Course Homepage

http://tinyurl.com/phy2013a

get this presentation from the "presentations" link from the above URL
Interpreting Phylogenies

Friday 26th April 2013

Department of Genetics
Cambridge University, UK

Sarah Parks
EMBL-EBI, Hinxton, UK

Aldia (Cilia) Antoniou
HCMR-IMBG
Heraklion, Crete, Greece

Aidan Budd
EMBL Heidelberg, Germany

Adrian Friday
Department of Zoology
Cambridge University, UK
Course Aims
Course Aims

• Help you interpreting phylogenetic trees:
  • more quickly
  • more effectively
• Get 1:1 help with specific questions from the trainers
• Find/meet/talk with others working on similar topics

but not...

• Help you better understand:
  • details of cutting-edge approaches to phylogenetic tree estimation
  • mathematics used in tree estimation
  • how to use commonly-used phylogeny estimation tools
Course Aims

• Help you interpreting phylogenetic trees:
  • more quickly presentations, discussions,
  • more effectively demonstrations, exercises

• Get 1:1 help with specific questions from the trainers
• Find/meet/talk with others working on similar topics
  introductions, coffee breaks, lunch, pub, LinkedIn group
Introductions
Introductions

Trainers:
- Names and where we work
- Outline our experience with these topics
- Three words describing our work
  - help highlight the kinds of questions we might be able to help you with

Trainees:
- Names and where you work
- Three words describing your work
  - help you find other trainees with similar interests
Introductions:
Experience using Phylogenetic Methods

Stand in a line, 0 experience at one end, Adrian at the other:
• silently/without discussing with each other
Introductions: Experience using Phylogenetic Methods

Stand in a line, 0 experience at one end, Adrian at the other:

• silently/without discussing with each other

Cerys
0 years experience

Gatty

Adrian
30+ years full time
Introductions: Experience using Phylogenetic Methods

Stand in a line, 0 experience at one end, Adrian at the other:

• silently/without discussing with each other

Cerys
0 years experience

Gatty

• then discuss your experience, and reposition accordingly

Adrian
30+ years full time
Introductions: Experience using Phylogenetic Methods

Stand in a line, 0 experience at one end, Adrian at the other:

• silently/without discussing with each other

• then discuss your experience, and reposition accordingly

Sit with someone from other end of the line

• more experienced people get practice explaining the topic
• less experienced people get help understanding the topic
• help you find out what each other are doing
Introductions: Process or Pattern?

Adrian will explain the difference between process and pattern in the context of evolution...

Stand in a line, most interested in pattern at one end, process at the other

- silently/without discussing with each other

pattern \[\xrightarrow{\text{Adrian}}\] process
Introductions:
Process or Pattern?

Adrian will explain the difference between process and pattern in the context of evolution...

Stand in a line, most interested in pattern at one end, process at the other

• silently/without discussing with each other

pattern  process

• then discuss your experience, and reposition accordingly
Before we start

• Mixture of presentations, demonstrations, and exercises
• Working in pairs is encouraged
• Please ask questions at any point
• This presentation is available from the “presentation” link from the course pages
  • http://tinyurl.com/phy2013a
Why do people care about phylogenetics?

and

How do people use phylogenetics?
Why Care About Phylogenies?
An Example

They investigated the basis for transmission dynamics of rabies virus in North African dogs.

Rabies is a very important disease (major economic impact).

They were interested in whether human activity played a role in influencing the dynamics of virus transmission.

Is this a tree you would expect to see, if human activity:

A. does
B. does not

significantly influence the dynamics of virus transmission?

Figure 1. MCC tree of 287 sequences of the Africa 1 clade, estimated from the N, P and G-L genes and intergenic regions of dog RABV, and showing the spatial structure of the viral lineages.

Why Care About Phylogenies?
An Example

They investigated the basis for transmission dynamics of rabies virus in North African dogs.

Rabies is a very important disease (major economic impact).

They were interested in whether human activity played a role in influencing the dynamics of virus transmission.

Is this a tree you would expect to see, if human activity:

A. does
B. does not

significantly influence the dynamics of virus transmission?

Figure 1. MCC tree of 287 sequences of the Africa 1 clade, estimated from the N, P and G-L genes and intergenic regions of dog RABV, and showing the spatial structure of the viral lineages.

http://www.plospathogens.org/article/info:doi/10.1371/journal.ppat.1001166
Why Care About Phylogenies?
An Example

They investigated the basis for transmission dynamics of rabies virus in North African dogs.

Rabies is a very important disease (major economic impact).

They were interested in whether human activities likely to play a role in influencing the dynamics of virus transmission.

Authors ask whether human activities likely to play a significant role in virus spread.

Is this a tree you would expect to see, if human activity:

A. does
B. does not

significantly influence the dynamics of virus transmission?


http://www.plospathogens.org/article/info:doi/10.1371/journal.ppat.1001166
Why Care About Phylogenies? An Example

They investigated the basis for transmission dynamics of rabies virus in North African dogs.

Rabies is a very important disease (major economic impact).

They were interested in whether human activity played a role influencing the dynamics of virus transmission.

Authors ask whether human activities likely to play a significant role in virus spread.

Does your observation of this tree make you think it more likely that humans play a role in promoting virus spread?

Figure 1. MCC tree of 287 sequences of the Africa 1 clade, estimated from the N, P and G-L genes and intergenic regions of dog RABV, and showing the spatial structure of the viral lineages.

http://www.plospathogens.org/article/info:doi/10.1371/journal.ppat.1001166
Why Care About Phylogenies?
An Example

Does your observation of this tree make you think it more likely that humans play a role in promoting virus spread?

Begin by making up your mind on your own, silently...

We'll take a vote to see who, after looking at the tree, considers that human intervention is:

A. more likely
B. not more likely
to contribute to spread of the virus

Then discuss this with your neighbours, we'll see if that changes the voting

Don't click to the next slide yet!!

Figure 1. MCC tree of 287 sequences of the Africa 1 clade, estimated from the N, P and G-L genes and intergenic regions of dog RABV, and showing the spatial structure of the viral lineages.

http://www.plospathogens.org/article/info:doi/10.1371/journal.ppat.1001166

http://www.plospathogens.org/article/info:doi/10.1371/journal.ppat.1001166

Aidan Budd, EMBL Heidelberg
Why Care About Phylogenies?
An Example

Does your observation of this tree make you think it more likely that humans play a role in promoting virus spread?

