Pairwise Alignment

Mark Voorhies

3/27/2012

Review: Tips and tricks

Making a file executable:

```
chmod ”a+x” pydotter.py
```

Handling file/directory names with spaces:

```
cd My\ Directory\ with\ Spaces
```

or

```
cd ”My␣Directory␣with␣Spaces”
```

Review: Tips and tricks

Killing a process on OS X:

- Try ctrl-c
- If that doesn't work:

  -
    ```
    ps -awx | grep name_of_process
    ```
  - First column in ps output is PID (process ID)

  -
    ```
    kill PID
    ```
  - If that doesn't work:

    ```
    kill -KILL PID
    ```
- On Linux:

  ```
  ps -ealf | grep name_of_process
  ```

Review: Content

- FASTA files

```
>Name Free-form annotation
MGCLLIMKEGGPGRKHKLIVMLYLDENQ
EHELPIMTRAPPEDINADNAMACHINEW
NQEDLYMNILKHGPPGEDEDRKHEDEDG
```

Review: Content

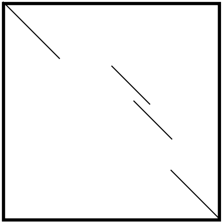
- FASTA files

```
>Name Free-form annotation
MGCLLIMKEGGPGRKHKLIVMLYLDENQ
EHELPIMTRAPPEDINADNAMACHINEW
NQEDLYMNILKHGPPGEDEDRKHEDEDG
```

- Dotplots: unbiased plot of all possible ungapped alignments of two sequences.

Pairwise Alignment

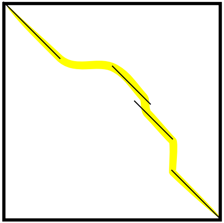
How can we automate our dotplot protocol to find the “best” gapped alignment of our sequences?



Pairwise Alignment

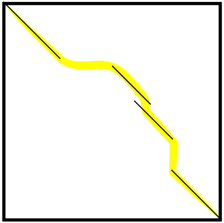
How can we automate our dotplot protocol to find the “best” gapped alignment of our sequences?

What do we mean by best?



Pairwise Alignment

How can we automate our dotplot protocol to find the “best” gapped alignment of our sequences?

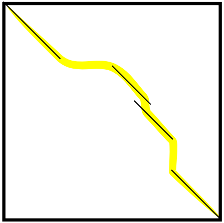


What do we mean by best?

- Residues with equivalent functional roles are paired

Pairwise Alignment

How can we automate our dotplot protocol to find the "best" gapped alignment of our sequences?

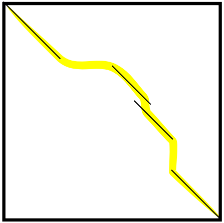


What do we mean by best?

- Residues with equivalent functional roles are paired
- Residues that derive from the same position in the common ancestor are paired (homology)

Pairwise Alignment

How can we automate our dotplot protocol to find the "best" gapped alignment of our sequences?



What do we mean by best?

- Residues with equivalent functional roles are paired
- Residues that derive from the same position in the common ancestor are paired (homology)
- The sequence alignment maximizes a similarity function

Deriving scores from alignments

Frequency of residue i:

$$p_i$$

Deriving scores from alignments

Frequency of residue i:

$$p_i$$

Frequency of residue i aligned to residue j:

$$q_{ij}$$

Deriving scores from alignments

Frequency of residue i:

$$p_i$$

Frequency of residue i aligned to residue j:

$$q_{ij}$$

Expected frequency if i and j are independent:

$$p_i p_j$$

Deriving scores from alignments

Frequency of residue i:

$$p_i$$

Frequency of residue i aligned to residue j:

$$q_{ij}$$

Expected frequency if i and j are independent:

$$p_i p_j$$

Ratio of observed to expected frequency:

$$\frac{q_{ij}}{p_i p_j}$$

Deriving scores from alignments

Frequency of residue *i*:

$$p_i$$

Frequency of residue *i* aligned to residue *j*:

$$q_{ij}$$

Expected frequency if *i* and *j* are independent:

$$p_i p_j$$

Ratio of observed to expected frequency:

$$\frac{q_{ij}}{p_i p_j}$$

Log odds (LOD) score:

$$s(i,j) = \log \frac{q_{ij}}{p_i p_j}$$

PAM (Dayhoff) and BLOSUM matrices

- PAM1 matrix originally calculated from manual alignments of highly conserved sequences (myoglobin, cytochrome C, etc.)

PAM (Dayhoff) and BLOSUM matrices

- PAM1 matrix originally calculated from manual alignments of highly conserved sequences (myoglobin, cytochrome C, etc.)
- We can think of a PAM matrix as evolving a sequence by one unit of time.

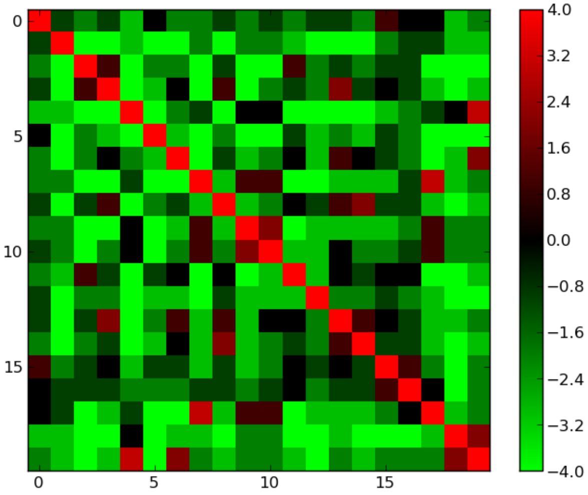
PAM (Dayhoff) and BLOSUM matrices

- PAM1 matrix originally calculated from manual alignments of highly conserved sequences (myoglobin, cytochrome C, etc.)
- We can think of a PAM matrix as evolving a sequence by one unit of time.
- If evolution is uniform over time, then PAM matrices for larger evolutionary steps can be generated by multiplying PAM1 by itself (so, higher numbered PAM matrices represent greater evolutionary distances).

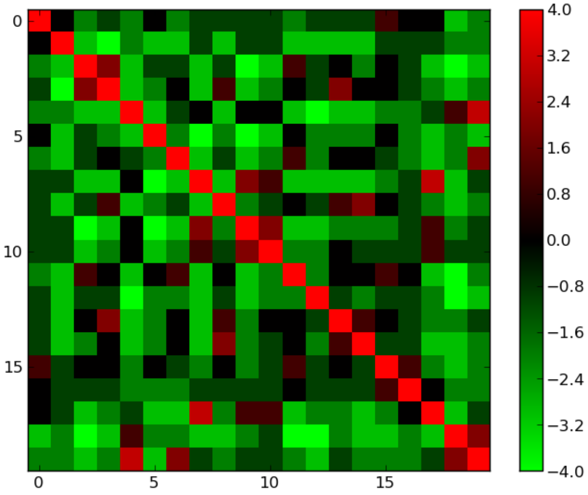
PAM (Dayhoff) and BLOSUM matrices

- PAM1 matrix originally calculated from manual alignments of highly conserved sequences (myoglobin, cytochrome C, etc.)
- We can think of a PAM matrix as evolving a sequence by one unit of time.
- If evolution is uniform over time, then PAM matrices for larger evolutionary steps can be generated by multiplying PAM1 by itself (so, higher numbered PAM matrices represent greater evolutionary distances).
- The BLOSUM matrices were determined from automatically generated ungapped alignments. Higher numbered BLOSUM matrices correspond to *smaller* evolutionary distances. BLOSUM62 is the default matrix for BLAST.

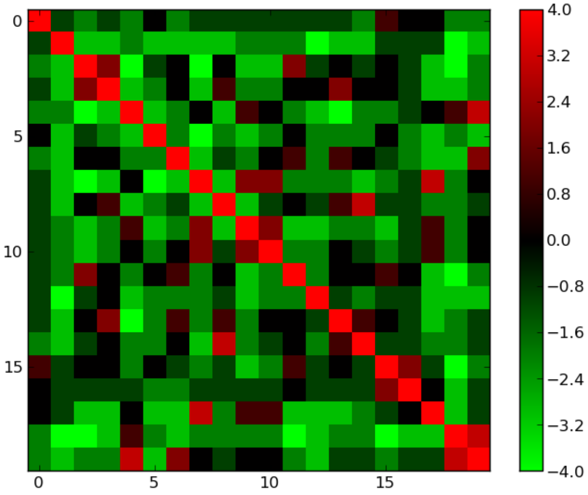
BLOSUM80



BLOSUM62



BLOSUM45



Fun with logarithms

In log space, multiplication and division become addition and subtraction:

$$\begin{aligned} \log(xy) &= \log(x) + \log(y) \\ \log(x/y) &= \log(x) - \log(y) \end{aligned}$$

Therefore, exponentiation becomes multiplication:

$$\log(x^y) = y \log(x)$$

Also, we can change of the base of a logarithm like so:

$$\log_A(x) = \log(x)/\log(A)$$

Scoring an alignment

Log odds (LOD) score:

$$s(i,j) = \log \frac{q_{ij}}{p_i p_j}$$

Scoring an alignment

Log odds (LOD) score:

$$s(i,j) = \log \frac{q_{ij}}{p_i p_j}$$

Multiplying independent probabilities is equivalent to adding independent log probabilities.

Scoring an alignment

Log odds (LOD) score:

$$s(i,j) = \log \frac{q_{ij}}{p_i p_j}$$

Multiplying independent probabilities is equivalent to adding independent log probabilities.

Therefore, for an ungapped alignment can be scored as:

$$S(x,y) = \log \prod_{i}^{N} \frac{q_{x_i y_i}}{p_{x_i} p_{y_i}} = \sum_{i}^{N} s(x_i, y_i)$$

Scoring an alignment

Log odds (LOD) score:

$$s(i,j) = \log \frac{q_{ij}}{p_i p_j}$$

Multiplying independent probabilities is equivalent to adding independent log probabilities.
Therefore, for an ungapped alignment can be scored as:

$$S(x,y) = \log \prod_{i}^{N} \frac{q_{x_i y_i}}{p_{x_i} p_{y_i}} = \sum_{i}^{N} s(x_i, y_i)$$

What about gaps?

