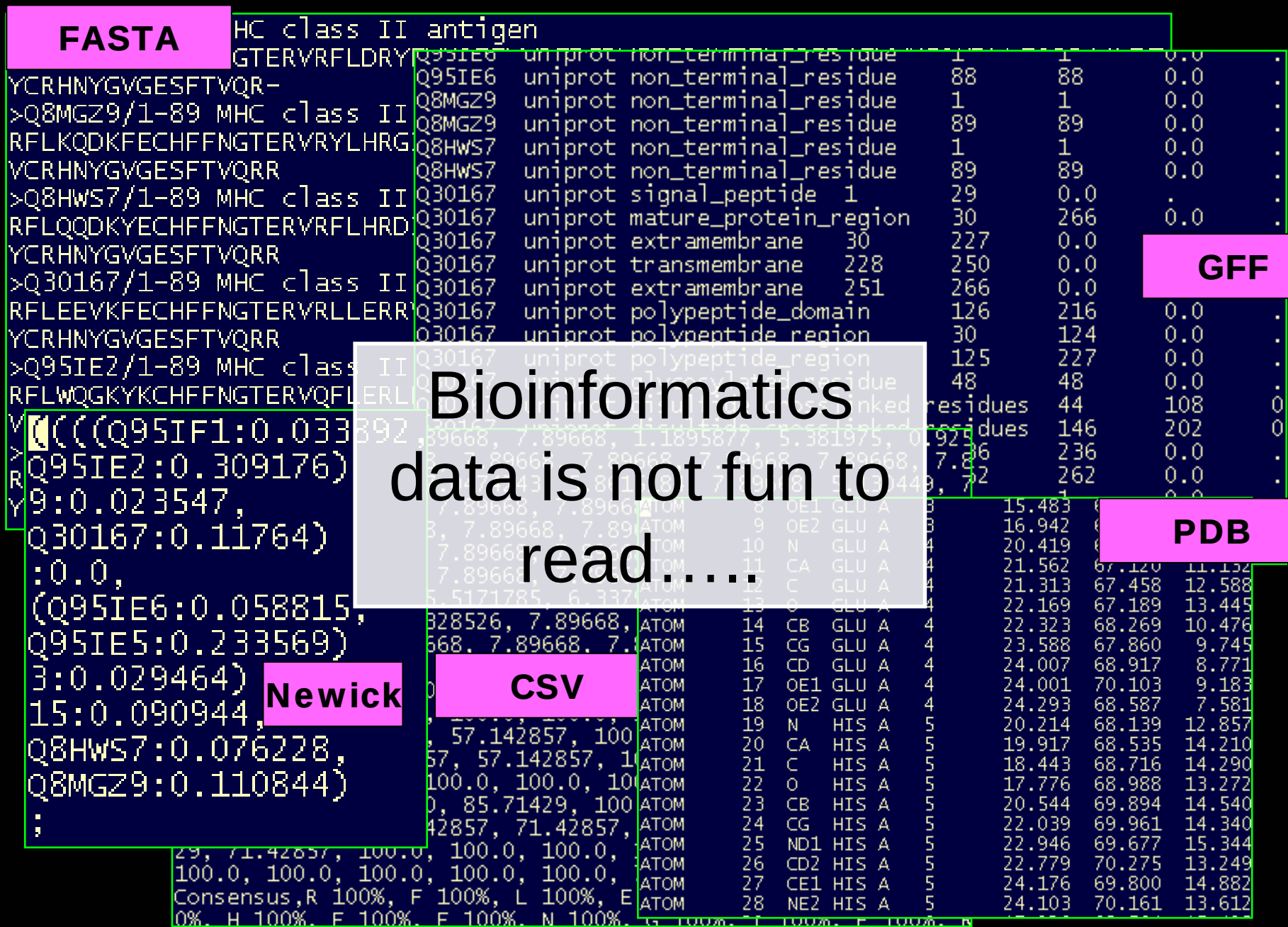
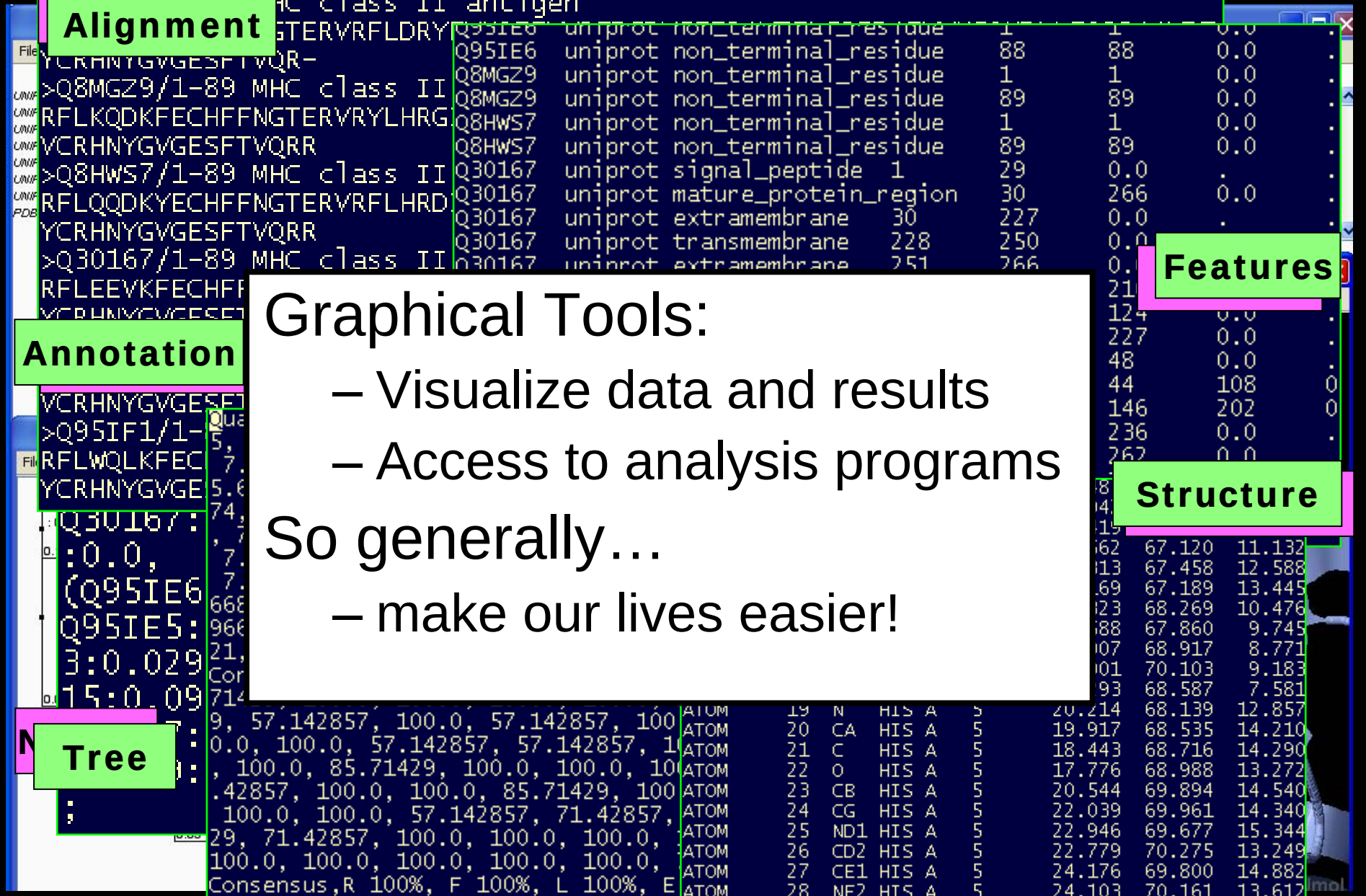
Practical Jalview

A guided tutorial and Jalview clinic

Jim Procter
Barton Group, College of Life Sciences
University of Dundee
j.procter@dundee.ac.uk

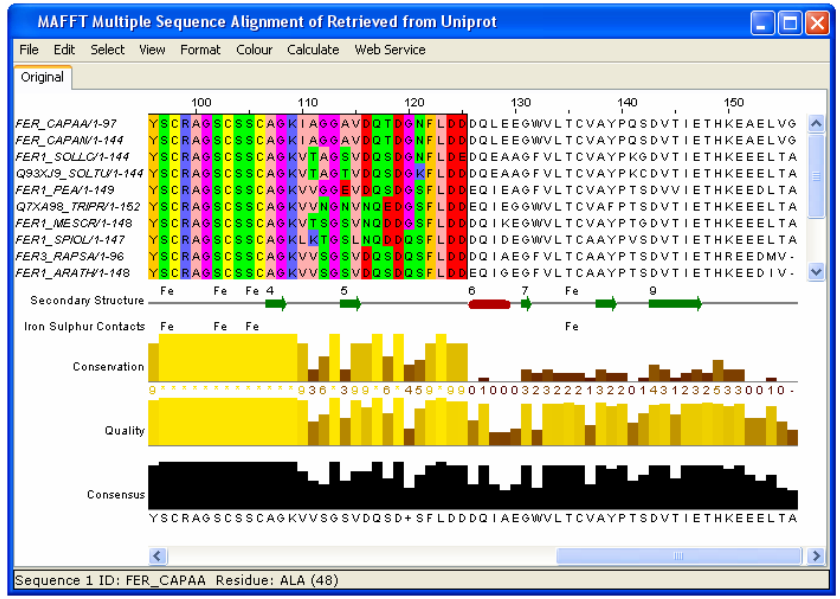

FASTA
HC class II antigen
GTERVRFLDRY
YCRHNYGVGESFTVQR-
>Q8MGZ9/1-89 MHC class II
RFLKQDKFECHFFNGTERVRYLHRG
VCRHNYGVGESFTVQRR
>Q8HWS7/1-89 MHC class II
RFLQQDKYECHFFNGTERVRFLHRD
YCRHNYGVGESFTVQRR
>Q30167/1-89 MHC class II
RFLEEVKFECHFFNGTERVRLLERR
YCRHNYGVGESFTVQRR
>Q95IE2/1-89 MHC class II
RFLWQGKYKCHFFNGTERVQFLERL
(((Q95IF1:0.033392,
Q95IE2:0.309176)
9:0.023547,
Q30167:0.11764)
:0.0,
(Q95IE6:0.058815,
Q95IE5:0.233569)
3:0.029464)
15:0.090944,
Q8HWS7:0.076228,
Q8MGZ9:0.110844)
;
Newick
Q95IE6 uniprot non_terminal_residue 1 1 0.0 .
Q95IE6 uniprot non_terminal_residue 88 88 0.0 .
Q8MGZ9 uniprot non_terminal_residue 1 1 0.0 .
Q8MGZ9 uniprot non_terminal_residue 89 89 0.0 .
Q8HWS7 uniprot non_terminal_residue 1 1 0.0 .
Q8HWS7 uniprot non_terminal_residue 89 89 0.0 .
Q30167 uniprot signal_peptide 1 29 0.0 . .
Q30167 uniprot mature_protein_region 30 266 0.0 .
Q30167 uniprot extramembrane 30 227 0.0
Q30167 uniprot transmembrane 228 250 0.0
Q30167 uniprot extramembrane 251 266 0.0
Q30167 uniprot polypeptide_domain 126 216 0.0 .
Q30167 uniprot polypeptide_region 30 124 0.0 .
Q30167 uniprot polypeptide_region 125 227 0.0 .
48 48 0.0 .
GFF
Bioinformatics data is not fun to read….
ATOM 8 OE1 GLU A 15.483
ATOM 9 OE2 GLU A 16.942
ATOM 10 N GLU A 20.419
ATOM 11 CA GLU A 21.562 67.120 11.152
ATOM 12 C GLU A 21.313 67.458 12.588
ATOM 13 O GLU A 22.169 67.189 13.445
ATOM 14 CB GLU A 4 22.323 68.269 10.476
ATOM 15 CG GLU A 4 23.588 67.860 9.745
ATOM 16 CD GLU A 4 24.007 68.917 8.771
ATOM 17 OE1 GLU A 4 24.001 70.103 9.183
ATOM 18 OE2 GLU A 4 24.293 68.587 7.581
ATOM 19 N HIS A 5 20.214 68.139 12.857
ATOM 20 CA HIS A 5 19.917 68.535 14.210
ATOM 21 C HIS A 5 18.443 68.716 14.290
ATOM 22 O HIS A 5 17.776 68.988 13.272
ATOM 23 CB HIS A 5 20.544 69.894 14.540
ATOM 24 CG HIS A 5 22.039 69.961 14.340
ATOM 25 ND1 HIS A 5 22.946 69.677 15.344
ATOM 26 CD2 HIS A 5 22.779 70.275 13.249
ATOM 27 CE1 HIS A 5 24.176 69.800 14.882
ATOM 28 NE2 HIS A 5 24.103 70.161 13.612
PDB
CSV
Consensus,R 100%, F 100%, L 100%, E


Alignment
Features
Annotation
Structure
Tree
Graphical Tools:
– Visualize data and results
– Access to analysis programs
So generally…
– make our lives easier!

