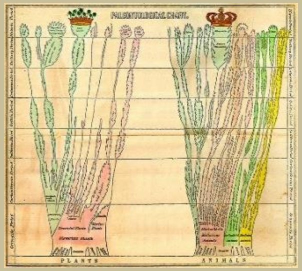
Mega

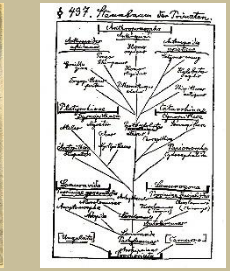
Phylogeny

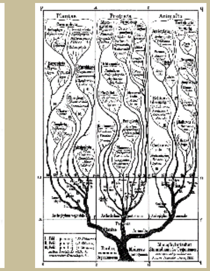
Approach

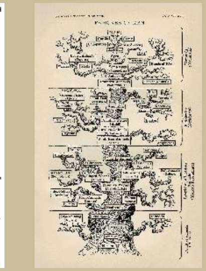
Background

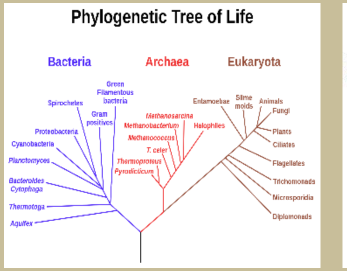
- **Reconstructing the "Tree of Life" is one of the enduring goals of evolutionary biology.**

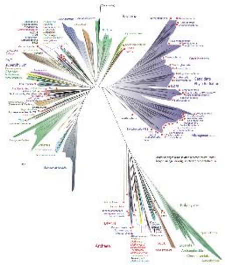


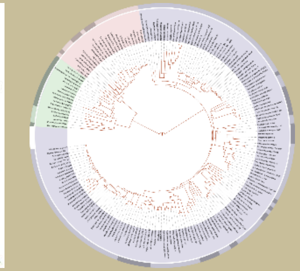








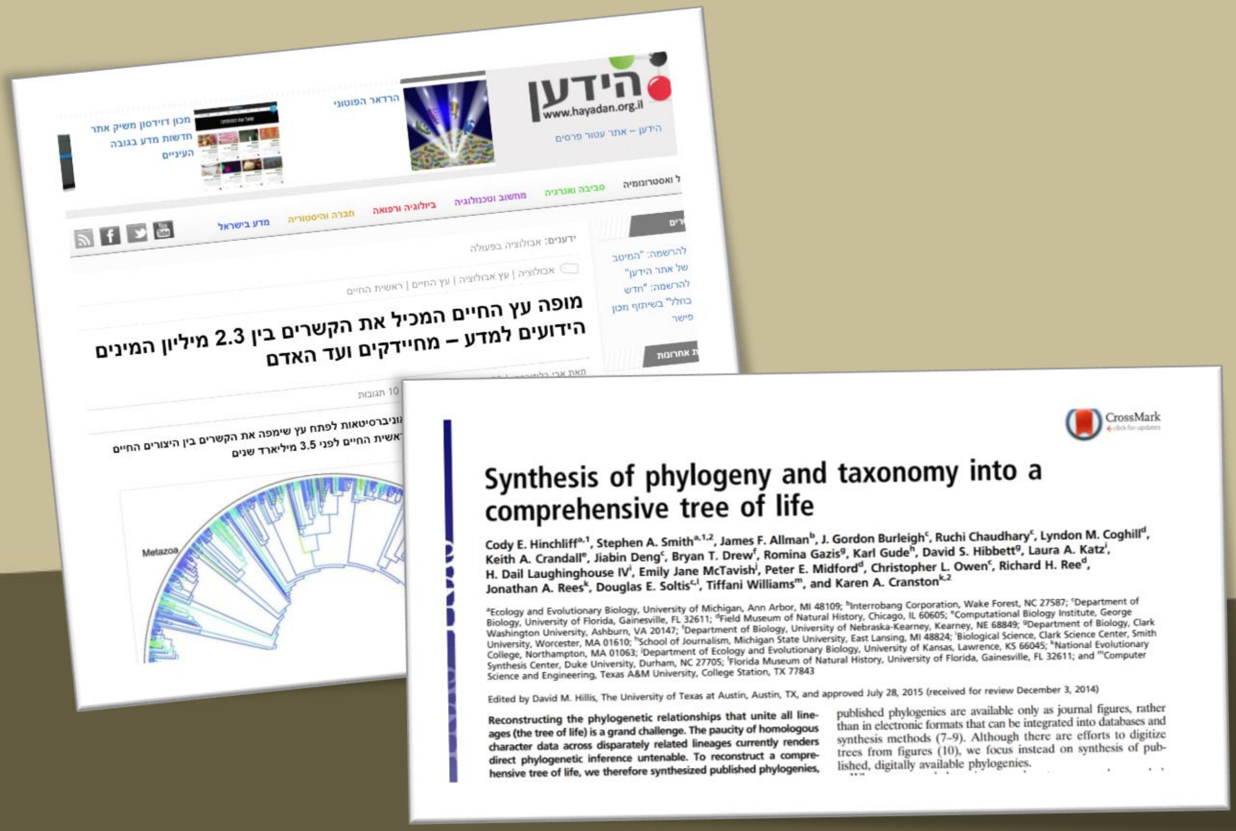




Since 19th

Background

- The massive accumulation of sequence data should provide more accurate phylogenies with the hope to resolve the tree of life.
- Recently, the Open Tree of Life was presented.