On the basis of this tree (and some other cunning analyses) the authors conclude that the data makes it is:

A. more likely that human intervention plays a key role in spread of the virus...

The tree supports this in by showing virus spreading strongly within political (national) boundaries, with very few infections crossing country borders

Obvious implications for public health policy

Figure 1. MCC tree of 287 sequences of the Africa 1 clade, estimated from the N, P and G-L genes and intergenic regions of dog RABV, and showing the spatial structure of the viral lineages.

http://www.plospathogens.org/article/info:doi/10.1371/journal.ppat.1001166

Aidan Budd, EMBL Heidelberg
Why Care About Phylogenies?
Important for Many of Publications

• Many molecular evolution/phylogenetics articles are published each year
• Number of articles published increases year-by-year
• Estimated 15 phylogenies published each day*

* Genomics. Genomics and the tree of life.
  Rokas A
  PMID: 17272702

** Source: ISI Web of Knowledge as of 28.03.2010

---

** Phylogenies**

<table>
<thead>
<tr>
<th>Year</th>
<th>1991</th>
<th>1993</th>
<th>1995</th>
<th>1997</th>
<th>1999</th>
<th>2001</th>
<th>2003</th>
<th>2005</th>
<th>2007</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Publications</td>
<td>0</td>
<td>2000</td>
<td>4000</td>
<td>6000</td>
<td>8000</td>
<td>10000</td>
<td>12000</td>
<td>+100,000 since 1975</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

query: ((phylogeny OR phylogenetic OR phylogenies OR phylogenetics))

---

** Molecular Evolution**

<table>
<thead>
<tr>
<th>Year</th>
<th>1991</th>
<th>1993</th>
<th>1995</th>
<th>1997</th>
<th>1999</th>
<th>2001</th>
<th>2003</th>
<th>2005</th>
<th>2007</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Publications</td>
<td>0</td>
<td>1000</td>
<td>2000</td>
<td>3000</td>
<td>4000</td>
<td>5000</td>
<td>6000</td>
<td>7000</td>
<td>72638 since 1975</td>
<td></td>
</tr>
</tbody>
</table>

query: (((evolution OR evolutionary OR evolves OR evolve) AND (molecule OR molecular OR molecules)))
Why Care About Phylogenies?
Many Highly Cited Articles in the Field

Some molecular-evolution-related articles are very highly cited

<table>
<thead>
<tr>
<th>Method/Software</th>
<th>Year</th>
<th>Original Citation</th>
<th># of Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEGA3</td>
<td>2004</td>
<td>Kumar et al. 2004 Brief Bioinform.;5(2):150-63. PMID: 15260895</td>
<td>6630</td>
</tr>
<tr>
<td>MRBAYES</td>
<td>2001</td>
<td>Huelsenbeck and Ronquist 2001 Bioinformatics;17(8):754-5 PMID: 11524383</td>
<td>5707</td>
</tr>
</tbody>
</table>

§ Source: ISI Web of Knowledge, as of 29.03.2010

as of 2006 in the ISI web of knowledge:
* most cited paper that year, 26th most cited in the entirety of science
** second most cited paper that year, 31st most cited paper in the entirety of science
Why Care About Phylogenies?
Many Important Applications

• Epidemiology
• Forensics
• Selecting conservation targets
• Monitoring trade in illegal organisms

• Central part of many bioinformatics tools inc:
  • predicting function (BLAST and many other methods)
  • building MSAs

• Key role across basic evolutionary research
  • characterising processes of evolutionary transformation
  • estimating patterns of transmission of genetic material

the large number, and high importance of these applications drives the large number of publications and citations
Why Care About Phylogenies?
Intellectual Curiosity

Phylogenies/evolutionary analyses can address questions that many humans find inherently interesting...

"becoming" questions e.g.

• How did humans become so different (intelligence etc.) from other great apes?

"where do we come from" questions e.g.

• What did our ancestors look like when they first emerged from the water to the land?
How/Why we use Phylogenies

We have a substantive question

Build a model involving parameters that, if known, could answer the question

- sequences cluster according to political boundaries?
- likely role for human activity in virus spread
- sequences cluster by geography rather than political boundaries?
- less evidence for role of human activity in virus spread

e.g. topology parameter of dog rabies phylogeny
How/Why we use Phylogenies

We have a substantive question
e.g. "Can we identify factors promoting disease transmission that could be addressed via public-health measures?"

Build a model involving parameters that, if known, could answer the question
e.g. topology parameter of dog rabies phylogeny
  • sequences cluster according to political boundaries?
    • likely role for human activity in virus spread
  • sequences cluster by geography rather than political boundaries?
    • less evidence for role of human activity in virus spread
How/Why we use Phylogenies

We have a substantive question

e.g. "Can we identify factors promoting disease transmission that could be addressed via public-health measures?"

e.g. identifying associations of the rabies virus with:
  • non-human vectors
  • specific human activities (e.g. trade, farming, migration)

Build a model involving parameters that, if known, could answer the question

e.g. topology parameter of dog rabies phylogeny
  • sequences cluster according to political boundaries?
    • likely role for human activity in virus spread
  • sequences cluster by geography rather than political boundaries?
    • less evidence for role of human activity in virus spread
How/Why we use Phylogenies

Collect observations informative about the model parameter(s)
  e.g. build multiple sequence alignment of dog rabies viruses taken from area of interest

Find best estimate(s) of the parameter(s), conditioned on these observations
  e.g. Bayesian estimation of tree topology

Estimate sampling/random error associated with parameter estimate
  e.g. internal branch posterior probabilities

Answer your question using these parameter estimates
  e.g. human trafficking in dogs likely to play a significant role in rabies virus transmission in North Africa
"Louisiana gastroenterologist" (Richard J. Schmidt) accused of attempted second degree murder for allegedly injecting a former lover with blood from one of his HIV+ patients

Phylogenetic analyses used as evidence in the trial

Few applications where we care this much (and the accused even more!) that the analysis is done as well as possible

i.e relevant parameters estimated as accurately and precisely as possible

Molecular evidence of HIV-1 transmission in a criminal case.
Metzker ML, Mindell DP, Liu XM, Ptak RG, Gibbs RA, Hillis DM.
PMID: 12388776
• victim (V)
• gastroenterologist's HIV+ patient (P)
• randomly-selected isolates of contemporary sequences from Louisiana (LA)

Note: several samples taken from patient (7) and victim (2)

Imagine you are a juror on this case....

How/would seeing this phylogeny influence the verdict you would choose in this case?