What is Jalview ?

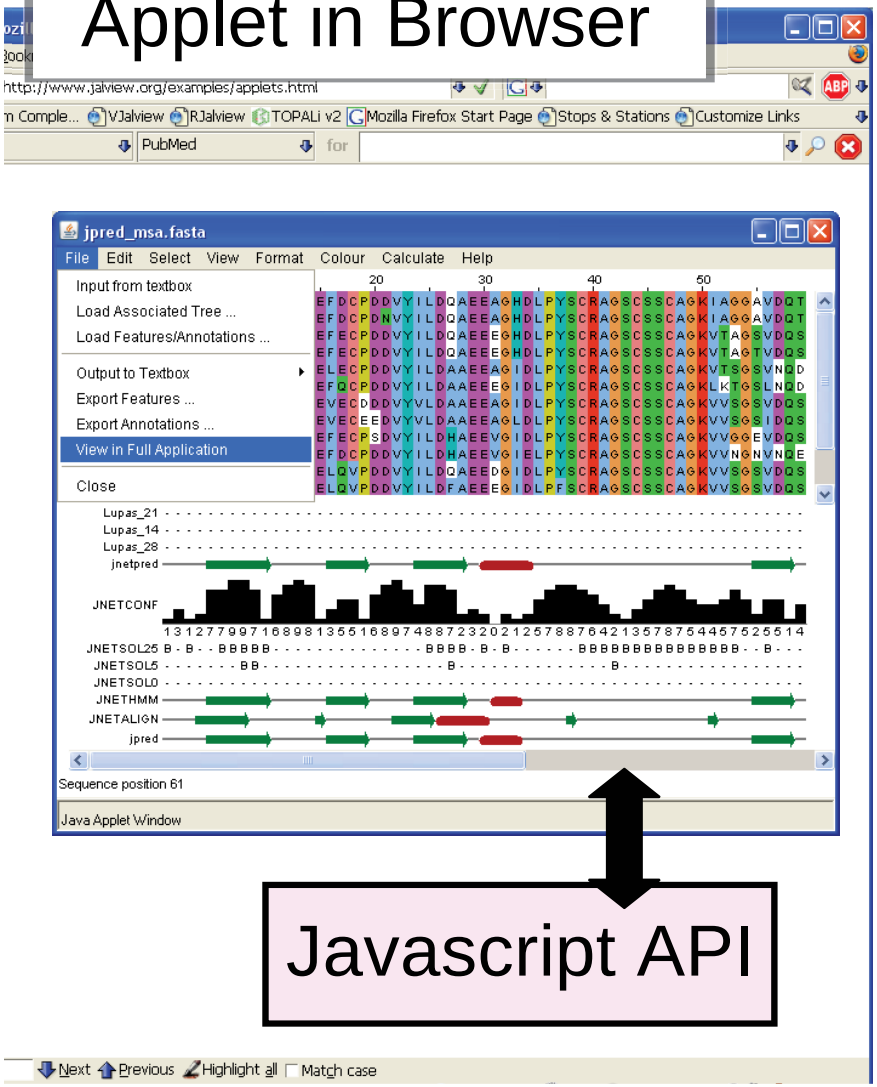
- A java alignment viewer
  - ***j***ava ***al***ignment ***view***er

    it's not just for viewing..
- Java ?
  - Programming language
    - Platform independence
    - Standalone or web based tool



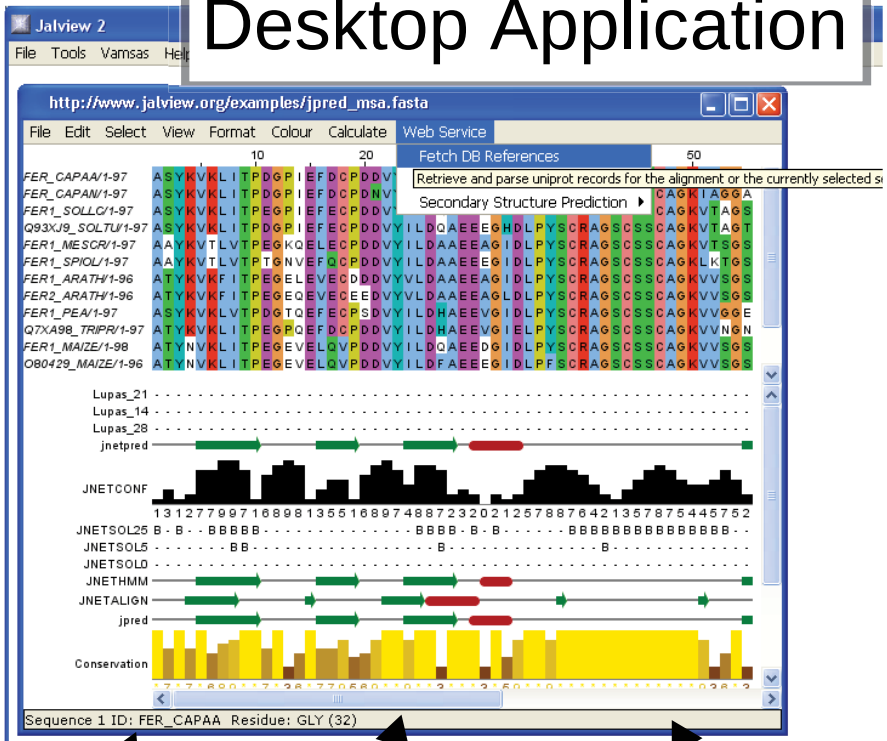
Jalview Flavours

Applet in Browser



Javascript API

Desktop Application



System Clipboard

Local File system

HTTP

http://www.jalview.org

Jalview

iple alignment editor

Get the latest stable release

Jalview Manual and Tutorial

Where to post bug reports and get help

- Home
- Overview
- Download
- Applet Versi
- Screenshots
- FAQ
- Documentation
- Release history
- Source Code
- Development Ver
- Links
- News Mailing List
- Discussion Mailing List

Please send problems and bug reports to the discussion list.

NEWS: 26th September 2010 - Jalview 2.6 release available on download page

NEWS: 26th September 2010 - New Jalview training course schedule

r written in Java. It is used widely in a variety of web
and the Pfam protein domain database) but is available
as a general purpose alignment editor.

Jalview Development is supported from 2009 to 2014 by the BBSRC, and coordinated by Geoff Barton at the University of Dundee. Version 2 arose from the "VAMSAS" Project (BBSRC eScience 2004-2007), with consultancy (blessing :-) from Michele Clamp; the originator of Jalview.

houses; Jim Procter; David Martin; Geoff Barton
; James Cuff; Stephen Searle; Geoff Barton

Thanks to Andreas Prlic for code and suggestions for DAS feature capabilities and Benjamin Schuster-Böckler for his Stockholm parsing code, both from the Wellcome Trust Sanger Institute, Cambridge.

If you use Jalview in your work, please quote this publication:

Waterhouse, A.M., Procter, J.B., Martin, D.M.A, Clamp, M. and Barton, G. J. (2009) "Jalview Version 2 - a multiple sequence alignment editor and analysis workbench" *Bioinformatics* **25** (9) 1189-1191 doi: 10.1093/bioinformatics/btp033

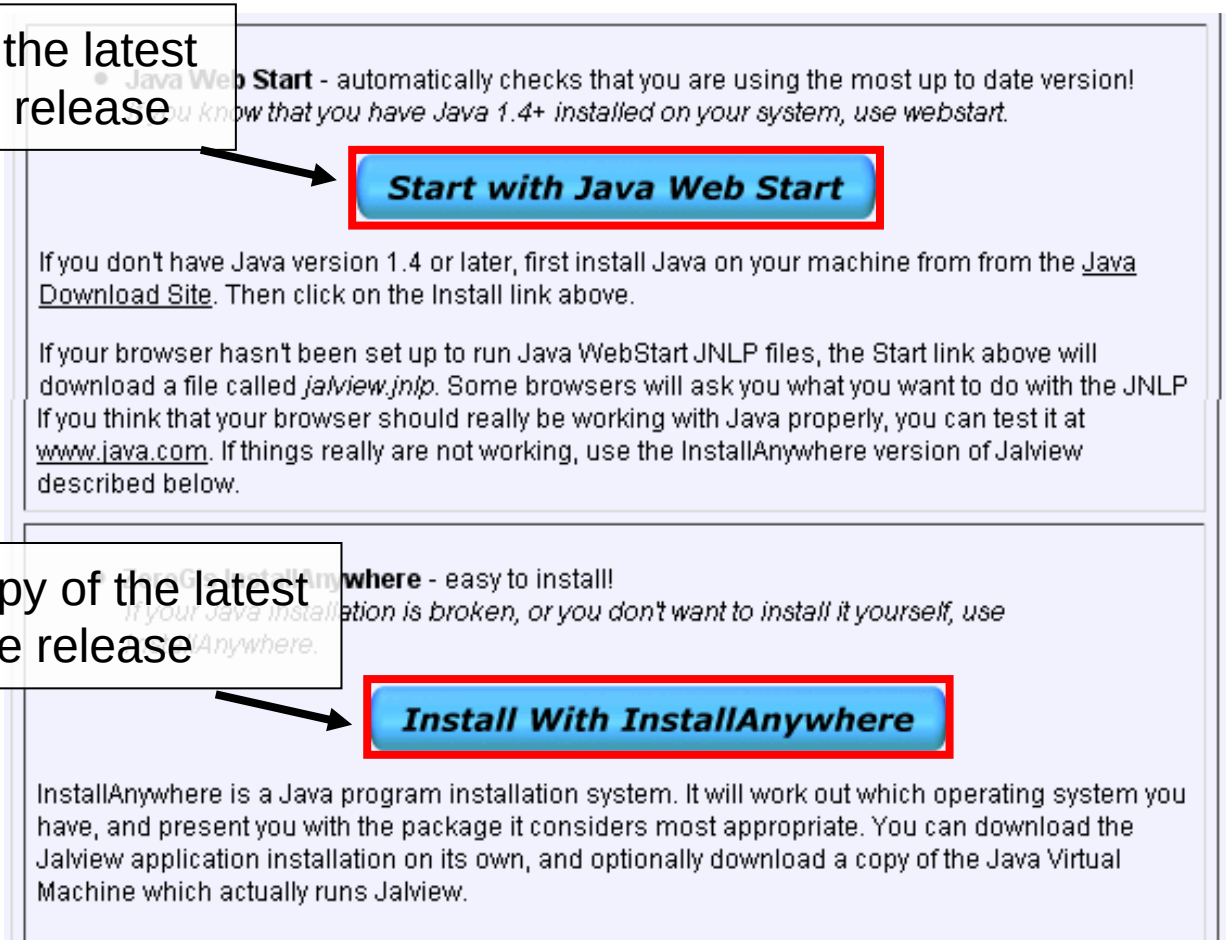
NEWS: 26th September 2010 - Jalview 2.6 release available on download page