Scoring an alignment

Log odds (LOD) score:

$$s(i,j) = \log \frac{q_{ij}}{p_i p_j}$$

Multiplying independent probabilities is equivalent to adding independent log probabilities.
Therefore, for an ungapped alignment can be scored as:

$$S(x,y) = \log \prod_{i}^{N} \frac{q_{x_i y_i}}{p_{x_i} p_{y_i}} = \sum_{i}^{N} s(x_i, y_i)$$

What about gaps?

- Probability of an insertion/deletion event (gap opening, G)
- Length distribution of insertions/deletions (gap extension, E)

Scoring an alignment

Log odds (LOD) score:

$$s(i,j) = \log \frac{q_{ij}}{p_i p_j}$$

Multiplying independent probabilities is equivalent to adding independent log probabilities.

Therefore, for an ungapped alignment can be scored as:

$$S(x,y) = \log \prod_i^N \frac{q_{x_i y_i}}{p_{x_i} p_{y_i}} = \sum_i^N s(x_i, y_i)$$

What about gaps?

- Probability of an insertion/deletion event (gap opening, G)
- Length distribution of insertions/deletions (gap extension, E)

$$S_{gapped}(x,y) = S(x,y) + \sum_i^{gaps} (G + E * L_i)$$

Scoring an alignment

Log odds (LOD) score:

$$s(i,j) = \log \frac{q_{ij}}{p_i p_j}$$

Multiplying independent probabilities is equivalent to adding independent log probabilities.
Therefore, for an ungapped alignment can be scored as:

$$S(x,y) = \log \prod_i^N \frac{q_{x_i y_i}}{p_{x_i} p_{y_i}} = \sum_i^N s(x_i, y_i)$$

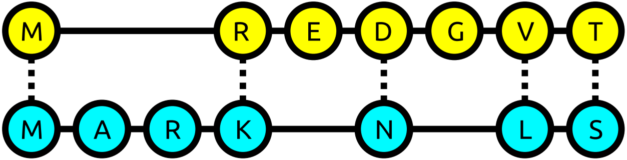
What about gaps?

- Probability of an insertion/deletion event (gap opening, G)
- Length distribution of insertions/deletions (gap extension, E)

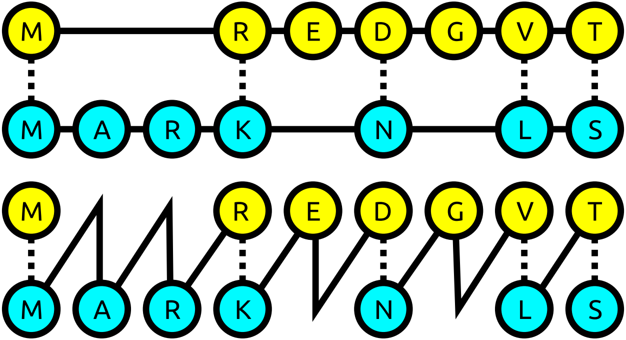
$$S_{gapped}(x,y) = S(x,y) + \sum_i^{gaps} (G + E * L_i)$$

We find an optimal alignment by finding x and y that maximize S.

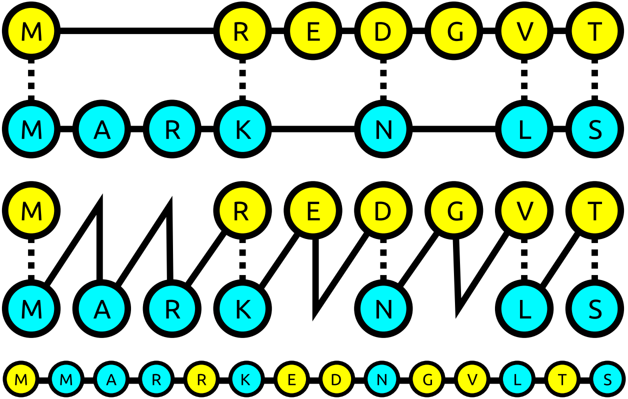
How many ways can we align two sequences?



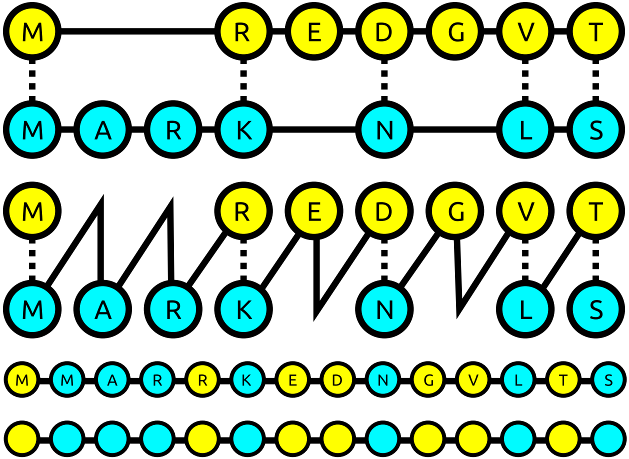
How many ways can we align two sequences?



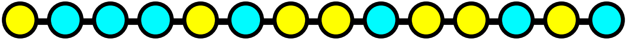
How many ways can we align two sequences?



How many ways can we align two sequences?



How many ways can we align two sequences?



Binomial formula:

$$\binom{k}{r} = \frac{k!}{(k-r)!r!}$$

$$\binom{2n}{n} = \frac{(2n)!}{n!n!}$$

Stirling's approximation:

$$x! \approx \sqrt{2\pi}\left(x^{x+\frac{1}{2}}\right)e^{-x}$$

$$\binom{2n}{n} \approx \frac{2^{2n}}{\sqrt{\pi n}}$$

Scoring an alignment quickly

$\frac{2^{2n}}{\sqrt{\pi n}}$ is too expensive.

Scoring an alignment quickly

$\frac{2^{2n}}{\sqrt{\pi n}}$ is too expensive.

$$S_{gapped}(x,y) = S(x,y) + \sum_{i}^{gaps} (G + E * L_i)$$

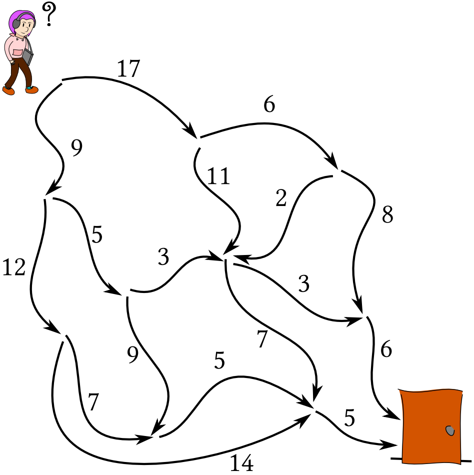
Scoring an alignment quickly

$\frac{2^{2n}}{\sqrt{\pi n}}$ is too expensive.

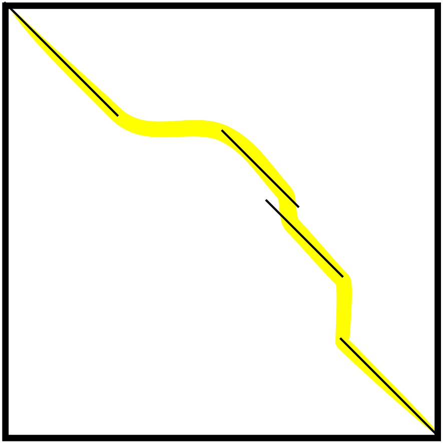
$$S_{gapped}(x,y) = S(x,y) + \sum_{i}^{gaps} (G + E * L_i)$$

The best alignment of any pair of subsequences is independent of the global alignment.

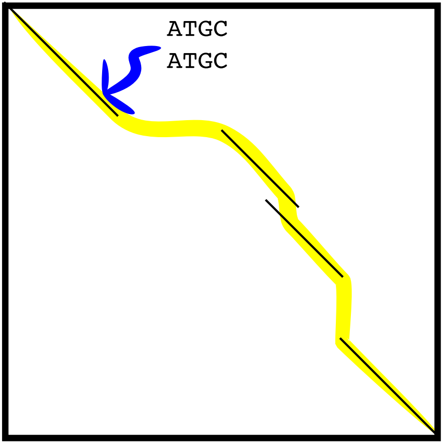
Dynamic Programming



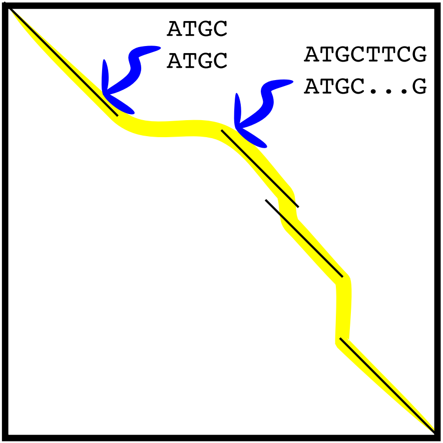
Needleman-Wunsch Global Alignment



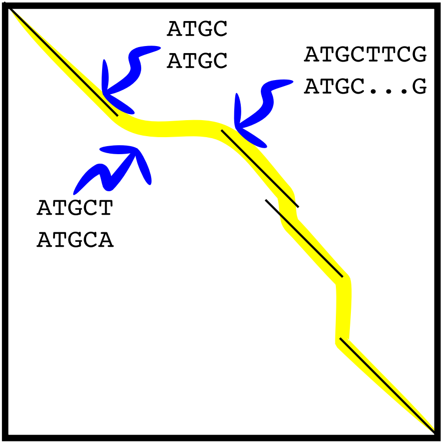
Needleman-Wunsch Global Alignment



Needleman-Wunsch Global Alignment



Needleman-Wunsch Global Alignment

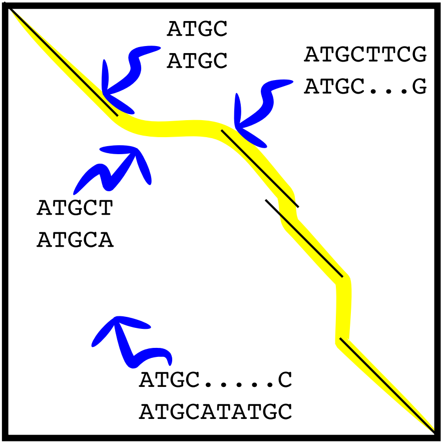


ATGC
ATGC

ATGCTTCG
ATGC...G

ATGCT
ATGCA

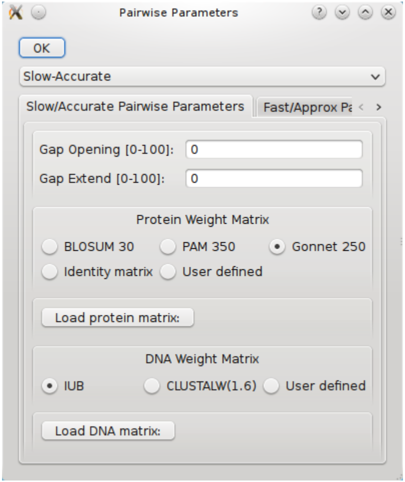
Needleman-Wunsch Global Alignment



Alignment speeds

- DOTTER: $O(n^2)$
- Exhaustive search: $\frac{2^{2n}}{\sqrt{\pi n}}$
- Dynamic programming: $O(n^2)$ to $O(n^3)$