1. More likely to choose "guilty"
2. More likely to choose "not guilty"
3. Would not influence your choice of verdict

Molecular evidence of HIV-1 transmission in a criminal case.
Metzker ML, Mindell DP, Liu XM, Ptak RG, Gibbs RA, Hillis DM.
PMID: 12388776
Applications of Phylogenetics: Selecting Conservation Targets

Prioritise organisms for inclusion in conservation programs, taking into account

- phylogenetic diversity
- conservation costs
- probability of extinction

Resource-aware taxon selection for maximizing phylogenetic diversity.
Pardi F, Goldman N.
MID: 17558965
Figure 4
Applications of Phylogenies: Monitoring Trade in Illegal Organisms


Determining the source animals for unlabelled meats

Used here to track illegal trade in protected whale and dolphin species.
Applications of Phylogenetics: Bioinformatics Tools

STRING 8—globally view on proteins and their functional interactions in 630 organisms.
PMID: 18940858

Further example
Chica C, Labarga A, Gould CM, López R, Gibson TJ.
PMID: 18460207
Applications of Phylogenetics: Building Progressive Multiple Alignments


Tuesday, 30 April 13
Rooted Phylogenies

Terminology and Concepts
Trees: Branches and Nodes

Trees consist of:
Trees: Branches and Nodes

Trees consist of:
- branches
Trees: Branches and Nodes

Trees consist of:
- branches
- nodes (ends of branches)
Try writing your own definition of:

- "phylogenetic tree"
- "node"
- "branch"
Try writing your own definition of:

- "phylogenetic tree"
- "node"
- "branch"

Writing makes you to be explicit about what you mean, and can help identify issues you are unclear about.
Try writing your own definition of:

- "phylogenetic tree"
- "node"
- "branch"

Writing makes you to be explicit about what you mean, and can help identify issues you are unclear about.

With your neighbour, try to write together definitions you are both happy with.
Phylogenetic Trees

Some possible definitions:
Some possible definitions:

**branch**

Branches represent lineages of taxonomic units that link nodes within a phylogenetic tree.
Phylogenetic Trees

Some possible definitions:

**branch**

Branches represent lineages of taxonomic units that link nodes within a phylogenetic tree.

Even with some very basic concepts there are still arguments about correct/good definitions/interpretations.
Phylogenetic Trees

Some possible definitions:

**phylogenetic tree:**
A geometric description representing the pathways of transmission of genetic information

**branch**
Branches represent lineages of taxonomic units that link nodes within a phylogenetic tree.

Even with some very basic concepts there are still arguments about correct/good definitions/interpretations
Phylogenetic Trees

Some possible definitions:

**phylogenetic tree:**
A geometric description representing the pathways of transmission of genetic information

**node**
In phylogenetic trees, nodes represent taxonomic units. Nodes attached to only a single terminal branch represent operational taxonomic units. Nodes attached to more than one branch represent hypothetical taxonomic units.

**branch**
Branches represent lineages of taxonomic units that link nodes within a phylogenetic tree.

Even with some very basic concepts there are still arguments about correct/good definitions/interpretations.
Internal/External Nodes/Branches

Branches and Nodes are either:

```
  a
 / \
 b   c
 /   /\n d   e
 /   / \n f   g
 /   /  \n h   i
```
Branches and Nodes are either:

- **internal/interior**
- **Node** - at the intersection of two or more branches
Internal/External Nodes/Branches

Branches and Nodes are either:

- **internal/interior**

- **Node** - at the intersection of two or more branches
- **Branch** - links two internal nodes
Branches and Nodes are either:

**Node** - at the intersection of two or more branches

**Branch** - links two internal nodes

**internal/interior**

**external/terminal**

**Node** - associated with an extant sequence/OTU (operational taxonomic unit)
Branches and Nodes are either:

**internal/interior**
- **Node** - at the intersection of two or more branches
- **Branch** - links two internal nodes

**external/terminal**
- **Node** - associated with an extant sequence/OTU (operational taxonomic unit)
- **Branch** - links an external and an internal node
Branches

- represent successive generations of “taxa”
- “later” taxa have “earlier” taxa as their ancestors
- i.e. a lineage
- time flows from the base of the tree to the tips
Branches

- represent successive generations of “taxa”
- “later” taxa have “earlier” taxa as their ancestors
- i.e. a lineage
- time flows from the base of the tree to the tips
Branches

- represent successive generations of “taxa”
- “later” taxa have “earlier” taxa as their ancestors
- i.e. a lineage
- time flows from the base of the tree to the tips
Branches

- represent successive generations of “taxa”
- “later” taxa have “earlier” taxa as their ancestors
- i.e. a lineage
- time flows from the base of the tree to the tips
Branches

- represent successive generations of “taxa”
- ‘later” taxa have “earlier” taxa as their ancestors
- i.e. a lineage
- time flows from the base of the tree to the tips
Internal Nodes

- represent hypothetical ancestral taxa/sequences/organisms
- i.e. **HTUs** - hypothetical taxonomic units

Time
Internal Nodes

- represent hypothetical ancestral taxa/sequences/organisms
- i.e. **HTUs** - hypothetical taxonomic units
Internal Nodes

- represent hypothetical ancestral taxa/sequences/organisms
- i.e. HTUs - hypothetical taxonomic units

Root (Root Node)
- A "special" internal node
- The most recent common ancestor of all OTUs
- Usually implies many other less recent common ancestors
Tree Representations

Most rooted tree figures use a “rectangular” rather than a “diagonal” representation.

Rectangular trees represent internal nodes with lines perpendicular to lines representing the branches.
Sequence Alignments
"Anatomy" of a Sequence Alignment

W K K L G S N V G
|   |   |   |   |   |
W G K V K N V D
"Anatomy" of a Sequence Alignment

Residues:
Monomers within a polymer (polypeptide or polynucleotide) chain
"Anatomy" of a Sequence Alignment

**Residues:**
Monomers within a polymer (polypeptide or polynucleotide) chain

**Sequences:**
List of residues in a polymer chain...
...listed in the same order they occur within the polymer

Tuesday, 30 April 13
"Anatomy" of a Sequence Alignment

I:1 residue correspondences/relationships

Correspondences between
- a single residue in one sequence and
- a single residue in another sequence
"Anatomy" of a Sequence Alignment

Residue has no equivalent in the top sequence
i.e. no residue in the top sequence has a 1:1 relationship with this residue
Could perhaps say there is a "1:2" relationship between this residue and these residues.
"Anatomy" of a Sequence Alignment

Residue has no equivalent in the top sequence i.e. no residue in the top sequence has a 1:1 relationship with this residue.

Could perhaps say there is a "1:2" relationship between this residue and these residues.