ידענים: אבולוציה בפעולה

אבולוציה | עץ אבולוציה | עץ החיים | ראשית החיים

מופה עץ החיים המכיל את הקשרים בין 2.3 מיליון המינים הידועים למדע – מחיידקים ועד האדם

10 תגובות

וניברסיטאות לפתח עץ שימפה את הקשרים בין היצורים החיים
ראשית החיים לפני 3.5 מיליארד שנים

Synthesis of phylogeny and taxonomy into a comprehensive tree of life

Cody E. Hinchliff[a,1], Stephen A. Smith[a,1,2], James F. Allman[b], J. Gordon Burleigh[c], Ruchi Chaudhary[c], Lyndon M. Coghill[d], Keith A. Crandall[e], Jiabin Deng[c], Bryan T. Drew[f], Romina Gazis[g], Karl Gude[h], David S. Hibbett[g], Laura A. Katz[i], H. Dail Laughinghouse IV[i], Emily Jane McTavish[j], Peter E. Midford[d], Christopher L. Owen[c], Richard H. Ree[d], Jonathan A. Rees[k], Douglas E. Soltis[c,l], Tiffani Williams[m], and Karen A. Cranston[k,2]

[a]Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109; [b]Interrobang Corporation, Wake Forest, NC 27587; [c]Department of Biology, University of Florida, Gainesville, FL 32611; [d]Field Museum of Natural History, Chicago, IL 60605; [e]Computational Biology Institute, George Washington University, Ashburn, VA 20147; [f]Department of Biology, University of Nebraska-Kearney, Kearney, NE 68849; [g]Department of Biology, Clark University, Worcester, MA 01610; [h]School of Journalism, Michigan State University, East Lansing, MI 48824; [i]Biological Science, Clark Science Center, Smith College, Northampton, MA 01063; [j]Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS 66045; [k]National Evolutionary Synthesis Center, Duke University, Durham, NC 27705; [l]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611; and [m]Computer Science and Engineering, Texas A&M University, College Station, TX 77843

Edited by David M. Hillis, The University of Texas at Austin, Austin, TX, and approved July 28, 2015 (received for review December 3, 2014)

Reconstructing the phylogenetic relationships that unite all lineages (the tree of life) is a grand challenge. The paucity of homologous character data across disparately related lineages currently renders direct phylogenetic inference untenable. To reconstruct a comprehensive tree of life, we therefore synthesized published phylogenies,

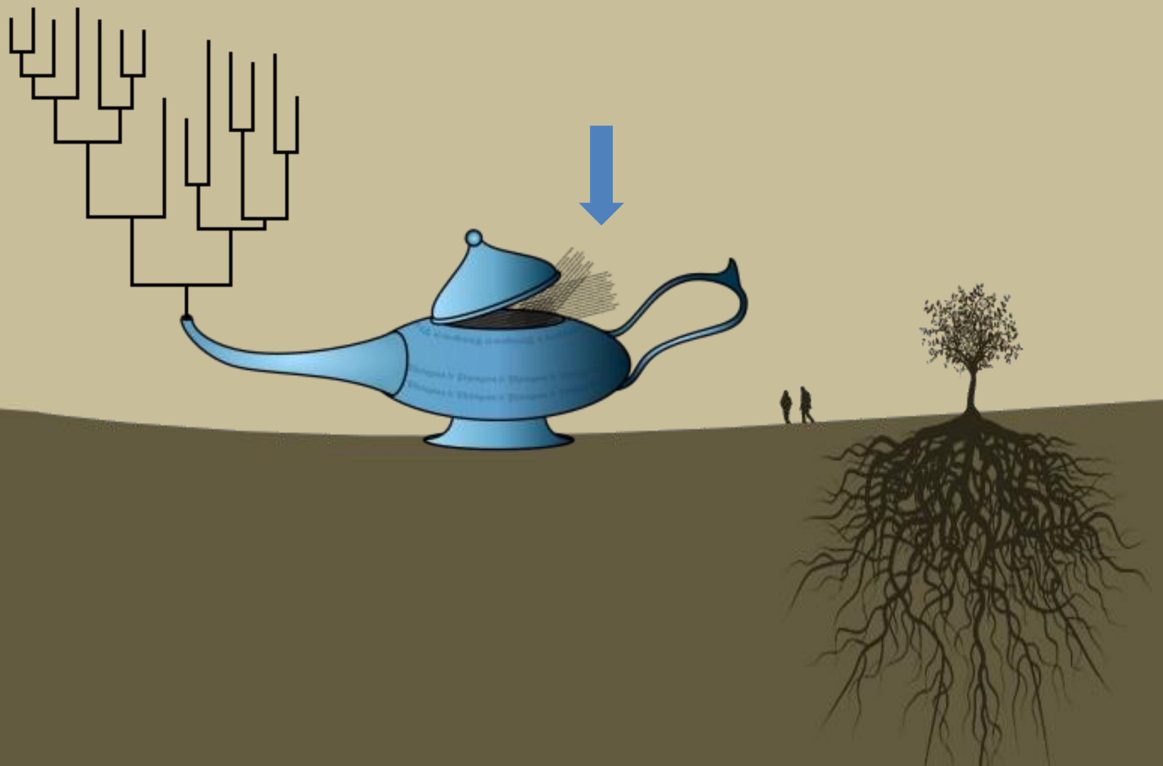
published phylogenies are available only as journal figures, rather than in electronic formats that can be integrated into databases and synthesis methods (7–9). Although there are efforts to digitize trees from figures (10), we focus instead on synthesis of published, digitally available phylogenies.