Setting gap penalties in CLUSTALX



Slow-Accurate

Slow/Accurate Pairwise Parameters | Fast/Approx Pa

Gap Opening [0-100]: 0

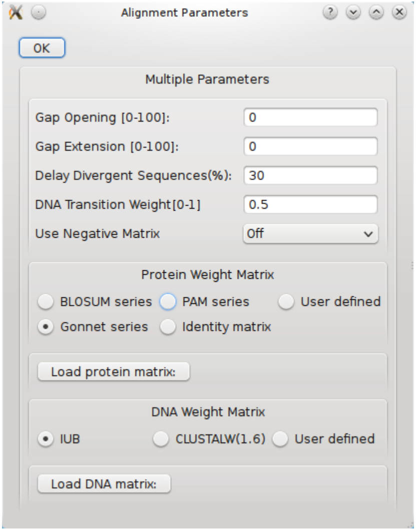
Gap Extend [0-100]: 0

Protein Weight Matrix

○ BLOSUM 30 ○ PAM 350 ● Gonnet 250
○ Identity matrix ○ User defined

DNA Weight Matrix

● IUB ○ CLUSTALW(1.6) ○ User defined



Multiple Parameters

Gap Opening [0-100]: 0

Gap Extension [0-100]: 0

Delay Divergent Sequences(%): 30

DNA Transition Weight[0-1] 0.5

Use Negative Matrix Off

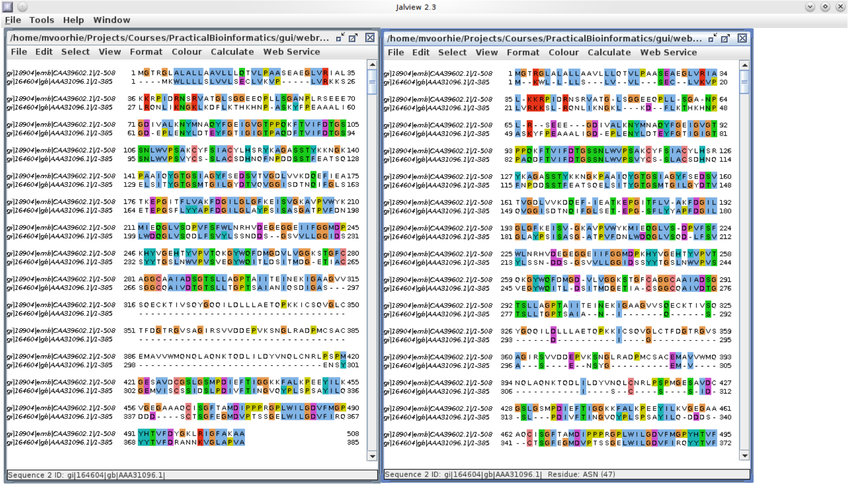
Protein Weight Matrix

○ BLOSUM series ○ PAM series ○ User defined
● Gonnet series ○ Identity matrix

DNA Weight Matrix

● IUB ○ CLUSTALW(1.6) ○ User defined

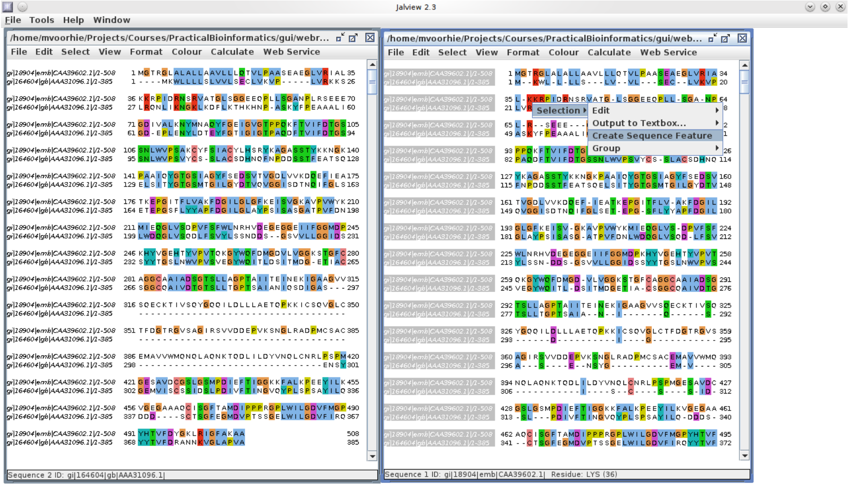
Annotating features in JALVIEW



```
/home/mvoorhie/Projects/Courses/PracticalBioinformatics/gui/webr...
File  Edit  Select  View  Format  Colour  Calculate  Web Service

gi|18904|emb|CAA39602.1|/1-508    1 MGTRGLALALLAAVLLLQTVLPAASEAEGLVRIAL 35
gi|164604|gb|AAA31096.1|/1-385    1 ----MKWLLLLSLVVLSECLVKVP-----LVRKKS 26

gi|18904|emb|CAA39602.1|/1-508   36 KKRPIDRNSRVATGLSGGEEQPLLSGANPLRSEEE 70
gi|164604|gb|AAA31096.1|/1-385   27 LRQNLIKNGKLKDFLKTHKHNP-ASKYFPEAAALI 60

gi|18904|emb|CAA39602.1|/1-508   71 GDIVALKNYMNAQYFGEIGVGTPPQKFTVIFDTGS 105
gi|164604|gb|AAA31096.1|/1-385   61 GD-EPLENYLDTEYFGTIGIGTPAQDFTVIFDTGS 94

gi|18904|emb|CAA39602.1|/1-508  106 SNLWVPSAKCYFSIACYLHSRYKAGASSTYKKNGK 140
gi|164604|gb|AAA31096.1|/1-385   95 SNLWVPSVYCS-SLACSDHNQFNPDDSSTFEATSQ 128

gi|18904|emb|CAA39602.1|/1-508  141 PAAIQYGTGSIAGYFSEDSVTVGDLVVKDQEFIEA 175
gi|164604|gb|AAA31096.1|/1-385  129 ELSITYGTGSMTGILGYDTVQVGGISDTNQIFGLS 163

gi|18904|emb|CAA39602.1|/1-508  176 TKEPGITFLVAKFDGILGLGFKEISVGKAVPVWYK 210
gi|164604|gb|AAA31096.1|/1-385  164 ETEPGSFLYYAPFDGILGLAYPSISASGATPVFDN 198

gi|18904|emb|CAA39602.1|/1-508  211 MIEQGLVSDPVFSFWLNRHVDEGEGGEIIFGGMDP 245
gi|164604|gb|AAA31096.1|/1-385  199 LWDQGLVSQDLFSVYLSSNDDS--GSVVLLGGIDS 231

gi|18904|emb|CAA39602.1|/1-508  246 KHYVGEHTYVPVTQKGYWQFDMGDVLVGGKSTGFC 280
gi|164604|gb|AAA31096.1|/1-385  232 SYYTGSLNWVPVSVEGYWQITLDSITMDG-ETIAC 265

gi|18904|emb|CAA39602.1|/1-508  281 AGGCAAIADSGTSLLAGPTAIITEINEKIGAAGVV 315
gi|164604|gb|AAA31096.1|/1-385  266 SGGCQAIVDTGTSLLTGPTSAIANIQSDIGAS--- 297

gi|18904|emb|CAA39602.1|/1-508  316 SQECKTIVSQYGQQILDLLLAETQPKKICSQVGLC 350
gi|164604|gb|AAA31096.1|/1-385      -----------------------------------

gi|18904|emb|CAA39602.1|/1-508  351 TFDGTRGVSAGIRSVVDDEPVKSNGLRADPMCSAC 385
gi|164604|gb|AAA31096.1|/1-385      -----------------------------------

gi|18904|emb|CAA39602.1|/1-508  386 EMAVVWMQNQLAQNKTQDLILDYVNQLCNRLPSPM 420
gi|164604|gb|AAA31096.1|/1-385  298 -------------------------------ENSY 301

gi|18904|emb|CAA39602.1|/1-508  421 GESAVDCGSLGSMPDIEFTIGGKKFALKPEEYILK 455
gi|164604|gb|AAA31096.1|/1-385  302 GEMVISCSSIDSLPDIVFTINGVQYPLSPSAYILQ 336

gi|18904|emb|CAA39602.1|/1-508  456 VGEGAAAQCISGFTAMDIPPPRGPLWILGDVFMGP 490
gi|164604|gb|AAA31096.1|/1-385  337 DDD----SCTSGFEGMDVPTSSGELWILGDVFIRQ 367

gi|18904|emb|CAA39602.1|/1-508  491 YHTVFDYGKLRIGFAKAA                  508
gi|164604|gb|AAA31096.1|/1-385  368 YYTVFDRANNKVGLAPVA                  385

Sequence 2 ID: gi|164604|gb|AAA31096.1|
```

```
/home/mvoorhie/Projects/Courses/PracticalBioinformatics/gui/web...
File  Edit  Select  View  Format  Colour  Calculate  Web Service

gi|18904|emb|CAA39602.1|/1-508    1 MGTRGLALALLAAVLLLQTVLPAASEAEGLVRIA 34
gi|164604|gb|AAA31096.1|/1-385    1 M--KWL-L-LLS---LV--VL---SEC--LVKVP 20

gi|18904|emb|CAA39602.1|/1-508   35 L-KKRPIDRNSRVATG-LSGGEEQPLL-SGA-NP 64
gi|164604|gb|AAA31096.1|/1-385   21 LVRKKSL-RQNLIKNGKL---KD--FLKTHKHNP 48

gi|18904|emb|CAA39602.1|/1-508   65 L-R--SEEE---GDIVALKNYMNAQYFGEIGVGT 92
gi|164604|gb|AAA31096.1|/1-385   49 ASKYFPEAAALIGD-EPLENYLDTEYFGTIGIGT 81

gi|18904|emb|CAA39602.1|/1-508   93 PPQKFTVIFDTGSSNLWVPSAKCYFSIACYLHSR 126
gi|164604|gb|AAA31096.1|/1-385   82 PAQDFTVIFDTGSSNLWVPSVYCS-SLACSDHNQ 114

gi|18904|emb|CAA39602.1|/1-508  127 YKAGASSTYKKNGKPAAIQYGTGSIAGYFSEDSV 160
gi|164604|gb|AAA31096.1|/1-385  115 FNPDDSSTFEATSQELSITYGTGSMTGILGYDTV 148

gi|18904|emb|CAA39602.1|/1-508  161 TVGDLVVKDQEF-IEATKEPGITFLV-AKFDGIL 192
gi|164604|gb|AAA31096.1|/1-385  149 QVGGISDTNQIFGLSET-EPG-SFLYYAPFDGIL 180

gi|18904|emb|CAA39602.1|/1-508  193 GLGFKEISV-GKAVPVWYKMIEQGLVS-DPVFSF 224
gi|164604|gb|AAA31096.1|/1-385  181 GLAYPSISASG-ATPVFDNLWDQGLVSQD-LFSV 212

gi|18904|emb|CAA39602.1|/1-508  225 WLNRHVDEGEGGEIIFGGMDPKHYVGEHTYVPVT 258
gi|164604|gb|AAA31096.1|/1-385  213 YLSSN-DDS-GSVVLLGGIDSSYYTGSLNWVPVS 244

gi|18904|emb|CAA39602.1|/1-508  259 QKGYWQFDMGD-VLVGGKSTGFCAGGCAAIADSG 291
gi|164604|gb|AAA31096.1|/1-385  245 VEGYWQITL-DSITMDGETIA-CSGGCQAIVDTG 276

gi|18904|emb|CAA39602.1|/1-508  292 TSLLAGPTAIITEINEKIGAAGVVSQECKTIVSQ 325
gi|164604|gb|AAA31096.1|/1-385  277 TSLLTGPTSAIA--N--I-------Q------S- 292

gi|18904|emb|CAA39602.1|/1-508  326 YGQQILDLLLAETQPKKICSQVGLCTFDGTRGVS 359
gi|164604|gb|AAA31096.1|/1-385  293 ------D----------I----G----------- 295

gi|18904|emb|CAA39602.1|/1-508  360 AGIRSVVDDEPVKSNGLRADPMCSACEMAVVWMQ 393
gi|164604|gb|AAA31096.1|/1-385  296 A---S----E--NSYG----------EM-V---- 305

gi|18904|emb|CAA39602.1|/1-508  394 NQLAQNKTQDLILDYVNQLCNRLPSPMGESAVDC 427
gi|164604|gb|AAA31096.1|/1-385  306 -----------I----S--C----S----S-ID- 312

gi|18904|emb|CAA39602.1|/1-508  428 GSLGSMPDIEFTIGGKKFALKPEEYILKVGEGAA 461
gi|164604|gb|AAA31096.1|/1-385  313 -SL---PDIVFTINGVQYPLSPSAYILQ-DDDS- 340

gi|18904|emb|CAA39602.1|/1-508  462 AQCISGFTAMDIPPPRGPLWILGDVFMGPYHTVF 495
gi|164604|gb|AAA31096.1|/1-385  341 --CTSGFEGMDVPTSSGELWILGDVFIRQYYTVF 372

Sequence 2 ID: gi|164604|gb|AAA31096.1|   Residue: ASN (47)
```

