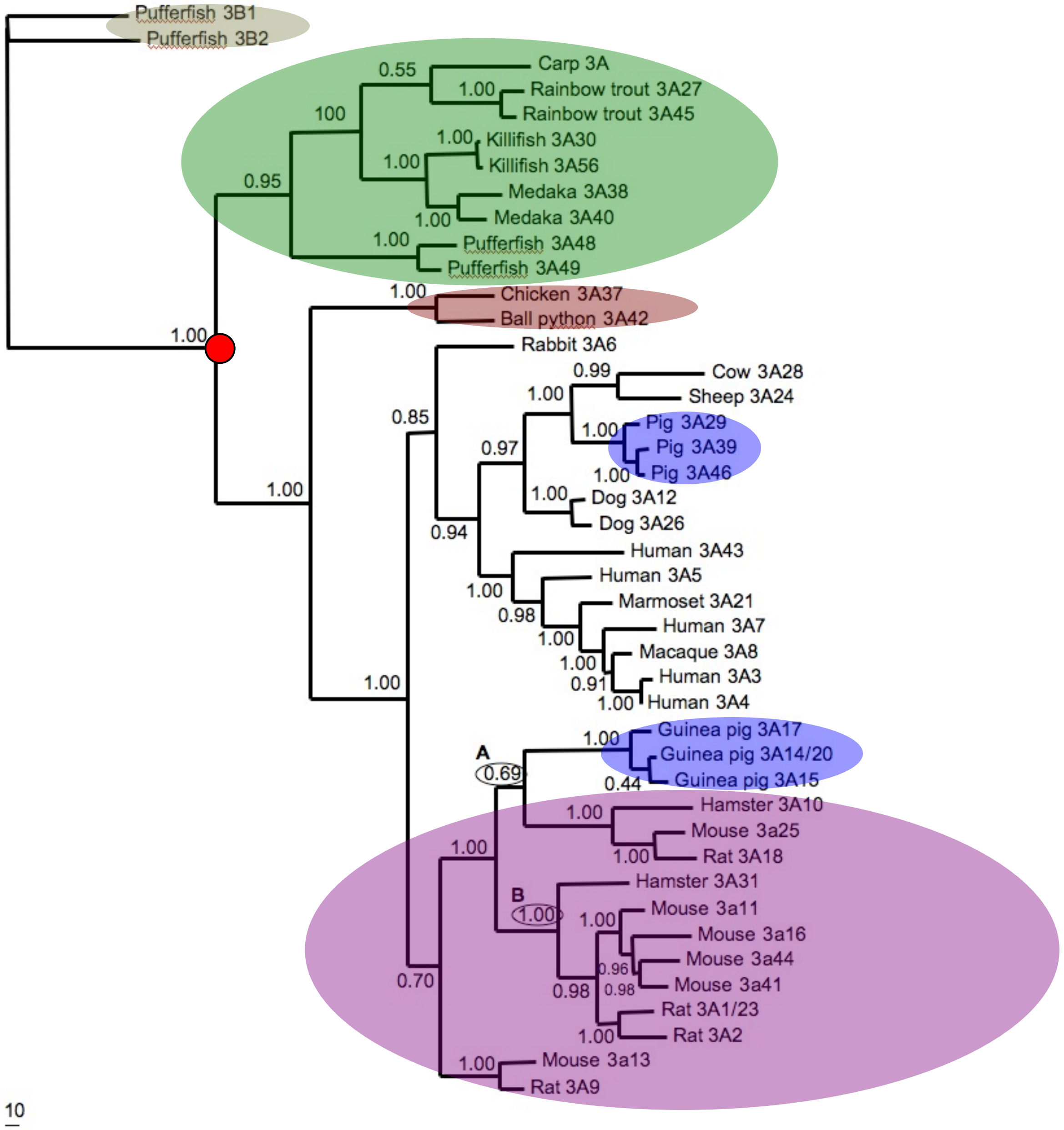
An Example: CYP3A Phylogeny

Outgroup choice is important

Trees can reflect gene duplication

Trees can reflect speciation

Trees can be a combination of gene trees and species trees



McArthur et al. 2003. J. Mol. Evol. 57: 200-211.

An Example: CYP3A Phylogeny

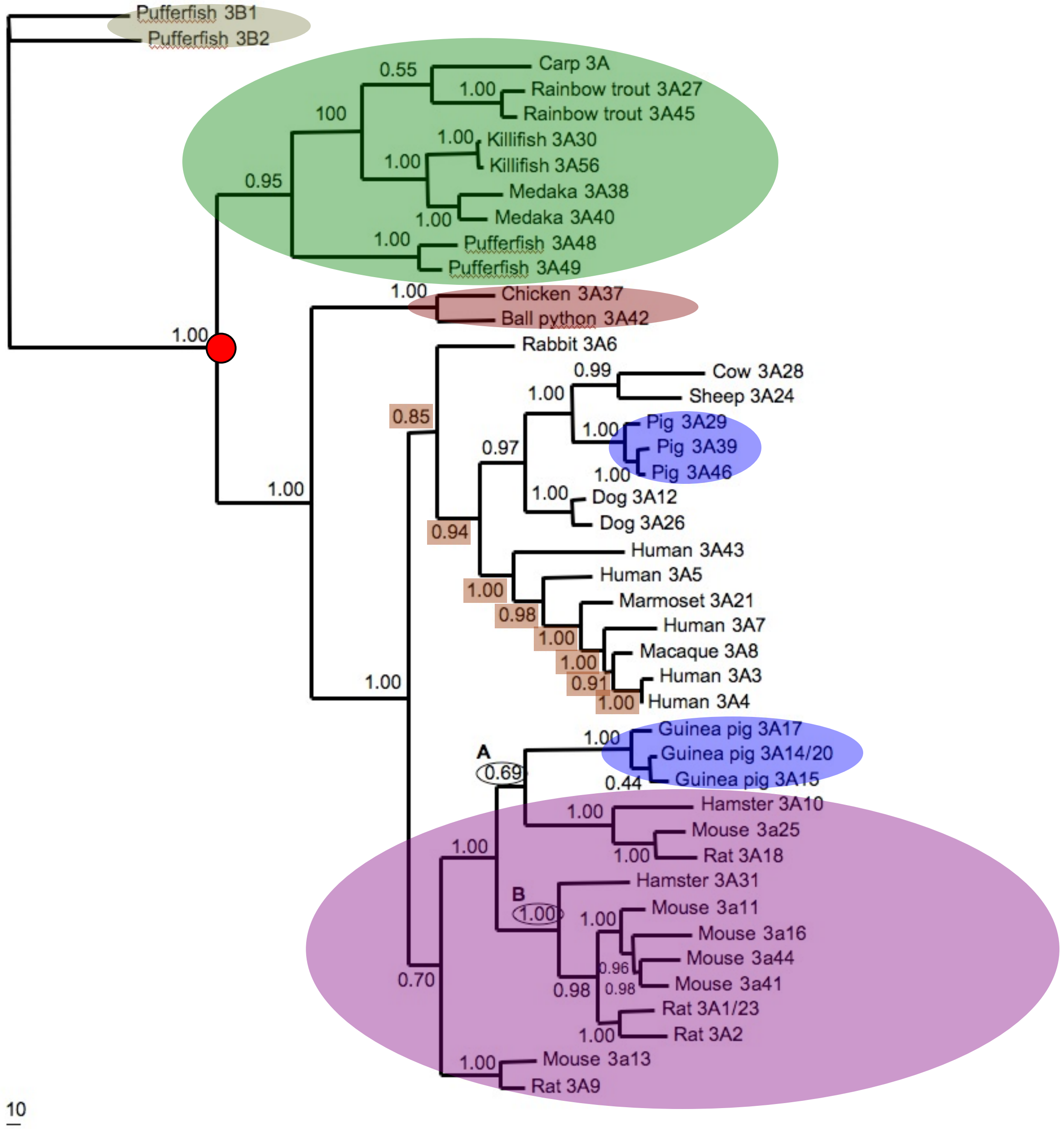
Outgroup choice is important

Trees can reflect gene duplication

Trees can reflect speciation

Trees can be a combination of gene trees and species trees

Trees should include confidence estimates



McArthur et al. 2003. J. Mol. Evol. 57: 200-211.

An Example: CYP3A Phylogeny

Outgroup choice is important

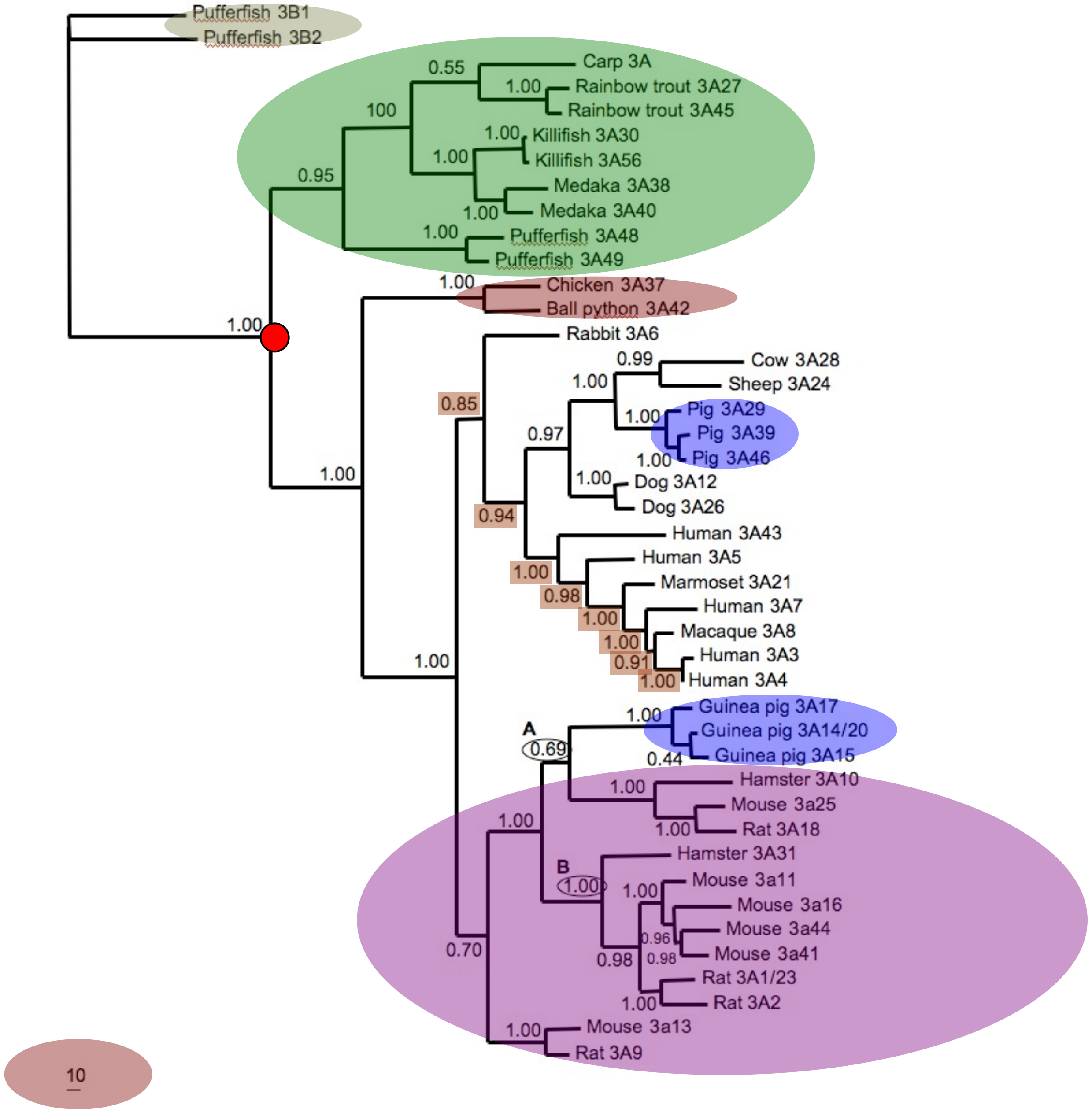
Trees can reflect gene duplication

Trees can reflect speciation

Trees can be a combination of gene trees and species trees

Trees should include confidence estimates

Trees include estimates of evolutionary distance



McArthur et al. 2003. J. Mol. Evol. 57: 200-211.

An Example: CYP3A Phylogeny

Outgroup choice is important

Trees can reflect gene duplication

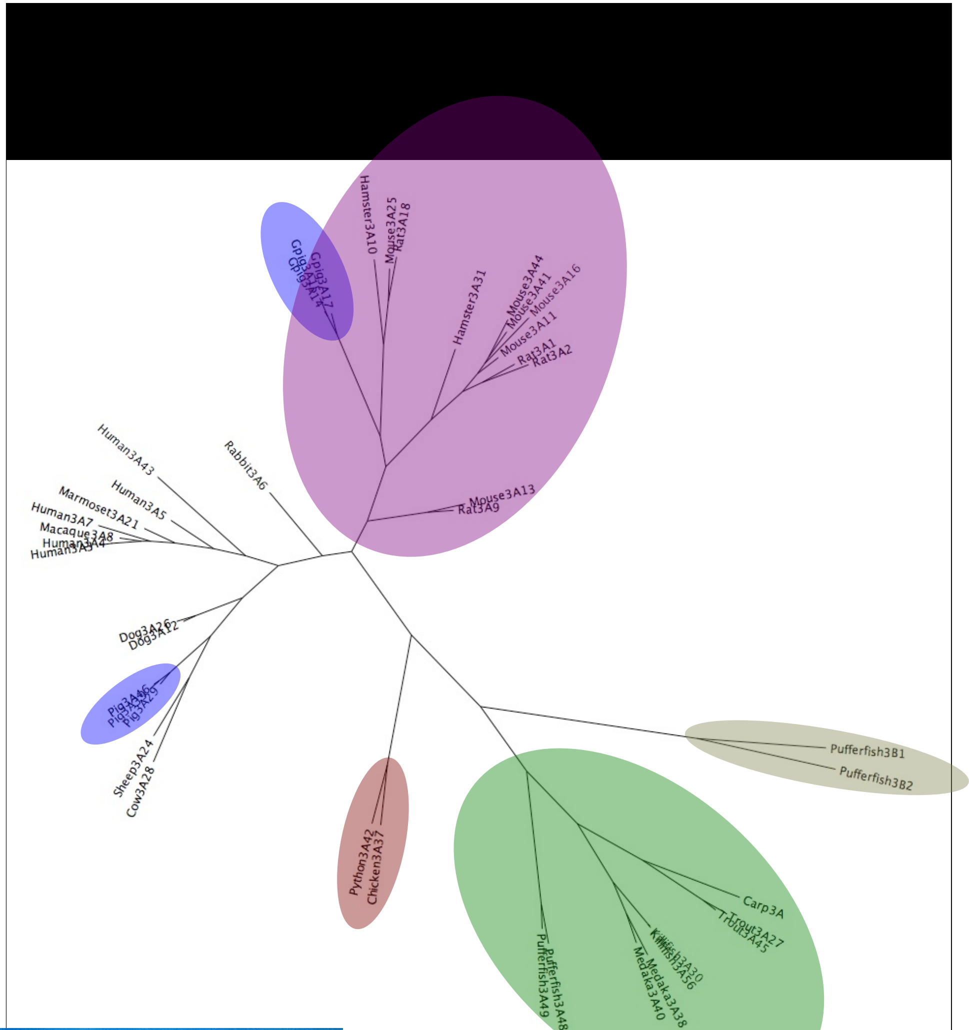
Trees can reflect speciation

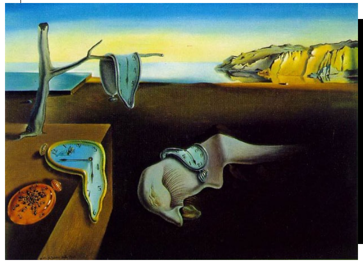
Trees can be a combination of gene trees and species trees

Trees should include confidence estimates

Trees include estimates of evolutionary distance

Branch lengths are a function of time and rate of evolution – evolutionary clocks are local (aka “soft”)