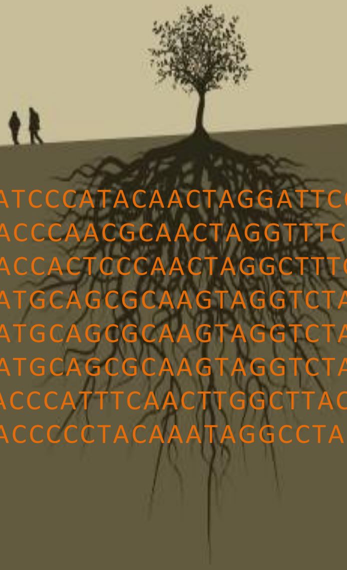
Current approaches for reconstructing mega phylogenies

- ***Supertree:*** **compile many individual trees with partially overlapping taxa into a single large phylogeny.**

- ***Supermatrix:*** **gather sequence data for many loci into a single multiple sequence alignment, which is then used to reconstruct a large phylogeny.**

+ twist



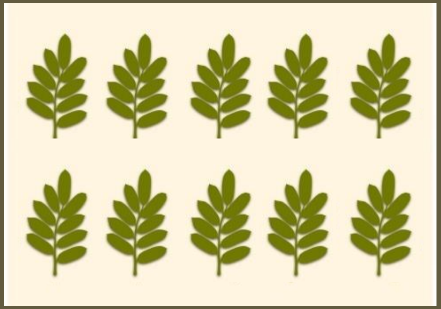


Species1:	ATGGCATATCCCATACAACTAGGATTCCAAGATGCAACC
Species2:	ATGGCACACCCAACGCAACTAGGTTTCAAGGACGCGGC
Species3:	ATGGCCAACCACTCCCAACTAGGCTTTCAAGACGCCTCC
Species4:	ATGGCACATGCAGCGCAAGTAGGTCTACAAGACGCTAC
Species5:	ATGGCACATGCAGCGCAAGTAGGTCTACAAGACGCTAC
Species6:	ATGGCACATGCAGCGCAAGTAGGTCTACAAGACGCTAC
Species7:	ATGGCTTACCCATTTCAACTTGGCTTACAAGACGCTACC
Species8:	ATGGCATACCCCCTACAAATAGGCCTACAAGATGCAAC

The reconstructed trees can only be as good as their underlying multiple sequence alignment (garbage in -> garbage out).

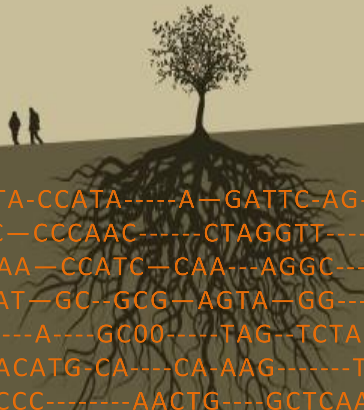
Slowly evolving regions: to align –

good for solving divergence events, cannot be used to solve events.