Annotating features in JALVIEW

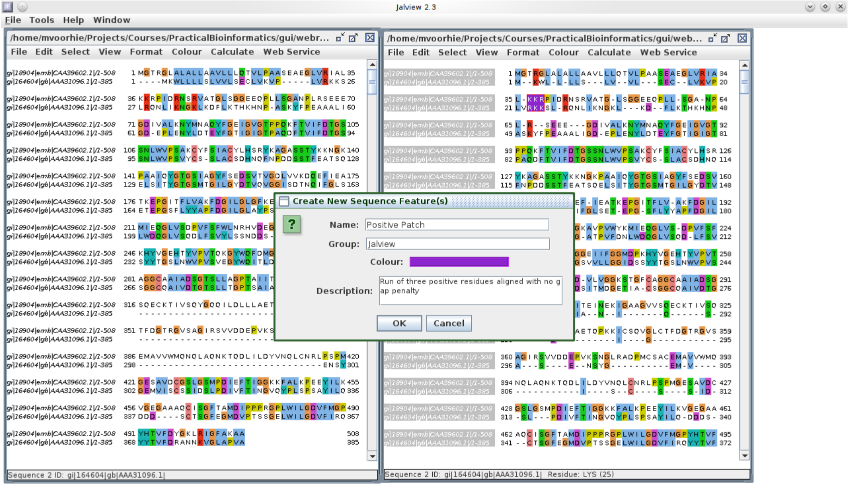


Jalview 2.3
File Tools Help Window
/home/mvoorhie/Projects/Courses/PracticalBioinformatics/gui/webr...
File Edit Select View Format Colour Calculate Web Service
/home/mvoorhie/Projects/Courses/PracticalBioinformatics/gui/web...
File Edit Select View Format Colour Calculate Web Service
Selection
Edit
Output to Textbox...
Create Sequence Feature
Group

```
gi|18904|emb|CAA39602.1|/1-508     1 MGTRGLALALLAAVLLLQTVLPAASEAEGLVRIAL 35
gi|164604|gb|AAA31096.1|/1-385     1 ----MKWLLLLSLVVLSECLVKVP-----LVRKKS 26

gi|18904|emb|CAA39602.1|/1-508    36 KKRPIDRNSRVATGLSGGEEQPLLSGANPLRSEEE 70
gi|164604|gb|AAA31096.1|/1-385    27 LRQNLIKNGKLKDFLKTHKHNP-ASKYFPEAAALI 60

gi|18904|emb|CAA39602.1|/1-508    71 GDIVALKNYMNAQYFGEIGVGTPPQKFTVIFDTGS 105
gi|164604|gb|AAA31096.1|/1-385    61 GD-EPLENYLDTEYFGTIGIGTPAQDFTVIFDTGS 94

gi|18904|emb|CAA39602.1|/1-508   106 SNLWVPSAKCYFSIACYLHSRYKAGASSTYKKNGK 140
gi|164604|gb|AAA31096.1|/1-385    95 SNLWVPSVYCS-SLACSDHNQFNPDDSSTFEATSQ 128

gi|18904|emb|CAA39602.1|/1-508   141 PAAIQYGTGSIAGYFSEDSVTVGDLVVKDQEFIEA 175
gi|164604|gb|AAA31096.1|/1-385   129 ELSITYGTGSMTGILGYDTVQVGGISDTNQIFGLS 163

gi|18904|emb|CAA39602.1|/1-508   176 TKEPGITFLVAKFDGILGLGFKEISVGKAVPVWYK 210
gi|164604|gb|AAA31096.1|/1-385   164 ETEPGSFLYYAPFDGILGLAYPSISASGATPVFDN 198

gi|18904|emb|CAA39602.1|/1-508   211 MIEQGLVSDPVFSFWLNRHVDEGEGGEIIFGGMDP 245
gi|164604|gb|AAA31096.1|/1-385   199 LWDQGLVSQDLFSVYLSSNDDS--GSVVLLGGIDS 231

gi|18904|emb|CAA39602.1|/1-508   246 KHYVGEHTYVPVTQKGYWQFDMGDVLVGGKSTGFC 280
gi|164604|gb|AAA31096.1|/1-385   232 SYYTGSLNWVPVSVEGYWQITLDSITMDG-ETIAC 265

gi|18904|emb|CAA39602.1|/1-508   281 AGGCAAIADSGTSLLAGPTAIITEINEKIGAAGVV 315
gi|164604|gb|AAA31096.1|/1-385   266 SGGCQAIVDTGTSLLTGPTSAIANIQSDIGAS--- 297

gi|18904|emb|CAA39602.1|/1-508   316 SQECKTIVSQYGQQILDLLLAETQPKKICSQVGLC 350
gi|164604|gb|AAA31096.1|/1-385       -----------------------------------

gi|18904|emb|CAA39602.1|/1-508   351 TFDGTRGVSAGIRSVVDDEPVKSNGLRADPMCSAC 385
gi|164604|gb|AAA31096.1|/1-385       -----------------------------------

gi|18904|emb|CAA39602.1|/1-508   386 EMAVVWMQNQLAQNKTQDLILDYVNQLCNRLPSPM 420
gi|164604|gb|AAA31096.1|/1-385   298 -------------------------------ENSY 301

gi|18904|emb|CAA39602.1|/1-508   421 GESAVDCGSLGSMPDIEFTIGGKKFALKPEEYILK 455
gi|164604|gb|AAA31096.1|/1-385   302 GEMVISCSSIDSLPDIVFTINGVQYPLSPSAYILQ 336

gi|18904|emb|CAA39602.1|/1-508   456 VGEGAAAQCISGFTAMDIPPPRGPLWILGDVFMGP 490
gi|164604|gb|AAA31096.1|/1-385   337 DDD----SCTSGFEGMDVPTSSGELWILGDVFIRQ 367

gi|18904|emb|CAA39602.1|/1-508   491 YHTVFDYGKLRIGFAKAA                  508
gi|164604|gb|AAA31096.1|/1-385   368 YYTVFDRANNKVGLAPVA                  385

Sequence 2 ID: gi|164604|gb|AAA31096.1|
```

```
gi|18904|emb|CAA39602.1|/1-508     1 MGTRGLALALLAAVLLQTVLPAASEAEGLVRIA 34
gi|164604|gb|AAA31096.1|/1-385     1 M--KWL-L-LLS---LV--VL---SEC--LVKVP 20

gi|18904|emb|CAA39602.1|/1-508    35 L-KKRPIDRNSRVATG-LSGGEEQPLL-SGA-NP 64
gi|164604|gb|AAA31096.1|/1-385    21 LVR                                 8

gi|18904|emb|CAA39602.1|/1-508    65 L-R--SEEE---                        2
gi|164604|gb|AAA31096.1|/1-385    49 ASKYFPEAAALI                        1

gi|18904|emb|CAA39602.1|/1-508    93 PPQKFTVIFDTG                       26
gi|164604|gb|AAA31096.1|/1-385    82 PAQDFTVIFDTGSSNLWVPSVYCS-SLACSDHNQ 114

gi|18904|emb|CAA39602.1|/1-508   127 YKAGASSTYKKNGKPAAIQYGTGSIAGYFSEDSV 160
gi|164604|gb|AAA31096.1|/1-385   115 FNPDDSSTFEATSQELSITYGTGSMTGILGYDTV 148

gi|18904|emb|CAA39602.1|/1-508   161 TVGDLVVKDQEF-IEATKEPGITFLV-AKFDGIL 192
gi|164604|gb|AAA31096.1|/1-385   149 QVGGISDTNQIFGLSET-EPG-SFLYYAPFDGIL 180

gi|18904|emb|CAA39602.1|/1-508   193 GLGFKEISV-GKAVPVWYKMIEQGLVS-DPVFSF 224
gi|164604|gb|AAA31096.1|/1-385   181 GLAYPSISASG-ATPVFDNLWDQGLVSQD-LFSV 212

gi|18904|emb|CAA39602.1|/1-508   225 WLNRHVDEGEGGEIIFGGMDPKHYVGEHTYVPVT 258
gi|164604|gb|AAA31096.1|/1-385   213 YLSSN-DDS-GSVVLLGGIDSSYYTGSLNWVPVS 244

gi|18904|emb|CAA39602.1|/1-508   259 QKGYWQFDMGD-VLVGGKSTGFCAGGCAAIADSG 291
gi|164604|gb|AAA31096.1|/1-385   245 VEGYWQITL-DSITMDGETIA-CSGGCQAIVDTG 276

gi|18904|emb|CAA39602.1|/1-508   292 TSLLAGPTAIITEINEKIGAAGVVSQECKTIVSQ 325
gi|164604|gb|AAA31096.1|/1-385   277 TSLLTGPTSAIA--N--I-------Q------S- 292

gi|18904|emb|CAA39602.1|/1-508   326 YGQQILDLLLAETQPKKICSQVGLCTFDGTRGVS 359
gi|164604|gb|AAA31096.1|/1-385   293 ------D----------I----G----------- 295

gi|18904|emb|CAA39602.1|/1-508   360 AGIRSVVDDEPVKSNGLRADPMCSACEMAVVWMQ 393
gi|164604|gb|AAA31096.1|/1-385   296 A---S----E--NSYG----------EM-V---- 305

gi|18904|emb|CAA39602.1|/1-508   394 NQLAQNKTQDLILDYVNQLCNRLPSPMGESAVDC 427
gi|164604|gb|AAA31096.1|/1-385   306 -----------I-----S--C----S-----S-ID- 312

gi|18904|emb|CAA39602.1|/1-508   428 GSLGSMPDIEFTIGGKKFALKPEEYILKVGEGAA 461
gi|164604|gb|AAA31096.1|/1-385   313 -SL---PDIVFTINGVQYPLSPSAYILQ-DDDS- 340

gi|18904|emb|CAA39602.1|/1-508   462 AQCISGFTAMDIPPPRGPLWILGDVFMGPYHTVF 495
gi|164604|gb|AAA31096.1|/1-385   341 --CTSGFEGMDVPTSSGELWILGDVFIRQYYTVF 372

Sequence 1 ID: gi|18904|emb|CAA39602.1|   Residue: LYS (36)
```