NEWS: 26th September 2010 - New Jalview training course schedule

Starting The Jalview Desktop

http://www.jalview.org/download.html

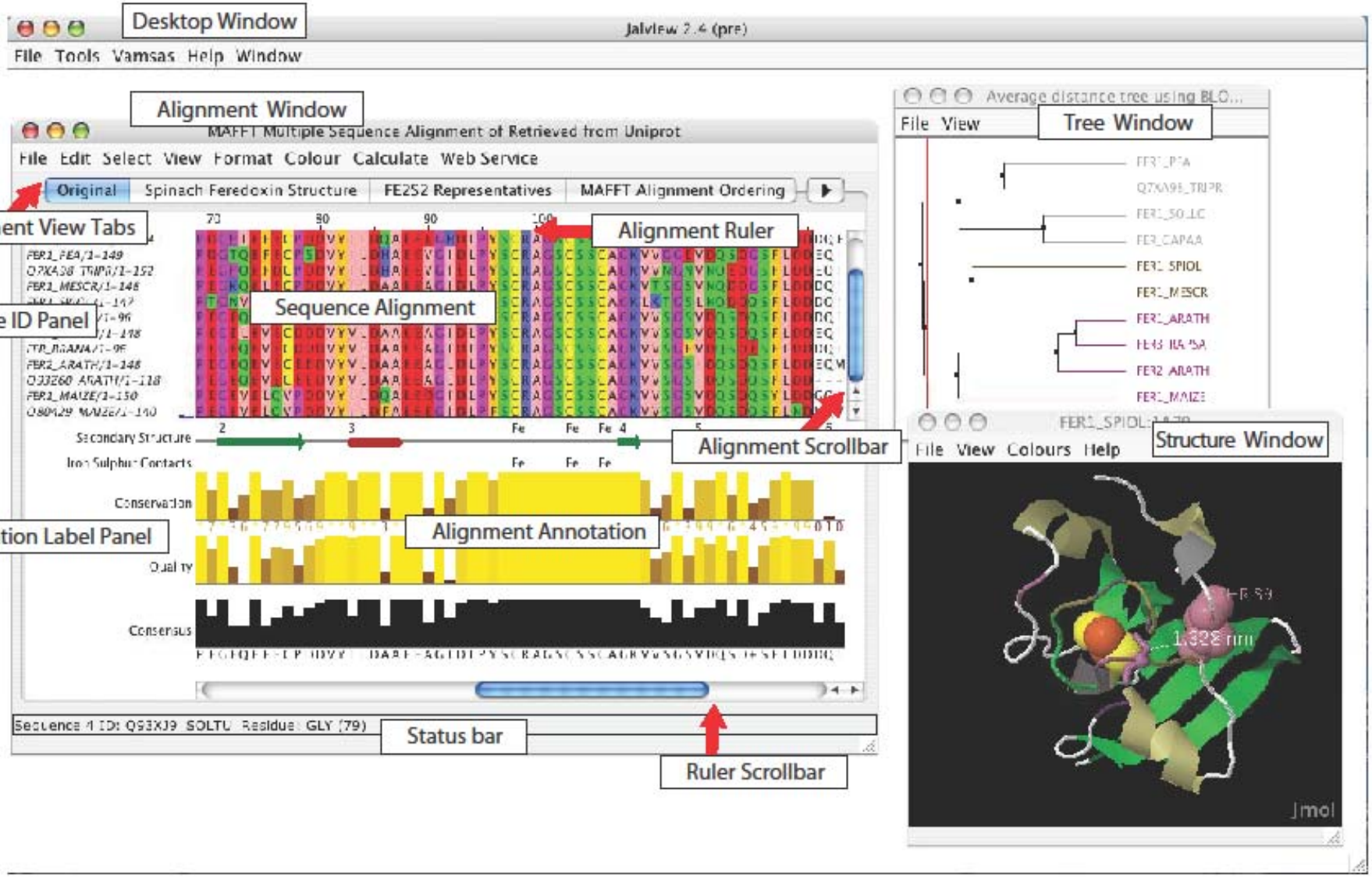
Launch the latest stable release



Install a copy of the latest stable release

Ex 1 – starting Jalview

Anatomy of Jalview: Figure 1.6



Ex 1 – starting Jalview

- Tasks
  - Modify user preferences
- Questions
  - Where to find help ?
  - How to report a bug ?

Ex 2 - Navigation

- Tasks
  - Open the overview window for a view
  - Jump to a specific row and column with keyboard mode
- Questions
  - How do you locate a sequence or sequence position if you don’t know its row/column ?
  - How do you find a sequence motif ?

Ex 3 Getting data into Jalview

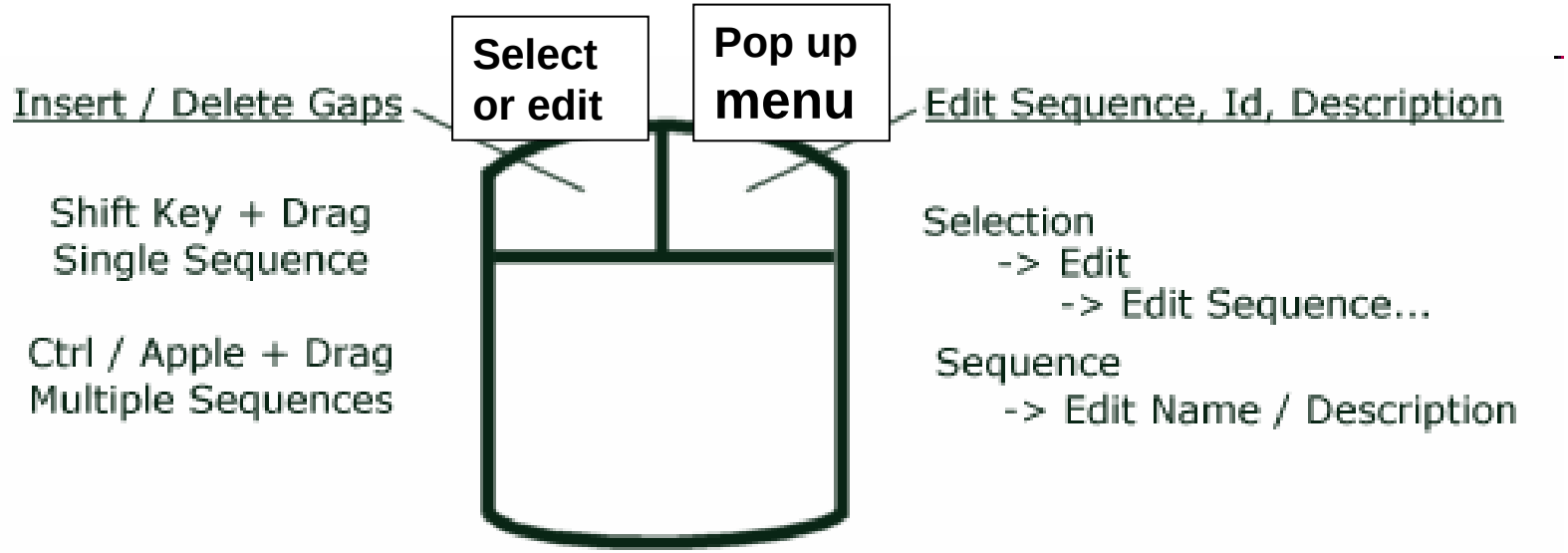
- Tasks
  - Importing an alignment via a url, local file, or cut’n’paste
  - Getting an alignment from Pfam
- Questions
  - What happens when you drag a file onto an existing alignment ?
  - What’s different about the alignment retrieved from Pfam ?
  - What if you want to load a *really* big alignment ?

Ex 4. Saving alignments

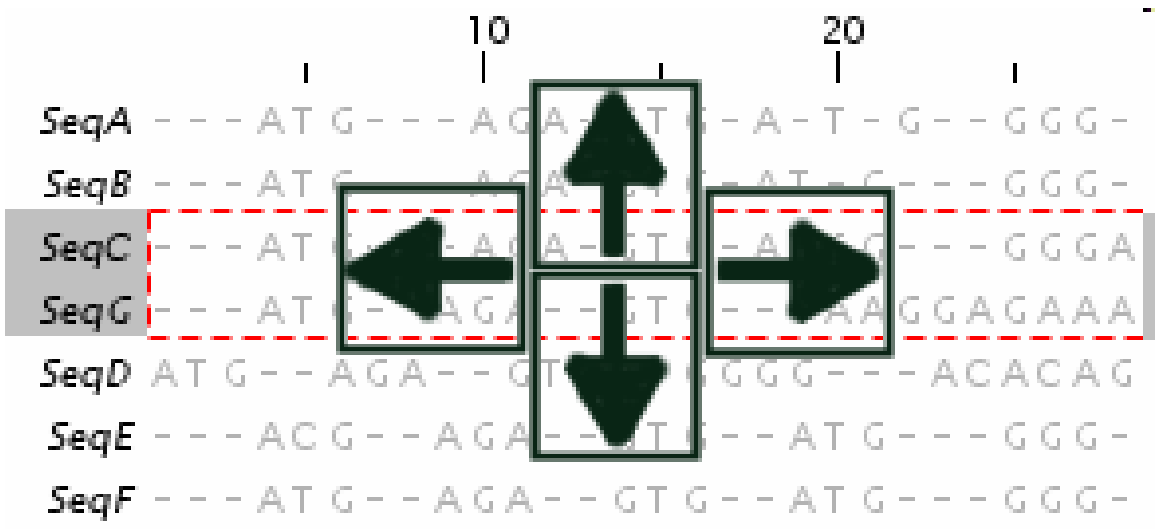
- Tasks
  - Save alignments in different formats
- Questions
  - What's the biggest difference between a BLC file and a pileup file ?
  - Why are Jalview projects useful ?

Ex 5,6,7,8 and 9 selecting, editing, hiding and showing

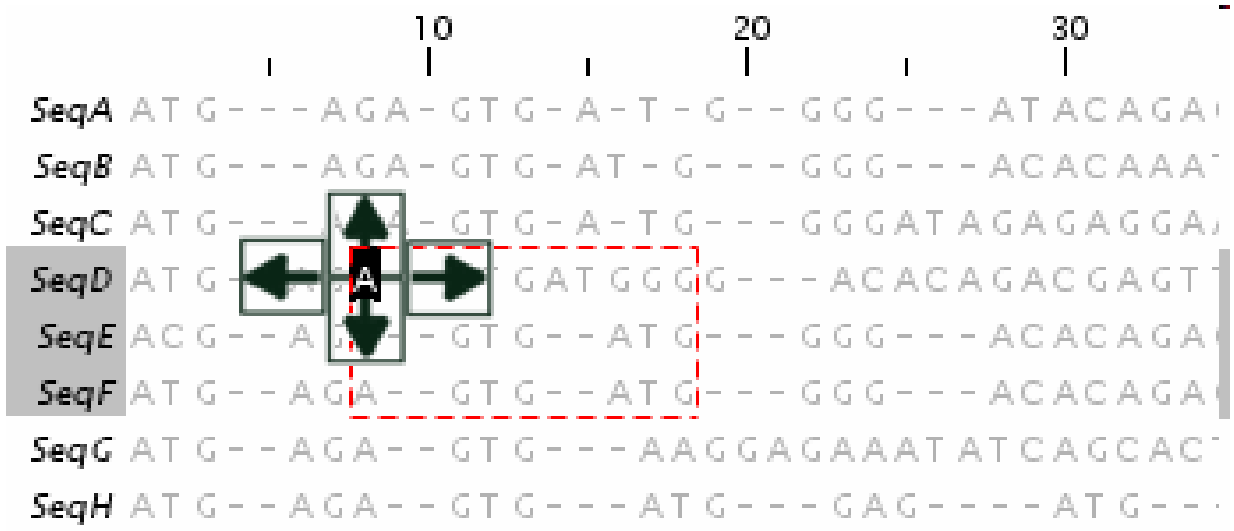
- Tasks
  - Get used to the mouse and keyboard based selection and alignment editing controls
  - Learn how to work on specific parts of an alignment
- Questions
  - Why is it useful to create representative sequences ?
  - How do you insert a gap in the middle of a sequence without affecting the rest of its alignment ?



Selected sequences can be moved up and down or slid from left to right



F2 enables/disables keyboard mode



Cursor Keys - Move Cursor
Alt + Cursor Keys - Move Sequence
[X] Space - Insert [X] gap(s)
[X] Delete / Backspace - Delete [X] gap(s)
8 C - Move to Column 8
4 S - Move to Sequence 4
8,4<return> - Move to column 8, sequence 4
6 P - Move to Position 6
Q - Define the top left corner of selection area
M - Define the bottom right corner of selection area

Ex 5,6,7,8 and 9 selecting, editing, hiding and showing

- Tasks
  - Get used to the mouse and keyboard based selection and alignment editing controls
  - Learn how to work on specific parts of an alignment
- Questions
  - Why is it useful to create representative sequences ?
  - How do you insert a gap in the middle of a sequence without affecting the rest of its alignment ?

Ex 10 & 11 : Colouring

- Tasks
  - Learn how to colour all, or part of the alignment by
    - Amino acid property
    - Annotation
- Questions
  - Why is colouring the alignment useful ?
  - How can you highlight the acidic residues ?