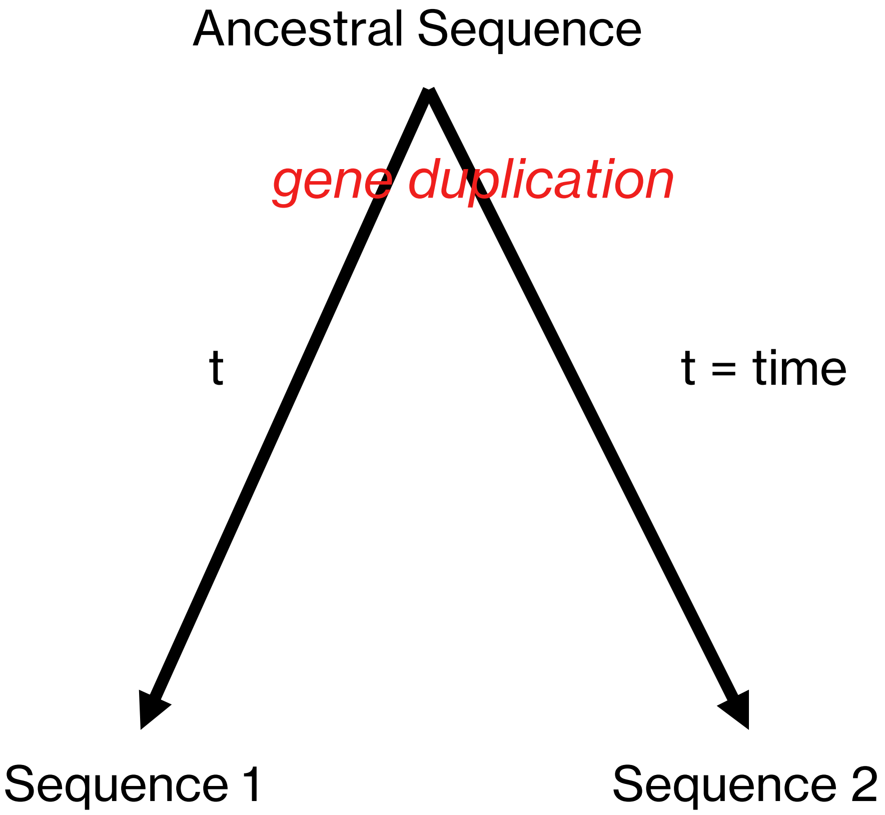


Dali, *Persistence of Time*

Substitution & Rates of Evolution

Ancestral Sequence

Substitution & Rates of Evolution



Ancestral Sequence

gene duplication

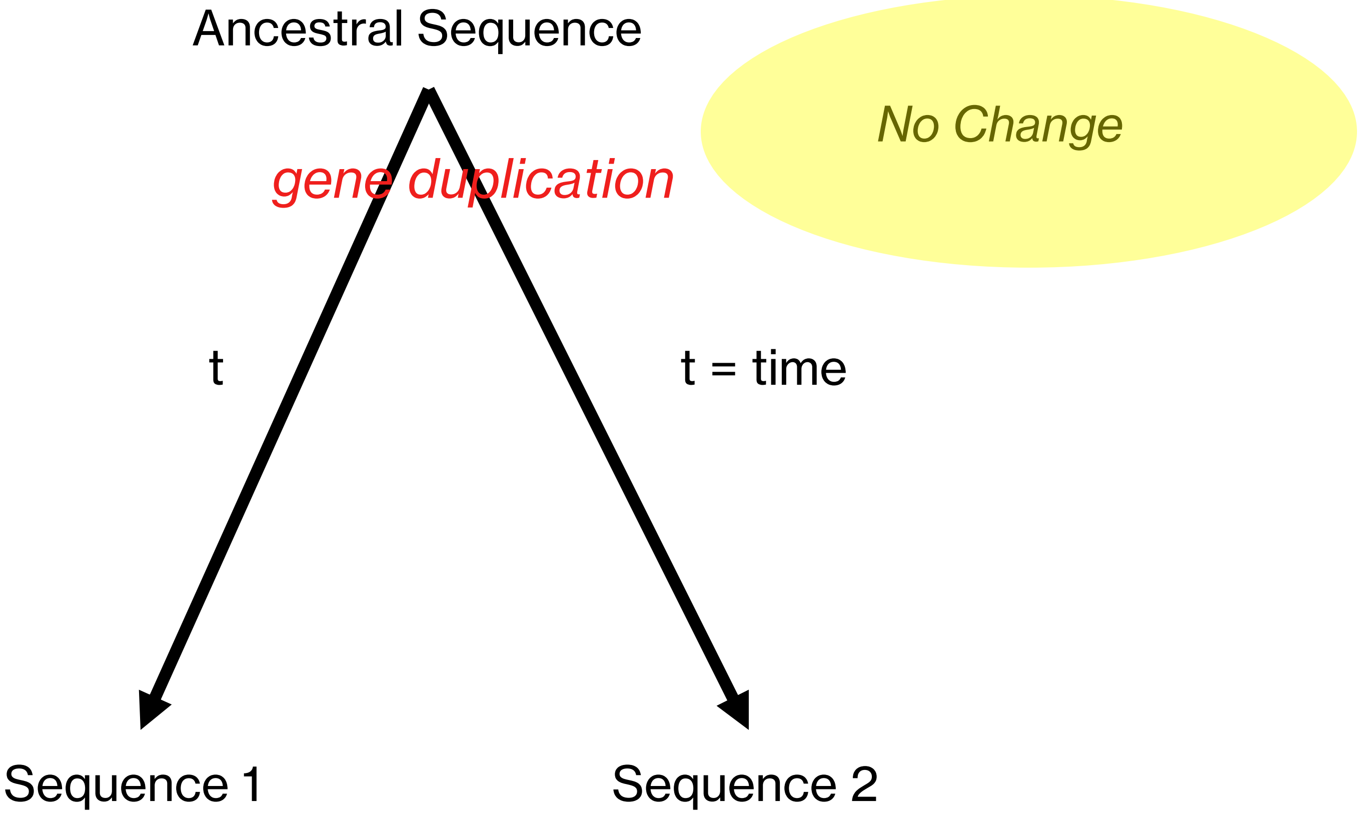
t

t = time

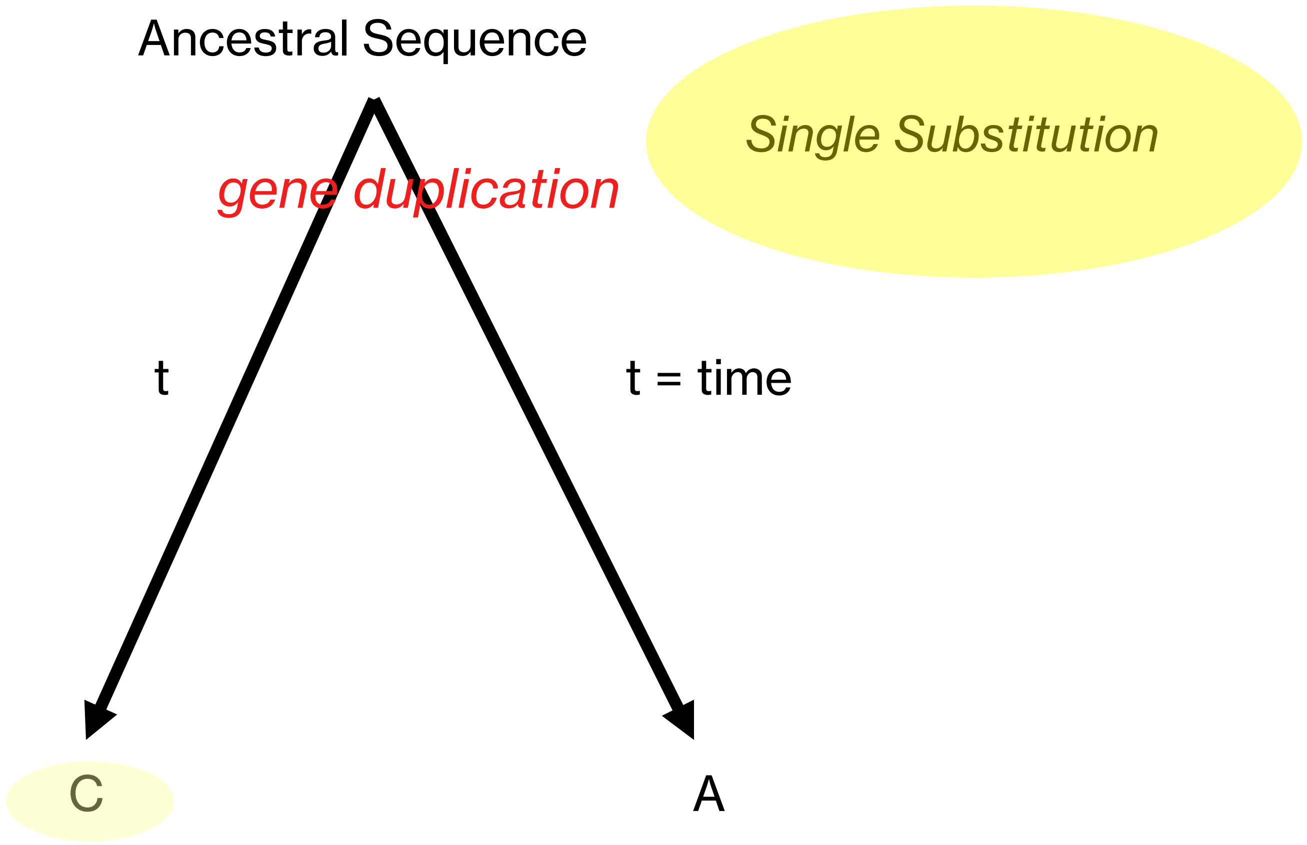
Sequence 1

Sequence 2

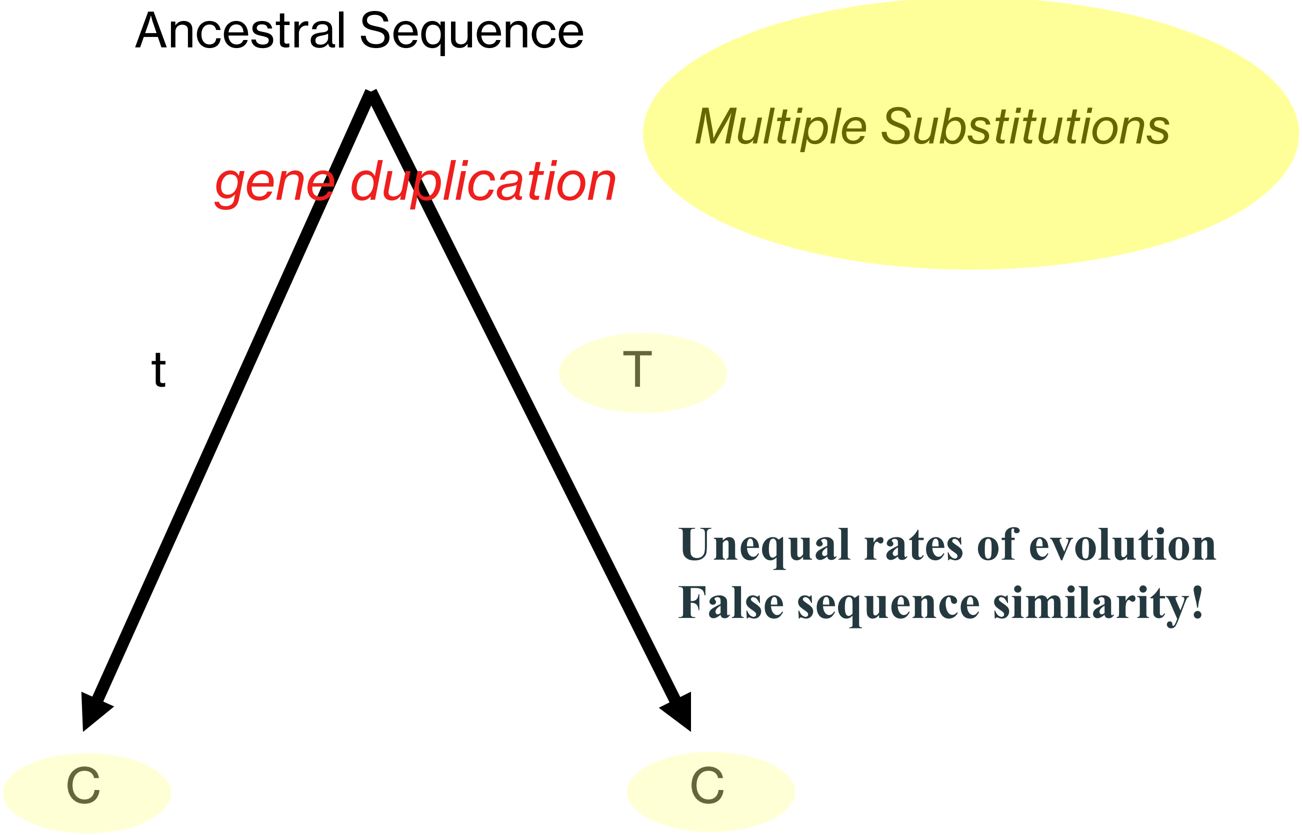
Substitution & Rates of Evolution



Substitution & Rates of Evolution



Substitution & Rates of Evolution



Unequal rates of evolution
False sequence similarity!

If rates of evolution were slow and equal, change would be rare and phylogenetics would be much easier. As they are not, one of the primary issues of phylogenetics is to determine history despite multiple substitution.

Key Issues of Phylogenetics

- How do we find homologous sites?
- How do we model substitution?
- How do we search for the best tree?

Key Issues of Phylogenetics

- How do we find homologous sites?
- How do we model substitution?
- How do we search for the best tree?

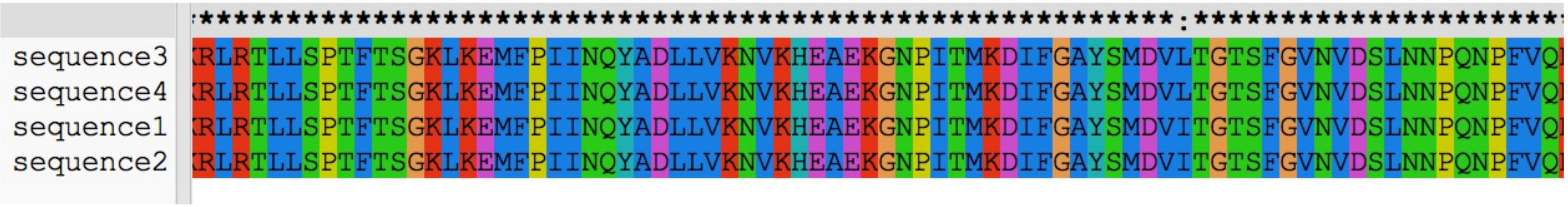
HOMOLOGOUS

A homologous trait is shared between two species because they inherited it from a common ancestor.

Information from homologous traits can be used to infer evolutionary relationships.

Non-homologous traits do not reflect evolutionary history but instead convergence. They can mislead inference of evolutionary relationships. Example: octopus eye versus human eye.

Multiple Sequence Alignment



```
          :********************************************************:*********************:
sequence3 KRLRTLLSPTFTSGKLKEMFPIINQYADLLVKNVKHEAEKGNPITMKDIFGAYSMDVLTGTSFGVNVDSLNNPQNPFVQ
sequence4 KRLRTLLSPTFTSGKLKEMFPIINQYADLLVKNVKHEAEKGNPITMKDIFGAYSMDVLTGTSFGVNVDSLNNPQNPFVQ
sequence1 KRLRTLLSPTFTSGKLKEMFPIINQYADLLVKNVKHEAEKGNPITMKDIFGAYSMDVITGTSFGVNVDSLNNPQNPFVQ
sequence2 KRLRTLLSPTFTSGKLKEMFPIINQYADLLVKNVKHEAEKGNPITMKDIFGAYSMDVITGTSFGVNVDSLNNPQNPFVQ
```

Alignment of highly similar sequences is easy – each column is assumed to reflect a homologous position in the protein

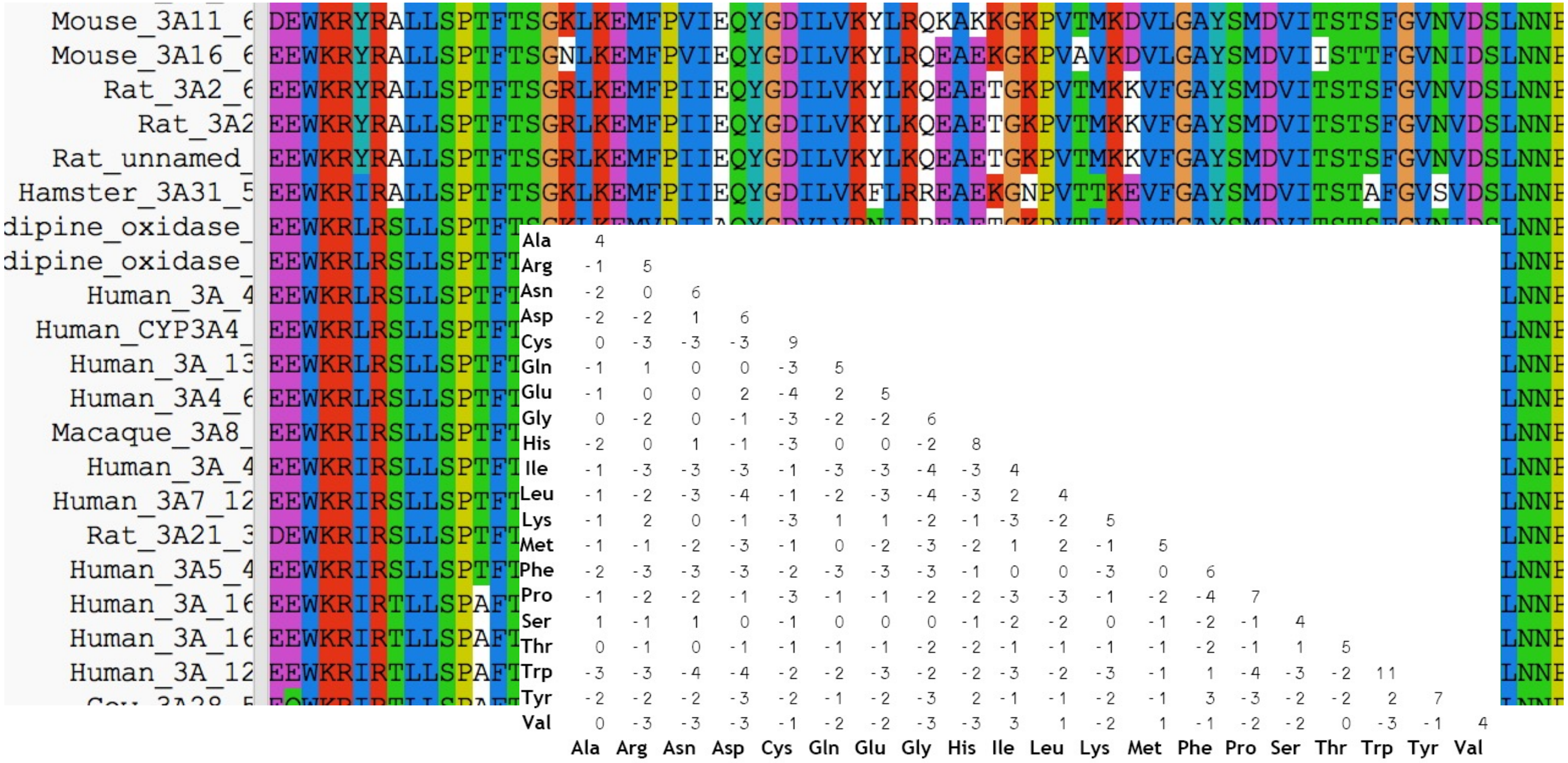
Blocks of identical sequence

Gaps infrequent

Substitutions conservative (i.e. leucine versus isoleucine)