Annotating features in JALVIEW



Jalview 2.3
File Tools Help Window
/home/mvoorhie/Projects/Courses/PracticalBioinformatics/gui/webr...
File Edit Select View Format Colour Calculate Web Service
/home/mvoorhie/Projects/Courses/PracticalBioinformatics/gui/web...
File Edit Select View Format Colour Calculate Web Service
gi|18904|emb|CAA39602.1|/1-508
gi|164604|gb|AAA31096.1|/1-385
Create New Sequence Feature(s)
OK
Cancel

Name: Positive Patch

Group: Jalview

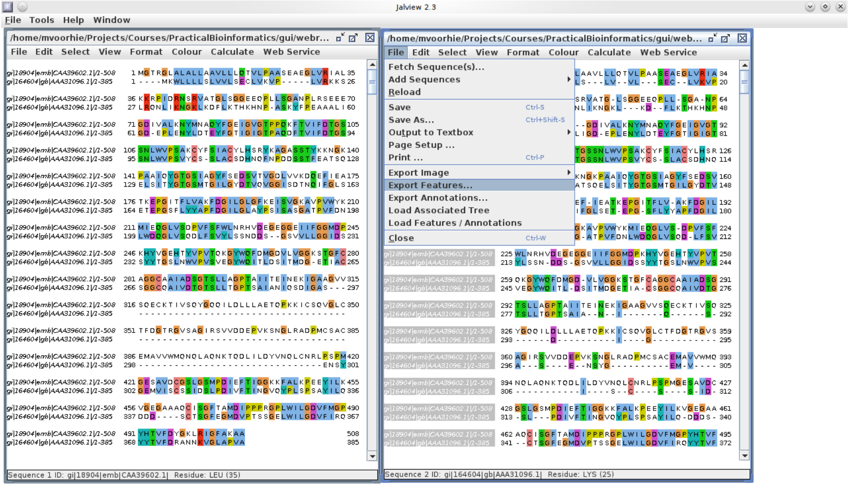
Colour:

Description: Run of three positive residues aligned with no gap penalty

Sequence 2 ID: gi|164604|gb|AAA31096.1|

Sequence 2 ID: gi|164604|gb|AAA31096.1| Residue: LYS (25)

Annotating features in JALVIEW



Jalview 2.3
File Tools Help Window
/home/mvoorhie/Projects/Courses/PracticalBioinformatics/gui/webr...
File Edit Select View Format Colour Calculate Web Service
/home/mvoorhie/Projects/Courses/PracticalBioinformatics/gui/web...
File Edit Select View Format Colour Calculate Web Service
Fetch Sequence(s)...
Add Sequences
Reload
Save Ctrl-S
Save As... Ctrl+Shift-S
Output to Textbox
Page Setup ...
Print ... Ctrl-P
Export Image
Export Features...
Export Annotations...
Load Associated Tree
Load Features / Annotations
Close Ctrl-W

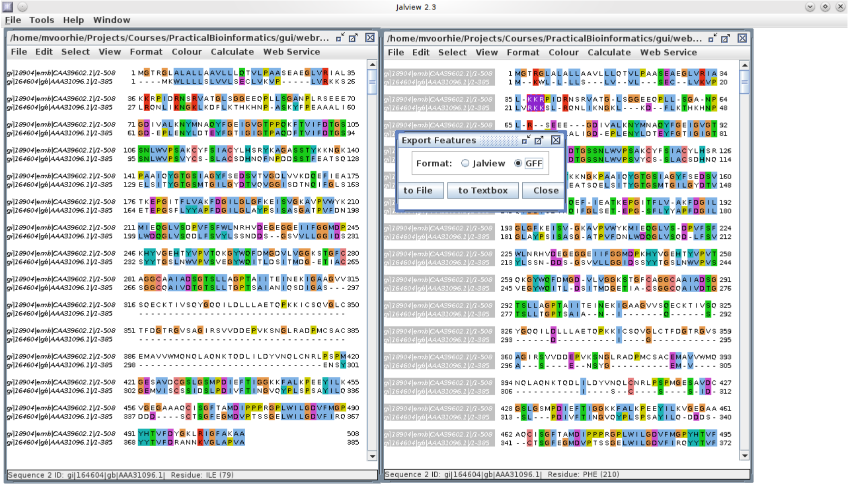
```
gi|18904|emb|CA439602.1|/1-508     1 MGTRGLALALLAAVLLLQTVLPAASEAEGLVRIAL 35
gi|164604|gb|AA431096.1|/1-385     1 ----MKWLLLLSLVVLSECLVKVP-----LVRKKS 26

gi|18904|emb|CA439602.1|/1-508    36 KKRPIDRNSRVATGLSGGEEQPLLSGANPLRSEEE 70
gi|164604|gb|AA431096.1|/1-385    27 LRQNLIKNGKLKDFLKTHKHNP-ASKYFPEAAALI 60

gi|18904|emb|CA439602.1|/1-508    71 GDIVALKNYMNAQYFGEIGVGTPPQKFTVIFDTGS 105
gi|164604|gb|AA431096.1|/1-385    61 GD-EPLENYLDTEYFGTIGIGTPAQDFTVIFDTGS 94

gi|18904|emb|CA439602.1|/1-508   106 SNLWVPSAKCYFSIACYLHSRYKAGASSTYKKNGK 140
gi|164604|gb|AA431096.1|/1-385    95 SNLWVPSVYCS-SLACSDHNQFNPDDSSTFEATSQ 128

gi|18904|emb|CA439602.1|/1-508   141 PAAIQYGTGSIAGYFSEDSVTVGDLVVKDQEFIEA 175
gi|164604|gb|AA431096.1|/1-385   129 ELSITYGTGSMTGILGYDTVQVGGISDTNQIFGLS 163

gi|18904|emb|CA439602.1|/1-508   176 TKEPGITFLVAKFDGILGLGFKEISVGKAVPVWYK 210
gi|164604|gb|AA431096.1|/1-385   164 ETEPGSFLYYAPFDGILGLAYPSISASGATPVFDN 198

gi|18904|emb|CA439602.1|/1-508   211 MIEQGLVSDPVFSFWLNRHVDEGEGGEIIFGGMDP 245
gi|164604|gb|AA431096.1|/1-385   199 LWDQGLVSQDLFSVYLSSNDDS--GSVVLLGGIDS 231

gi|18904|emb|CA439602.1|/1-508   246 KHYVGEHTYVPVTQKGYWQFDMGDVLVGGKSTGFC 280
gi|164604|gb|AA431096.1|/1-385   232 SYYTGSLNWVPVSVEGYWQITLDSITMDG-ETIAC 265

gi|18904|emb|CA439602.1|/1-508   281 AGGCAAIADSGTSLLAGPTAIITEINEKIGAAGVV 315
gi|164604|gb|AA431096.1|/1-385   266 SGGCQAIVDTGTSLLTGPTSAIANIQSDIGAS--- 297

gi|18904|emb|CA439602.1|/1-508   316 SQECKTIVSQYGQQILDLLLAETQPKKICSQVGLC 350
gi|164604|gb|AA431096.1|/1-385       -----------------------------------

gi|18904|emb|CA439602.1|/1-508   351 TFDGTRGVSAGIRSVVDDEPVKSNGLRADPMCSAC 385
gi|164604|gb|AA431096.1|/1-385       -----------------------------------

gi|18904|emb|CA439602.1|/1-508   386 EMAVVWMQNQLAQNKTQDLILDYVNQLCNRLPSPM 420
gi|164604|gb|AA431096.1|/1-385   298 ------------------------------ENSY 301

gi|18904|emb|CA439602.1|/1-508   421 GESAVDCGSLGSMPDIEFTIGGKKFALKPEEYILK 455
gi|164604|gb|AA431096.1|/1-385   302 GEMVISCSSIDSLPDIVFTINGVQYPLSPSAYILQ 336

gi|18904|emb|CA439602.1|/1-508   456 VGEGAAAQCISGFTAMDIPPPRGPLWILGDVFMGP 490
gi|164604|gb|AA431096.1|/1-385   337 DDD----SCTSGFEGMDVPTSSGELWILGDVFIRQ 367

gi|18904|emb|CA439602.1|/1-508   491 YHTVFDYGKLRIGFAKAA                  508
gi|164604|gb|AA431096.1|/1-385   368 YYTVFDRANNKVGLAPVA                  385
```