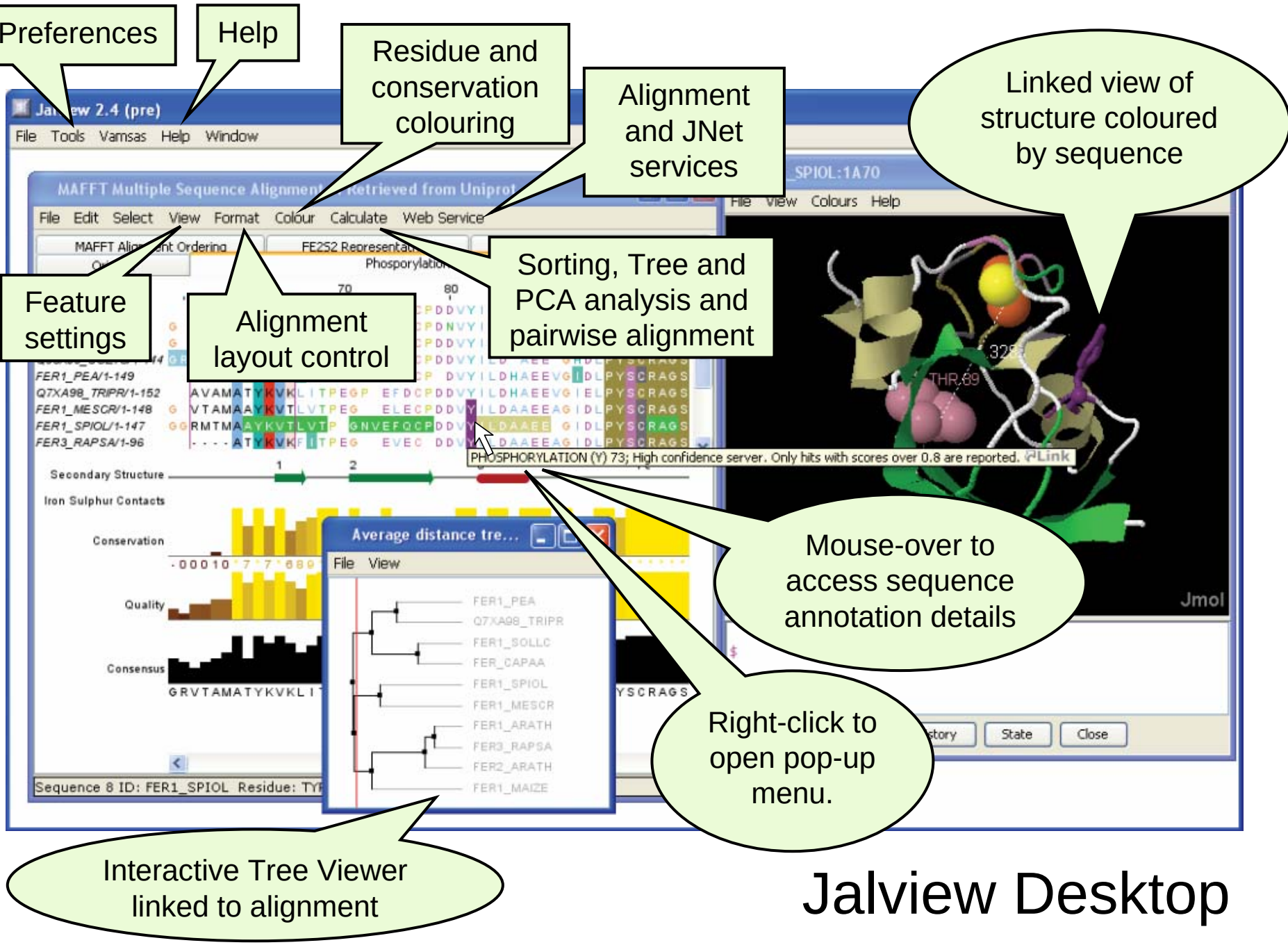
Ex 12,13 – alignment layout and export

- Tasks
  - Adjust the alignment formatting options
    - Wrap
    - Sequence id margin
  - Export the alignment as a figure
    - HTML, EPS and PNG
- Questions
  - How do you control the number of columns shown in wrapped mode ?
  - How can you easily experiment with different alignment figure layouts ?
  - Do you know how to edit EPS files ?

Part 2
Alignment, annotation and Analysis

Topics

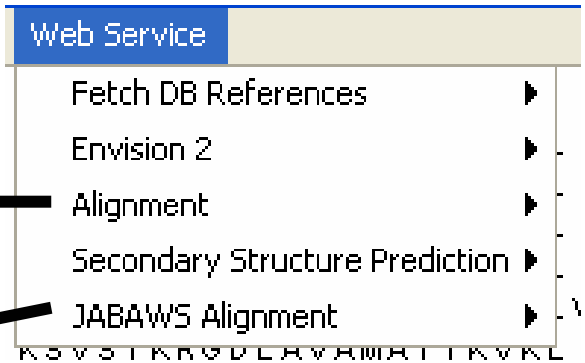
- creating your own alignments
- protein secondary structure prediction
  - Section 2.3.4
- Alignment annotation
  - Sect. 2.4.4
- alignment analysis with phylogenetic trees and principal component analysis
  - Section 2.2
- working with sequence annotation
  - Section 2.4.1-3 and Section 2.5
- DNA and Protein sequences and Jalview
  - Section 2.6
- working with PDB structures
  - Section 2.1


Preferences
Help
Residue and conservation colouring
Alignment and JNet services
Linked view of structure coloured by sequence
Feature settings
Alignment layout control
Sorting, Tree and PCA analysis and pairwise alignment
Mouse-over to access sequence annotation details
Right-click to open pop-up menu.
Interactive Tree Viewer linked to alignment
Jalview Desktop
File Tools Vamsas Help Window
File Edit Select View Format Colour Calculate Web Service
File View Colours Help
Secondary Structure
Iron Sulphur Contacts
Conservation
Quality
Consensus
Average distance tre...
File View
Sequence 8 ID: FER1_SPIOL Residue: TYR
PHOSPHORYLATION (Y) 73; High confidence server. Only hits with scores over 0.8 are reported.
Jmol
State
Close

Jalview's Alignment Methods

- Needleman and Wunsch Pairwise Alignment
  - Global alignment of pairs of sequences
    - Mostly used internally (described in section 2.2.7)
- Multiple Sequence Alignment Services

– ClustalW – Muscle – MAFFT	Available as Jalview and Jaba Service.
– ProbCons – T-COFFEE	Only provided by Jaba.



Jalview 2.5 alignment exercise 20 (sect. 2.3.3)

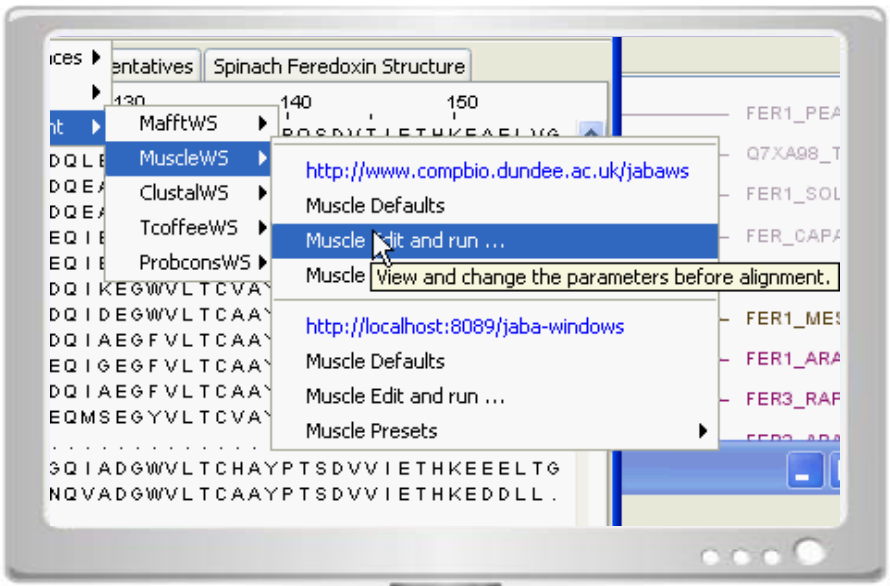
- Tasks
  - Align sequences using different methods
    - Use the ‘alignment’ submenu
  - Explore how hidden regions affect alignment jobs.
- Questions
  - Why does jalview run several jobs if the input includes hidden regions ?
    - Is this useful ?
  - What does ‘re-alignment’ mean ?

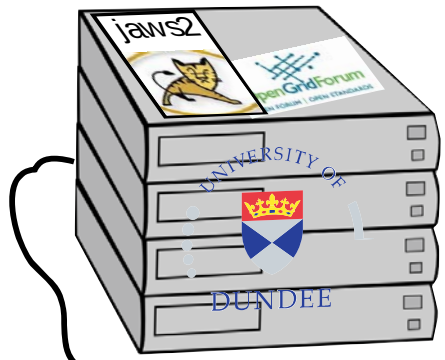
New in Jalview 2.6

Java Bioinformatics Analysis (Jaba) Webservices

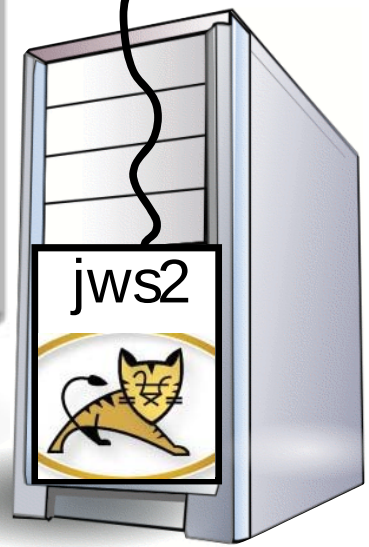
JABAWS replaces original Jalview 2 services:

- Extensible framework for wrapping command line programs
- Can be installed on user's own machine/cluster





HTTP



Peter Troshin

Alignment Job Parameter Settings

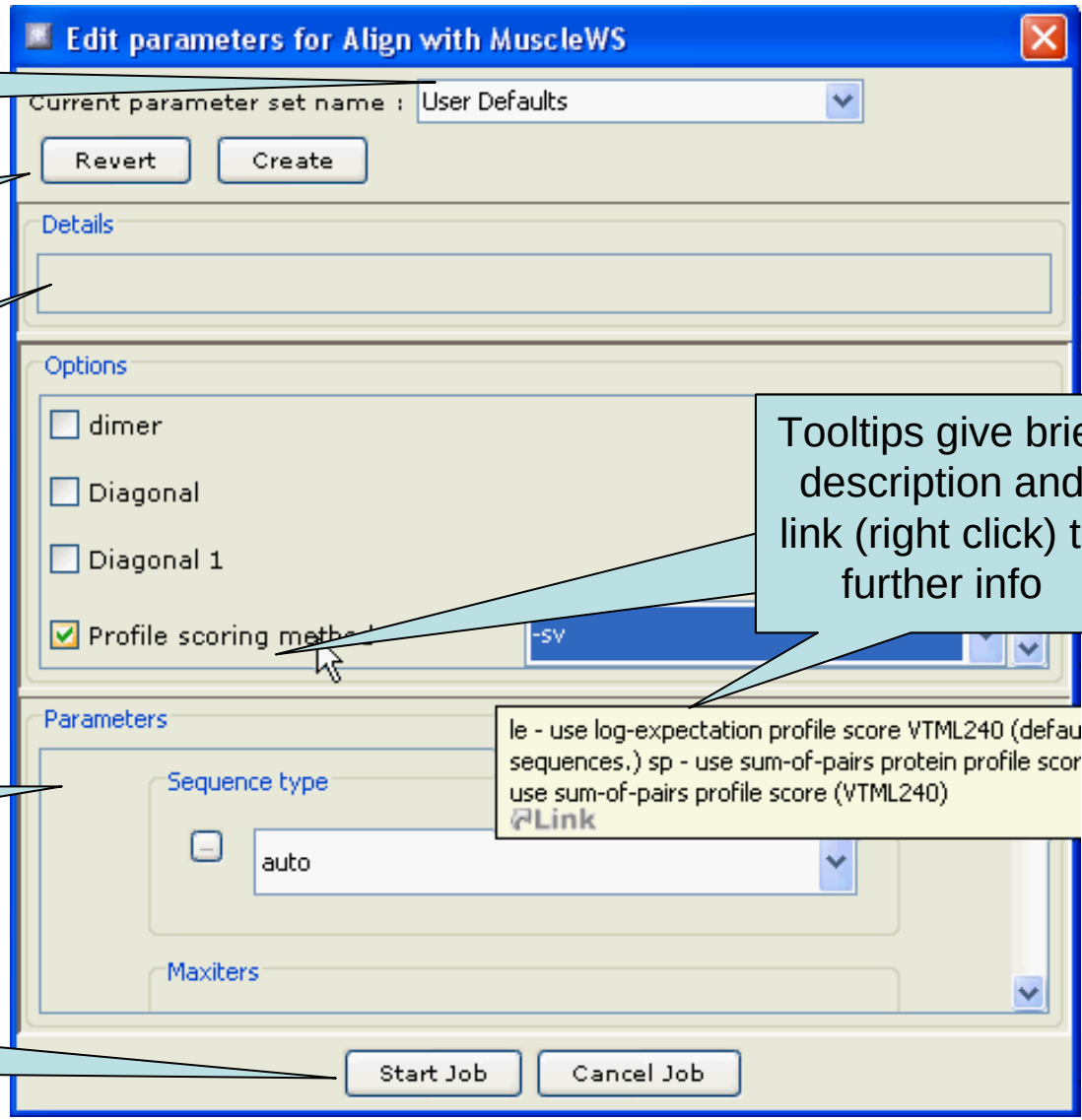
Browse or edit to change name of set

Buttons appear to create, update, rename or delete user settings.

text box to add notes for the parameter set

Parameters contains more complex settings

Start job with current settings or cancel.