However, alignments focus on 1:1 relationships.
# Sequence Alignment Within a Grid

Often represented using a grid/matrix:

<table>
<thead>
<tr>
<th>WK KL GS NV G</th>
<th>WK KL LS GS NV V</th>
<th>W Ground/K N V D</th>
</tr>
</thead>
<tbody>
<tr>
<td>WG KV KN V D</td>
<td>WG KV KN V D</td>
<td>WG KV KN V D</td>
</tr>
</tbody>
</table>

Tuesday, 30 April 13
Sequence Alignment Within a Grid

Often represented using a grid/matrix:

One sequence per row
Sequence Alignment Within a Grid

Often represented using a grid/matrix:

One sequence per row

Residues in the same column are 'equivalent'
**Sequence Alignment Within a Grid**

Often represented using a **grid/matrix**:

One sequence per row

Residues in the same column are 'equivalent'

Gap characters (usually "-") indicate that the sequence contains no residues 'equivalent' to other residues in that column
Build an Automatic MSA

Search Internet for "EBI Muscle"

http://www.ebi.ac.uk/Tools/msa/muscle/
Build an Automatic MSA

Copy and paste sequences in FASTA format
Click "Submit"
Build an Automatic MSA

Wait for result to be returned

Click "Download Alignment File" to reach plain-text version of alignment
Build an Automatic MSA

Download file or copy-paste text into text editor to store alignment on local computer

View alignment in MSA viewer (e.g. JalView) etc.

> TBA_ENVRU Q8R16 Tubulin alpha chain OS=Encephalitozoon cuniculi GN=TUB1 PE=3 SV=1
MREIISLHICAGVQIGNACWELYCKEHGILPNGQLDQNKMD-----DESAEFSPTSVG
TYVPRTLMVLEPGVLSKTKGYRELYHPCQLISGKEDANAYRCHYTVGKE11EPAM
EOIIRRMADSDCGLQGFLYHSFGGTGSGSFAAMLMDRLAEEFGKSKLEFSVYPARKIA
AVVEPYNSTLLTHTTLDSDCSFLVDNEAIYDME----NLG1QRPYTDINRVI4QQVSSIT
ASLRFPGSLNVLDTLEFTQNLVPYRIRHPFLVAYSPMLESKEAHAHEKLQVEITNACEFENQ
QMVRCDTRKGYACCLLFRGDVNPKEANGATANVKAXKTNQFVEWCPTFKOVGINSRK
PTLVDGMAEAVSRACVLSNNTA1S4154WKL1A4MNFSKRAFVHYVYEGMEEGEFS
EAREDMLLEDYERISSNA------EPVDEY

> TBA1A_HUMAN Q71U36 Tubulin alpha-1A chain OS=Homo sapiens GN=TUBA1A PE=1 SV=1
MRCGSIHGVCAGVQIGNACWELYCKEHGILPNGQLDQNKMD-----DESAEFSPTSVG
KHPVPRAVFDLEPTVDEIRC7YQLFHPQELIQTGKEKANAYRCHYTVGKE11IDVL
DR1RKLADQCTLGQGFLYHSFGGTGSGTSSLMLRSVDGCKSKLEFSIYPAQST
AVVEPYNSTLLTHTTLDSDCSFLVDNEAIYD1CRNLDIERPTYTNLNLQ1IVSSIT
ASLRFDGALNLVDEFTQNLVPYRIRHNPFLVAYSPMLESKEAHAHEKLQVEITNACEFENQ
QMVRCDTRKGYACCLLFRGDVNPKEANGATANVKAXKTNQFVEWCPTFKOVGINSRK
PTLVDGMAEAVSRACVLSNNTA1S4154WKL1A4MNFSKRAFVHYVYEGMEEGEFS
EAREDMLLEDYERISSNA------EPVDEY

> TBA3_YEAST P09734 Tubulin alpha-3 chain OS=Saccharomyces cerevisiae GN=TUB3 PE=1 SV=1
MREVISINVRGACAGVQIGNACWELYCKEHGILPNGQLDQNKMD-----DESAEFSPTSVG
KFPYPRAIYVDFENPVIDEHEYNLQIKNFXAANAYRCHYTVGKE11VDSVE
ER1RKMADCDCGLQGFLYHSFGGTGSGSFAAMLMDRLAEEFGKSKLEFSVYPARKIA
SVVEPYNSTLLTHTTLDSDCSFMDNEAIYD1CRNLDIERPTYTNLNLQ1IVSSIT
ASLRFDGASNLVDNLEFTQNLVPYRIRHNPFLVAYSPMLESKEAHAHEKLQVEITNACEFENQ
QMVRCDTRKGYACCLLFRGDVNPKEANGATANVKAXKTNQFVEWCPTFKOVGINSRK
PTLVDGMAEAVSRACVLSNNTA1S4154WKL1A4MNFSKRAFVHYVYEGMEEGEFS
EAREDMLLEDYERISSNA------YAEF

I, EMBL Heidelberg

Tuesday, 30 April 13
Which Automatic Aligner to Use?

There are many different automatic MSA tools

Different automatic MSA tools are designed for different tasks

- CLUSTALX, MUSCLE, PROBCONS
divergent protein sequences

- NAST
multiple alignment of 16S rRNA genes

- PRANK
multiple alignment of relatively similar DNA sequences in an evolutionary context

- EXPRESSO(3DCoffee)
multiple alignment of protein sequences, some of which have 3D structural information

- MAUVE, Enredo
multiple alignment of genomes

- and many others...

Begin your analysis using a tool designed for your question of interest
e.g. working on 16S rRNA genes? try out NAST
“Equivalence”/similarity of residues

Residues in the same column either:

- Structurally equivalent/similar
- Evolutionary equivalent/related/homologous

Different applications assume different types of equivalence

Different types of similarity not necessarily equivalent
Structural Similarity

Unaligned

Bacterial toxins 1ji6 and 1i5p

Structurally Aligned
Residues with a similar structural context may lie almost on top of each other within a structural alignment. Clearly, the dark green and red side chains have more similar structural contexts than they do with the adjacent light-coloured side chains.
Structural Similarity

Chain 1: 16KVGS\_LIGKR---IL\_SELWGIIFPSGST
1111111111111111111

Chain 2: 16VVGP\_FAGALTSFY\_SFLNTIWP-SDA
Some regions of the structures do not have structurally equivalent residues in the other structure.

Alignment gaps are a sure sign of such residues.

Placing such residues in the same column as residues from other sequences is a misalignment - to be avoided!