Sequence 1 ID: gi|18904|emb|CAA39602.1| Residue: LEU (35)

```
LAAVLLLQTVLPAASEAEGLVRIA 34
LS---LV--VL---SEC--LVKVP 20

SRVATG-LSGGEEQPLL-SGA-NP 64
NLIKNGKL---KD--FLKTHKHNP 48

--GDIVALKNYMNAQYFGEIGVGT 92
LIGD-EPLENYLDTEYFGTIGIGT 81

TGSSNLWVPSAKCYFSIACYLHSR 126
TGSSNLWVPSVYCS-SLACSDHNQ 114

KNGKPAAIQYGTGSIAGYFSEDSV 160
ATSQELSITYGTGSMTGILGYDTV 148

EF-IEATKEPGITFLV-AKFDGIL 192
IFGLSET-EPG-SFLYYAPFDGIL 180

GKAVPVWYKMIEQGLVS-DPVFSF 224
G-ATPVFDNLWDQGLVSQD-LFSV 212

gi|18904|emb|CA439602.1|/1-508   225 WLNRHVDEGEGGEIIFGGMDPKHYVGEHTYVPVT 258
gi|164604|gb|AA431096.1|/1-385   213 YLSSN-DDS-GSVVLLGGIDSSYYTGSLNWVPVS 244

gi|18904|emb|CA439602.1|/1-508   259 QKGYWQFDMGD-VLVGGKSTGFCAGGCAAIADSG 291
gi|164604|gb|AA431096.1|/1-385   245 VEGYWQITL-DSITMDGETIA-CSGGCQAIVDTG 276

gi|18904|emb|CA439602.1|/1-508   292 TSLLAGPTAIITEINEKIGAAGVVSQECKTIVSQ 325
gi|164604|gb|AA431096.1|/1-385   277 TSLLTGPTSAIA--N--I--------Q------S- 292

gi|18904|emb|CA439602.1|/1-508   326 YGQQILDLLLAETQPKKICSQVGLCTFDGTRGVS 359
gi|164604|gb|AA431096.1|/1-385   293 ------D-----------I----G---------- 295

gi|18904|emb|CA439602.1|/1-508   360 AGIRSVVDDEPVKSNGLRADPMCSACEMAVVWMQ 393
gi|164604|gb|AA431096.1|/1-385   296 A---S----E--NSYG-----------EM-V---- 305

gi|18904|emb|CA439602.1|/1-508   394 NQLAQNKTQDLILDYVNQLCNRLPSPMGESAVDC 427
gi|164604|gb|AA431096.1|/1-385   306 -----------I----S--C----S----S-ID- 312

gi|18904|emb|CA439602.1|/1-508   428 GSLGSMPDIEFTIGGKKFALKPEEYILKVGEGAA 461
gi|164604|gb|AA431096.1|/1-385   313 -SL---PDIVFTINGVQYPLSPSAYILQ-DDDS- 340

gi|18904|emb|CA439602.1|/1-508   462 AQCISGFTAMDIPPPRGPLWILGDVFMGPYHTVF 495
gi|164604|gb|AA431096.1|/1-385   341 --CTSGFEGMDVPTSSGELWILGDVFIRQYYTVF 372
```

Sequence 2 ID: gi|164604|gb|AAA31096.1| Residue: LYS (25)

Annotating features in JALVIEW


Jalview 2.3
File Tools Help Window
/home/mvoorhie/Projects/Courses/PracticalBioinformatics/gui/webr...
File Edit Select View Format Colour Calculate Web Service

```
gi|18904|emb|CAA39602.1|/1-508     1 MGTRGLALALLAAVLLLQTVLPAASEAEGLVRIAL 35
gi|164604|gb|AAA31096.1|/1-385     1 ----MKWLLLLSLVVLSECLVKVP-----LVRKKS 26

gi|18904|emb|CAA39602.1|/1-508    36 KKRPIDRNSRVATGLSGGEEQPLLSGANPLRSEEE 70
gi|164604|gb|AAA31096.1|/1-385    27 LRQNLIKNGKLKDFLKTHKHNP-ASKYFPEAAALI 60

gi|18904|emb|CAA39602.1|/1-508    71 GDIVALKNYMNAQYFGEIGVGTPPQKFTVIFDTGS 105
gi|164604|gb|AAA31096.1|/1-385    61 GD-EPLENYLDTEYFGTIGIGTPAQDFTVIFDTGS 94

gi|18904|emb|CAA39602.1|/1-508   106 SNLWVPSAKCYFSIACYLHSRYKAGASSTYKKNGK 140
gi|164604|gb|AAA31096.1|/1-385    95 SNLWVPSVYCS-SLACSDHNQFNPDDSSTFEATSQ 128

gi|18904|emb|CAA39602.1|/1-508   141 PAAIQYGTGSIAGYFSEDSVTVGDLVVKDQEFIEA 175
gi|164604|gb|AAA31096.1|/1-385   129 ELSITYGTGSMTGILGYDTVQVGGISDTNQIFGLS 163

gi|18904|emb|CAA39602.1|/1-508   176 TKEPGITFLVAKFDGILGLGFKEISVGKAVPVWYK 210
gi|164604|gb|AAA31096.1|/1-385   164 ETEPGSFLYYAPFDGILGLAYPSISASGATPVFDN 198

gi|18904|emb|CAA39602.1|/1-508   211 MIEQGLVSDPVFSFWLNRHVDEGEGGEIIFGGMDP 245
gi|164604|gb|AAA31096.1|/1-385   199 LWDQGLVSQDLFSVYLSSNDDS--GSVVLLGGIDS 231

gi|18904|emb|CAA39602.1|/1-508   246 KHYVGEHTYVPVTQKGYWQFDMGDVLVGGKSTGFC 280
gi|164604|gb|AAA31096.1|/1-385   232 SYYTGSLNWVPVSVEGYWQITLDSITMDG-ETIAC 265

gi|18904|emb|CAA39602.1|/1-508   281 AGGCAAIADSGTSLLAGPTAIITEINEKIGAAGVV 315
gi|164604|gb|AAA31096.1|/1-385   266 SGGCQAIVDTGTSLLTGPTSAIANIQSDIGAS--- 297

gi|18904|emb|CAA39602.1|/1-508   316 SQECKTIVSQYGQQILDLLLAETQPKKICSQVGLC 350
gi|164604|gb|AAA31096.1|/1-385       -----------------------------------

gi|18904|emb|CAA39602.1|/1-508   351 TFDGTRGVSAGIRSVVDDEPVKSNGLRADPMCSAC 385
gi|164604|gb|AAA31096.1|/1-385       -----------------------------------

gi|18904|emb|CAA39602.1|/1-508   386 EMAVVWMQNQLAQNKTQDLILDYVNQLCNRLPSPM 420
gi|164604|gb|AAA31096.1|/1-385   298 -------------------------------ENSY 301

gi|18904|emb|CAA39602.1|/1-508   421 GESAVDCGSLGSMPDIEFTIGGKKFALKPEEYILK 455
gi|164604|gb|AAA31096.1|/1-385   302 GEMVISCSSIDSLPDIVFTINGVQYPLSPSAYILQ 336

gi|18904|emb|CAA39602.1|/1-508   456 VGEGAAAQCISGFTAMDIPPPRGPLWILGDVFMGP 490
gi|164604|gb|AAA31096.1|/1-385   337 DDD----SCTSGFEGMDVPTSSGELWILGDVFIRQ 367

gi|18904|emb|CAA39602.1|/1-508   491 YHTVFDYGKLRIGFAKAA                  508
gi|164604|gb|AAA31096.1|/1-385   368 YYTVFDRANNKVGLAPVA                  385
```

Sequence 2 ID: gi|164604|gb|AAA31096.1| Residue: ILE (79)