Tooltips give brief description and link (right click) to further info

Jaba Alignment Exercise

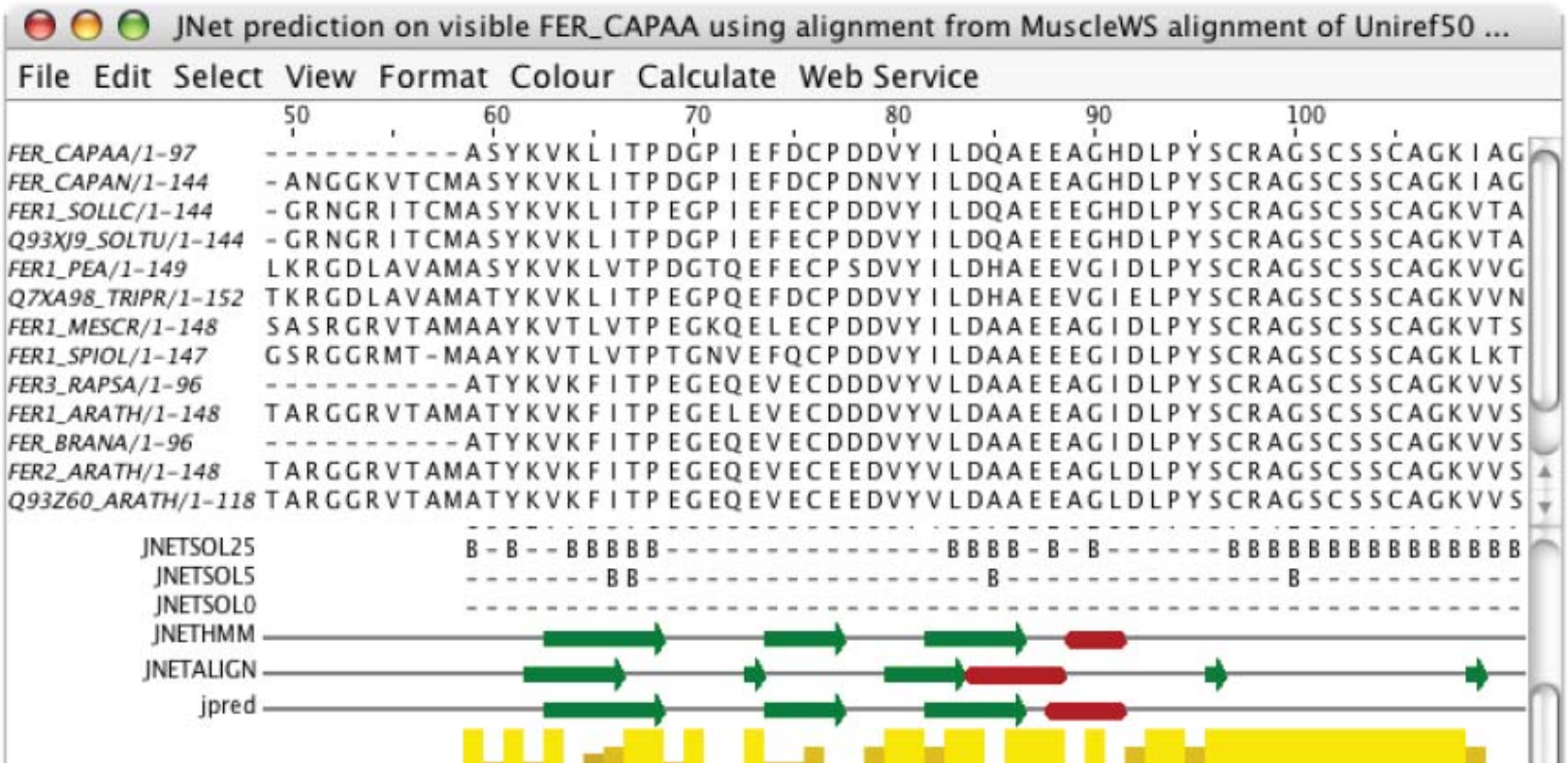
- Task
  - Run the alignment from step ***b*** of ex. 20 using the JABA clustalW service
    1. Run with default settings
    2. Use the ‘Edit parameters’ dialog to run an alignment with the following:
       - Gap opening (internal and end gaps) = 3
       - Gap Extension = 0.05
  - Compare the two alignments. You may want to save them for later, too.

- Questions
  - What effect has modifying the gap penalties had on the feredoxin alignment ?

Protein Secondary Structure Prediction Sect 2.3.4

- Jalview interfaces with the Jpred protein secondary structure predictor
- Prediction is based on
  - Neural net which can recognise helical, coil or beta strand using amino acid patterns
  - Amino acid profile for a sequence
    - Multiple sequence alignment
    - Profile from sequence database search
      - **P**osition **S**pecific **S**ubstitution **M**atrix

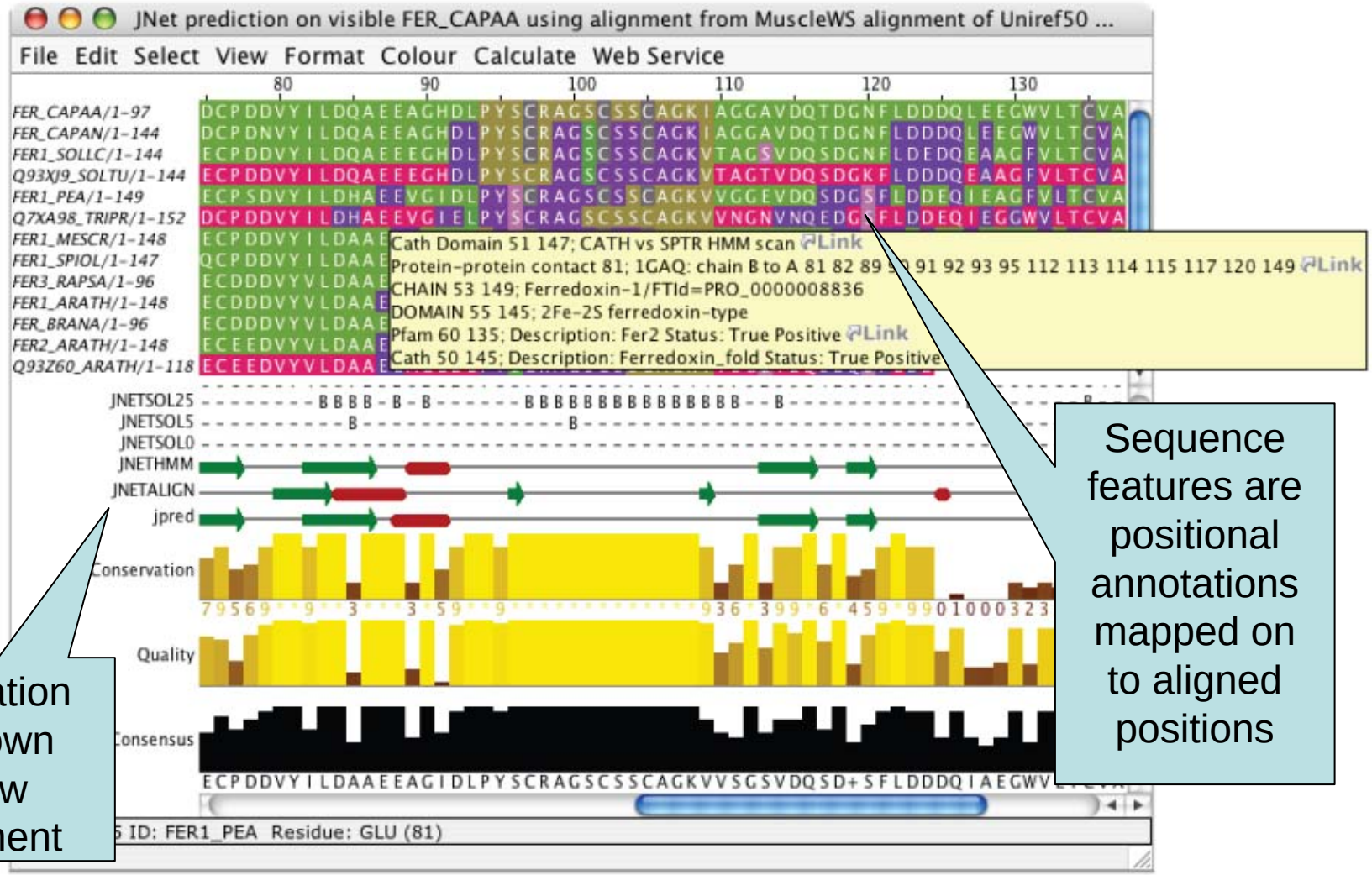
Protein Secondary Structure Prediction Sect 2.3.4



Exercise 21

- Tasks
  - Perform a variety of Jnet predictions
    - Note the effect of hidden regions
    - Learn about sequence associated annotation
  - Save your results for the next exercise
- Questions
  - What other data does Jnet provide ?
  - Which is better – a PSI blast prediction or an MSA based prediction ?

Alignment Annotation and sequence features.



Annotation Is shown below alignment

Sequence features are positional annotations mapped on to aligned positions

Creating, editing and using annotation. Exercise 23

- Tasks
  - Manually annotate some columns using the interactive editing functions
  - Learn about jalview annotation files
    - How to change the appearance of quantitative data.
    - Understand how to create sequence associated annotation
- Questions
  - What other things can be defined in jalview annotation files ?

Alignment Analysis

Using jalview to analyse the relationships between aligned sequences.

Comparative Sequence Analysis

1. Identify homologs of interest

   Query sequence databases, identify similar sequences with BLAST, etc.

2. Create a reliable alignment

   a. Apply automated alignment method

   b. Verify alignment using known information

      - Functional or biological characterisation

   c. Realign or manually curate if required.

3. Apply clustering methods to:

   - Investigate sequence/function variation*
   - Infer evolutionary history

Jalview's original role

PCA and Phylogeny Exercises Section 2.2 - Exercise 15 and 16

- Tasks
  - Calculate Principal component analyses (PCAs) and trees on the feredoxin alignment
  - Explore the use of the interactive tree viewer
    - Use it to select subgroups on the alignment.
- Questions
  - What is the role of BLOSUM62 or Percentage identity in the tree building process ?