Clustal software very effective in generating alignments

Multiple Sequence Alignment

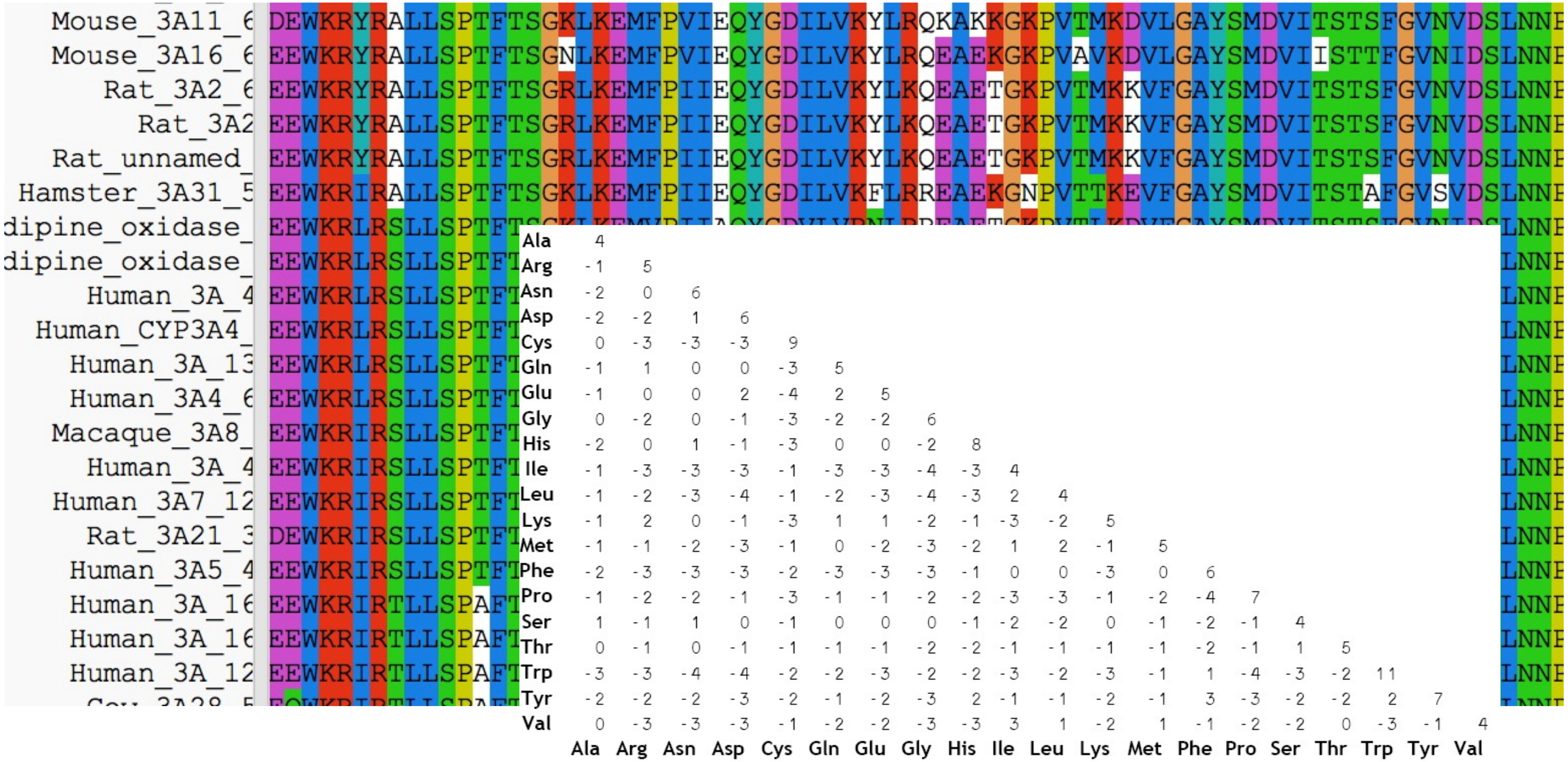


```
    Mouse_3A11_6 DEWKRYRALLSPTFTSGKLKEMFPVIEQYGDILVKYLRQKAKKGKPVTMKDVLGAYSMDVITSTSFGVNVDSLNNP
    Mouse_3A16_6 EEWKRYRALLSPTFTSGNLKEMFPVIEQYGDILVKYLRQEAEKGKPVAVKDVLGAYSMDVIISTTFGVNIDSLNNP
       Rat_3A2_6 EEWKRYRALLSPTFTSGRLKEMFPIIEQYGDILVKYLKQEAETGKPVTMKKVFGAYSMDVITSTSFGVNVDSLNNP
         Rat_3A2 EEWKRYRALLSPTFTSGRLKEMFPIIEQYGDILVKYLKQEAETGKPVTMKKVFGAYSMDVITSTSFGVNVDSLNNP
    Rat_unnamed_ EEWKRYRALLSPTFTSGRLKEMFPIIEQYGDILVKYLKQEAETGKPVTMKKVFGAYSMDVITSTSFGVNVDSLNNP
  Hamster_3A31_5 EEWKRIRALLSPTFTSGKLKEMFPIIEQYGDILVKFLRREAEKGNPVTTKEVFGAYSMDVITSTAFGVSVDSLNNP
dipine_oxidase_ EEWKRLRSLLSPTFT                                                            LNNP
dipine_oxidase_ EEWKRLRSLLSPTFT                                                            LNNP
      Human_3A_4 EEWKRLRSLLSPTFT                                                            LNNP
   Human_CYP3A4_ EEWKRLRSLLSPTFT                                                            LNNP
     Human_3A_13 EEWKRLRSLLSPTFT                                                            LNNP
     Human_3A4_6 EEWKRLRSLLSPTFT                                                            LNNP
    Macaque_3A8_ EEWKRIRSLLSPTFT                                                            LNNP
      Human_3A_4 EEWKRIRSLLSPTFT                                                            LNNP
     Human_3A7_12 EEWKRIRSLLSPTFT                                                           LNNP
      Rat_3A21_3 DEWKRIRSLLSPTFT                                                            LNNP
     Human_3A5_4 EEWKRIRSLLSPTFT                                                            LNNP
     Human_3A_16 EEWKRIRTLLSPAFT                                                            LNNP
     Human_3A_16 EEWKRIRTLLSPAFT                                                            LNNP
     Human_3A_12 EEWKRIRTLLSPAFT                                                            LNNP
```

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Just like BLAST, alignment software like Clustal uses BLOSUM, PAM, or other empirical substitution matrices to model conservative amino acid substitutions. Different matrices can be used for different evolutionary timescales.

Multiple Sequence Alignment



```
Mouse_3A11_6 DEWKRYRALLSPTFTSGKLKEMFPVIEQYGDILVKYLRQKAKKGKPVTMKDVLGAYSMDVITSTSFGVNVDSLNNP
Mouse_3A16_6 EEWKRYRALLSPTFTSGNLKEMFPVIEQYGDILVKYLRQEAEKGKPVAVKDVLGAYSMDVIISTTFGVNIDSLNNP
   Rat_3A2_6 EEWKRYRALLSPTFTSGRLKEMFPIIEQYGDILVKYLKQEAETGKPVTMKKVFGAYSMDVITSTSFGVNVDSLNNP
     Rat_3A2 EEWKRYRALLSPTFTSGRLKEMFPIIEQYGDILVKYLKQEAETGKPVTMKKVFGAYSMDVITSTSFGVNVDSLNNP
Rat_unnamed_ EEWKRYRALLSPTFTSGRLKEMFPIIEQYGDILVKYLKQEAETGKPVTMKKVFGAYSMDVITSTSFGVNVDSLNNP
Hamster_3A31_5 EEWKRIRALLSPTFTSGKLKEMFPIIEQYGDILVKFLRREAEKGNPVTTKEVFGAYSMDVITSTAFGVSVDSLNNP
dipine_oxidase_ EEWKRLRSLLSPTFT
dipine_oxidase_ EEWKRLRSLLSPTFT
  Human_3A_4 EEWKRLRSLLSPTFT
Human_CYP3A4_ EEWKRLRSLLSPTFT
 Human_3A_13 EEWKRLRSLLSPTFT
 Human_3A4_6 EEWKRLRSLLSPTFT
Macaque_3A8_ EEWKRIRSLLSPTFT
  Human_3A_4 EEWKRIRSLLSPTFT
Human_3A7_12 EEWKRIRSLLSPTFT
  Rat_3A21_3 DEWKRIRSLLSPTFT
 Human_3A5_4 EEWKRIRSLLSPTFT
 Human_3A_16 EEWKRIRTLLSPAFT
 Human_3A_16 EEWKRIRTLLSPAFT
 Human_3A_12 EEWKRIRTLLSPAFT
```

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Clustal and related methods can handle up to ~70% sequence divergence when BLOSUM or other matrices are used. Insertion and deletions of amino acids (INDELs) are the challenge.

Divergence beyond 70% takes special algorithms, often using protein structure prediction to guide alignment. Important for generating seed alignments for HMMs!

. = same amino acid, * = conservative change, : = semi-conservative change
- = gap; SRS = substrate recognition site

```
Pufferfish 3B1   MFEFVLFSGTTWALLALFFALLLLYGVWPYHHFKKLGIRGPRPLPFMGSTFYYRKGIIPFESWCQAEYGDVWGMFEGRTPVLMVSDPEILKTVLVKECYS
Killifish 3A30   .G-YFYLTAE..T..VA.VT...V.AY...GT..R...S..K.V..F.TMLH..R.FFT.DEE.KKK..K...IYD..Q...C.T....I.A......L.
Pig 3A29         .DLIPG..TE..V...TSLV..Y...TYSHGL.....P......YF.NILG....VDH.DKK.FQQ..KM..VYD..Q.L.A.T..NMI.S........
Human 3A4        .ALIPDLAME..L...VSLV..Y...THSHGL.....P..T....L.NILS.H..FCM.DME.HKK..K...FYD.QQ...AIT..DMI..........
Rat 3A18         .EIIPNL.IE..V...TSLM.FYI..TYSHGL.....P..K.V.LF.TI.N.GD.MWK.DDD.YKK..KI..FY..PQ.F.AIM....I.M........
Rat 3A1          .DLLSALTLE..V...VVLV..YGF.TRTHGL...Q..P..K....F.TVLN.YM.LWK.DVE.HKK..KI..L.D.QM.LFAIT.T.MI.N......F.
                 *     ::  *  ::     *:          * : *: ** * *  *       *    :  *  :*   ** ::   *   : :   :: :*:*:*

                SRS 1
Pufferfish 3B1   VFTNRRD-SFAGPLEDSVSAVKDERWKRIRSTISPCFTSGRLKNAFPIVARYADRITKKLE-QSNLDEPINVKEFLAPYSLDAVTSVSFSVEADSINNPN
Killifish 3A30   F.....NFRLN...Y.A..IAE.DQ......VL..S.......EM.E.MKNHSANLIRSMKKKADK...LDL...FGS..M.V...TA...DI..L...S
Pig 3A29         ......SFGPL.AMRNAL.LAE..E.....TLL..T....K..EM...ISH.G.LLVSN.RKEAEKGK.VTM.DIFGA..M.VI..TA.G.NI..L...Q
Human 3A4        ......PFGPV.FMKSAI.IAE..E...L..LL..T....K..EMV..I.Q.G.VLVRN.RREAETGK.VTL.DVFGA..M.VI..T..G.NI..L...Q
Rat 3A18         ......CFGPM.FMKKAITMSE..E...L.TIL..T....K..EM..LMRQ.G.TLL.N.RREEAKG....M.DIFGA..M.VI.GT..G.NV..L...Q
Rat 3A1          .......FGPV.IMGKA..VA...E...Y.ALL..T.......EM...IEQ.G.ILV.Y.KQEAETGK.VTM.KVFGA..M.VI..T..G.NV..L...K
                  ***:      * :  ::    :: *:* *  :*  **:  :::    :      :   :            *  :  * :* :    * *  **: :*

                   SRS 2                            SRS 3                                                   SRS 4
Pufferfish 3B1   DPLIVNLKKVFK-FNFVVFFLVAFFPFCARLFQFLGIDPIPRSSVNYFYNVIKNFKDQHHADTRG---DFLQVLIQSEIP--QSEIKSPKGLTEHEILSQ
Killifish 3A30   ..FVT.I..ML.D.LNPL.LA......LGPILEKFELSFF.K.VTDF..ASLEKI.SNRE.SQQKSRV....LM.D.Q--KNS-GAQQD.S..D......
Pig 3A29         ..FVE.S..LL.S.FDPFLLSLI....LTPI.EV.N.TLF.K....F.TKSV.RM.ESRLT.QQKRRV.L..LM.N.Q---NSK.MDPH.S.SNE.LVA.
Human 3A4        ..FVE.T..LLRD.LDPF.LSITV...LIPILEV.N.CVF..EVT.FLRKSV.RM.ESRLE..QKHRV....LM.D.QK--NSK.TE.H.A.SDL.LVA.
Rat 3A18         ..FVQKA..IL.QIFDPFLLS.VL...LTPIYEM.NFSIF..Q.M.F.KKFV.TM.KNRLDSNQKNRV....LMMNTQ---NSKGQE.Q.A.SDL.MAA.
Rat 3A1          ..FVEKT..LLRD.FDPL.LS.VL...LTPIYEM.N.CMF.KD.IEF.KKFVYRM.ETRLDSVQKHRV....LMMNAHN--DSKDKE.HTA.SDM..TA.
                   ::                     *:   :   :    :*     ::   :      :         :::  ::                :  :  *:  *
                 ----F helix-----            ----------G helix----------              --H helix-          --I helix--

                                                                                     SRS 5
Pufferfish 3B1   AFIFIFGGYETTTTTLTNVLYGLAINPDVLQVLHKEIDTNIPSDAPISYEDLMGLQYLDQVLNESQRLYPTAPRLERACKKTVQIHGLTILEGTIVGIPV
Killifish 3A30   SM....A....SSSS..FLA.N..T..E.MKK.QE...ATF.NK..VH.QP..EME...C.I...L..F.I.A....VA.AA.E.N.VV.PKDMV.M..T
Pig 3A29         GI....A.....SSA.SLLA.E..TH...Q.K.QE..EATF.NK..PT.DA.AQME...M.V..TL....I.A........D.E...VFVPK..V.VV..
Human 3A4        SI....A.....SSV.SFIM.E..TH...Q.K.QE...AVL.NK..PT.DTVLQME...M.V..TL..F.I.M....V...D.E.N.MF.PK.WV.M..S
Rat 3A18         .I.......DA.S.SISFIM.E..TR.N.QKK.QN...RAL.NK..VT.DA..EME...M.V...L....I.T..D.VS..D.E.N.VF.PK..V.T..I
Rat 3A1          SI....A...P.SS..SF..HS..TH..TQKK.QE...RAL.NK..PT.DTV.EME...M....TL....IGN....V...D.E.N.VFMPK.SV.M..S
                    ::   *: ::  :    : *:        :  **:  :*         :  ::***  :*   *:      : *  *   :: *    :   * :*
                 ---------------I helix---------------

                                                                                              SRS 6
Pufferfish 3B1   HLLHKDPRFWSSPEEFRPERFSKDSTEEVNPYAFMPFGLGPRNCVGMRYAILVMKMLIVRLLQSYTVETCKDTMIPLEFDWK--SQPLKPIKLSFIPRQK
Killifish 3A30   WP..R..EI.PE..A.K......KNKDNID..IY....S.....I...F.LVLI.LAV.EI..Q.SFSV..E.EV.F.M.IQGLLA.KR..Q.KLV..S
Pig 3A29         FV..R..DL.PE...........KHKDTI...TYL...T.....I...F.LMN..LAL..V..NFSFKP..E.Q...KLTTQGLT..E..VV.KIL..DG
Human 3A4        YA..R..KY.TE..K.L......KNKDNID..IYT...S.....I...F.LMN..LALI.V..NFSFKP..E.Q...KLSLGGLL..E..VV.KVES.DG
Rat 3A18         YP..RN.EY.LE....N......ENKGSID..VYL...N.....I...F.LIS..LAVIGV..NFNIQP.EK.Q...KISRQPIF..EG..I.KLVS.D
Rat 3A1          YA..R..QH.PE...........ENKGSID..VYL...N.....I...F.LMN..LALTKV..NFSFQP..E.Q...KLSRQGLL..T...I.KVV..DE
                   ::     *  *: * *****:      : * : *** *****:* * *:  :*: :   :* :    *  *   : :           : *    *
                                                    --Heme binding--
```

Conservative & semi-conservative labels based on alignment of 50+ sequences

Multiple Sequence Alignment

- Alignment involves appropriate substitution models
- Alignment of divergent sequences may require structural constraints (e.g. rRNA folding) or special algorithms (e.g. MUSCLE)
- Any alignment may have sub-sections that are poorly aligned and should be removed from phylogenetic analyses as HOMOLOGY is uncertain (i.e. not sure if all amino acids in the column reflect the same ancestral position in the protein).

Key Issues of Phylogenetics

- How do we find homologous sites?
- How do we model substitution?
- How do we search for the best tree?

Modeling substitution and finding the best tree are intertwined in a concept called the “optimality criteria” – the philosophical / mathematical / computational framework you use to find the best (aka optimal) phylogenetic tree

Optimality Criteria

PARSIMONY

DISTANCE METHODS
(aka neighbour-joining, minimum evolution)

MAXIMUM LIKELIHOOD

BAYESIAN INFERENCE

Optimality Criteria

DISTANCE METHODS
(aka neighbour-joining, minimum evolution) →

Simplify a multiple sequence alignment to a distance matrix

Loss of information; unreliable method

But it is very fast, so often used as a first-pass estimate (e.g. our BLAST detection of a pmr contaminant)

BEWARE publications with neighbour-joining trees!

Optimality Criteria

PARSIMONY

DISTANCE METHODS
(aka neighbour-joining, minimum evolution)

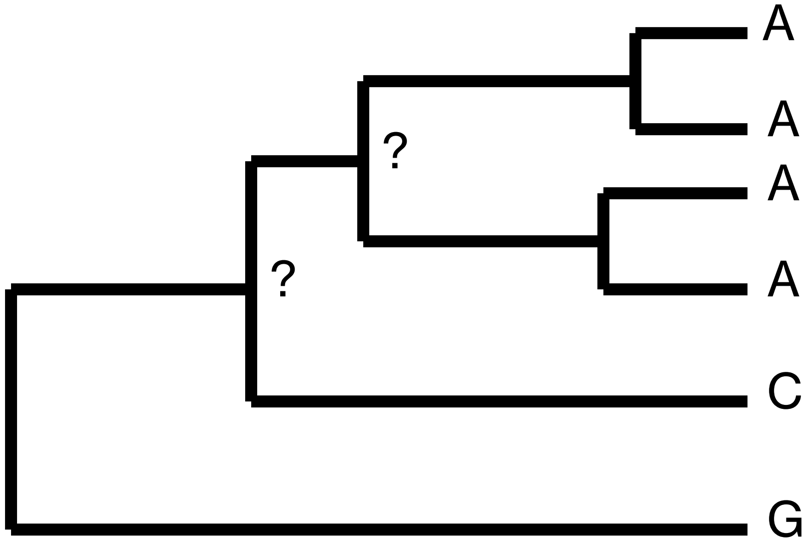
MAXIMUM LIKELIHOOD

BAYESIAN INFERENCE

Ignoring Multiple Substitutions: Parsimony

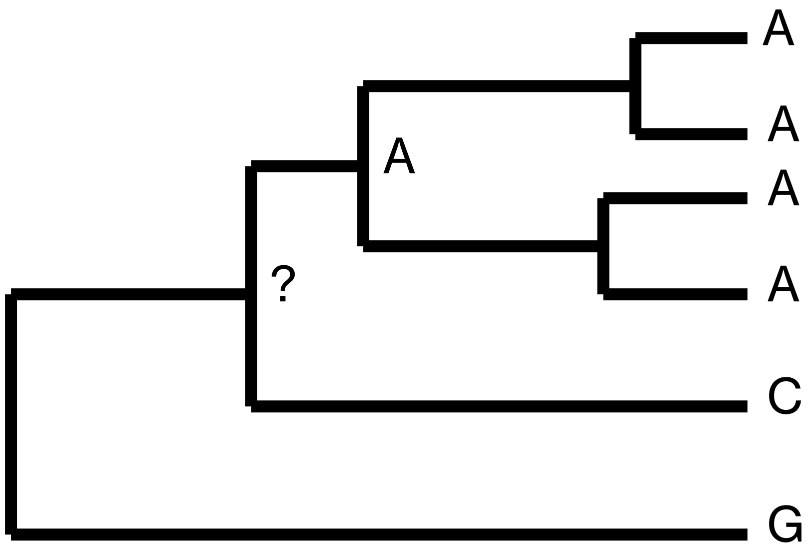
Parsimony is Occam’s Razor, i.e. the simplest explanation is the easiest