Rapidly evolving regions: to align –

good for solving event, cannot be used to solve divergence events.



```
Species1:     AT-----CATA-CCATA-----A—GATTC-AG-AT--GC-AA-CC
Species2:     ATGG-CAC—CCCAAC------CTAGGTT----GGA-CGC-GGC-
Species3:     ATGG-----AA—CCATC—CAA---AGGC---------------G------
Species4:     --ATG—--AT—GC--GCG—AGTA—GG-----A—G----CTAC
Species5:     ATGGC-------A----GC00-----TAG--TCTACAAGAC-----TAC
Species6:     AT—G—CACATG-CA----CA-AAG-------TCCA--GACGTAC-
Species7:     AG-----TACCC--------AACTG----GCTCAA--GA---GCTACC
Species8:     A-GGC-T-----CCC----ACA-A—AGGC-CTACA-GATGCAAC-
Species9:     ---TGGCAAC-CC--CCTACAAAT---------TAC---AGA--CAAC
Species10:    A---GCAT--------CCCTA---TA----CCT----ACA------GC---C
Species11:    -----CATATC-----TACAACTAG----TTCA---------TGCA--CC
Species12:    ATGGC-------------C—GCAAC--TATTTCA-AGGACG---GC
Species13:    AT---CCA-CCA-CT-CA-CTAG-----CTTT-----GAC—GCC-CC
Species14:    ATG----GCCAT-----GCGC----GTAG------------------CTAC
Species15:    ---TGG--ACAT-AG--GC—AGTAG-GTC--CAAGA—CGC--C
Species16:    A--GGCC------GCA-AGA--GGT--TA--CAAG---ACGCTAC--
Species17:    ---GGCT---AC—CCA---CAA-----GGCTTA---AG--GCTA-CC
Species18:    AT--GG--CATA---C---RAAAT—AG-GC-C—---GATG--AC-
Species19:    AT---ATA------ATA----ACTAG----TTCCAA---TGC-A----C--
Species20:    A—TGG—CAC--CC---GCAACTA---GTCA---GAC—CG-GC-
Species21:    AT---C-----CAA--CACT----CAACTA—T-T-CA-C-------TCC-
Species22:    ---GGC------ATGCA----------GTAG----TACAA--ACGCTAC
Species23:    --ATGGC--TGCA-----GCA-AGTA------CTAC-----CG--TAC-
Species24:    ---GA--C—AT-G—G-C----------GGTCTAA---AC--CTA—C-
Species25:    ATG—C----C-CAT-----TTAA---TCTT—ACAAG-C---TAC-C
Species26:    --TGG--CA-----CCC------A-AT—AGG-CCTACA-----TGCAC
Species27:    -------CATA--TCCC----------------TTCCA----ATG-—AACC-
Species28:    A--TGGC-ACAC---AA--GC---CTGGT---C-AGGAC—G-GGC
Species29:    ---G---CC--AC----CTCCC—AAC-A----CTTC-----ACGC-TCC
Species30:    A-T-GG-C---ATGCA--GC--GCAAGG----TAC--GA--GCTAC

...

Species9999:  ATGGCA-TAC—-----ACAA—TAG-CACAA--GATGC-AAC-
Species10000:ATGG----ATATCCCATACAACTAGGATTCCAA----GCAACC
```

The problem

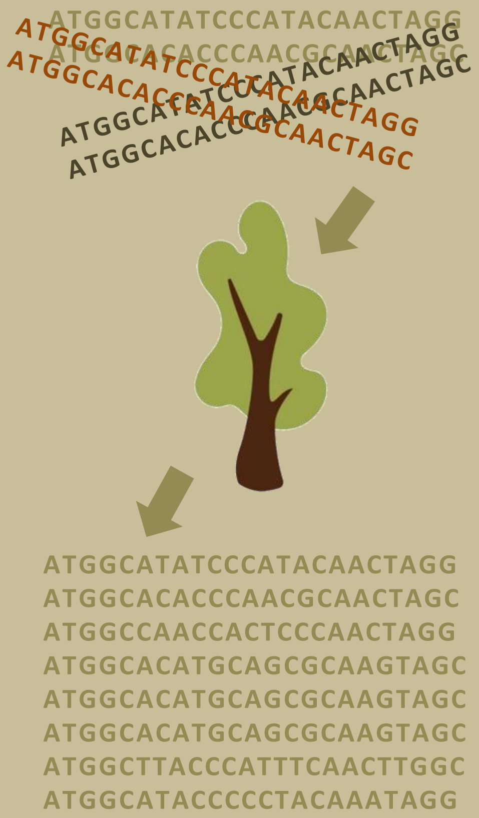
Aligning rapidly evolving loci for highly diverged taxa leads to poor alignments and inaccurate trees.

The problem

- **Almost all MSA algorithms build a phylogeny during the estimation procedures.**