Phylogenetic analysis and Jalview

- Built in tree methods
  - UPGMA
    - Fast, simple, but not reliable for phylogenetic inferrence
  - Neighbour joining
    - Slower than UPGMA
    - Useful for a first approximation
      - NJ does not work well for very divergent sequence sets
        - » Need to add in close relatives to get an idea of topology
- Import trees from elsewhere
  - Load a Newick format tree file onto an alignment from another program

Issues to consider for accurate phylogenetic inferrence

- Evolutionary Model selection
  - Different distance measures
    - %age identity, BLOSUM or other substitution matrix
    - Evolutionary rate models
- Phylogenetic inferrence method
- Reliability (bootstrap, Max. likelihood)
- Appropriate visualisation

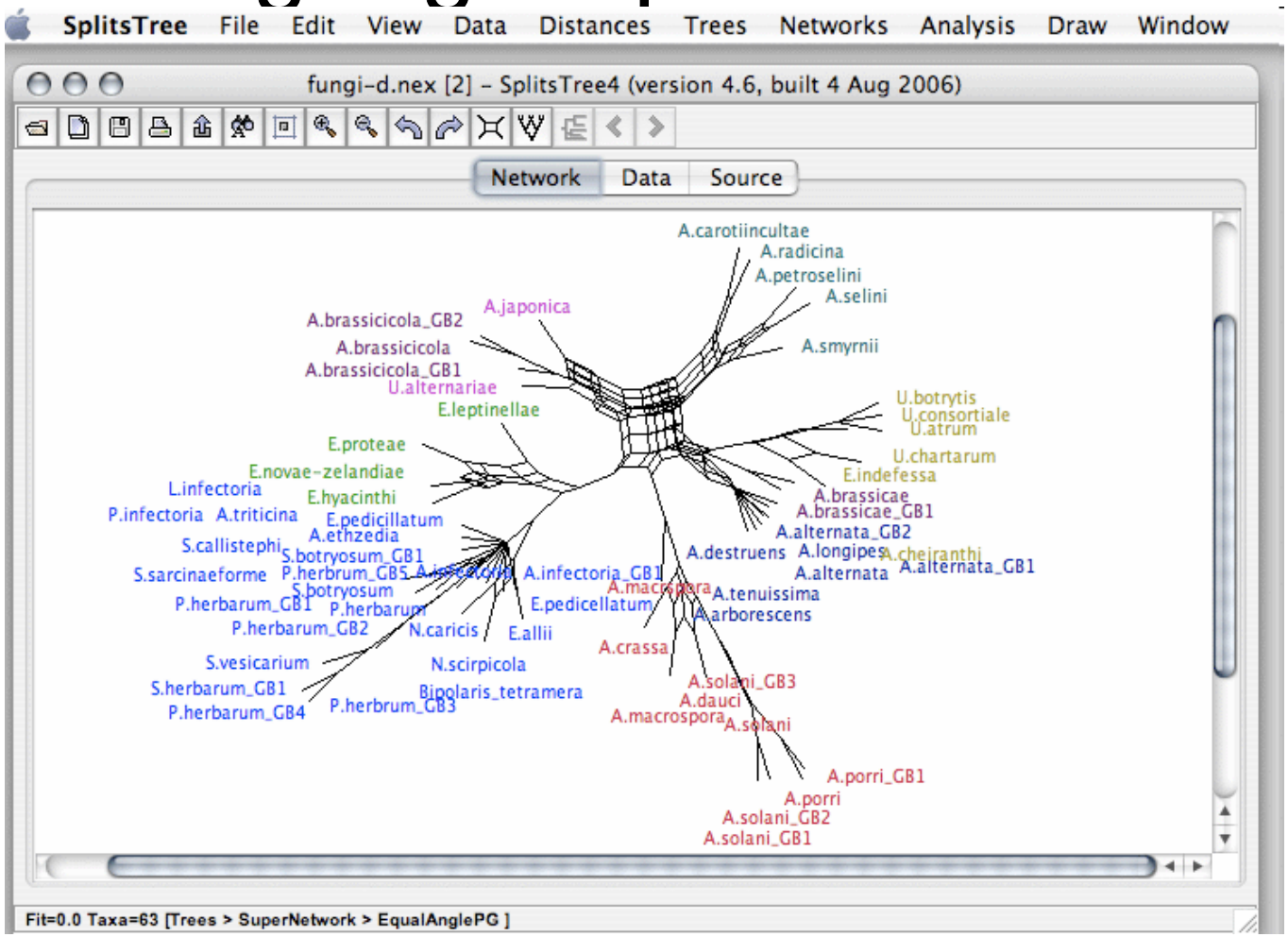
Classes of Phylogenetic Methods

- Parsimony
  - Infer traits inherited/lost at each evolutionary event in the ancestry of related organisms
- Distance based
  - Estimate evolutionary distance between two species and their ***most recent common ancestor***
- Maximum Parsimony Approaches
  - search all tree topologies to find smallest tree + 'trait' labelling that explains observed organism traits
- Bayesian & Maximum Likelihood Approaches
  - Determine most likely tree + evolutionary distance

Bootstrapping

- Way to measure reliability of tree
  - Only usually needed for trees calculated with simple heuristics
    - Bootstrapping is implicitly performed in Maximum likelihood and bayesian approaches.
- Approach
  - Randomly sample the data used for tree calculation
    - E.g. take random subsets of alignment
  - Construct a new tree and compare with original
  - Annotate branches in original tree with proportion they appeared in all bootstrap trees.
- Interpretation
  - More reliable topologies should have higher ‘Support’
  - Test is confounded when rate of evolution is heterogeneous
    - Usual 95% reliability assumption no longer holds

SplitsTree: Bootstrap visualisation google:splitstree



Daniel Huson and David Bryant, 2001.

http://www-ab.informatik.uni-tuebingen.de/software/splitstree4

Common phylogenetic programs

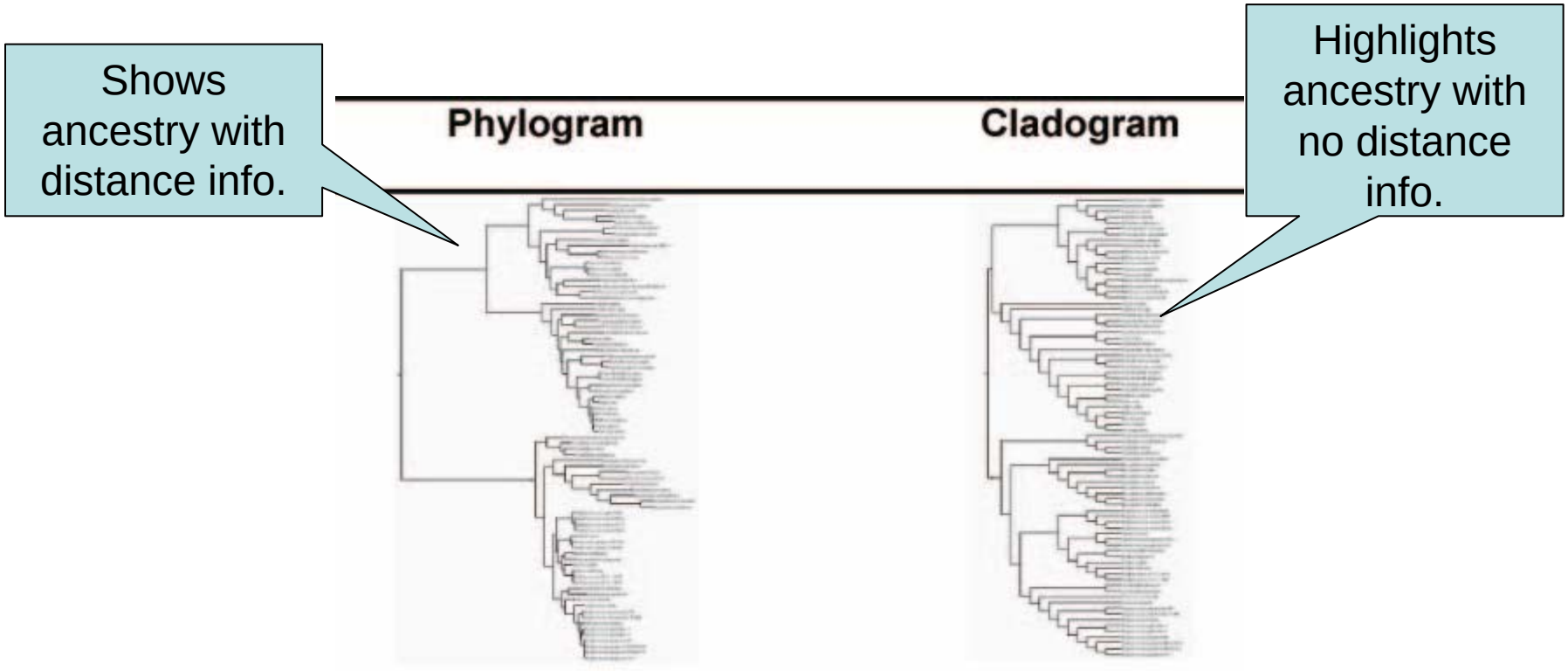
- Simple distance based methods
  - Neighbour joining + UPGMA
    - Jalview and many others.
- Parsimony methods
  - PAUP, PHYLIP's MIX program
- Maximum Likelihood methods
  - MrBayes
- GUI tools
  - SplitsTree 4: google:splitstree
  - MEGA: www.megasoftware.net
  - TOPALi: www.topali.org

Tree visualisations

- Formal terminology
  - Trees
    - Most tree plots are ***dendrograms***
  - Trees showing taxonomic lineage
    - ***Cladogram***
  - Trees where branch length equals:
    - number of mutations (Percent ID, BLOSUM, etc)
      - ***Phylogram***
    - Time
      - ***Chronogram***

Types of tree visualization

- Traditional rectangular layout

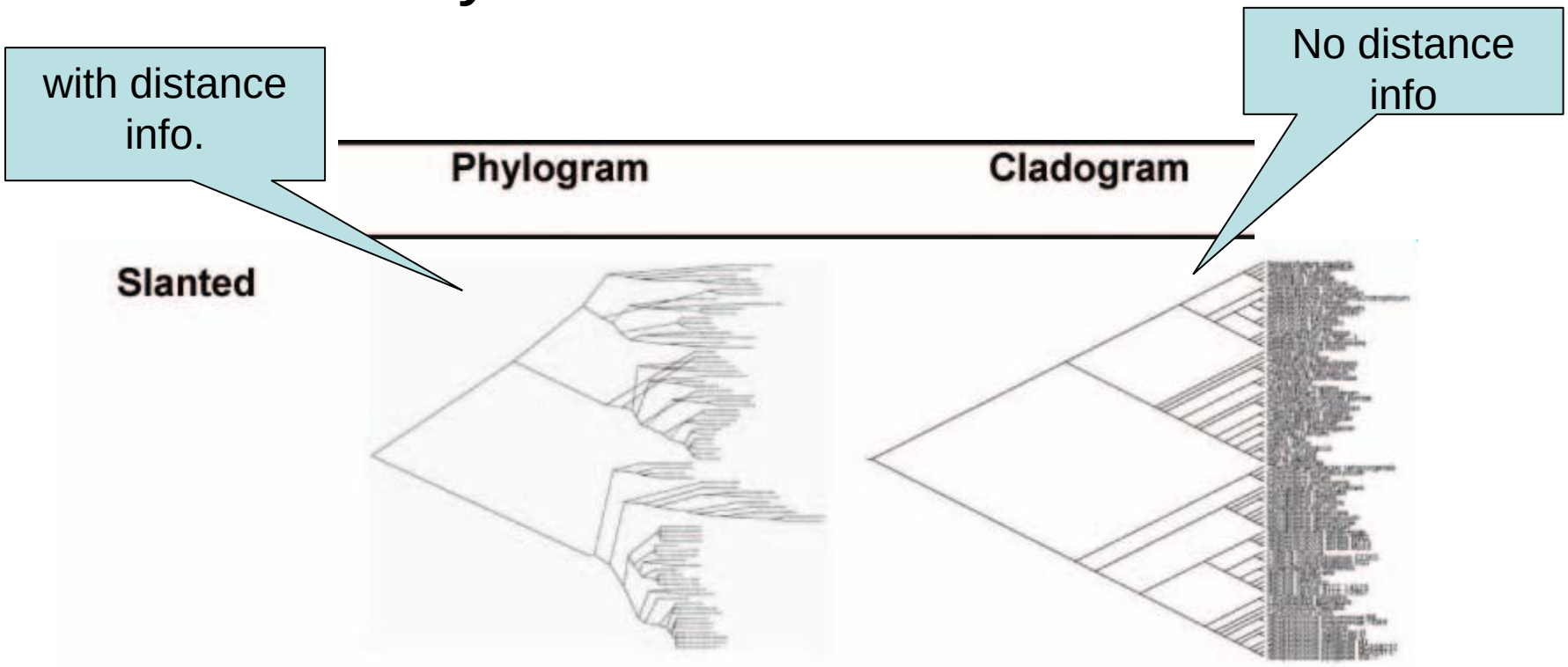


Rectangular plots are more difficult to navigate with very large sequence sets.