Parsimony seeks to minimize the number of substitutions



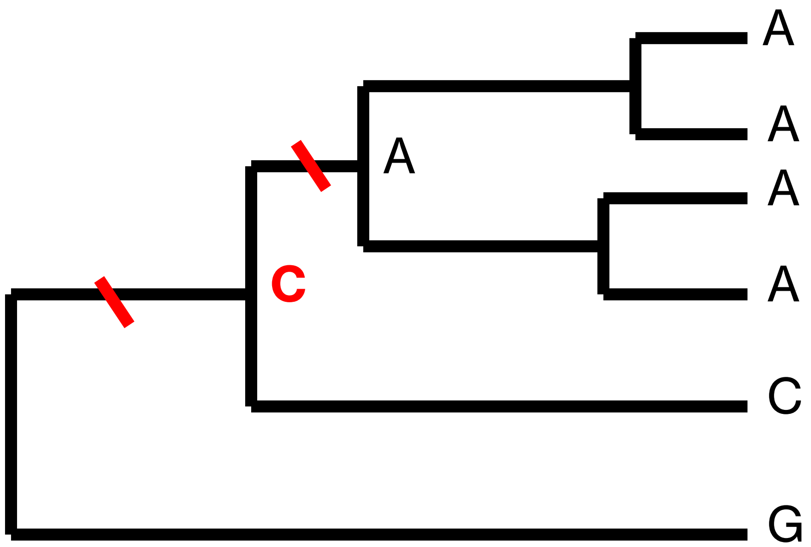
Ignoring Multiple Substitutions: Parsimony

No changes needed to explain the adenosines



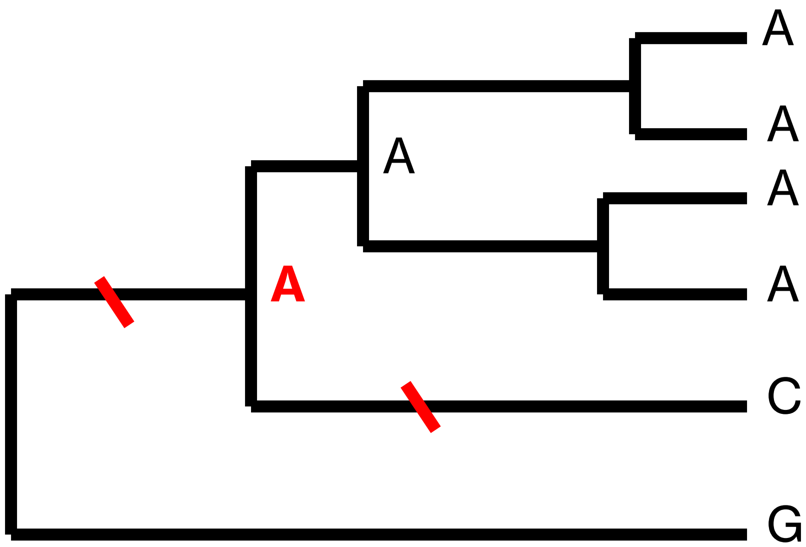
Ignoring Multiple Substitutions: Parsimony

Score: two changes



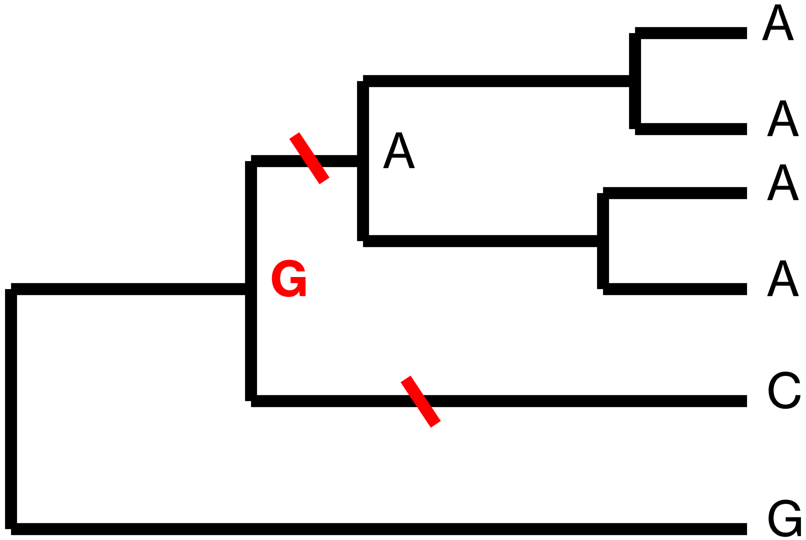
Ignoring Multiple Substitutions: Parsimony

Score: two changes



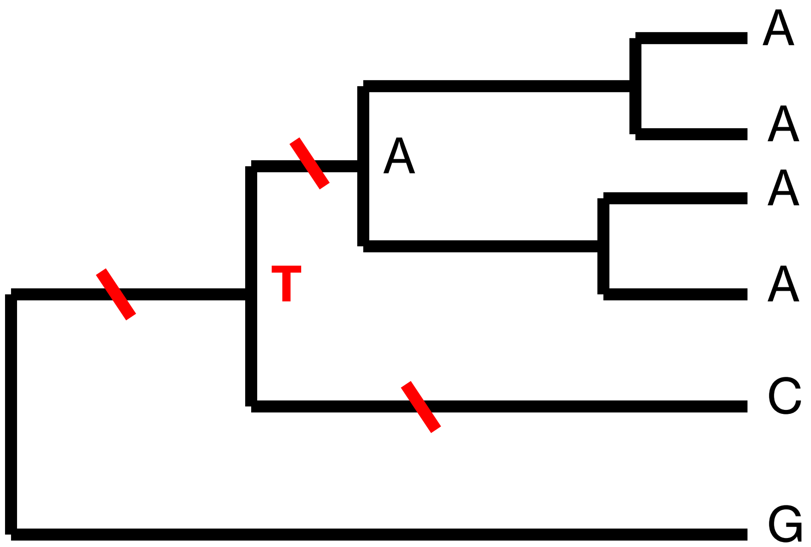
Ignoring Multiple Substitutions: Parsimony

Score: two changes



Ignoring Multiple Substitutions: Parsimony

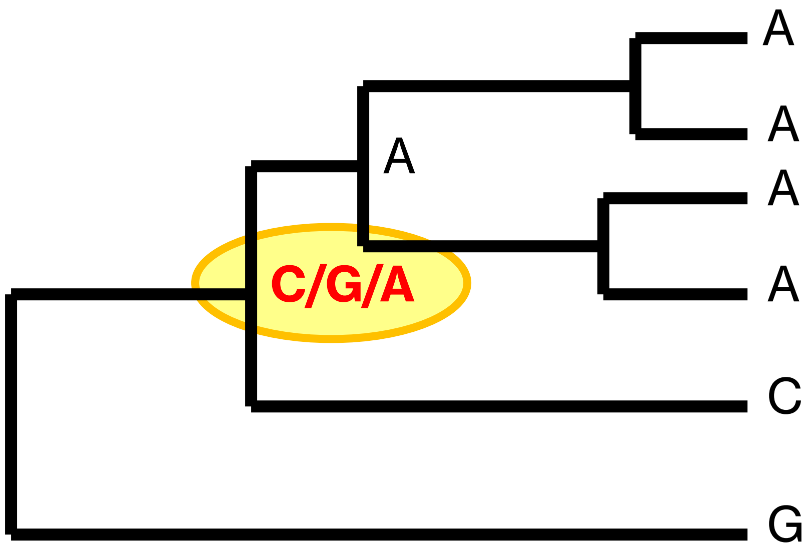
Score: three changes



Ignoring Multiple Substitutions: Parsimony

NO SUBSTITUTION MODEL!
3 of 4 possible bases are equally parsimonious!
LACK OF RESOLUTION!

Parsimony is fast since it only needs to count change but it deals poorly with multiple substitutions or unequal rates of evolution – it often lacks resolution



Optimality Criteria

PARSIMONY

DISTANCE METHODS
(aka neighbour-joining, minimum evolution)

MAXIMUM LIKELIHOOD

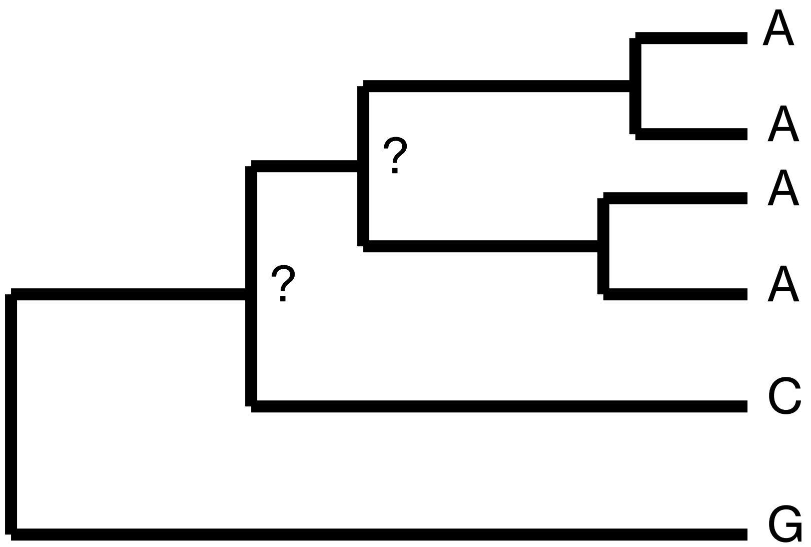
BAYESIAN INFERENCE

Accounting for Multiple Substitutions: Likelihood

Maximum likelihood (ML) does not apply Occam's Razor

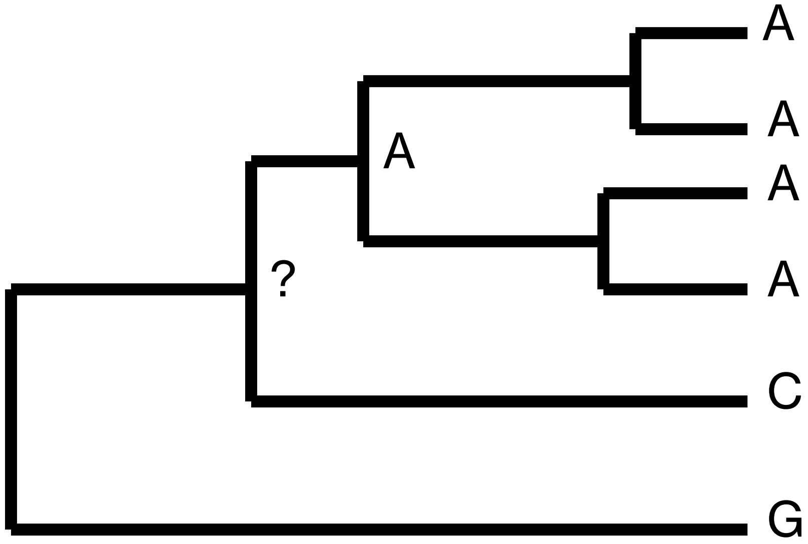
ML uses substitution models to better predict which changes have occurred

In ML, the length of the branches is as important as the shape of the tree



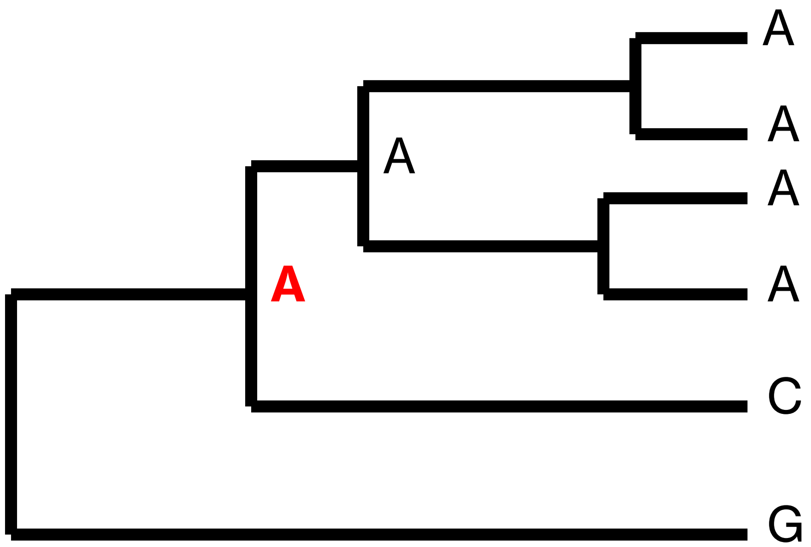
Accounting for Multiple Substitutions: Likelihood

No changes needed to explain the adenosines



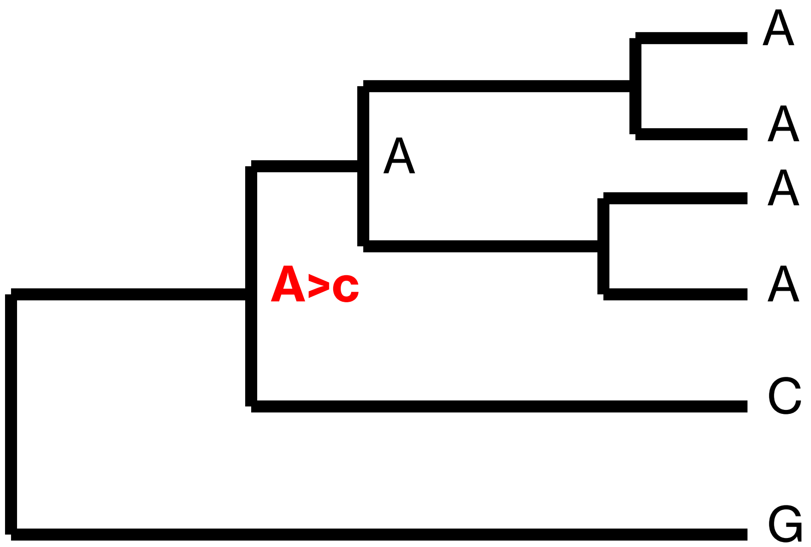
Accounting for Multiple Substitutions: Likelihood

A is the most likely as it is the shortest route



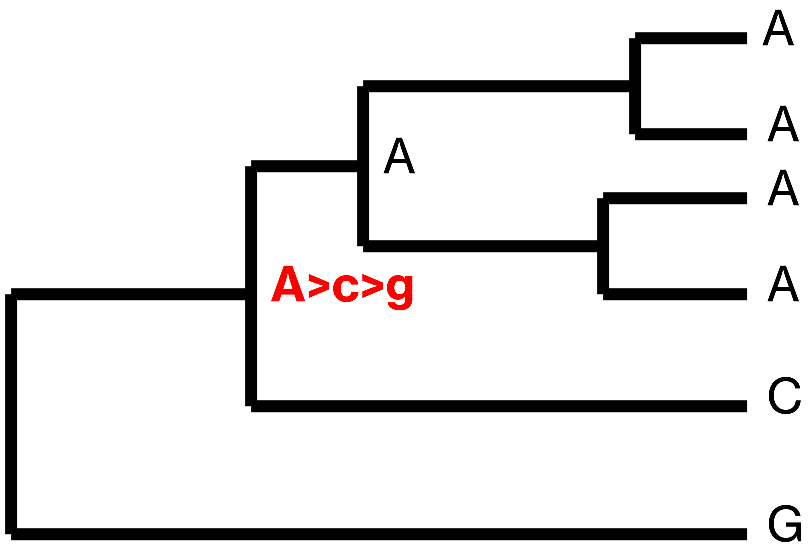
Accounting for Multiple Substitutions: Likelihood

C is the next most likely as it is the next shortest route



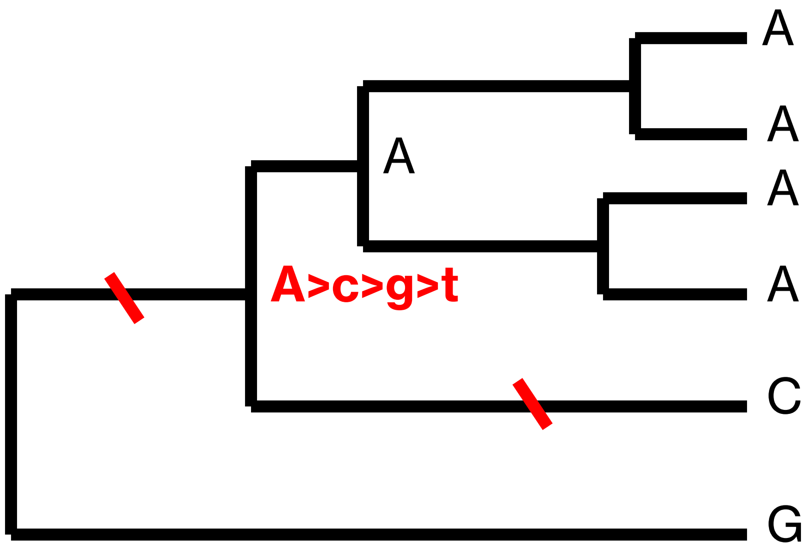
Accounting for Multiple Substitutions: Likelihood

G is the third most likely as it is the third shortest route



Accounting for Multiple Substitutions: Likelihood

T is the least likely as there is no route

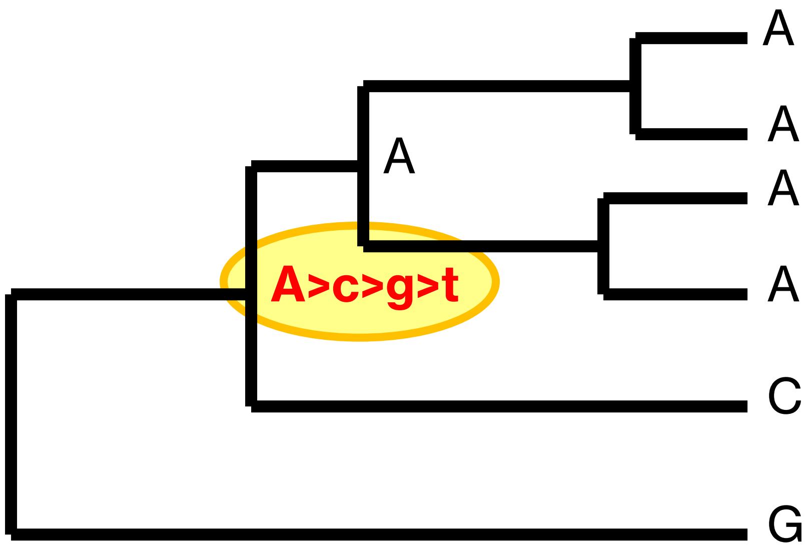


Accounting for Multiple Substitutions: Likelihood

BRANCH LENGTHS (e.g. shortest route)
COME FROM A SUBSTITUTION MODEL

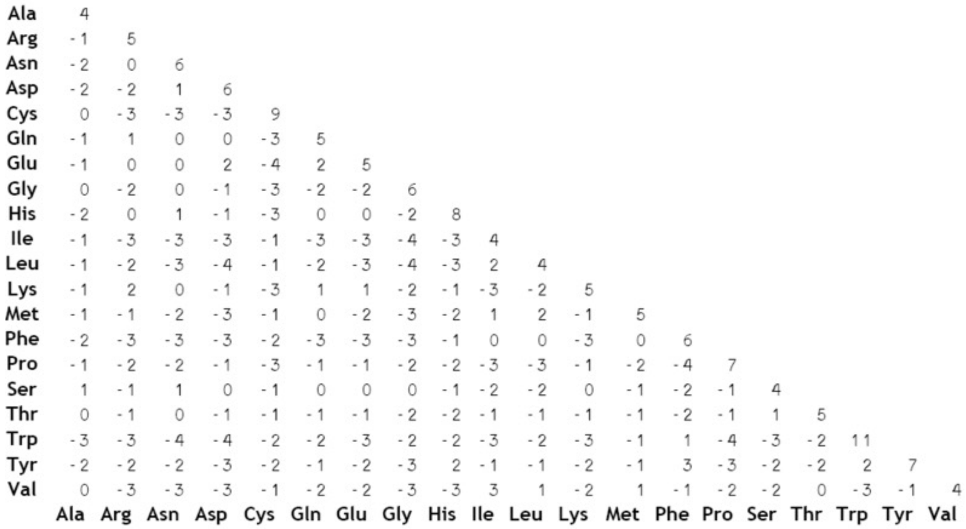
A SUBSTITUTION MODEL ADDS
RESOLUTION!