- **Those phylogenies are often based on pairwise distances.**

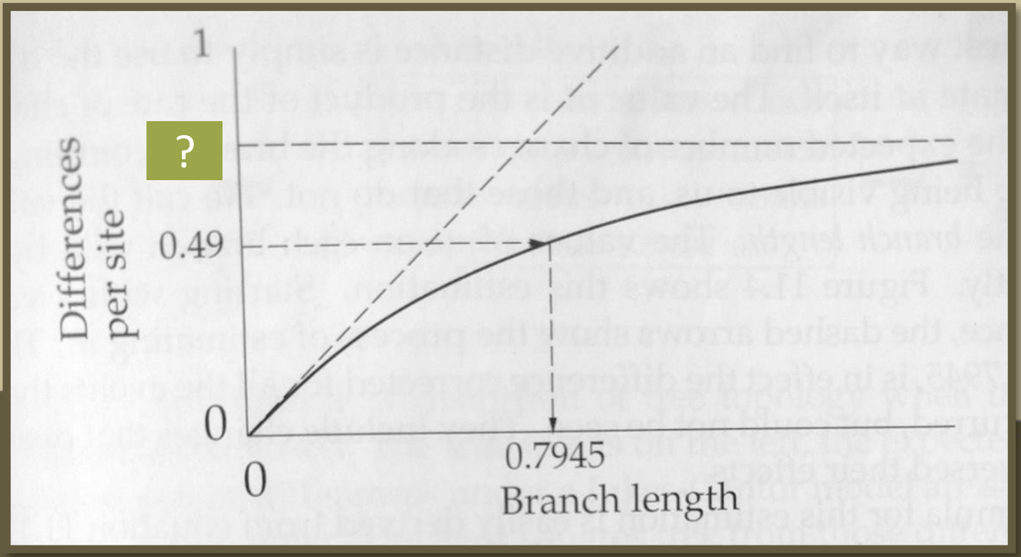
- **Pairwise distances:**
  - **raw pairwise distances (uncorrected).**
  - **model-corrected (e.g. Jukes-Cantor).**



ATGGCATATCCCATACAACTAGG
ATGGCACACCCAACGCAACTAGC
ATGGCCAACCACTCCCAACTAGG
ATGGCACATGCAGCGCAAGTAGC
ATGGCACATGCAGCGCAAGTAGC
ATGGCACATGCAGCGCAAGTAGC
ATGGCTTACCCATTTCAACTTGGC
ATGGCATACCCCCTACAAATAGG

The problem

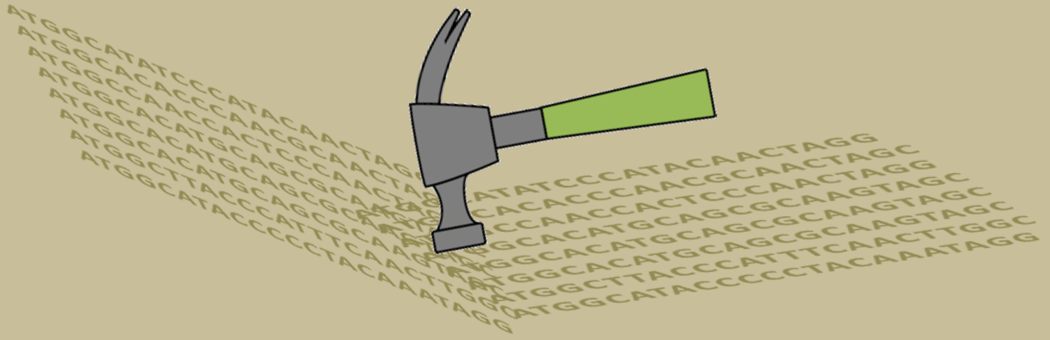
- **These methods are susceptible to problems with saturation (i.e. multiple mutations at the same site for the same organism)**



ATGGCATATCCCATACAACTAGG
ATGGCACACCCAACGCAACTAGC
ATGGCCAACCACTCCCAACTAGG
ATGGCACATGCAGCGCAAGTAGC
ATGGCACATGCAGCGCAAGTAGC
ATGGCACATGCAGCGCAAGTAGC
ATGGCTTACCCATTTCAACTTGGC
ATGGCATACCCCCTACAAATAGG

Mega phylogeny approach

- **Separate sequences into subgroups of aligned sequences based on the degree of sequence saturation.**