Procter, *et. al.* 2010, *Nature Methods*.

Types of tree visualization

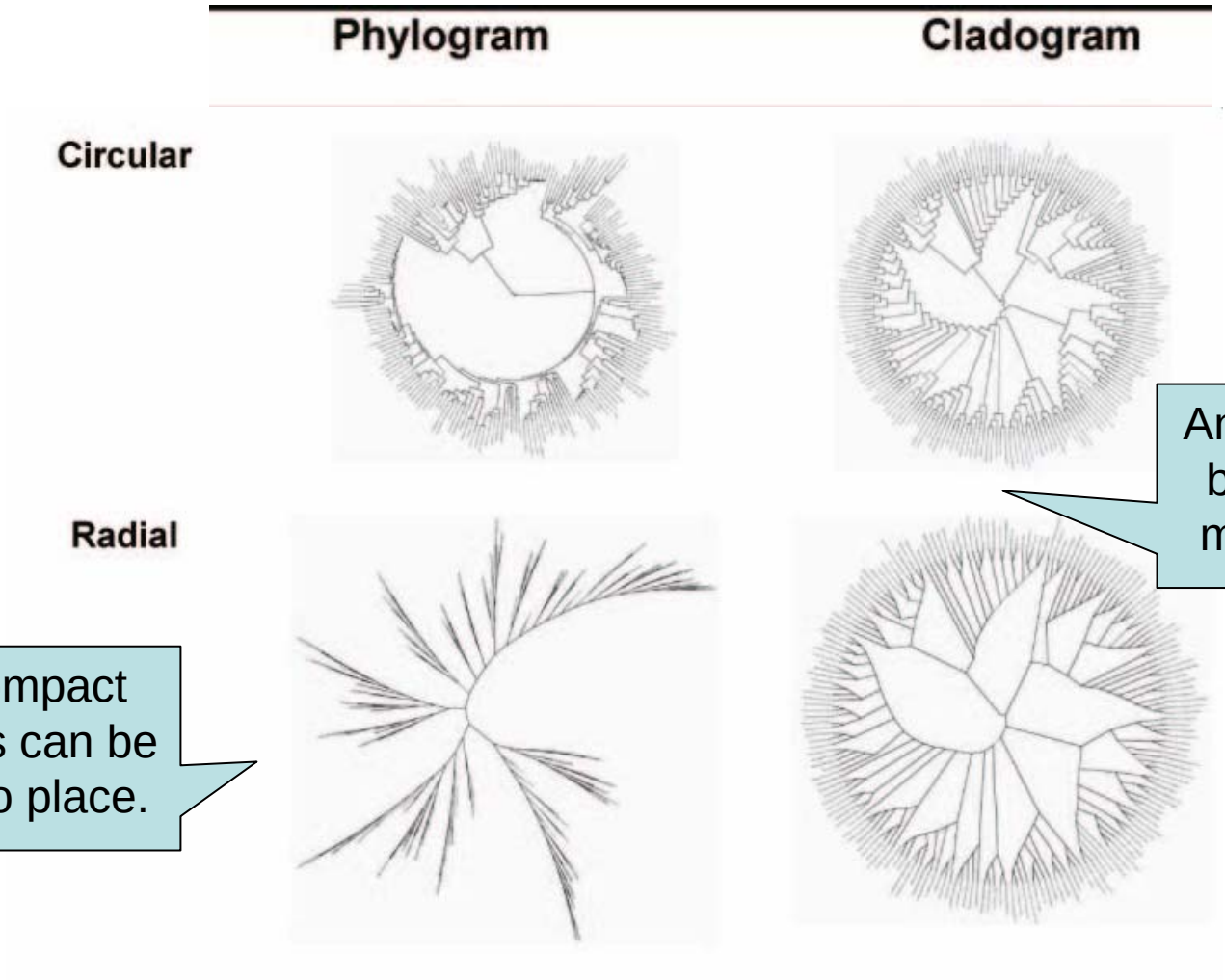
- Slanted layout



Slanted plots make it easier to compare the number of ancestors present in different branches.

Procter, *et. al.* 2010, *Nature Methods*.

Types of tree visualization



Ancestry clear in both, choice is matter of taste.

most compact but labels can be difficult to place.

Large trees are best portrayed as circular and radial projections.

Procter, *et. al.* 2010, *Nature Methods*.

To root or not to root.

- Rectangular, Slanted and Circular plots ***imply*** ancestry
  - ⇒Oldest organism should appear at root of the tree
  - – Usually called the ***Outgroup***
- Options if you don't know the root
  - ⇒**Mid-point rooting** provides a 'balanced tree'
    - Root is placed midway between most distal taxa
      - – Jalview does this
  - ⇒Show a **radial phylogram** if absolute root is not known

Back to Jalview…

Tree based conservation analysis Sect. 2.2.3 Exercise 17

- “Poor man’s” character inference analysis
  - Compare conservation patterns within and between branches of a tree
- Task
    - Use interactive tree viewer to subdivide alignment and identify difference in conservation pattern
- Questions
  - How can you tell which differences are important ?
  - How can you navigate the sub-groups of a large alignment ?

Sub group annotation Exercise 19

- Task
  - Use the group consensus sequence logos to more easily compare tree subgroups
  - Use ‘Make groups for selection’ to subdivide groups by specific mutation
- Questions
  - How can you work out which group is associated with which annotation row ?

Getting and working with sequence features and annotation

- Sequence Databases
- Sequence feature sources
  - DAS Sequence feature retrieval
  - GFF and Jalview annotation files
- Visualizing features
  - Highlighting annotated regions
  - Shading and reordering based on scores and labels

Jalview and Sequence Databases Sec 2.5.1 Ex. 24

- Can retrieve new sequences or match against existing records using IDs
- Task
  - Recover the Uniprot annotation for the ferredoxin sequences using their IDs
  - Verify retrieval by examining annotation
- Question
  - What happens if only a subsequence is present in the alignment ?
  - Does database annotation get shared between alignments ?

Sequence Features
Section 2.4.1-3 & Ex 22

  - Annotate the whole or part of a sequence
  - Database refs are special case.
- Tasks
  - Visualise, create, modify, import and export features.
- Questions
  - What are the different types of file formats available for import and export
  - Are there any mechanisms for discovering sequence annotation ?

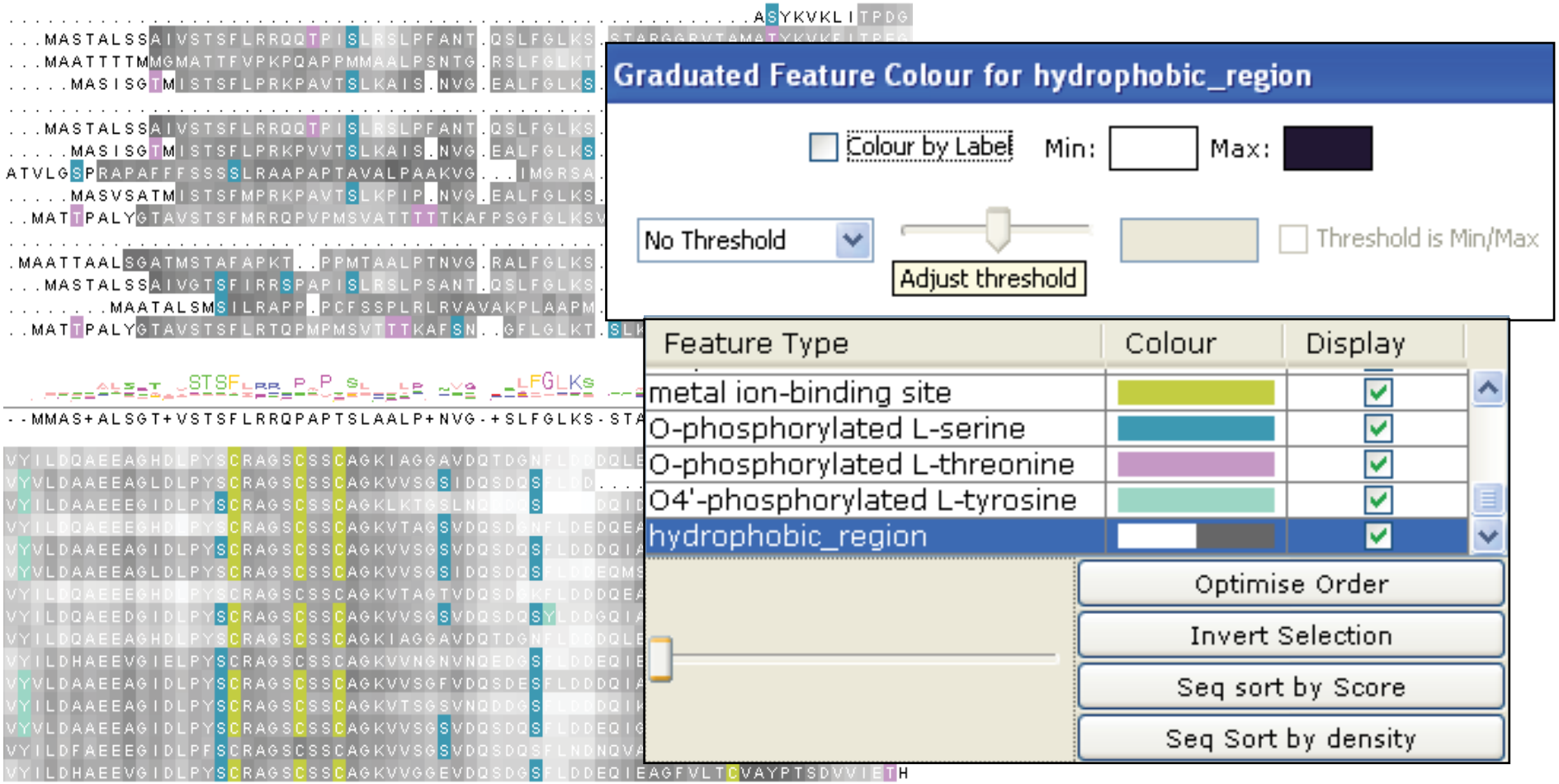
Features and the Distributed Annotation System Section 2.5.2, Exercise 25

  - Web servers that jalview can use to discover annotation for a sequence
- Task
  - Browse available DAS sources for protein sequences
  - Retrieve annotation for the ferredoxin alignment.
- Question
  - What does ‘optimise order’ do ?

Working with sequence features

- Task
  - Shading features using labels and scores
  - Sorting alignment using feature scores
- Questions
  - What kinds of annotation are best displayed with a ‘label’ colourscheme ?
  - How would you display only the highest or lowest scoring features ?

Shading, thresholding, colour by label.