Each possible solution has a likelihood and must be considered – ML is more complex, slower, but more accurate



Substitution models

Empirical



	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164-5.

jModelTest: phylogenetic model averaging. Molecular Biology and Evolution 25: 1253-1256.

Theoretical

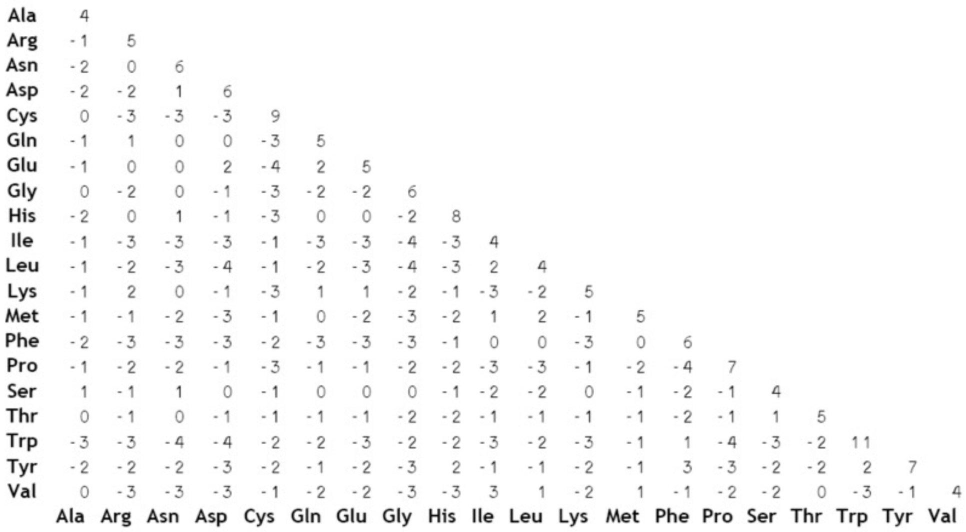
	A	C	G	T
A		$\mu\pi_C a$	$\mu\pi_G b$	$\mu\pi_T c$
C			$\mu\pi_G d$	$\mu\pi_T e$
G				$\mu\pi_T f$
T				

Estimation based on the following parameters

- A measure of mutation rates
- The relative amount of each base available
- Bias

Substitution models

Empirical



	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Theoretical

	A	C	G	T
A		$\mu\pi_C a$	$\mu\pi_G b$	$\mu\pi_T c$
C			$\mu\pi_G d$	$\mu\pi_T e$
G				$\mu\pi_T f$
T				

There are wide diversity of substitution models that can be used and you can design your own. Key aspects:

- Unequal rates of substitution
- Unequal amino acid / nucleotide frequencies

Key Issues of Phylogenetics

- How do we find homologous sites?
- How do we model substitution?
- How do we search for the best tree?

The “best tree” is a function of the data (i.e. multiple sequence alignment) and substitution model.

bad alignment = bad tree

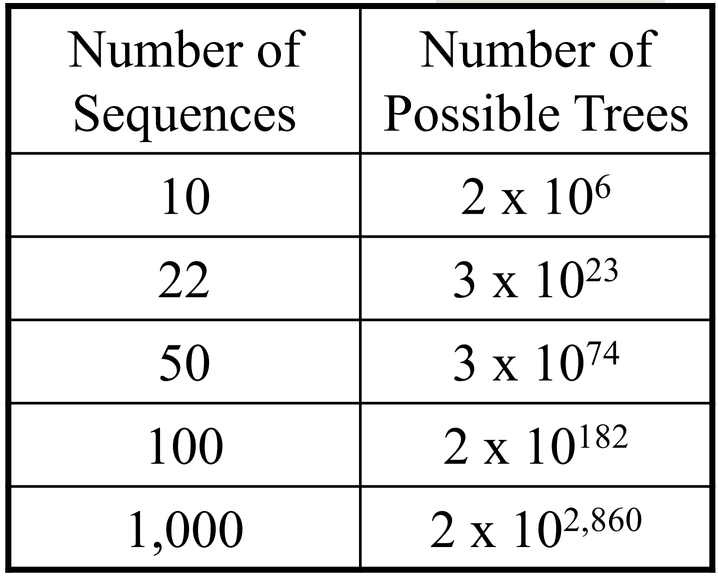
bad model = bad tree

Key Issues of Phylogenetics

- The “search space” in phylogenetics is huge! We cannot search all of it.

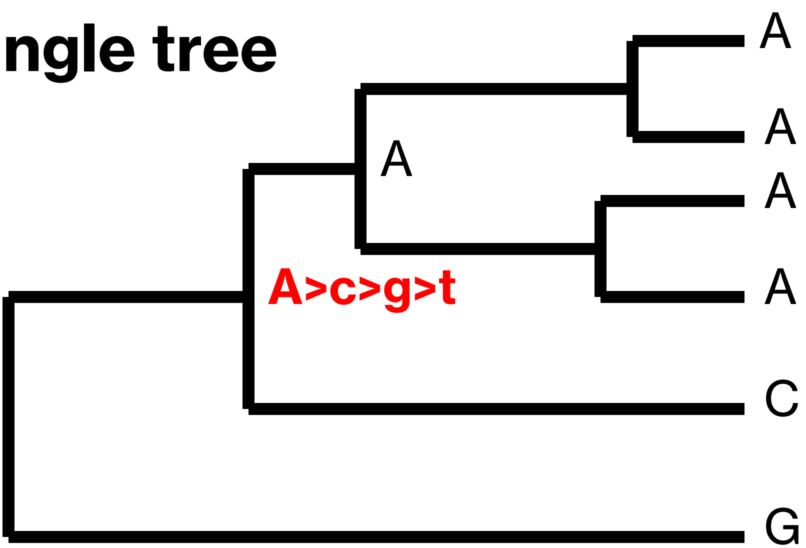
- Yet we need to find the tree with the best “score”

- Two tasks:
- Scoring trees
- Searching trees



Number of Sequences	Number of Possible Trees
10	2 x 10^{6}
22	3 x 10^{23}
50	3 x 10^{74}
100	2 x 10^{182}
1,000	2 x $10^{2,860}$

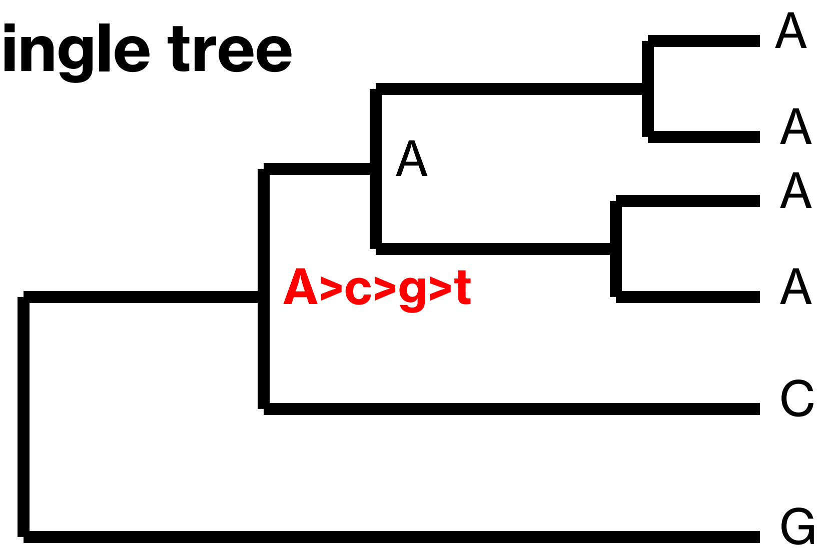
Scoring a single tree



	A	C	G	T
A		$\mu\pi_C a$	$\mu\pi_G b$	$\mu\pi_T c$
C			$\mu\pi_G d$	$\mu\pi_T e$
G				$\mu\pi_T f$
T				

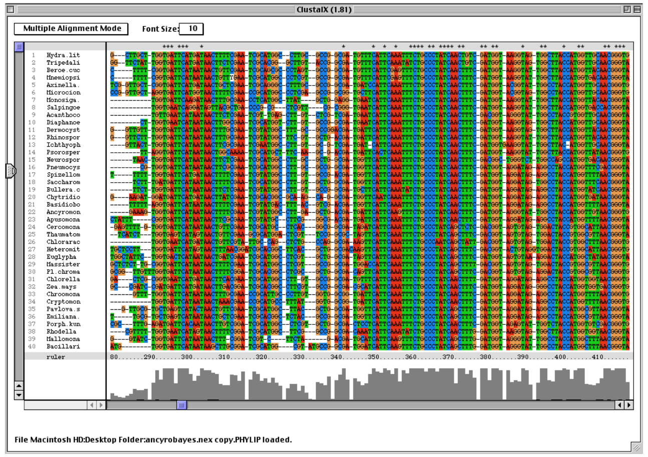
SCORE

Scoring a single tree



SCORE

	A	C	G	T
A		$\mu\pi_C a$	$\mu\pi_G b$	$\mu\pi_T c$
C			$\mu\pi_G d$	$\mu\pi_T e$
G				$\mu\pi_T f$
T				

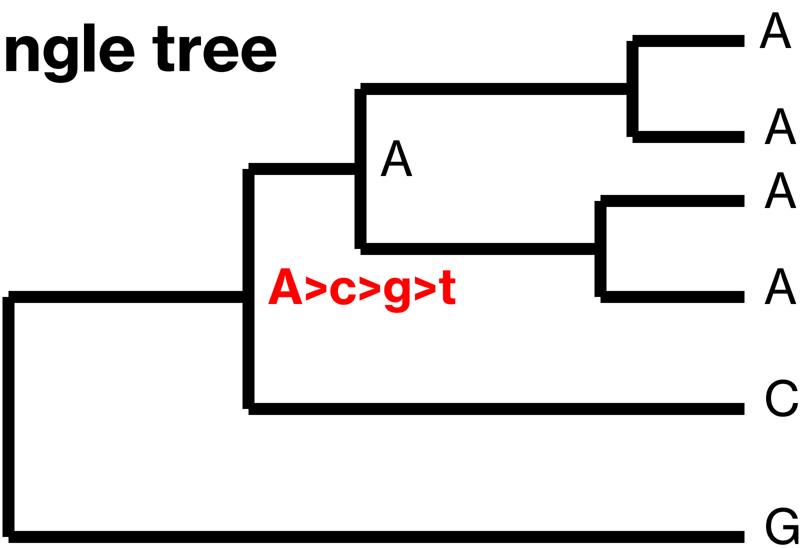


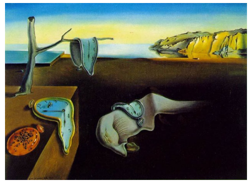
```
 1  Hydra.lit
 2  Tripedali
 3  Beroe.cuc
 4  Mnemiopsi
 5  Axinella.
 6  Microcion
 7  Monosiga.
 8  Salpingoe
 9  Acanthoco
10  Diaphanoe
11  Dermocyst
12  Rhinospor
13  Ichthyoph
14  Psorosper
15  Neurospor
16  Pneumocys
17  Spizellom
18  Saccharom
19  Bullera.c
20  Chytridio
21  Basidiobo
22  Ancyromon
23  Apusomona
24  Cercomona
25  Thaumatom
26  Chlorarac
27  Heteromit
28  Euglypha
29  Massister
30  Pl.chroma
31  Chlorella
32  Zea.mays
33  Chroomona
34  Cryptomon
35  Pavlova.s
36  Emiliana.
37  Porph.kun
38  Rhodella
39  Mallomona
40  Bacillari
    ruler     80.......290.......300.......310.......320.......330.......340.......350.......360.......370.......380.......390.......400.......410.......
```

File Macintosh HD:Desktop Folder:ancyrobayes.nex copy.PHYLIP loaded.

Each site (i.e. column) contributes to the score given the substitution model

Scoring a single tree



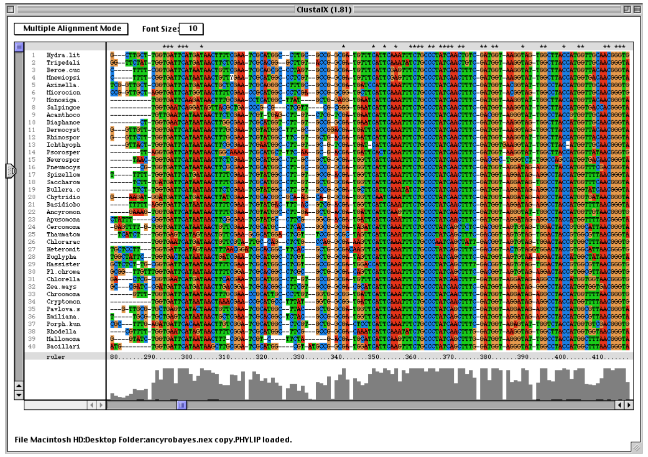


Among Site Rate Variation

Each site (i.e. column) does not contribute equally

SCORE

	A	C	G	T
A		$\mu\pi_C a$	$\mu\pi_G b$	$\mu\pi_T c$
C			$\mu\pi_G d$	$\mu\pi_T e$
G				$\mu\pi_T f$
T				

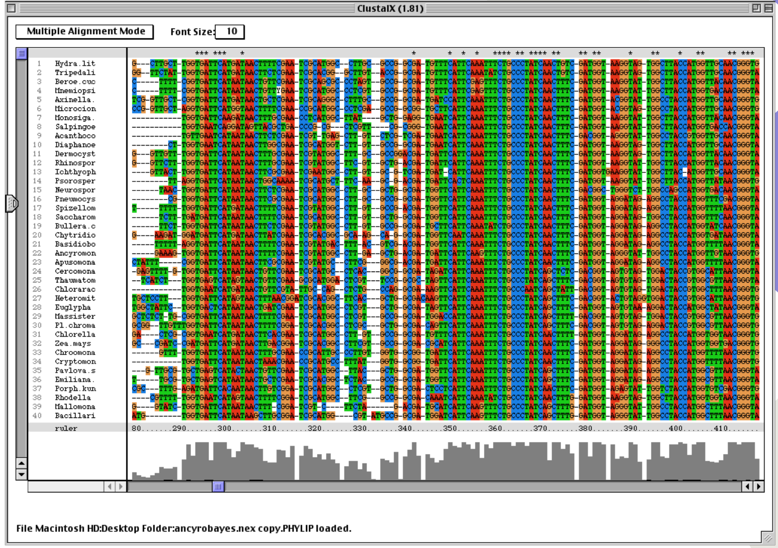


File Macintosh HD:Desktop Folder:ancyrobayes.nex copy.PHYLIP loaded.

Each site (i.e. column) contributes to the score given the substitution model

Among-site rate variation

- Different parts of a protein are under different evolutionary pressure will evolve at different rates
- For example a region with an important function (e.g. binding site, globular domains, structural regions, functional domains) will change more slowly than less essential regions
- Each column in the multiple sequence alignment thus has a faster or slower version of the substitution model



File Macintosh HD:Desktop Folder:ancyrobayes.nex copy.PHYLIP loaded.

Scoring Requires:

1. A tree topology to score
2. Multiple sequence alignment
3. Substitution model
    - Unequal rates of substitution among amino acids or nucleotides
    - Unequal frequencies of amino acids or nucleotides
4. Incorporation of among-site rate variation

If you can do all of this then you will end up with the correct tree

(e.g., with ML or Bayesian)

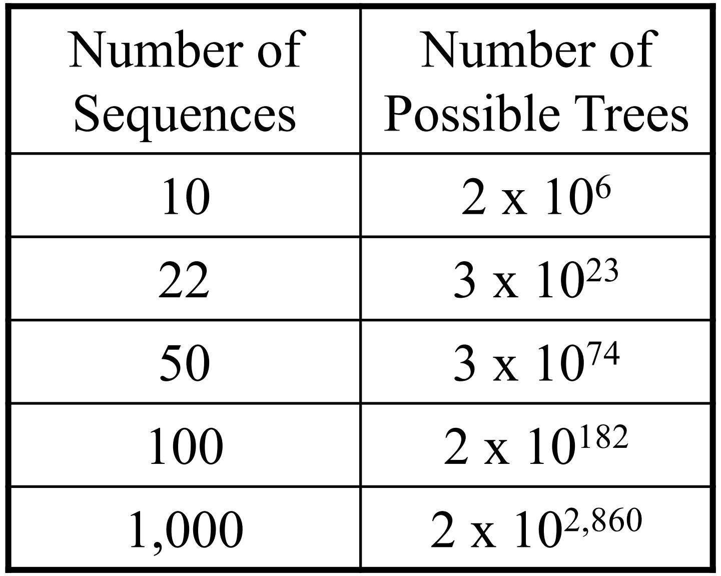
Scoring Requires:

1. A tree topology to score
2. Multiple sequence alignment
3. Substitution model
    - Unequal rates of substitution among amino acids or nucleotides
    - Unequal frequencies of amino acids or nucleotides
4. Incorporation of among-site rate variation

Caveats

- Phylogenetic models generally work well even though they are an over-simplification
- All methods assume neutrality or near-neutrality; positive selection is problematic
- The methods break down when your data has extremes, e.g. fundamentally different rules of evolution in different parts of your protein

How Do We Find the Best Tree?



Number of Sequences	Number of Possible Trees
10	2 x 10^{6}
22	3 x 10^{23}
50	3 x 10^{74}
100	2 x 10^{182}
1,000	2 x $10^{2,860}$

We can’t score them all – it would take far too long!