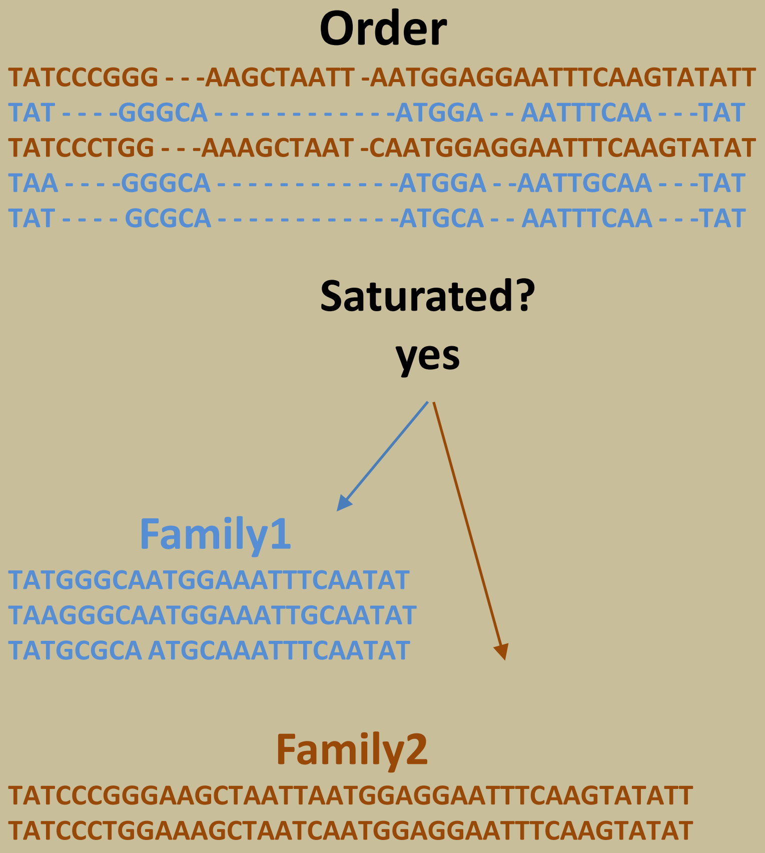

ATGGCATATCCCCATACAACTAGG
ATGGCACACCCCAACGCAACTAGC
ATGGCCAACCCACTCCCCAACTAGG
ATGGCACATGCAGCGCGCAAGTAGC
ATGGCACATGCAGCGCGCAAGTAGC
ATGGCACATGCAGCGCGCAAGTAGC
ATGGCTTACCCATTTCAACTTGGC
ATGGCATACCCCCCTACAAATAGG

Mega phylogeny approach

- If the most inclusive group of sequences is saturated, then the group is broken up into less inclusive groups using the next level in the taxonomic hierarchy.

- For example, if an "order" is found to be saturated, it would be broken into "families". Each smaller subset of sequences is then re-aligned and the saturation reassessed.

- This process continues iteratively.



Order

TATCCCGGG - - -AAGCTAATT -AATGGAGGAATTTCAAGTATATT
TAT - - - -GGGCA - - - - - - - - - - - - -ATGGA - - AATTTCAA - - -TAT
TATCCCTGG - - -AAAGCTAAT -CAATGGAGGAATTTCAAGTATAT
TAA - - - -GGGCA - - - - - - - - - - - - -ATGGA - -AATTGCAA - - -TAT
TAT - - - - GCGCA - - - - - - - - - - - - -ATGCA - - AATTTCAA - - -TAT

Saturated?
yes

Family1

TATGGGCAATGGAAATTTCAATAT
TAAGGGCAATGGAAATTGCAATAT
TATGCGCA ATGCAAATTTCAATAT

Family2

TATCCCGGGAAGCTAATTAATGGAGGAATTTCAAGTATATT
TATCCCTGGAAAGCTAATCAATGGAGGAATTTCAAGTATAT

Detect saturation across a set of sequence data

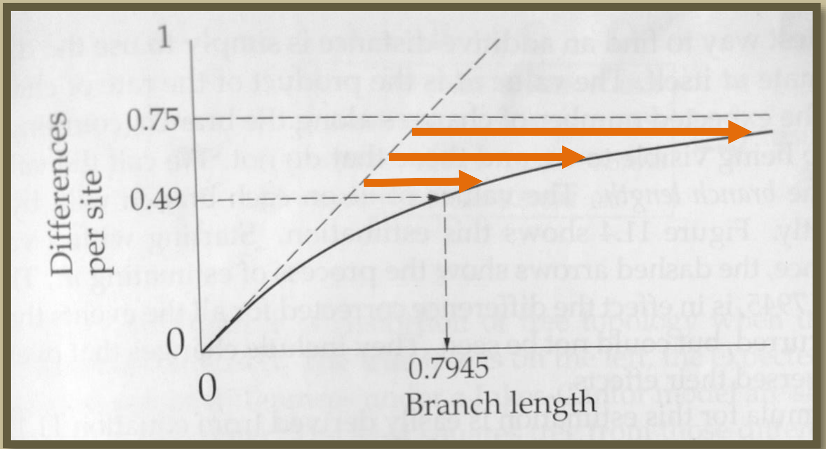
d1 =
raw pair-wise distance

ACGGTCATGTATAC
TCTCTCCTGTGAAT
d1 = 0.5

d2 =
Jukes-Cantor distance

ACGGTCATGTATAC
TCTCTCCTGTGAAT
d2 = 0.8

- **x = |d1 – d2|** *(x = 0.3)*
- **Measure of dispersion is based on the median absolute deviation (MAD):**