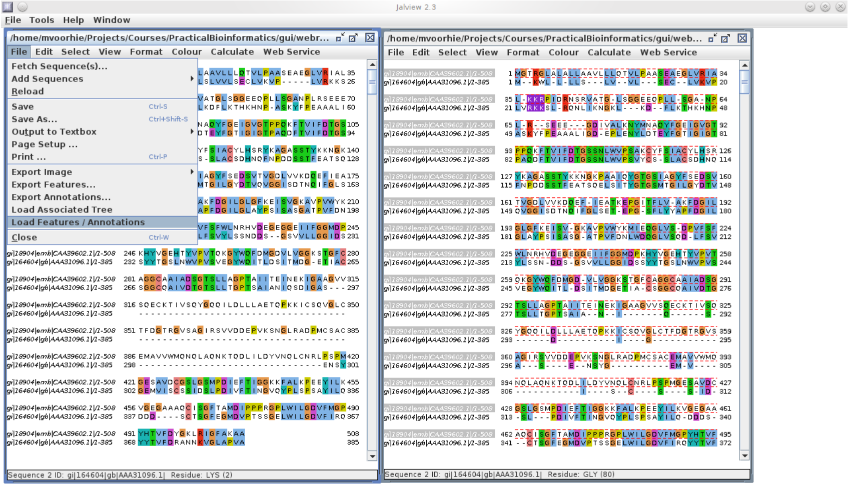
/home/mvoorhie/Projects/Courses/PracticalBioinformatics/gui/web...
File Edit Select View Format Colour Calculate Web Service

```
gi|18904|emb|CAA39602.1|/1-508     1 MGTRGLALALLAAVLLLQTVLPAASEAEGLVRIA 34
gi|164604|gb|AAA31096.1|/1-385     1 M--KWL-L-LLS---LV--VL---SEC--LVKVP 20

gi|18904|emb|CAA39602.1|/1-508    35 L-KKRPIDRNSRVATG-LSGGEEQPLL-SGA-NP 64
gi|164604|gb|AAA31096.1|/1-385    21 LVRKKSL-RQNLIKNGKL---KD--FLKTHKHNP 48

gi|18904|emb|CAA39602.1|/1-508    65 L-R--SEEE---GDIVALKNYMNAQYFGEIGVGT 92
                                              ALIGD-EPLENYLDTEYFGTIGIGT 81

                                              DTGSSNLWVPSAKCYFSIACYLHSR 126
                                              DTGSSNLWVPSVYCS-SLACSDHNQ 114

                                              KKNGKPAAIQYGTGSIAGYFSEDSV 160
                                              EATSQELSITYGTGSMTGILGYDTV 148

                                              QEF-IEATKEPGITFLV-AKFDGIL 192
                                              QIFGLSET-EPG-SFLYYAPFDGIL 180

gi|18904|emb|CAA39602.1|/1-508   193 GLGFKEISV-GKAVPVWYKMIEQGLVS-DPVFSF 224
gi|164604|gb|AAA31096.1|/1-385   181 GLAYPSISASG-ATPVFDNLWDQGLVSQD-LFSV 212

gi|18904|emb|CAA39602.1|/1-508   225 WLNRHVDEGEGGEIIFGGMDPKHYVGEHTYVPVT 258
gi|164604|gb|AAA31096.1|/1-385   213 YLSSN-DDS-GSVVLLGGIDSSYYTGSLNWVPVS 244

gi|18904|emb|CAA39602.1|/1-508   259 QKGYWQFDMGD-VLVGGKSTGFCAGGCAAIADSG 291
gi|164604|gb|AAA31096.1|/1-385   245 VEGYWQITL-DSITMDGETIA-CSGGCQAIVDTG 276

gi|18904|emb|CAA39602.1|/1-508   292 TSLLAGPTAIITEINEKIGAAGVVSQECKTIVSQ 325
gi|164604|gb|AAA31096.1|/1-385   277 TSLLTGPTSAIA--N--I-------Q------S- 292

gi|18904|emb|CAA39602.1|/1-508   326 YGQQILDLLLAETQPKKICSQVGLCTFDGTRGVS 359
gi|164604|gb|AAA31096.1|/1-385   293 ------D----------I----G----------- 295

gi|18904|emb|CAA39602.1|/1-508   360 AGIRSVVDDEPVKSNGLRADPMCSACEMAVVWMQ 393
gi|164604|gb|AAA31096.1|/1-385   296 A---S----E--NSYG----------EM-V---- 305

gi|18904|emb|CAA39602.1|/1-508   394 NQLAQNKTQDLILDYVNQLCNRLPSPMGESAVDC 427
gi|164604|gb|AAA31096.1|/1-385   306 -----------I----S--C----S----S-ID- 312

gi|18904|emb|CAA39602.1|/1-508   428 GSLGSMPDIEFTIGGKKFALKPEEYILKVGEGAA 461
gi|164604|gb|AAA31096.1|/1-385   313 -SL---PDIVFTINGVQYPLSPSAYILQ-DDDS- 340

gi|18904|emb|CAA39602.1|/1-508   462 AQCISGFTAMDIPPPRGPLWILGDVFMGPYHTVF 495
gi|164604|gb|AAA31096.1|/1-385   341 --CTSGFEGMDVPTSSGELWILGDVFIRQYYTVF 372
```

Sequence 2 ID: gi|164604|gb|AAA31096.1| Residue: PHE (210)

Format: ○ Jalview ● GFF

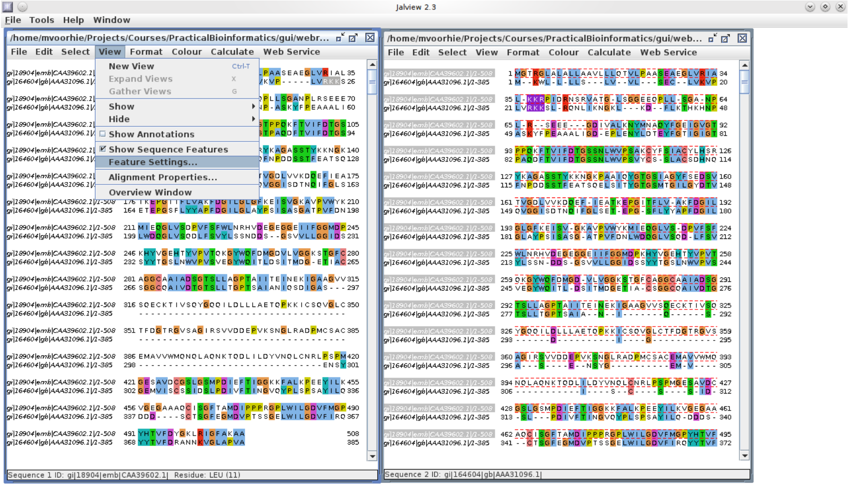
Annotating features in JALVIEW


Jalview 2.3
File Tools Help Window
/home/mvoorhie/Projects/Courses/PracticalBioinformatics/gui/webr...
File Edit Select View Format Colour Calculate Web Service
Fetch Sequence(s)...
Add Sequences
Reload
Save Ctrl-S
Save As... Ctrl+Shift-S
Output to Textbox
Page Setup ...
Print ... Ctrl-P
Export Image
Export Features...
Export Annotations...
Load Associated Tree
Load Features / Annotations
Close Ctrl-W
/home/mvoorhie/Projects/Courses/PracticalBioinformatics/gui/web...
File Edit Select View Format Colour Calculate Web Service

```
LAAVLLLQTVLPAASEAEGLVRIAL35
LSLVVLSECLVKVP-----LVRKKS26

VATGLSGGEEQPLLSGANPLRSEEE70
LKDFLKTHKHNP-ASKYFPEAAALI60

NAQYFGEIGVGTPPQKFTVIFDTGS105
DTEYFGTIGIGTPAQDFTVIFDTGS94

YFSIACYLHSRYKAGASSTYKKNGK140
S-SLACSDHNQFNPDDSSTFEATSQ128

IAGYFSEDSVTVGDLVVKDQEFIEA175
MTGILGYDTVQVGGISDTNQIFGLS163

AKFDGILGLGFKEISVGKAVPVWYK210
APFDGILGLAYPSISASGATPVFDN198

VFSFWLNRHVDEGEGGEIIFGGMDP245
LFSVYLSSNDDS--GSVVLLGGIDS231

gi|18904|emb|CA439602.1|/1-508   246 KHYVGEHTYVPVTQKGYWQFDMGDVLVGGKSTGFC280
gi|164604|gb|AA431096.1|/1-385   232 SYYTGSLNWVPVSVEGYWQITLDSITMDG-ETIAC265

gi|18904|emb|CA439602.1|/1-508   281 AGGCAAIADSGTSLLAGPTAIITEINEKIGAAGVV315
gi|164604|gb|AA431096.1|/1-385   266 SGGCQAIVDTGTSLLTGPTSAIANIQSDIGAS---297

gi|18904|emb|CA439602.1|/1-508   316 SQECKTIVSQYGQQILDLLLAETQPKKICSQVGLC350
gi|164604|gb|AA431096.1|/1-385       -----------------------------------

gi|18904|emb|CA439602.1|/1-508   351 TFDGTRGVSAGIRSVVDDEPVKSNGLRADPMCSAC385
gi|164604|gb|AA431096.1|/1-385       -----------------------------------

gi|18904|emb|CA439602.1|/1-508   386 EMAVVWMQNQLAQNKTQDLILDYVNQLCNRLPSPM420
gi|164604|gb|AA431096.1|/1-385   298 -------------------------------ENSY301

gi|18904|emb|CA439602.1|/1-508   421 GESAVDCGSLGSMPDIEFTIGGKKFALKPEEYILK455
gi|164604|gb|AA431096.1|/1-385   302 GEMVISCSSIDSLPDIVFTINGVQYPLSPSAYILQ336

gi|18904|emb|CA439602.1|/1-508   456 VGEGAAAQCISGFTAMDIPPPRGPLWILGDVFMGP490
gi|164604|gb|AA431096.1|/1-385   337 DDD----SCTSGFEGMDVPTSSGELWILGDVFIRQ367

gi|18904|emb|CA439602.1|/1-508   491 YHTVFDYGKLRIGFAKAA                 508
gi|164604|gb|AA431096.1|/1-385   368 YYTVFDRANNKVGLAPVA                 385

Sequence 2 ID: gi|164604|gb|AAA31096.1|   Residue: LYS (2)
```

```
gi|18904|emb|CA439602.1|/1-508     1 MGTRGLALALLAAVLLLQTVLPAASEAEGLVRIA 34
gi|164604|gb|AA431096.1|/1-385     1 M--KWL-L-LLS---LV--VL---SEC--LVKVP 20

gi|18904|emb|CA439602.1|/1-508    35 L-KKRPIDRNSRVATG-LSGGEEQPLL-SGA-NP 64
gi|164604|gb|AA431096.1|/1-385    21 LVRKKSL-RQNLIKNGKL---KD--FLKTHKHNP 48

gi|18904|emb|CA439602.1|/1-508    65 L-R--SEEE---GDIVALKNYMNAQYFGEIGVGT 92
gi|164604|gb|AA431096.1|/1-385    49 ASKYFPEAAALIGD-EPLENYLDTEYFGTIGIGT 81

gi|18904|emb|CA439602.1|/1-508    93 PPQKFTVIFDTGSSNLWVPSAKCYFSIACYLHSR 126
gi|164604|gb|AA431096.1|/1-385    82 PAQDFTVIFDTGSSNLWVPSVYCS-SLACSDHNQ 114

gi|18904|emb|CA439602.1|/1-508   127 YKAGASSTYKKNGKPAAIQYGTGSIAGYFSEDSV 160
gi|164604|gb|AA431096.1|/1-385   115 FNPDDSSTFEATSQELSITYGTGSMTGILGYDTV 148

gi|18904|emb|CA439602.1|/1-508   161 TVGDLVVKDQEF-IEATKEPGITFLV-AKFDGIL 192
gi|164604|gb|AA431096.1|/1-385   149 QVGGISDTNQIFGLSET-EPG-SFLYYAPFDGIL 180

gi|18904|emb|CA439602.1|/1-508   193 GLGFKEISV-GKAVPVWYKMIEQGLVS-DPVFSF 224
gi|164604|gb|AA431096.1|/1-385   181 GLAYPSISASG-ATPVFDNLWDQGLVSQD-LFSV 212

gi|18904|emb|CA439602.1|/1-508   225 WLNRHVDEGEGGEIIFGGMDPKHYVGEHTYVPVT 258
gi|164604|gb|AA431096.1|/1-385   213 YLSSN-DDS-GSVVLLGGIDSSYYTGSLNWVPVS 244

gi|18904|emb|CA439602.1|/1-508   259 QKGYWQFDMGD-VLVGGKSTGFCAGGCAAIADSG 291
gi|164604|gb|AA431096.1|/1-385   245 VEGYWQITL-DSITMDGETIA-CSGGCQAIVDTG 276

gi|18904|emb|CA439602.1|/1-508   292 TSLLAGPTAIITEINEKIGAAGVVSQECKTIVSQ 325
gi|164604|gb|AA431096.1|/1-385   277 TSLLTGPTSAIA--N--I---------Q-------S- 292

gi|18904|emb|CA439602.1|/1-508   326 YGQQILDLLLAETQPKKICSQVGLCTFDGTRGVS 359
gi|164604|gb|AA431096.1|/1-385   293 ------D-----------I-----G------------ 295

gi|18904|emb|CA439602.1|/1-508   360 AGIRSVVDDEPVKSNGLRADPMCSACEMAVVWMQ 393
gi|164604|gb|AA431096.1|/1-385   296 A---S-----E--NSYG------------EM-V----- 305

gi|18904|emb|CA439602.1|/1-508   394 NQLAQNKTQDLILDYVNQLCNRLPSPMGESAVDC 427
gi|164604|gb|AA431096.1|/1-385   306 -------------I-----S--C-----S-----S-ID- 312

gi|18904|emb|CA439602.1|/1-508   428 GSLGSMPDIEFTIGGKKFALKPEEYILKVGEGAA 461
gi|164604|gb|AA431096.1|/1-385   313 -SL---PDIVFTINGVQYPLSPSAYILQ-DDDS- 340

gi|18904|emb|CA439602.1|/1-508   462 AQCISGFTAMDIPPPRGPLWILGDVFMGPYHTVF 495
gi|164604|gb|AA431096.1|/1-385   341 --CTSGFEGMDVPTSSGELWILGDVFIRQYYTVF 372

Sequence 2 ID: gi|164604|gb|AAA31096.1|   Residue: GLY (80)
```