Instead, we use “tree space” search algorithms

Optimality Criteria

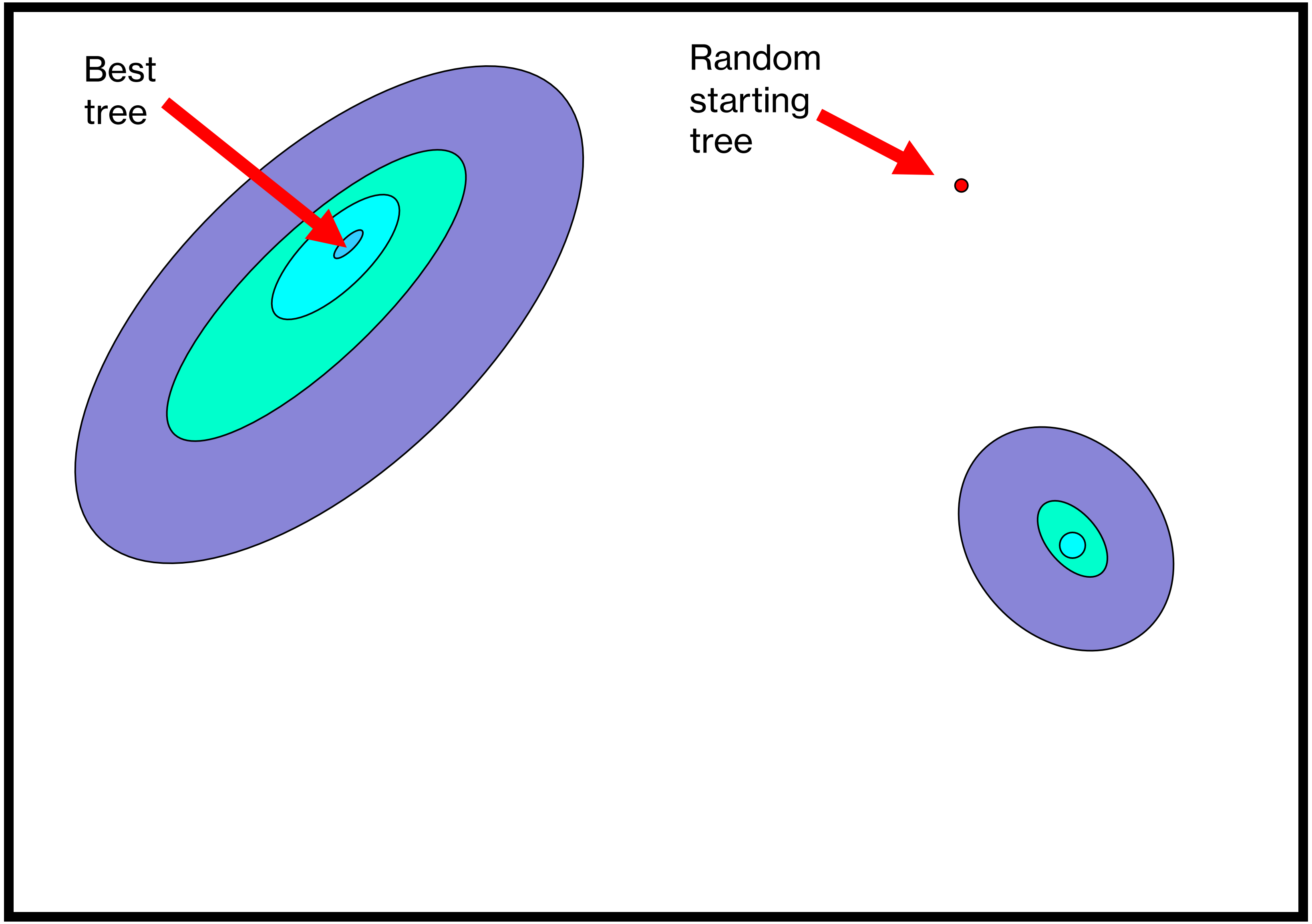
PARSIMONY

DISTANCE METHODS
(aka neighbour-joining, minimum evolution)

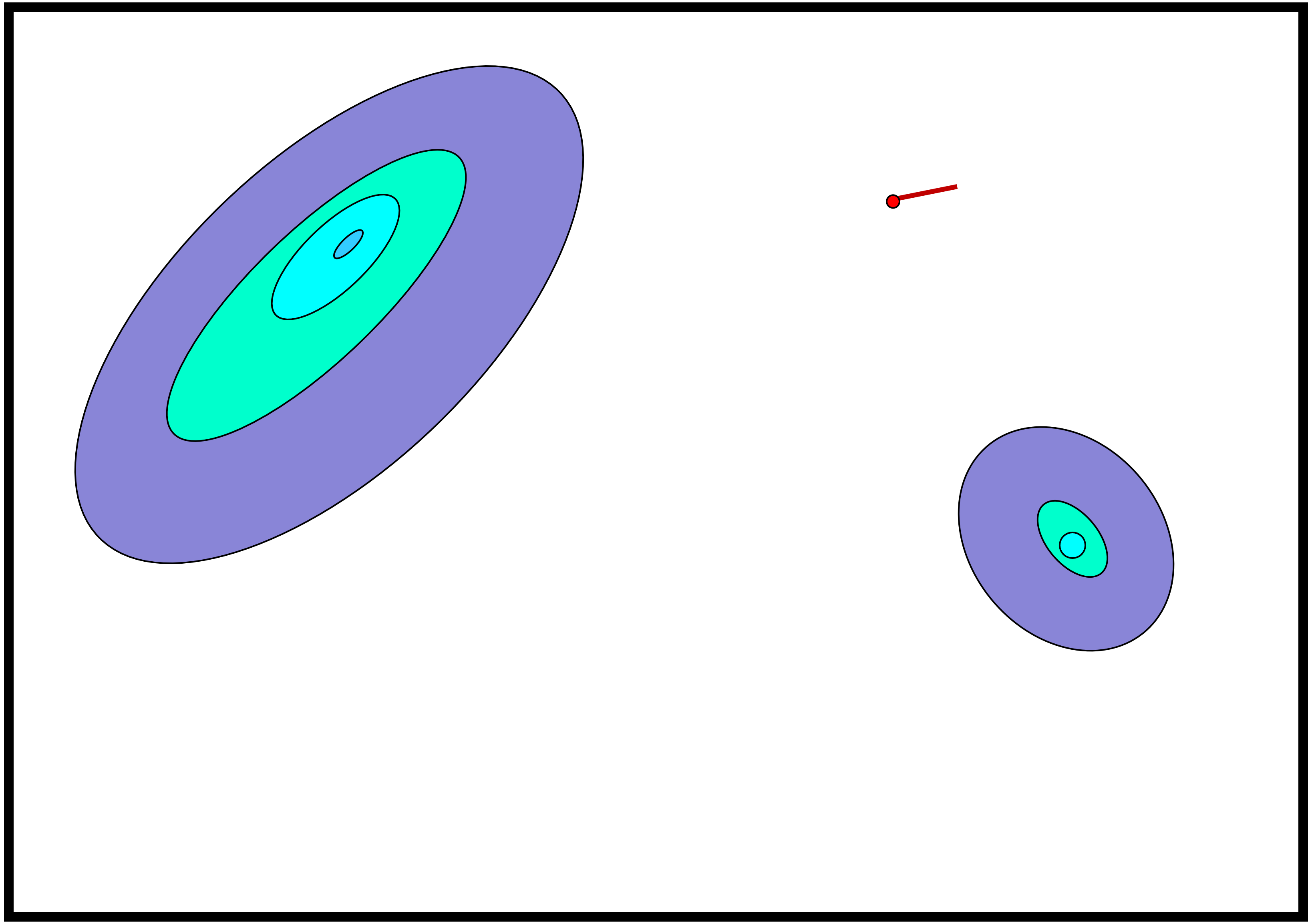
MAXIMUM LIKELIHOOD – branch swapping, find the best tree

BAYESIAN INFERENCE
– MC^3, sample the cloud of best trees
– Bayesian is an extension of maximum likelihood
– not going to cover this in Biochem 3BP3