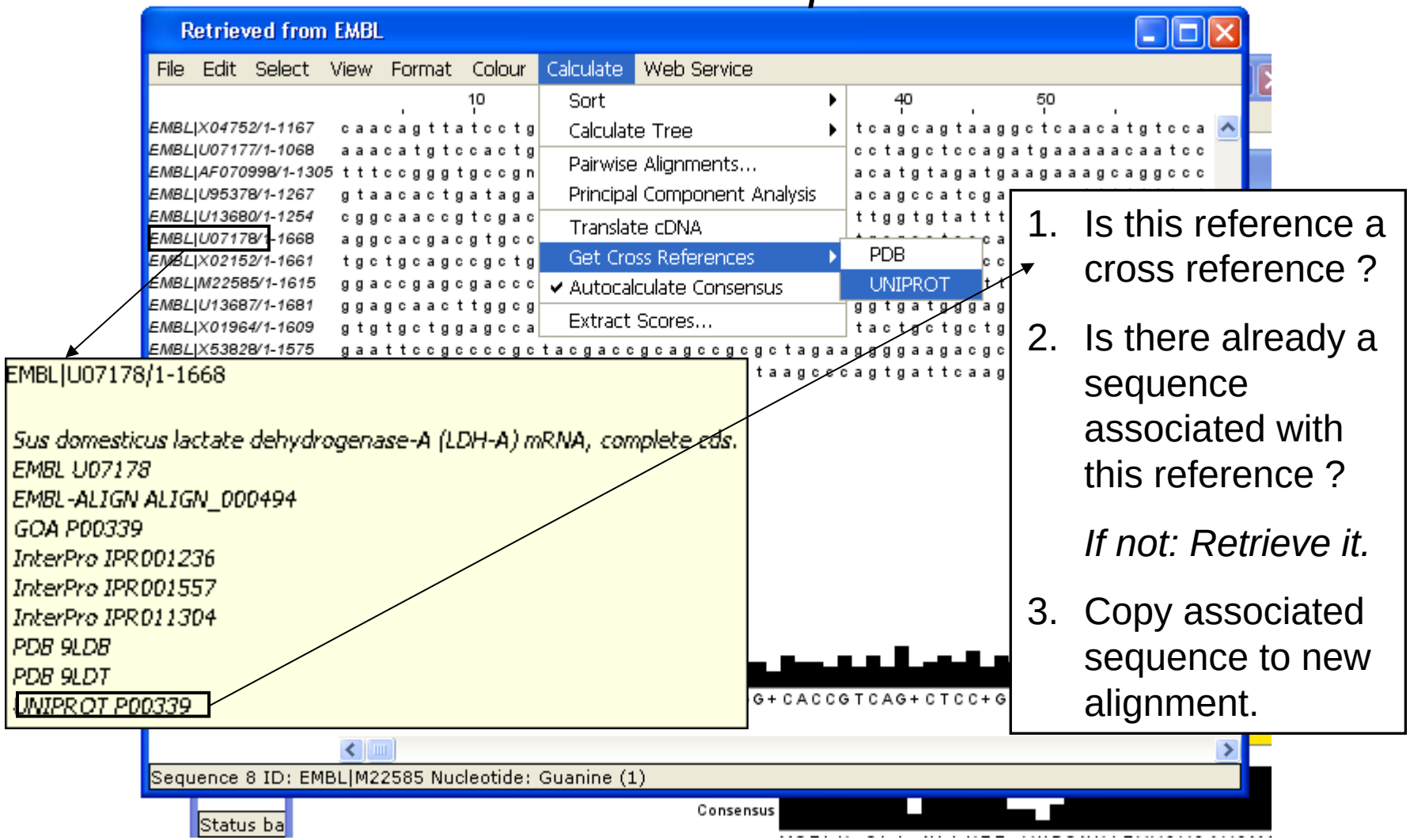


DNA and Protein in Jalview

- From DNA to Protein
  - Calculations => Translate cDNA
  - View protein annotation on exons using EMBL records
- From protein to DNA
  - Recover DNA for proteins using EMBL cross references

Semantic Processing: Database Reference Tracing

'get me the sequences from database **blah** *for the selected sequences'*



1. Is this reference a cross reference ?
2. Is there already a sequence associated with this reference ?

   If not: Retrieve it.
3. Copy associated sequence to new alignment.

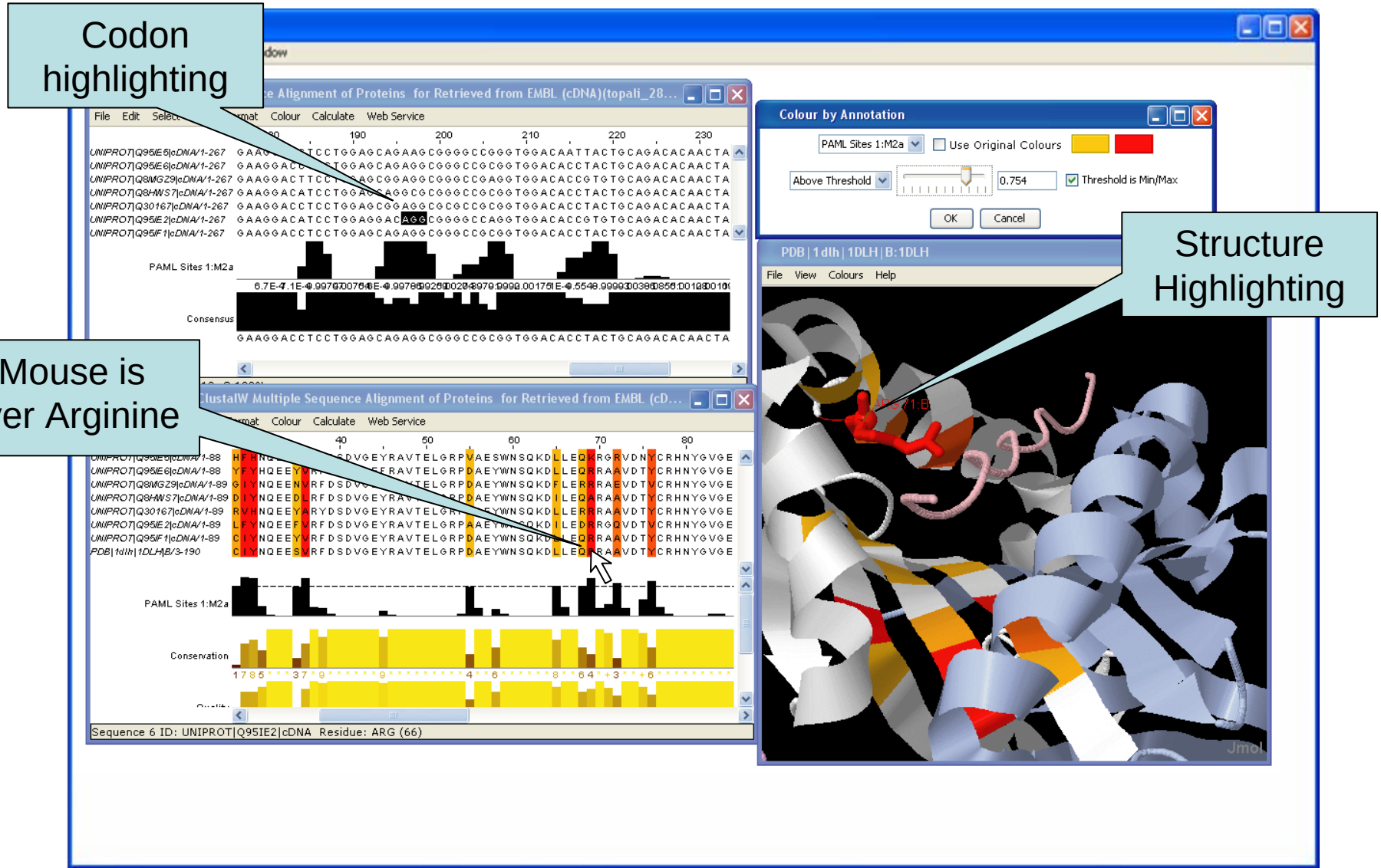
Protein Feature visualization on DNA
Section 2.6, exercise 28

- Task
  - Retrieve a DNA contig and visualize features from UNIPROT at their coding positions.
- Question
  - What information that Jalview can use is carried by EMBL sequence records ?

Protein Structure and Jalview Section 2.1

- Jalview includes the Jmol molecular graphics viewer
  - Structures can be coloured by their aligned sequences
  - Position of mouse highlighted in sequence or structure

Structure shaded by sequence

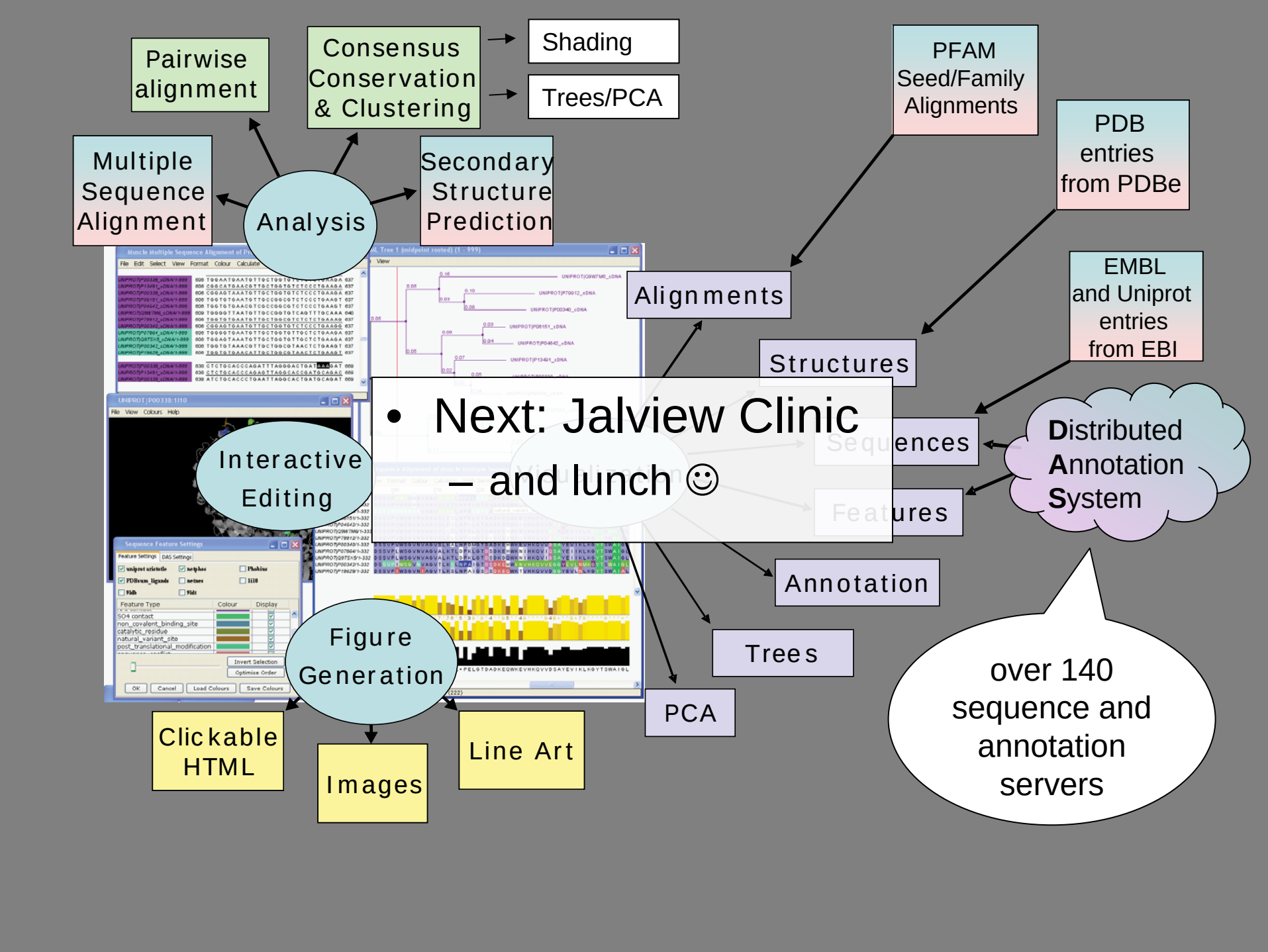


Protein Structures in Jalview
Sec 2.1. Exercise 14

- Task
  - Discover PDB structures for ferredoxin sequence(s)
  - Save and load structures and manipulate colouring
- Questions
  - How does Jalview match up sequence data to structural data

Final Exercise:
Superposing Structures with Jalview 2.6

- Task
  - Align structures using the ferredoxin alignment
    1. Make sure the structure associated with FER1_SPIOL is shown.
    2. Discover PDB ids for the MAIZE ferredoxin sequence
    3. View the structure and say ‘yes’ when asked to add it to the existing FER1_SPIOL structure.
    4. The structures will be retrieved and superimposed and aligned regions rendered as cartoons.
- Questions
  - What colourscheme would highlight the conserved parts of the structures ?
  - What if you only wanted to superimpose using just part of the alignment ?


Pairwise alignment
Consensus Conservation & Clustering
Shading
Trees/PCA
PFAM Seed/Family Alignments
PDB entries from PDBe
Multiple Sequence Alignment
Analysis
Secondary Structure Prediction
EMBL and Uniprot entries from EBI
Alignments
Structures
Sequences
Features
Annotation
Trees
PCA
Distributed Annotation System
Interactive Editing
Figure Generation
Clickable HTML
Images
Line Art
over 140 sequence and annotation servers
• Next: Jalview Clinic
– and lunch ☺

Jalview Clinic

- Try out the exercise/examples with your own data
- Identify things you can’t do but want to
- Use Jalview with other analysis programs
- Two way process
  - You learn more about your data
  - We learn what Jalview needs to be able to do better.