<table>
<thead>
<tr>
<th>l15p:</th>
<th>DNFLNPTQN---PVPLSITSSVN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>111111 11111111111</td>
</tr>
</tbody>
</table>

| lji6:   | NSWKKTPLSLRSKRSQDRIRELFS |
Evolutionary "Equivalence"

Residues are "evolutionarily equivalent" when:
Residues are "evolutionarily equivalent" when:

- they are derived from the same residue in an ancestral sequence
Evolutionary "Equivalence"

Residues are "evolutionarily equivalent" when:

- they are derived from the same residue in an ancestral sequence
- the only mutations experienced during divergence from this ancestral residue were **point substitutions**
Residues are "evolutionarily equivalent" when:

• they are derived from the same residue in an ancestral sequence
• the only mutations experienced during divergence from this ancestral residue were **point substitutions**

![Diagram showing evolutionary relationships between amino acid sequences](image)
Evolutionary "Equivalence"

Residues are "evolutionarily equivalent" when:

- they are derived from the same residue in an ancestral sequence
- the only mutations experienced during divergence from this ancestral residue were **point substitutions**

```
AGWYTI  →  AGWWTI  →  AGWWTI  →  AGWWTI
AGWYTI  →  AGWYTI  →  AAWYTI  →  AAQQQWYTI
```

*Point Substitution*  
Y-W  

*Point Substitution*  
G-A  

*Insertion*  
QQQ  

*Two copies of gene generated*

A misalignment

```
AGWYTI  AGWWTI  AGWWTI  AGWWTI  AG---WWTI
AGWYTI  AGWYTI  AGWYTI  AAWYTI  AA---WWTI
AGWYTI  AGWYTI  AGWYTI  AAWYTI  AAQQQWYTI
```

Aidan Budd, EMBL Heidelberg
Quiz

Evolutionary Interpretation of Alignments

Which alignment of the final sequences (X, Y or Z) only places residues in the same column if they are related by substitution events?

X

KGE PG--- IGLPG
KGI PGD PA FGD PG
RG I PGE VL GA Q PG

Y

KGE PG----- IGL----- PG
KGI PG----- DPA FGD PG
RG I PGE VL GA Q PG

Z

KGE------- PG I GL------- PG
KGI PG------- DPA FGD PG
RG I PGE VL GA Q------- PG

Aidan Budd, EMBL Heidelberg
"True" alignment given history described above

KGE--------PGIGL--------PG
KGIPG------------DPAFGDGP
RGIPGEVLGAQ--------PG

PRANK
RGIPGEVLGAQPG
KGIPGDPAFGDPG
---KGEPGIGLPG
QuizEvolutionary Interpretation of Alignments

Different automatic MSA software gives different results

All are different from the "true" alignment (assuming the scenario of transformation on the previous slide is true)...

... because that scenario is very unlikely under the models of evolutionary transformation incorporated within these tools
Non-Equivalence of Evolutionary and Structural Alignments

Demonstration 1:
Structural equivalence without evolutionary equivalence
Structural alignment of SH3-interaction motifs from nef and ncf1

aligned ncf1/nef1 SH3 interaction motifs
Non-Equivalence of Evolutionary and Structural Alignments

Demonstration 2b:
Sequences differ by ONE amino acid residue and have different folds

A minimal sequence code for switching protein structure and function.
Alexander PA, He Y, Chen Y, Orban J, Bryan PN.
PMID: 19923431

Aidan Budd, EMBL Heidelberg
Quiz - Numbers of Insertions

The **minimum** number of insertion events required to account for the section of haemoglobin alignment shown above is?

(a) 2  (b) 1  (c) 0  (d) 3
Quiz - Numbers of Insertions

The **minimum** number of insertion events required to account for the section of haemoglobin alignment shown above is?

If all sequences are the same length, we can explain their diversity without inferring ANY insertions or deletions

If and alignment contains sequences that are all either length $x$ or $y$, then we can explain their diversity by inferring just one insertion or deletion
Quiz - Numbers of Insertions

The minimum number of insertion events required to account for the section of haemoglobin alignment shown above is?

We can ALWAYS explain observed sequence length diversity with:

• 0 insertions (all length variation due to deletion)
• 0 deletions (all length variation due to insertion)
• a combination of insertions and deletions

Perhaps we should instead focus on inferring the most likely scenario?

(Although if this is not particularly relevant for our analysis, perhaps we should focus instead on something completely different!)
Branch Lengths
Scaled and Unscaled Trees
Unscaled Trees

Branch lengths provide no information

Branch lengths usually chosen to align OTU labels

Re-rooting the tree typically changes the choice of branch lengths

Same unscaled unrooted tree
Branch length usually represents some measure of the difference/distance between TUs at ends of the branch.
Scaled Trees

Branch length usually represents some measure of the difference/distance between TUs at ends of the branch.

Tree should be presented together with a scale bar.

Same **unscaled** unrooted tree

Same **scaled** unrooted tree
Scaled Trees

Branch length usually represents some measure of the difference/distance between TUs at ends of the branch.

Tree should be presented together with a scale bar.

For rectangular trees, “node lines” are NOT branches! Their length provides no indication of intertaxa difference/distance!

i.e. distance between taxa C and G is the sum of the green and cyan lines (it does NOT include the length of the red line!)

Same unscaled unrooted tree

Same scaled unrooted tree
Usually an ESTIMATE of the EXPECTED/AVERAGE number of substitutions per site between two sequences:

SeqA: IKTILKWWSP
SeqB: IKTIVKWDSP
Branch Lengths

**Usually** an ESTIMATE of the EXPECTED/AVERAGE number of substitutions per site between two sequences

<table>
<thead>
<tr>
<th>SeqA</th>
<th>IKTILKWWSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeqB</td>
<td>IKTIVKWDSAP</td>
</tr>
</tbody>
</table>

If we assume:
- All identical residues between two sequences have not experienced substitutions
- All different residues have experienced one substitution

Mean/Average No. Substitutions $= \frac{2}{10} = 0.2$
Branch Lengths

**Usually** an ESTIMATE of the EXPECTED/AVERAGE number of substitutions per site between two sequences

SeqA  IKTILKWWSP
SeqB  IKTIVKWDSP

If we assume:
- All identical residues between two sequences have not experienced substitutions
- All different residues have experienced one substitution

Mean/Average No. Substitutions = 2/10 = 0.2

SeqA  0.2  SeqB
Branch Lengths

**Usually** an ESTIMATE of the EXPECTED/AVERAGE number of substitutions per site between two sequences

SeqA: IKTILKWSP  
SeqB: IKTIVKWDSP

Branch-length estimate depends on SUBSTITUTION MODEL
Usually an ESTIMATE of the EXPECTED/AVERAGE number of substitutions per site between two sequences

`SeqA  IKTILKWWSP`

`SeqB  IKTIVKWDSP`

Branch-length estimate depends on SUBSTITUTION MODEL.
**Branch Lengths**

**Usually** an ESTIMATE of the EXPECTED/AVERAGE number of substitutions per site between two sequences

<table>
<thead>
<tr>
<th>SeqA</th>
<th>IKTILKWWSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>SeqB</td>
<td>IKTIVKWDSP</td>
</tr>
</tbody>
</table>

Branch-length estimate depends on SUBSTITUTION MODEL

Further assumptions of this model

- All alignment positions/residues evolve (are substituted at) the same rate
- All residues substitute to all other residues at the same rate
  i.e. $A \rightarrow G$ at same frequency as $A \rightarrow W$

<table>
<thead>
<tr>
<th>SeqA</th>
<th>0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SeqB</td>
</tr>
</tbody>
</table>
More Rooted Tree Terminology
Relatedness
(in the context of phylogenetic trees)

Often, we estimate phylogenies, to inform our ideas about the patterns of relatedness between taxa.