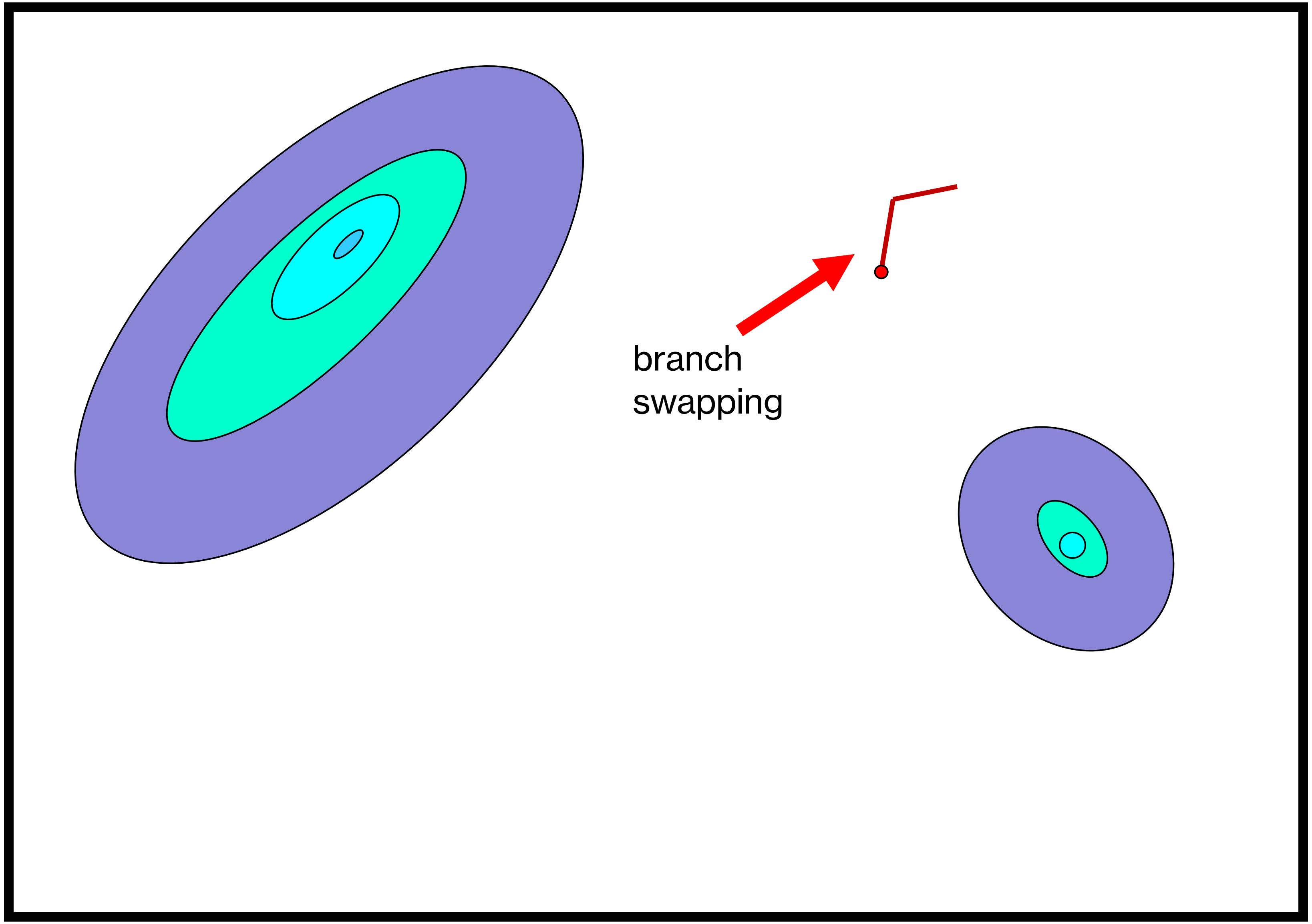
Branch Swapping



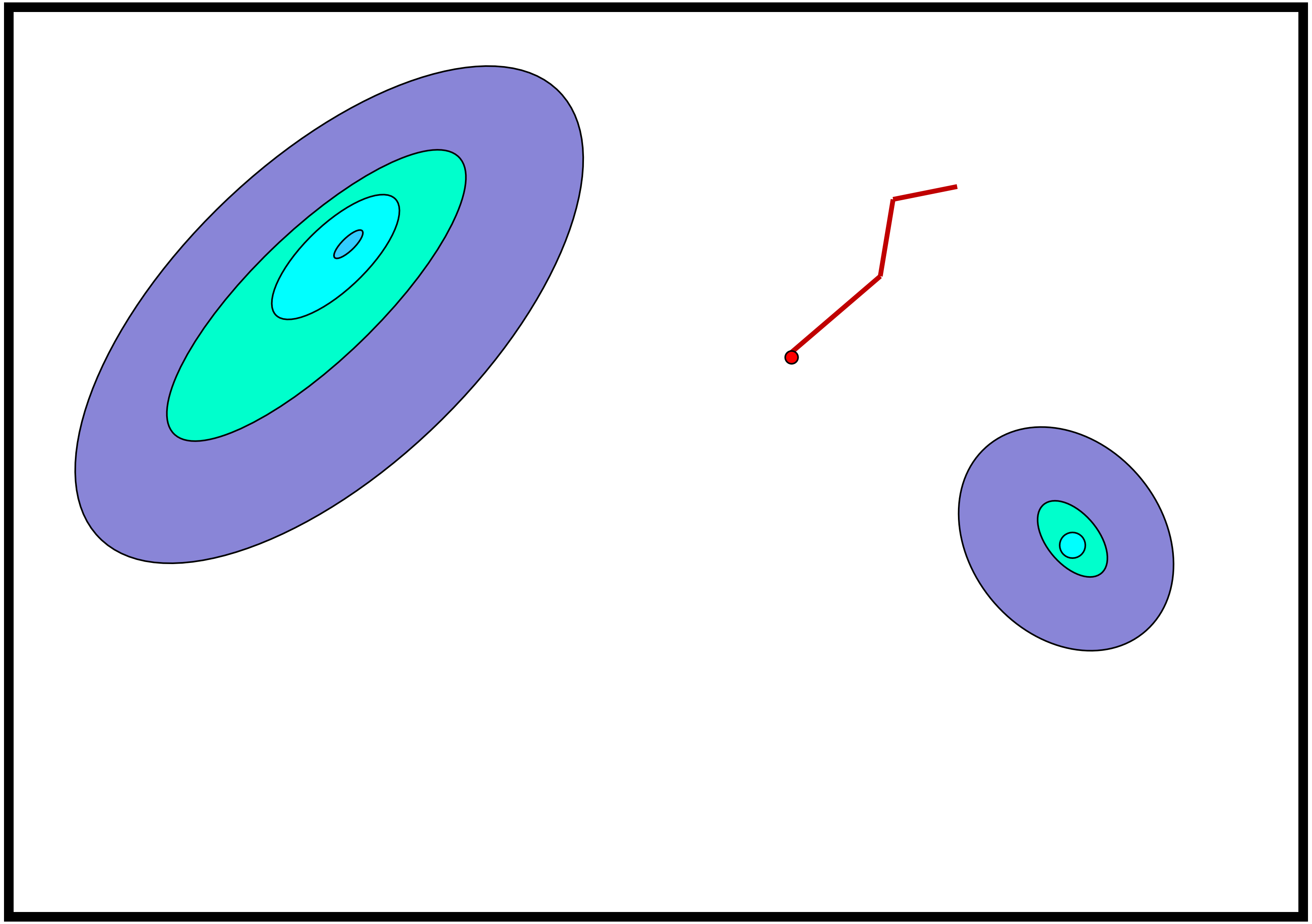
Branch Swapping



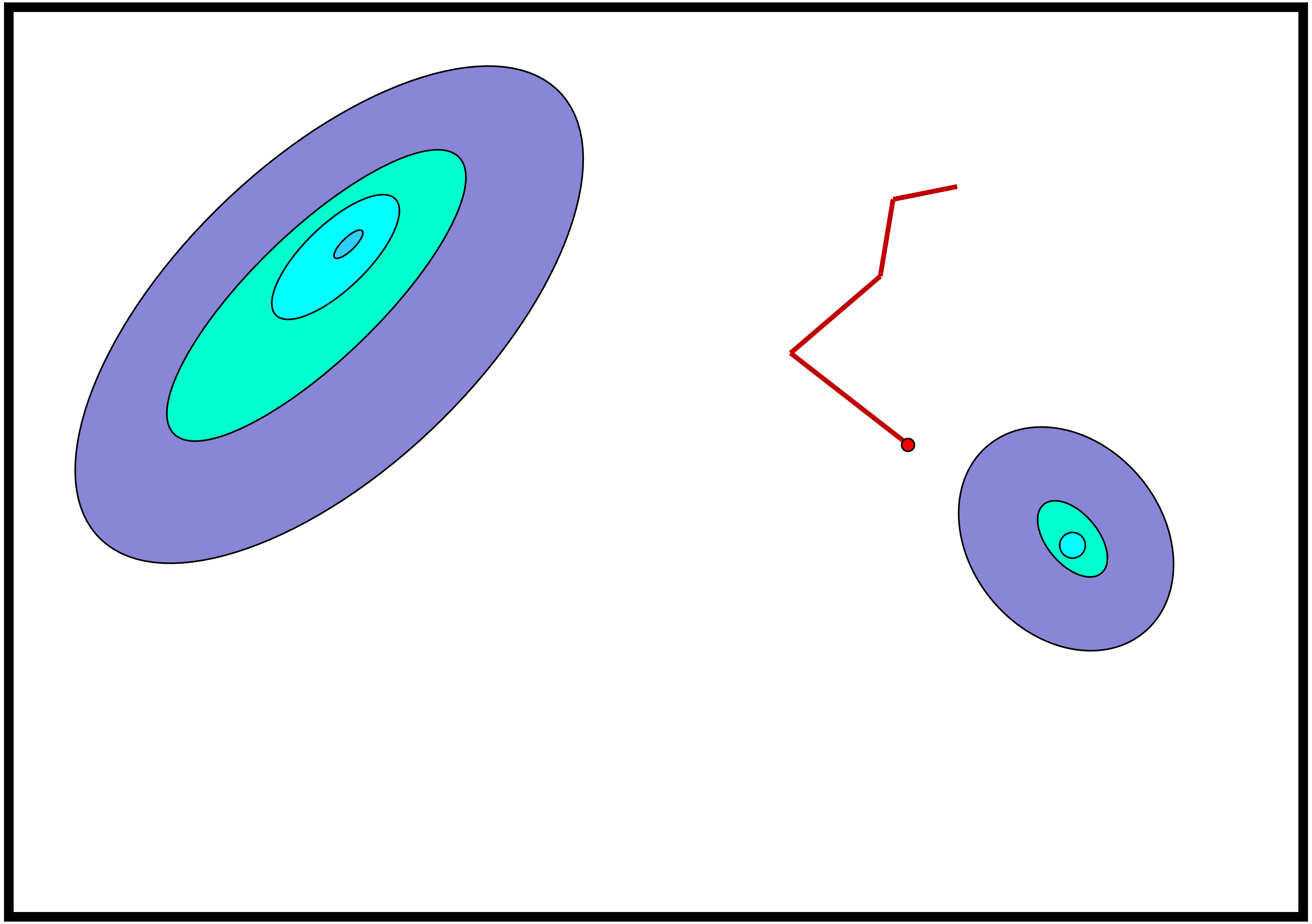
Branch Swapping



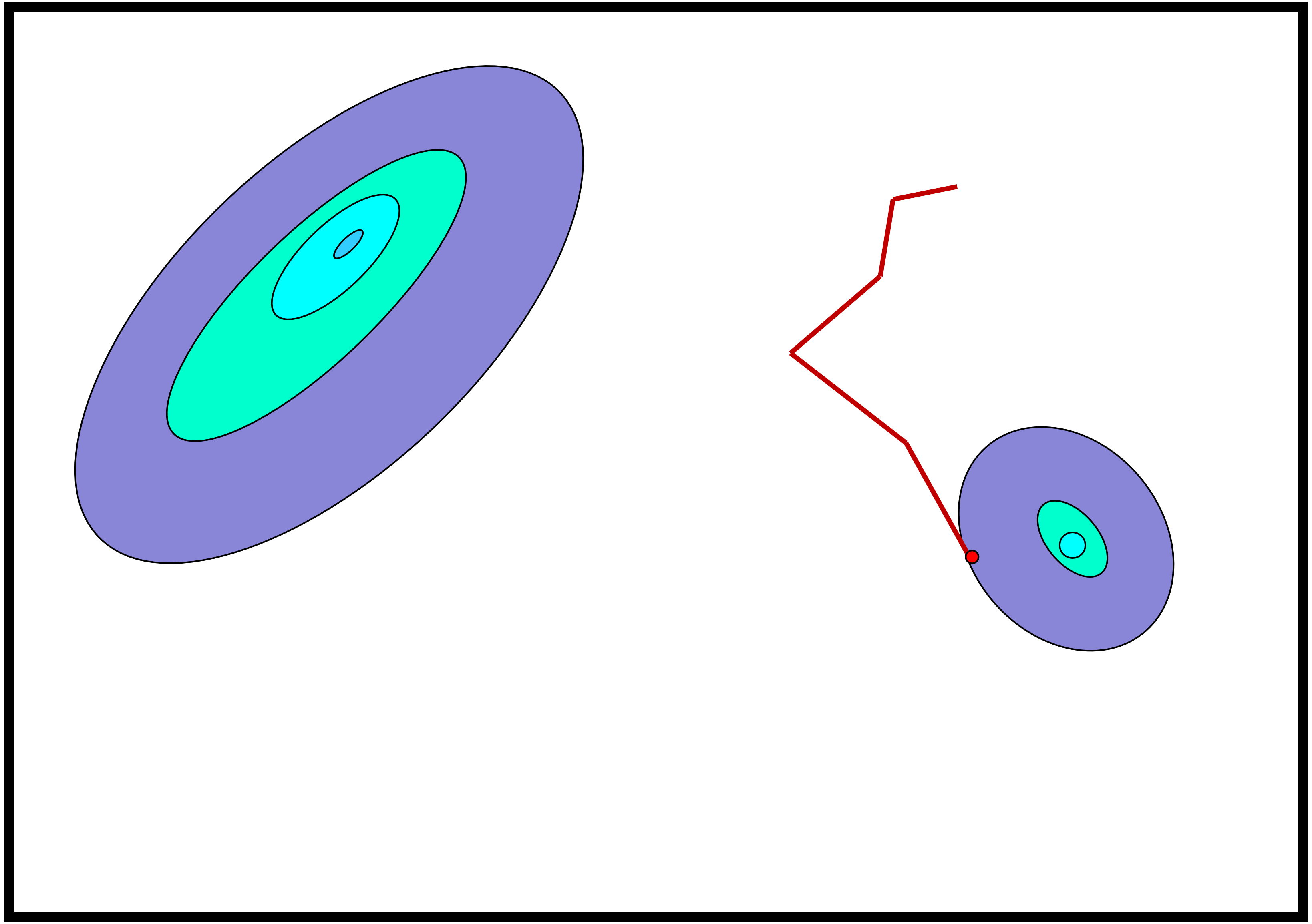
Branch Swapping



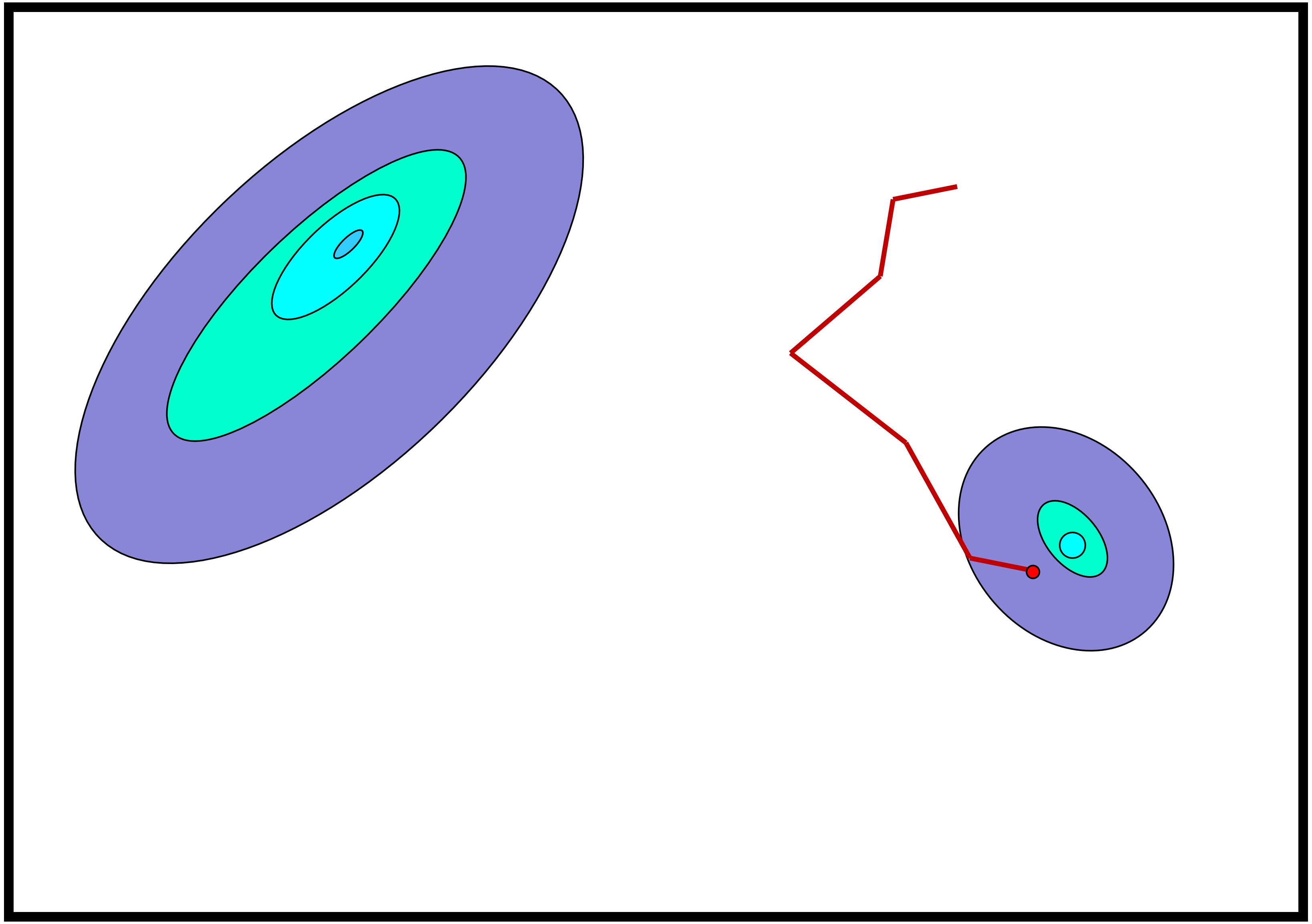
Branch Swapping



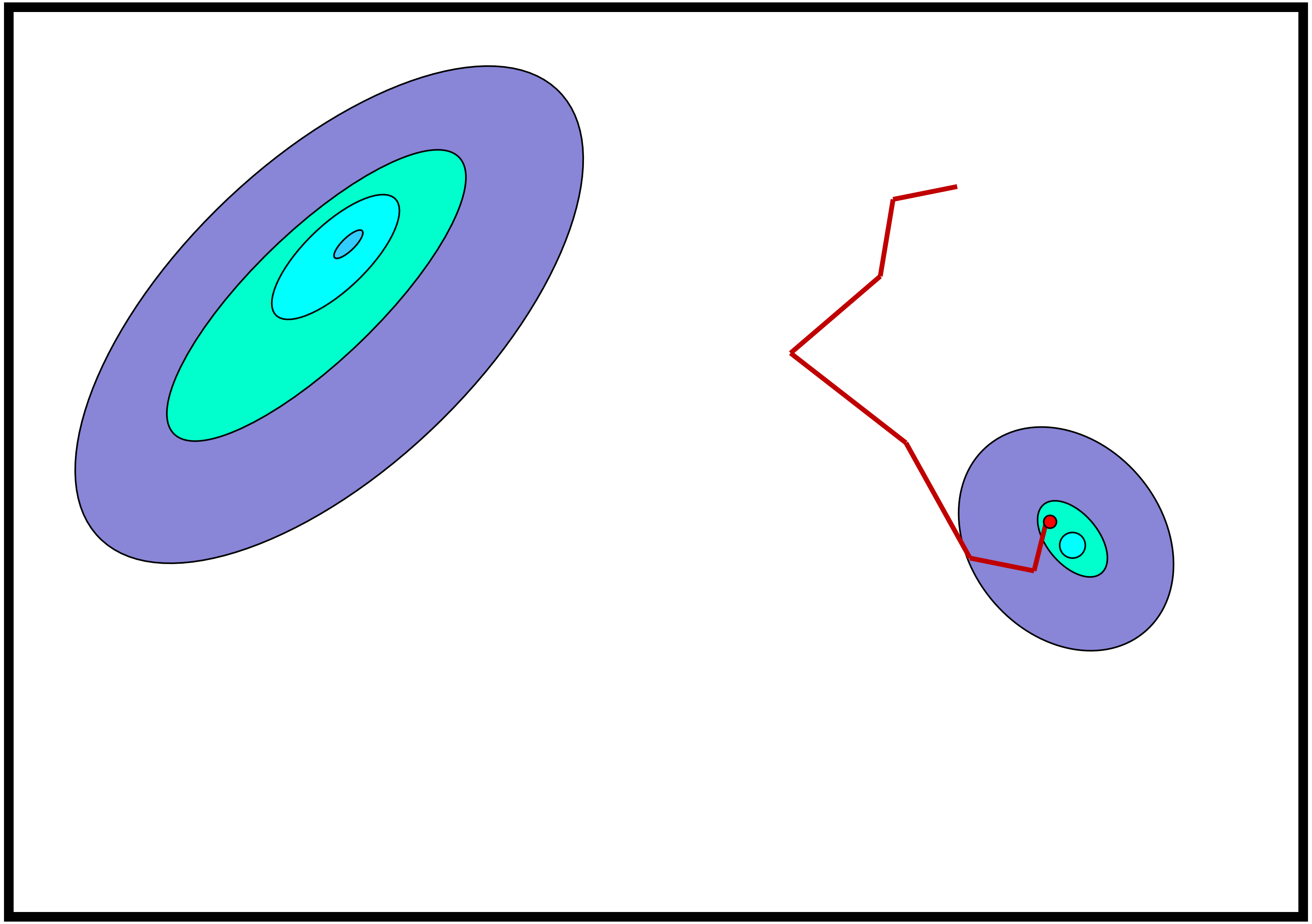
Branch Swapping



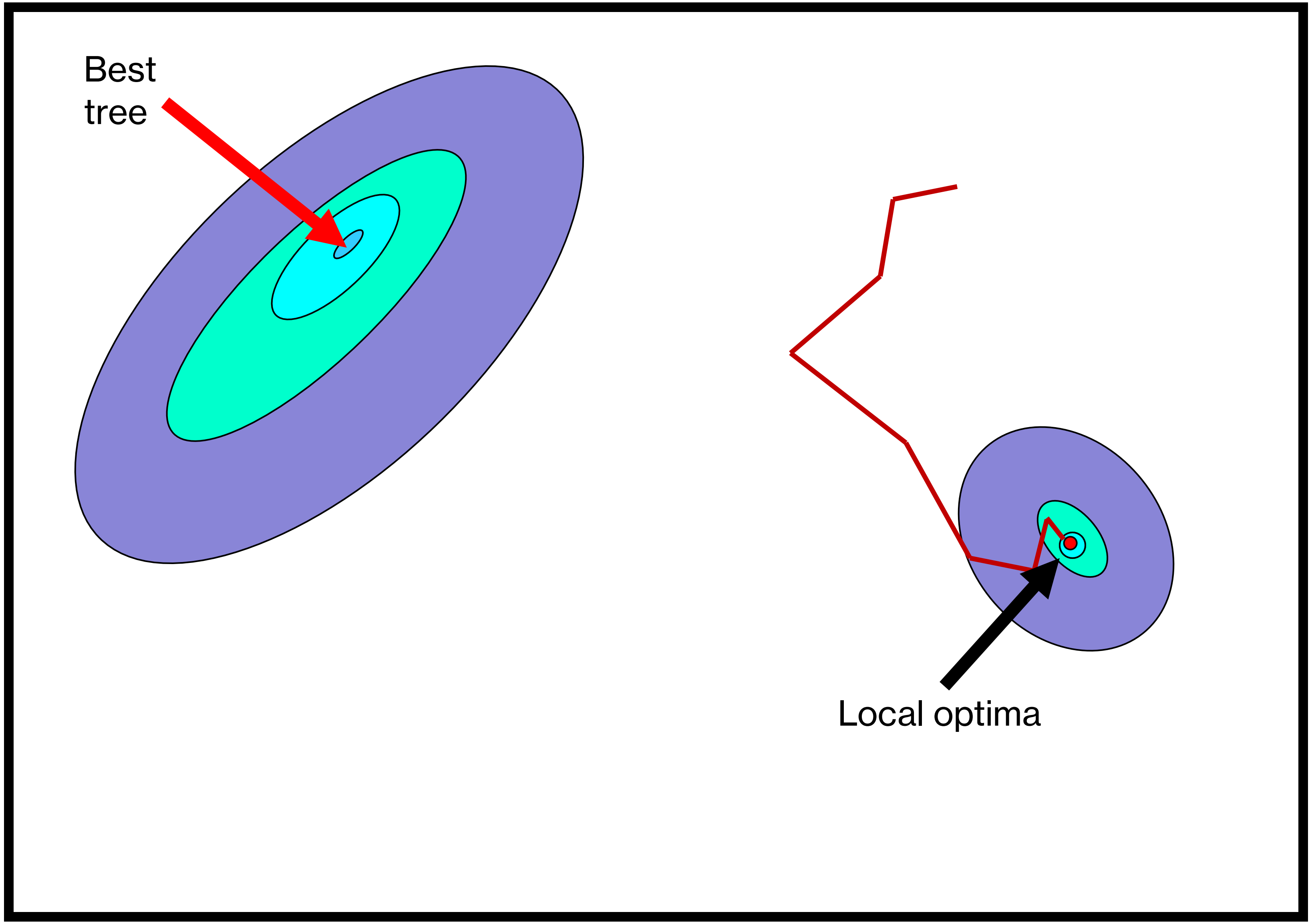
Branch Swapping



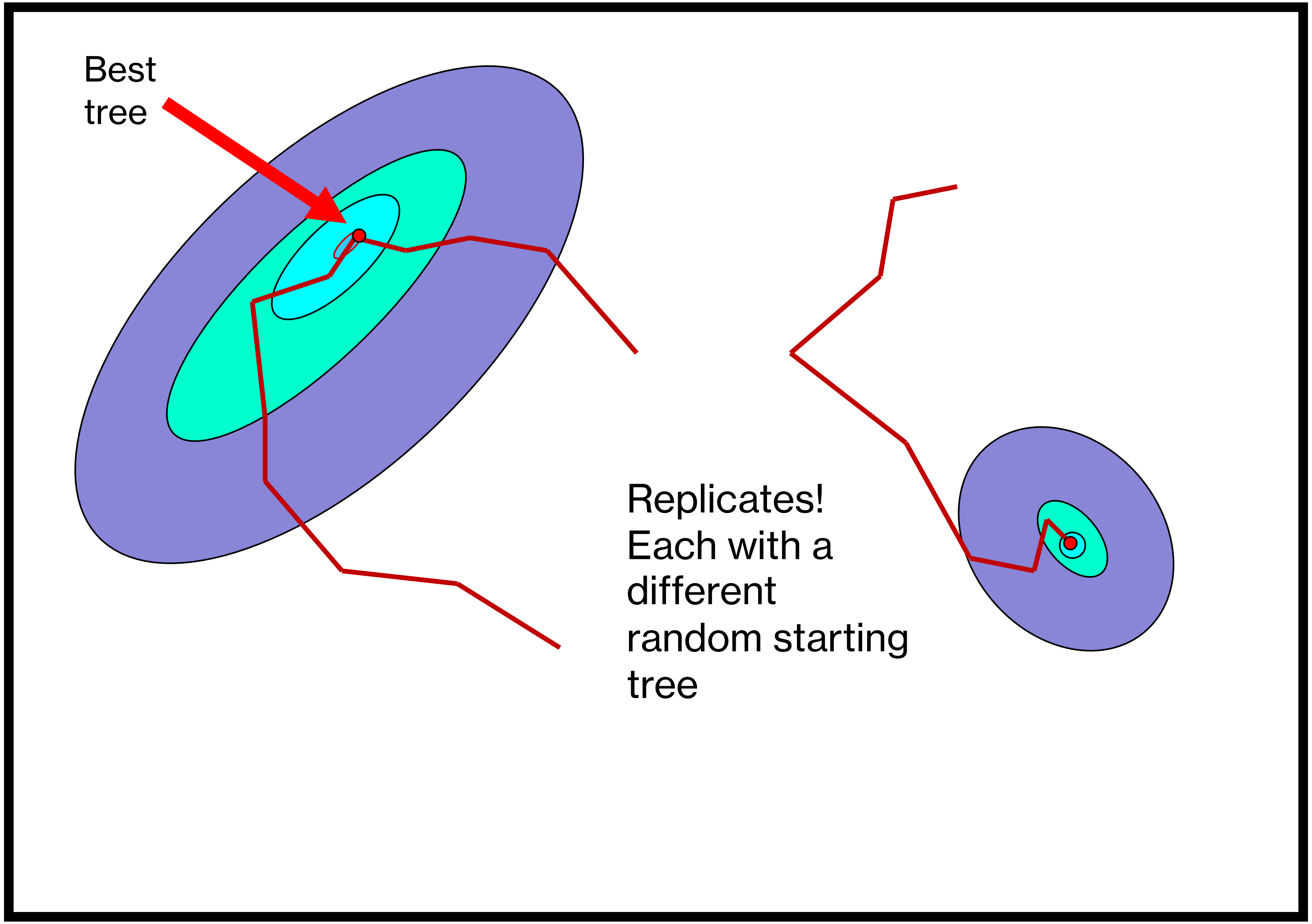
Branch Swapping



Branch Swapping



Branch Swapping



Best tree

Replicates! Each with a different random starting tree

Optimality Criteria

PARSIMONY

DISTANCE METHODS
(aka neighbour-joining, minimum evolution)

MAXIMUM LIKELIHOOD – branch swapping, find the best tree

BAYESIAN INFERENCE – branch swapping, find the best tree
– never examine the majority of trees!
– how many branch swapping replicates to avoid local optima?