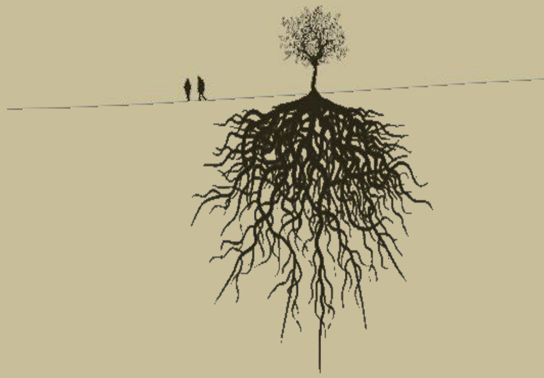
MAD = Med (|xi - Med (x)|)

e.g.:
(9 ,6 ,4 ,**2** ,2 ,1 ,1)

absolute deviations:

(7 ,4 ,2 ,0 ,0 ,1 ,1) => (7 ,4 ,2 ,**1** ,1 ,0 ,0)

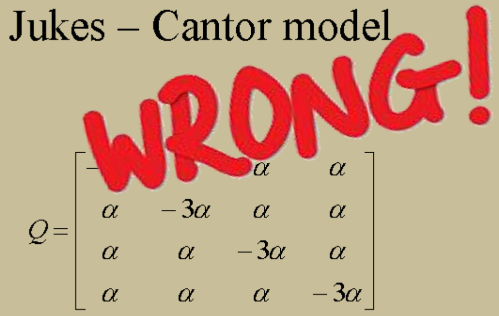
Detect saturation across a set of sequence data



- The larger the MAD -> the larger the overall spread of distances.

- That is, above a certain value the assumed nucleotide substitution model is no longer adequately accounting for the rate variation exhibited by pair-wise distances among species.

Large MAD ----->



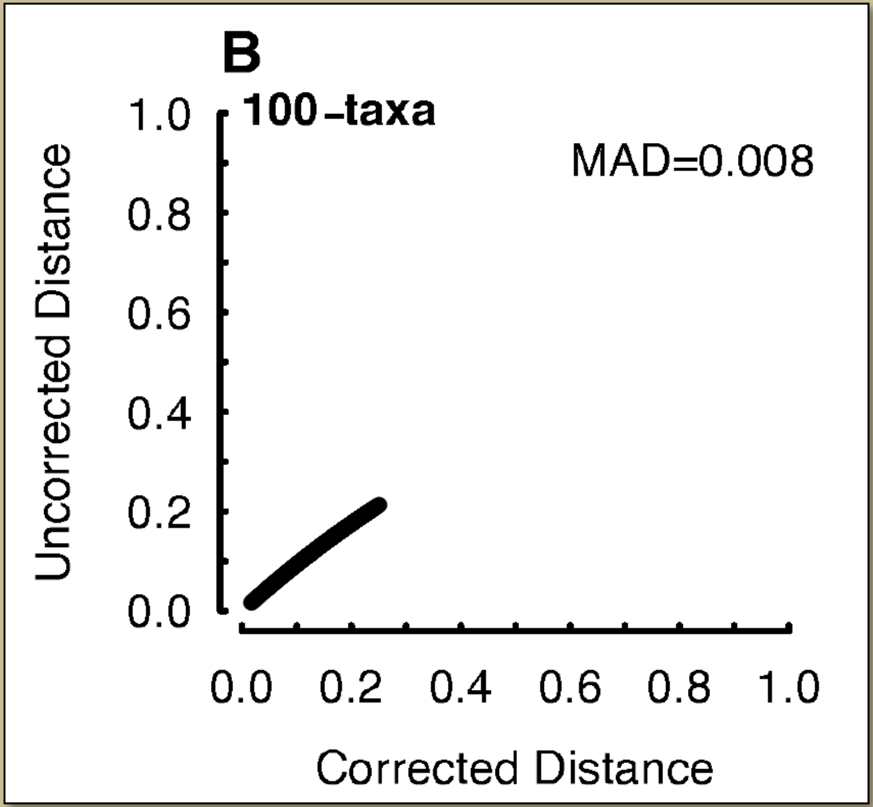
Jukes – Cantor model

WRONG!