"relatedness" in this context has a specific meaning, as exemplified here:

"the more recently species share a common ancestor, the more closely related they are" *

As "relatedness" has other meanings in other contexts, there can be some confusion about its meaning in a phylogenetic context.

As many analyses focus on patterns of relatedness, it is important to have a clear understanding of how the term is used in this context.

Thus, in the next slides, we will look at several examples of how the word is used when describing phylogenetic relationships.

* Evolution. The tree-thinking challenge.
  Baum DA, Smith SD, Donovan SS.
  PMID: 16284166
Relatedness
(in the context of phylogenetic trees)

"the more recently species share a common ancestor, the more closely related they are" *

of all the OTUs represented in this tree

Relatedness
(in the context of phylogenetic trees)

"the more recently species share a common ancestor, the more closely related they are" *

of all the OTUs represented in this tree

\[
g \text{ is more closely related to } h
\]

* Evolution. The tree-thinking challenge.
Baum DA, Smith SD, Donovan SS.
PMID: 16284166
"the more recently species share a common ancestor, the more closely related they are" *

of all the OTUs represented in this tree

\(\textcolor{red}{g}\) is more closely related to \(\textcolor{red}{h}\)

than \(\textcolor{red}{g}\) is to \(\textcolor{blue}{i}\)

* Evolution. The tree-thinking challenge.
Baum DA, Smith SD, Donovan SS.
PMID: 16284166
Relatedness
(in the context of phylogenetic trees)

"the more recently species share a common ancestor, the more closely related they are" *

of all the OTUs represented in this tree

g is more closely related to h

than g is to i

because g and h share common ancestors

that neither share with i

* Evolution. The tree-thinking challenge. 
Baum DA, Smith SD, Donovan SS.
PMID: 16284166
Relatedness
(in the context of phylogenetic trees)

of all the OTUs represented in this tree

\( g \) is equally distantly related to \( b \)

as it is to \( e \)
Relatedness
(in the context of phylogenetic trees)

of all the OTUs represented in this tree

\( g \) is equally distantly related to \( b \) as it is to \( e \)

because \( g \) shares the same most recent common ancestor with \( b \) as it does with \( e \)
Relatedness
(in the context of phylogenetic trees)

of all the OTUs represented in this tree

\( i \) is **most closely related** to \( g \) and \( h \)

(i.e. \( i \) is the **sister group** of \( g \) and \( h \)...
which is equivalent to saying \( g \) and \( h \) are the sister group of \( i \))
Relatedness
(in the context of phylogenetic trees)

of all the OTUs represented in this tree

\[ \text{i is most closely related to g and h} \]

(i.e. \(\text{i is the sister group of g and h} \)

which is equivalent to saying \(\text{g and h are the sister group of i}\)

because \(\text{i shares common ancestors with g and h that it does not share with any other OTUs in the tree}\)
Relatedness
(in the context of phylogenetic trees)

"the more recently species share a common ancestor, the more closely related they are" *

Which of the following taxa is more closely related to?

* Evolution. The tree-thinking challenge.
  Baum DA, Smith SD, Donovan SS.
  PMID: 16284166
Tree Topology - Unscaled Trees

Trees with identical topologies...
Trees with identical topologies...

... describe the same set of "relatedness statements" between taxa
Tree Topology - Unscaled Trees

Trees with identical topologies...
... describe the same set of "relatedness statements" between taxa

i.e. any (true!) statement such as

"c is more closely related to a than c is to e"

is true for all trees with identical topologies
Trees with \textit{identical topologies}...

... describe \textbf{the same set} of "relatedness statements" between taxa

i.e. any (true!) statement such as

"\textbf{c is more closely related to a than c is to e}\"

is true for all trees with identical topologies

Trees with \textit{different topologies}...
Tree Topology - Unscaled Trees

Trees with **identical topologies**...

... describe **the same set** of "relatedness statements" between taxa

i.e. any (true!) statement such as

"**c is more closely related to a than c is to e**"

is true for all trees with identical topologies

Trees with **different topologies**...

... describe **different sets** of "relatedness statements" between taxa

Tuesday, 30 April 13
Tree Topology - Scaled Trees

Trees with **identical topologies**...

Trees with **different topologies**...

**identical** topologies

**different** topologies
Tree Topology - Scaled Trees

Trees with **identical topologies**...

... describe the same set of "relatedness statements" between taxa

Trees with **different topologies**...

**identical** topologies

**different** topologies
Tree Topology - Scaled Trees

Trees with identical topologies...

... describe the same set of "relatedness statements" between taxa.

i.e. any (true!) statement such as

"c is more closely related to a than c is to e"

is true for all trees with identical topologies.

Trees with different topologies...

identical topologies

different topologies
Trees with **identical topologies**...

...describe **the same set** of "relatedness statements" between taxa

i.e. any (true!) statement such as

"**c** is more closely related to **a** than **c** is to **e**"

is true for all trees with identical topologies

Trees with **different topologies**...

...describe **different sets** of "relatedness statements" between taxa
Unrooted Phylogenies
There's no root on the tree...

...which is usually interpreted as meaning that these taxa are related by a rooted tree but we don't know where the root is.

Many applications of phylogenies require a rooted tree

But many tree estimation tools yield only unrooted trees!
There are multiple **rooted tree topologies** for a given unrooted tree topology.
There are multiple \textbf{rooted tree topologies} for a given unrooted tree topology

Unrooted trees can be rooted on their:
There are multiple **rooted tree topologies** for a given unrooted tree topology.

Unrooted trees can be rooted on their:
- **branches**
There are multiple rooted tree topologies for a given unrooted tree topology.

Unrooted trees can be rooted on their:

- branches
There are multiple rooted tree topologies for a given unrooted tree topology.

Unrooted trees can be rooted on their:

- branches
There are multiple **rooted tree topologies** for a given unrooted tree topology.

Unrooted trees can be rooted on their:
- branches
- interior nodes
There are multiple rooted tree topologies for a given unrooted tree topology.

Unrooted trees can be rooted on their:
- branches
- interior nodes
Unrooted $\rightarrow$ Rooted

There are multiple **rooted tree topologies** for a given unrooted tree topology.