Optimality Criteria

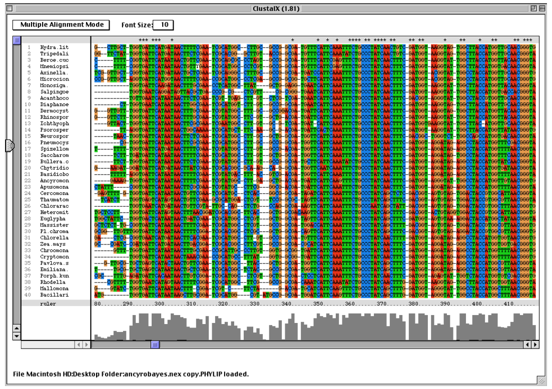
PARSIMONY

DISTANCE METHODS
(aka neighbour-joining, minimum evolution)

MAXIMUM LIKELIHOOD – branch swapping, find the best tree

BAYESIAN INFERENCE – branch swapping, find the best tree
– never examine the majority of trees!
– how many branch swapping replicates to avoid local optima?
– **what about confidence? bootstrap!**

Maximum Likelihood & Bootstrapping

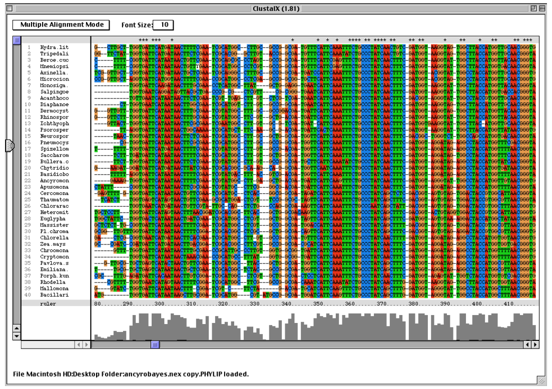


File Macintosh HD:Desktop Folder:ancyrobayes.nex copy.PHYLIP loaded.

branch swapping,
with replication

Best Tree!

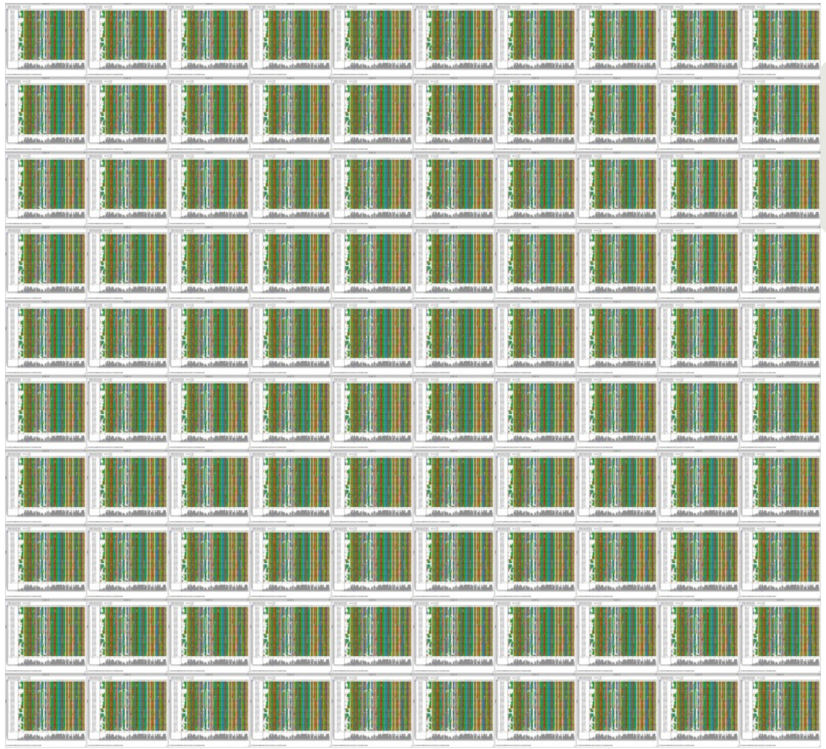
Maximum Likelihood & Bootstrapping



```
1   Hydra.lit
2   Tripedali
3   Beroe.cuc
4   Mnemiopsi
5   Axinella.
6   Microcion
7   Monosiga.
8   Salpingoe
9   Acanthoco
10  Diaphanoe
11  Dermocyst
12  Rhinospor
13  Ichthyoph
14  Psorosper
15  Neurospor
16  Pneumocys
17  Spizellom
18  Saccharom
19  Bullera.c
20  Chytridio
21  Basidiobo
22  Ancyromon
23  Apusomona
24  Cercomona
25  Thaumatom
26  Chlorarac
27  Heteromit
28  Euglypha
29  Massister
30  Pl.chroma
31  Chlorella
32  Zea.mays
33  Chroomona
34  Cryptomon
35  Pavlova.s
36  Emiliana.
37  Porph.kun
38  Rhodella
39  Mallomona
40  Bacillari
    ruler      80.......290.......300.......310.......320.......330.......340.......350.......360.......370.......380.......390.......400.......410......
```

File Macintosh HD:Desktop Folder:ancyrobayes.nex copy.PHYLIP loaded.

resampling with replication

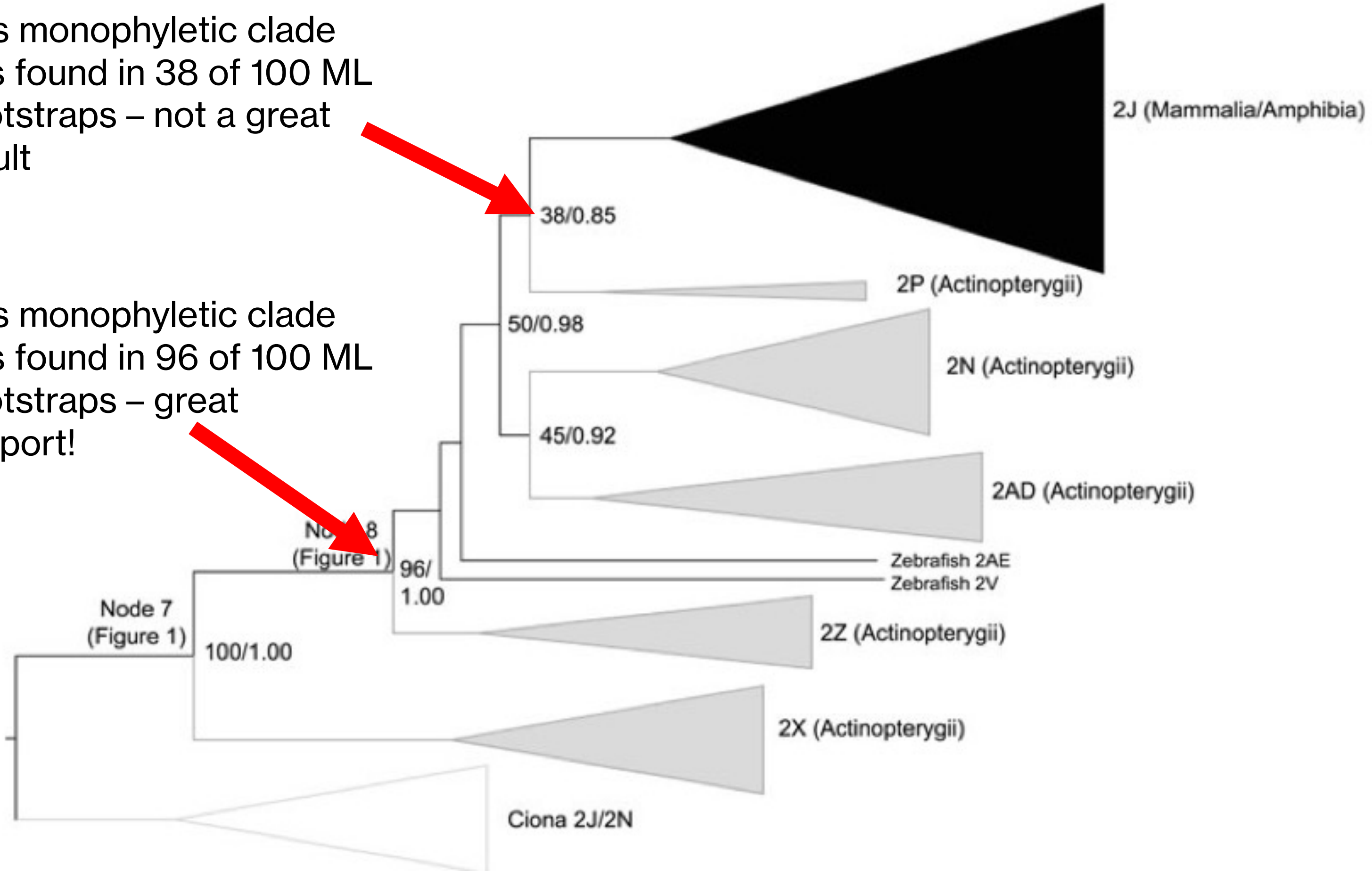


branch swapping, with replication

Best Tree!

Consensus Tree of 100 bootstraps

Maximum Likelihood & Bootstrapping



This monophyletic clade was found in 38 of 100 ML bootstraps – not a great result

This monophyletic clade was found in 96 of 100 ML bootstraps – great support!

Optimality Criteria

PARSIMONY

DISTANCE METHODS
(aka neighbour-joining, minimum evolution)

MAXIMUM LIKELIHOOD – branch swapping, find the best tree

BAYESIAN INFERENCE – branch swapping, find the best tree
– never examine the majority of trees!
– how many branch swapping replicates to avoid local optima?
– **what about confidence? bootstrap!**
– **bootstrap is computationally expensive!**

Optimality Criteria

PARSIMONY

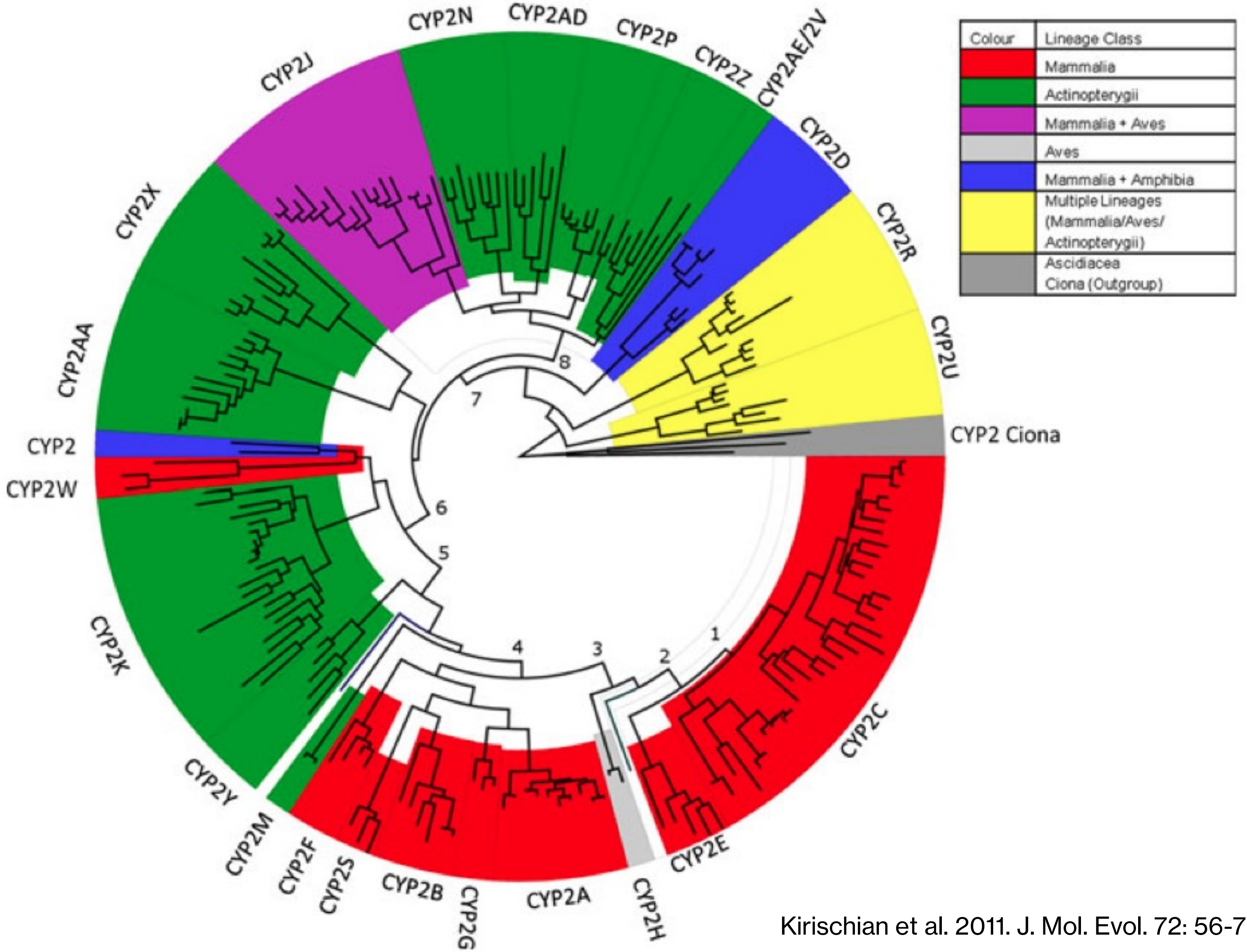
DISTANCE METHODS
(aka neighbour-joining, minimum evolution)

MAXIMUM LIKELIHOOD – branch swapping, find the best tree

BAYESIAN INFERENCE – branch swapping, find the best tree
– never examine the majority of trees!
– how many branch swapping replicates to avoid local optima?
– **what about confidence? bootstrap!**
– **bootstrap is computationally expensive!**

WHAT ABOUT BIG TREES?

RAxML – fast tree space searching, but local or sub-optima more likely



Colour	Lineage Class
🟥	Mammalia
🟩	Actinopterygii
🟪	Mammalia + Aves
⬜	Aves
🟦	Mammalia + Amphibia
🟨	Multiple Lineages (Mammalia/Aves/ Actinopterygii)
▨	Ascidiacea Ciona (Outgroup)

Kirischian et al. 2011. J. Mol. Evol. 72: 56-71.

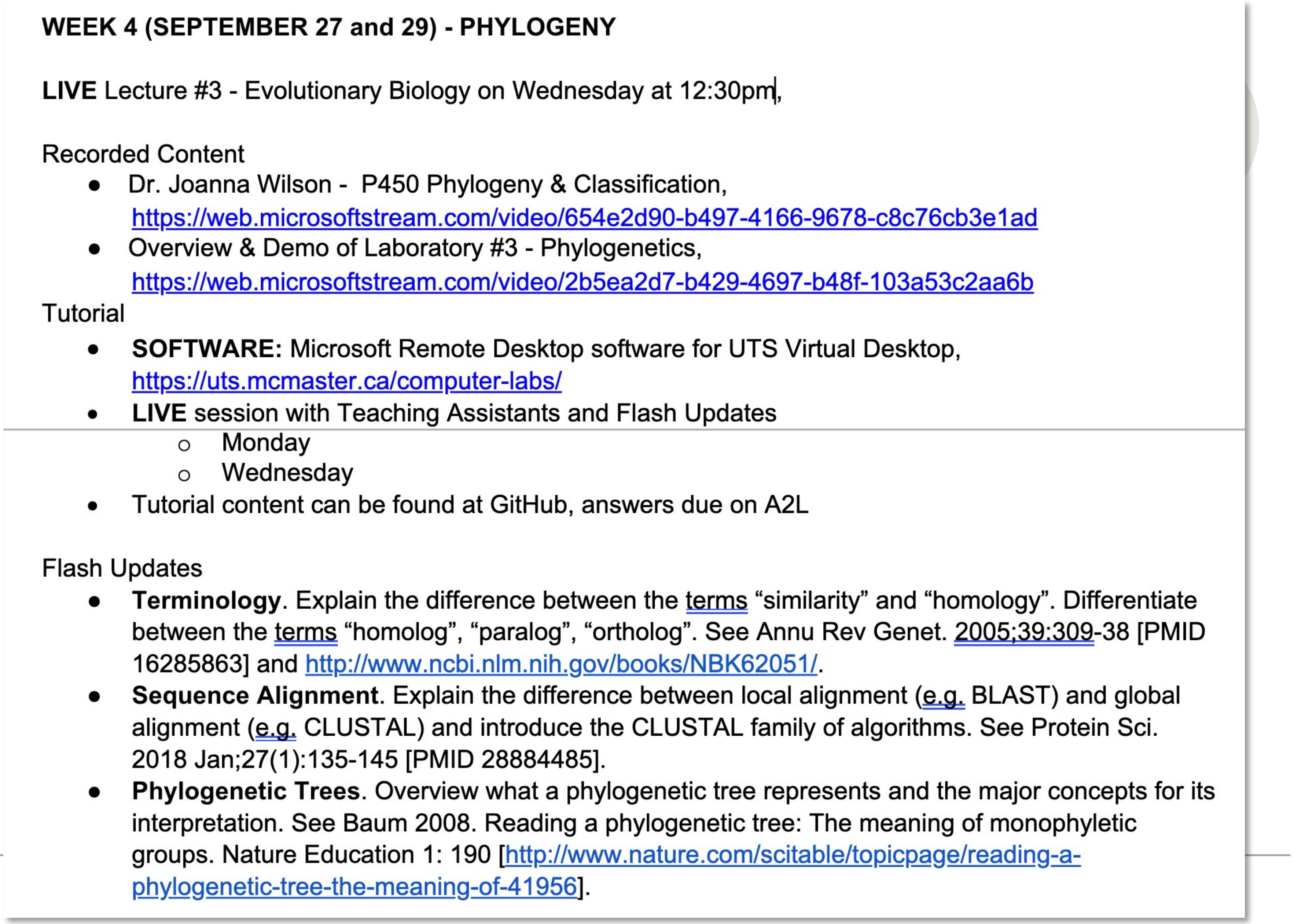
Stamatakis (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690.

Conclusions

- Phylogenetics is not a black box method
- Be skeptical of distance (NJ) and parsimony trees
- Be aware of the assumptions and pitfalls

- how good is the alignment?
- all sites homologous?
- poorly aligned regions removed?
- how was the substitution model selected?
- which optimality criteria is being used? why?
- was the search of tree space robust?
- was confidence properly assessed?
- any known biases / extremes?
- long branch attraction?
- composition bias?

This week



WEEK 4 (SEPTEMBER 27 and 29) - PHYLOGENY

LIVE Lecture #3 - Evolutionary Biology on Wednesday at 12:30pm,

Recorded Content

- Dr. Joanna Wilson - P450 Phylogeny & Classification, https://web.microsoftstream.com/video/654e2d90-b497-4166-9678-c8c76cb3e1ad
- Overview & Demo of Laboratory #3 - Phylogenetics, https://web.microsoftstream.com/video/2b5ea2d7-b429-4697-b48f-103a53c2aa6b

Tutorial

- **SOFTWARE:** Microsoft Remote Desktop software for UTS Virtual Desktop, https://uts.mcmaster.ca/computer-labs/
- **LIVE** session with Teaching Assistants and Flash Updates
  - Monday
  - Wednesday
- Tutorial content can be found at GitHub, answers due on A2L

Flash Updates

- **Terminology**. Explain the difference between the terms “similarity” and “homology”. Differentiate between the terms “homolog”, “paralog”, “ortholog”. See Annu Rev Genet. 2005;39:309-38 [PMID 16285863] and http://www.ncbi.nlm.nih.gov/books/NBK62051/.
- **Sequence Alignment**. Explain the difference between local alignment (e.g. BLAST) and global alignment (e.g. CLUSTAL) and introduce the CLUSTAL family of algorithms. See Protein Sci. 2018 Jan;27(1):135-145 [PMID 28884485].
- **Phylogenetic Trees**. Overview what a phylogenetic tree represents and the major concepts for its interpretation. See Baum 2008. Reading a phylogenetic tree: The meaning of monophyletic groups. Nature Education 1: 190 [http://www.nature.com/scitable/topicpage/reading-a-phylogenetic-tree-the-meaning-of-41956].