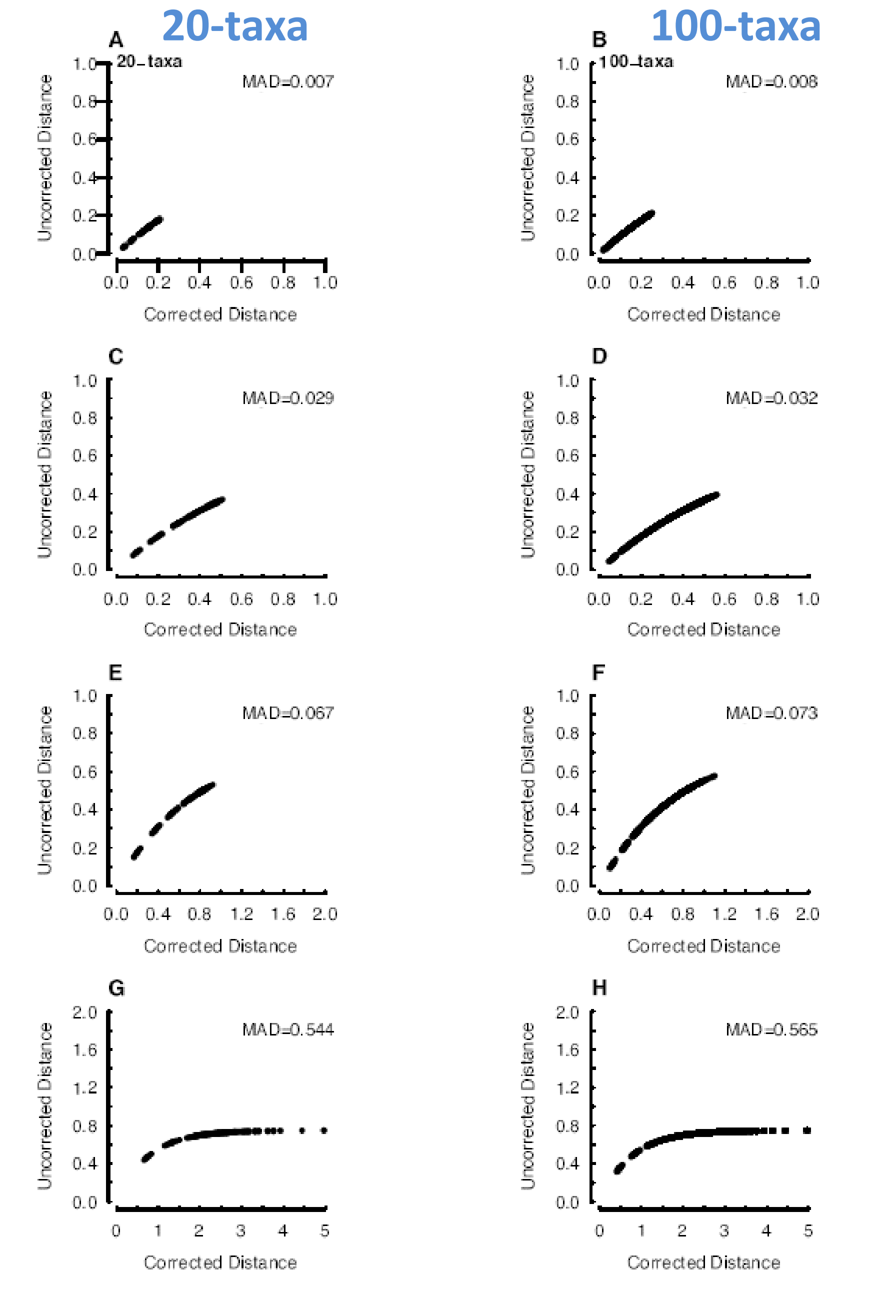
$$Q=\begin{bmatrix} - & & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix}$$

Assessing saturation

- **Determine a threshold for subdividing sequences into smaller alignments.**



Assessing saturation



20-taxa

100-taxa

Tree length

One more slide:
Our mega phylogeny approach

Locus 1 ->

```
ATGGCATATCCCATACAACTAGGATTCCAAGT
ATGGCTTACCCATTTCAACTTGGCTTACAAGA
ATGGCCAACCACTCCCAACTAGGCTTTCAAGT
ATGGCACATCCCACACAATTAGGATTCCAAGA
ATGGCCTACCCATTCCAACTTGGTCTACAAGA
```

Locus 2 ->

```
TATCCCGGG - - -AAGCTAATT -AATGGAGGAATTTCAAGTATATT
TAT - - - -GGGCA - - - - - - - - - - - - -ATGGA - - AATTTCAA - - -TAT
TATCCCTGG - - -AAAGCTAAT -CAATGGAGGAATTTCAAGTATAT
TAA - - - -GGGCA - - - - - - - - - - - - -ATGGA - -AATTGCAA - - -TAT
TAT - - - - GCGCA - - - - - - - - - - - - -ATGCA - - AATTTCAA - - -TAT
```

Locus 1

```
ATGGCATATCCCATACAACTAGGATTCCAAGT
ATGGCTTACCCATTTCAACTTGGCTTACAAGA
ATGGCCAACCACTCCCAACTAGGCTTTCAAGT
ATGGCACATCCCACACAATTAGGATTCCAAGA
ATGGCCTACCCATTCCAACTTGGTCTACAAGA
```

Locus 2a   Locus 2b

```
TATCCCGGGAAGCTAATTAATGGAGGAATTTCAAGTATATT-----------------------------------------------
TATCCCTGGAAAGCTAATCAATGGAGGAATTTCAAGTATAT- ----------------------------------------------
--------------------------------------------------------------------------------TATGGGCAATGGAAATTTCAATAT
--------------------------------------------------------------------------------TAAGGGCAATGGAAATTGCAATAT
--------------------------------------------------------------------------------TATGCGCAATGCAAATTTCAATATT
```

Thanks.