Unrooted trees can be rooted on their:
- branches
- interior nodes
- terminal nodes
There are multiple rooted tree topologies for a given unrooted tree topology.

Unrooted trees can be rooted on their:
- branches
- interior nodes
- terminal nodes
There are multiple rooted tree topologies for a given unrooted tree topology.

Unrooted trees can be rooted on their:
- branches
- interior nodes
- terminal nodes
This **unrooted** tree can be rooted to yield several different **rooted topologies**.

How many different rooted topologies exist by placing the root on a/an:

- branch? Draw all of these.
- terminal node? Draw one of these.
- internal node? Draw one of these.
add slide(s) showing the answers to the previous question
Quiz

Is OTU d more closely related to:

1. a than it is to b or c
2. b than it is to a or c
3. c than it is to a or b

Is OTU d more closely related to
Is OTU d more closely related to:

1. a than it is to b or c
2. b than it is to a or c
3. c than it is to a or b

We can't answer this question without making some assertion/inference/assumption about the position of the root of the tree.
Quiz

Is OTU d more closely related to:

1. a than it is to b or c
2. b than it is to a or c
3. c than it is to a or b

We can't answer this question without making some assertion/inference/assumption about the position of the root of the tree, as the asserted patterns of relatedness tree are different when the tree is rooted in different positions.
Quiz

Is OTU d more closely related to:

1. a than it is to b or c
2. b than it is to a or c
3. c than it is to a or b

We can't answer this question without making an assertion/inference/assumption about the position of the root of the tree as the asserted patterns of relatedness indicated by a tree are different when the tree is rooted in different positions.

However, we can make some statements about relatedness from an unrooted tree e.g.

Is OTU d more closely related to
Is OTU d more closely related to:

1. a than it is to b or c
2. b than it is to a or c
3. c than it is to a or b

We can't answer this question without making some assertion/inference/assumption about the position of the root of the tree, as the asserted patterns of relatedness indicated by a tree are different when the tree is rooted in different positions.

However, we can make some statements about relatedness from an unrooted tree e.g.

"d is certainly not most closely related to a or to b"

i.e. d is certainly not the sister group of either b or a

Is OTU d more closely related to
Quiz: recognise identical topologies II

Which of the trees has the same TOPOLOGY and ROOT as tree A?

Aidan Budd, EMBL Heidelberg

Tuesday, 30 April 13
Demonstration and Exercise

Viewing and manipulating unscaled trees with NJplot
- Rotating around internal branches
- Re-rooting
Branch Lengths and Visualising Trees

Demonstration and Exercise

• Working with scaled trees in Dendroscope
  • Formatting trees for figures
  • Working with large trees
More Introductions: Which Organisms do you Work on?

Stand within a circle:
- **scale**: (linear) time
- **Aidan**: an extant archaeon
- **Adrian**: an extant bacterium
- **Sarah** and **Cilia**: extant eukaryotes
- **centre**: LUCA (last universal common ancestor)
- **you** stand in at node representing the most recent common ancestor of organisms you work with

at first, **don't** discuss with each other
then **do** discuss with each other and readjust as needed
- help you get to know what each other are doing better
- thinking again about how we interpret trees
NEWICK/PHYLIP Tree Files/Format
Why bother reviewing the NEWICK format...

when it’s not designed to be easy to read by humans?
Why bother reviewing the NEWICK format...

when it’s not designed to be easy to read by humans?

1. Sometimes we need to read it ourselves
   • e.g. if software doesn’t accept our tree as input

2. Provides a different perspective on the structure of phylogenies
   • helps reinforce our understanding of phylogeny structure
NEWICK string includes the terminal node OTU labels

\[(\,(A,B)\,,C)\,;\]
NEWICK Format Tree Strings - Basics

NEWICK string includes the terminal node OTU labels

Paired parentheses represent internal nodes

Parentheses are paired if they contain/enclose the same number of open as close parentheses (i.e. the same number of '(' and ')' characters

\[(a, ((b, c), (d, e)))\]
NEWICK Format Tree Strings - Basics

NEWICK string includes the terminal node OTU labels

Paired parentheses represent internal nodes

Parentheses are paired if they contain/enclose the same number of open as close parentheses (i.e. the same number of '(' and ')' characters

\[(a, (((b, c), (d, e))))\]
NEWICK Format Tree Strings - Basics

NEWICK string includes the terminal node OTU labels

Paired parentheses represent internal nodes

Parentheses are paired if they contain/enclose the same number of open as close parentheses (i.e. the same number of '(' and ')' characters

```plaintext
((A,B),C);
```

```plaintext
(a,((b,c),(d,e)));
```
NEWICK Format Tree Strings - Basics

Comma-separated elements within paired parentheses...

( ((A, B), C) ; )
NEWICK Format Tree Strings - Basics

Comma-separated elements within paired parentheses...

((A,B), C);
Comma-separated elements within paired parentheses...

...are nodes (HTU or OTU) that are direct descendants of the 'enclosing parenthesis pair' node
Comma-separated elements within paired parentheses...

...are nodes (HTU or OTU) that are direct descendants of the 'enclosing parenthesis pair' node

Newick string is terminated by a semicolon

Tuesday, 30 April 13
NEWICK Format - Polytomies

More than two comma-separated nodes in a set of parentheses indicates a polytomy/multifurcation

\[( ((A,B,C),D) ); \]
NEWICK Format - Indicating Rooted Trees

Does the outermost parentheses enclose ONLY TWO NODES?

This indicates a rooted tree

\[ ((A,B,C),D); \]
Do the outermost parentheses enclose THREE NODES?

Convention assumes this is an unrooted tree
(although it could be a rooted tree with a polytomy at the root...)

((A, B), C, D);
NEWICK Format - Branch Lengths

Branch lengths are assigned to nodes using a number preceded by a “:”

These described the lengths of the edges...
from the (descendant) node attached to the branch length ...
to the enclosing/parental node

((A:1,B:1):2,C:3)

Aidan Budd, EMBL Heidelberg
NEWICK Format - Branch Lengths

Branch lengths are assigned to nodes using a number preceded by a “:”

These described the lengths of the edges...
from the (descendant) node attached to the branch length ...
to the enclosing/parental node

\[((A:1,B:1):2,C:3)\]
NEWICK Format - Branch Lengths

Branch lengths are assigned to nodes using a number preceded by a “:”

These described the lengths of the edges...
from the (descendant) node attached to the branch length ...
to the enclosing/parental node

((A:I,B:I):2,C:3)
NEWICK Format - Branch Lengths

Branch lengths are assigned to nodes using a number preceded by a “:”

These described the lengths of the edges...
from the (descendant) node attached to the branch length ...
to the enclosing/parental node

((A:1,B:1):2,C:3)
NEWICK Format - Branch Lengths

Branch lengths are assigned to nodes using a number preceded by a “:”

These described the lengths of the edges...
from the (descendant) node attached to the branch length ... to the enclosing/parental node

(((A:1,B:1):2,C:3)

Aidan Budd, EMBL Heidelberg
NEWICK Format: Compatibility Trouble-Shooting

If tree string not accepted as input then check/try the following points:

• string should contain the same number of open and close parentheses i.e. the same number of '(' and ')' characters
• substitute 'special' characters in the node labels for any of
  • A-Z
  • a-z
  • 0-9
  • _
• remove gaps/whitespace from taxon names
• remove all whitespace from the string
• include the entire string on a single line of the input file
• remove branch length from the root node
• remove internal node labels
• change hard polytomies to soft - i.e. insert zero-length or very short internal branches to yield a bifurcating tree
NEWICK Format - Branch Lengths

Branch lengths assigned using “:”s and associated with

• terminal branches/OTUs
• the “interior” branch associated with enclosing parentheses

((A:1,B:1):2,C:3)
Visualising Trees

Demonstration/Exercises

• Tree (NEWICK/PHYLIP format) Data
• Editing Trees Using MESQUITE
Which of A, B, C, and D is NOT equivalent to E? (i.e. has different topology)
Where does the root go?
Where does the root go?

A B and C are all from the same unrooted tree

Which rooted tree has the root in the most plausible position?
Example Phylogeny Estimation Workflow
Example Phylogeny Estimation Workflow

1. Pose a substantive question
2. Build a model involving parameters that, if known, could answer the question
3. Collect observations informative about the model parameter(s)
4. Find best estimate(s) of the parameter(s), conditioned on these observations
5. Estimate sampling/random error associated with parameter estimate
6. Answer your question using these parameter estimates
Example Phylogeny Estimation Workflow

1. **Pose a substantive question**

2. **Build a model involving parameters that, if known, could answer the question**

3. **Collect observations informative about the model parameter(s)**

4. **Find best estimate(s) of the parameter(s), conditioned on these observations**

5. **Estimate sampling/random error associated with parameter estimate**

6. **Answer your question using these parameter estimates**
Example Phylogeny Estimation Workflow

1. **Pose a substantive question**

   Can we identify factors promoting disease transmission that could be addressed via public-health measures?
Example Phylogeny Estimation Workflow

1. Pose a substantive question
2. Build a model involving parameters that, if known, could answer the question
3. Collect observations informative about the model parameter(s)
4. Find best estimate(s) of the parameter(s), conditioned on these observations
5. Estimate sampling/random error associated with parameter estimate
6. Answer your question using these parameter estimates
2. Build a model involving parameters that, if known, could answer the question

As we saw before, tree topology parameter can help us address this question.

The full probabilistic model

- A tree topology (to be estimated, $n^n$)
- Branch lengths (to be estimated, $2n-3$)
- A substitution model (to be (partly) estimated, 1, 3, 4, ...208 ...)
- A distribution of site rates (to be estimated, 1, 2, ...)

Thanks to Olivier Gascuel for these slides.
Example Phylogeny Estimation Workflow

1. Pose a substantive question

2. Build a model involving parameters that, if known, could answer the question

3. Collect observations informative about the model parameter(s)

4. Find best estimate(s) of the parameter(s), conditioned on these observations

5. Estimate sampling/random error associated with parameter estimate

6. Answer your question using these parameter estimates
3. Collect observations informative about the model parameter(s)

We aim at explaining the data (alignment) using a probabilistic scenario of the evolution of each of the sites along a phylogeny.

Modeling sequence evolution: standard assumptions

- Evolution is independent among lineages
- Evolution is memory-less (Markovian)
- The sites evolve independently and identically
- Most models are time reversible
- Most models are time homogeneous and stationary

Thanks to Olivier Gascuel for this slide
3. Collect observations informative about the model parameter(s)

Acquire Sequences

Sequencing Machine → Assemble Contigs → Unaligned Sequences

Database Searches
- Keyword
- Sequence Similarity
- Data-Base Cross-References

Pre-Calculated Alignment

Possible MSA tools:
- PRANK
- PROBCONS
- MAFFT
- CLUSTAL

Automatically Align Sequences

No

Yes → Initial MSA

Specific Question → Acquire Sequences

Initial Sequences Aligned?
Example Phylogeny Estimation Workflow

Unusual Sequences

Short/fragmented sequences

With CLUSTALX "Quality"->"Show Low-Scoring Segments" switched on

Unusual pattern of "conservation"
Example Phylogeny Estimation Workflow

3. Collect observations informative about the model parameter(s)

Demonstration: Building an initial alignment of selected rabies phosphoproteins

Demonstration: Visualise your alignment using JalView
Example Phylogeny Estimation Workflow

1. Pose a substantive question
2. Build a model involving parameters that, if known, could answer the question
3. Collect observations informative about the model parameter(s)
4. Find best estimate(s) of the parameter(s), conditioned on these observations
5. Estimate sampling/random error associated with parameter estimate
6. Answer your question using these parameter estimates
4. Find best estimate(s) of the parameter(s), conditioned on these observations

5. Estimate sampling/random error associated with parameter estimate

Demonstration: Choose a substitution model using ProtTest

Demonstration: Estimate a phylogeny using this model, with aLRT
Example Phylogeny Estimation Workflow

1. Pose a substantive question
2. Build a model involving parameters that, if known, could answer the question
3. Collect observations informative about the model parameter(s)
4. Find best estimate(s) of the parameter(s), conditioned on these observations
5. Estimate sampling/random error associated with parameter estimate
6. Answer your question using these parameter estimates
6. Answer your question using these parameter estimates
File Formats

Software only accepts data in particular format(s)
Format sometimes not very precisely specified

Common problems: **taxon labels**

- contain any characters other than A-Z and 0-9 (e.g. white space "", slashes \\ or /", pipes | etc.)
- are not unique (e.g. two sequences labeled HumanA)
- are the wrong length (often a maximum of 10, sometimes exactly 10 is required)
File Formats

Software only accepts data in particular format(s)

Format sometimes not very precisely specified

Common problems: **sequence representation**

- if gaps allowed - wrong character used to represent them (e.g. '.' instead of '-')
- if gaps not allowed - the presence of gaps in the alignment
- sequence is of the wrong kind of molecule (DNA instead of protein etc.)
- sequence contains any characters other than the "alphabet" describing the sequences e.g. 'X' for protein alignments, 'N' for DNA alignments
- all sequences not the same length (check using JalView)