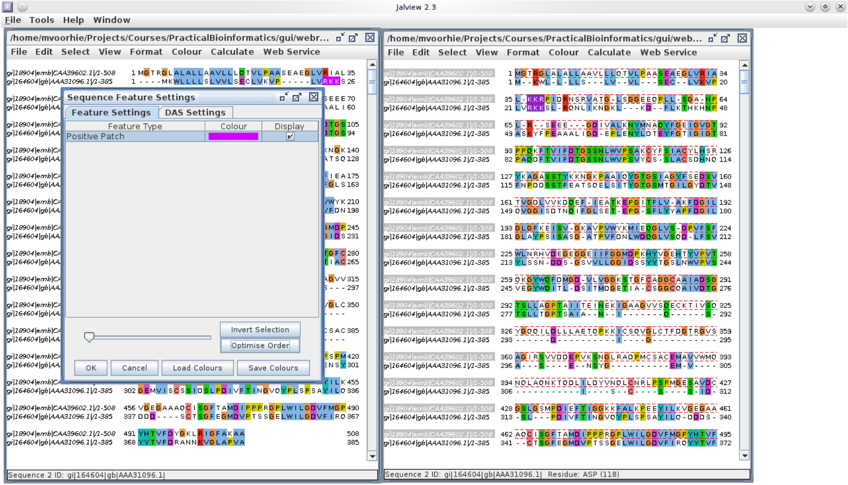
Annotating features in JALVIEW


Jalview 2.3
File Tools Help Window
/home/mvoorhie/Projects/Courses/PracticalBioinformatics/gui/webr...
File Edit Select View Format Colour Calculate Web Service
New View Ctrl-T
Expand Views X
Gather Views G
Show
Hide
□ Show Annotations
☑ Show Sequence Features
Feature Settings...
Alignment Properties...
Overview Window
gi|18904|emb|CAA39602.1|/1-508
gi|164604|gb|AAA31096.1|/1-385
/home/mvoorhie/Projects/Courses/PracticalBioinformatics/gui/web...
File Edit Select View Format Colour Calculate Web Service

Sequence 1 ID: gi|18904|emb|CAA39602.1| Residue: LEU (11)

Sequence 2 ID: gi|164604|gb|AAA31096.1|

Annotating features in JALVIEW


Jalview 2.3
File Tools Help Window
/home/mvoorhie/Projects/Courses/PracticalBioinformatics/gui/webr...
/home/mvoorhie/Projects/Courses/PracticalBioinformatics/gui/web...
File Edit Select View Format Colour Calculate Web Service
Sequence Feature Settings
Feature Settings
DAS Settings
Invert Selection
Optimise Order
OK
Cancel
Load Colours
Save Colours

Feature Type	Colour	Display
Positive Patch	🟪	☑

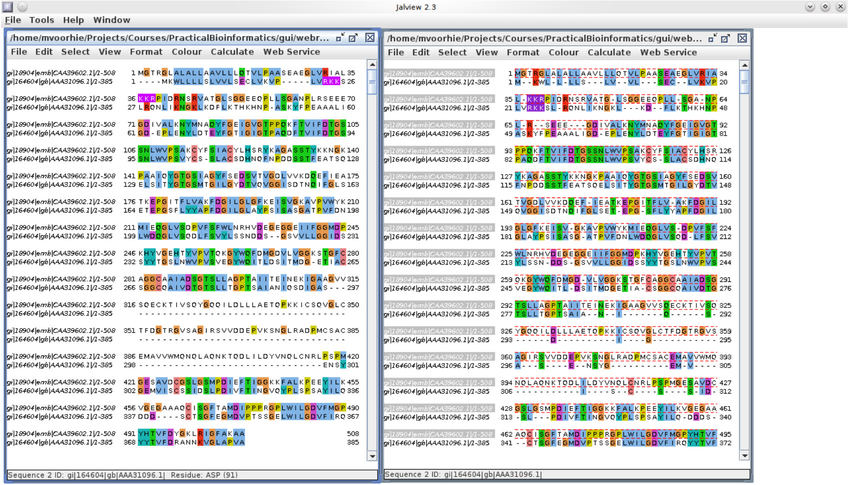
gi|18904|emb|CA439602.1|/1-508

gi|164604|gb|AAA31096.1|/1-385

Sequence 2 ID: gi|164604|gb|AAA31096.1|

Sequence 2 ID: gi|164604|gb|AAA31096.1| Residue: ASP (118)

Annotating features in JALVIEW



Jalview 2.3
File Tools Help Window
/home/mvoorhie/Projects/Courses/PracticalBioinformatics/gui/webr...
File Edit Select View Format Colour Calculate Web Service
/home/mvoorhie/Projects/Courses/PracticalBioinformatics/gui/web...
File Edit Select View Format Colour Calculate Web Service

```
gi|18904|emb|CAA39602.1|/1-508    1 MGTRGLALALLAAVLLLQTVLPAASEAEGLVRIAL 35
gi|164604|gb|AAA31096.1|/1-385    1 ----MKWLLLLSLVVLSECLVKVP-----LVRKKS 26

gi|18904|emb|CAA39602.1|/1-508   36 KKRPIDRNSRVATGLSGGEEQPLLSGANPLRSEEE 70
gi|164604|gb|AAA31096.1|/1-385   27 LRQNLIKNGKLKDFLKTHKHNP-ASKYFPEAAALI 60

gi|18904|emb|CAA39602.1|/1-508   71 GDIVALKNYMNAQYFGEIGVGTPPQKFTVIFDTGS 105
gi|164604|gb|AAA31096.1|/1-385   61 GD-EPLENYLDTEYFGTIGIGTPAQDFTVIFDTGS 94

gi|18904|emb|CAA39602.1|/1-508  106 SNLWVPSAKCYFSIACYLHSRYKAGASSTYKKNGK 140
gi|164604|gb|AAA31096.1|/1-385   95 SNLWVPSVYCS-SLACSDHNQFNPDDSSTFEATSQ 128

gi|18904|emb|CAA39602.1|/1-508  141 PAAIQYGTGSIAGYFSEDSVTVGDLVVKDQEFIEA 175
gi|164604|gb|AAA31096.1|/1-385  129 ELSITYGTGSMTGILGYDTVQVGGISDTNQIFGLS 163

gi|18904|emb|CAA39602.1|/1-508  176 TKEPGITFLVAKFDGILGLGFKEISVGKAVPVWYK 210
gi|164604|gb|AAA31096.1|/1-385  164 ETEPGSFLYYAPFDGILGLAYPSISASGATPVFDN 198

gi|18904|emb|CAA39602.1|/1-508  211 MIEQGLVSDPVFSFWLNRHVDEGEGGEIIFGGMDP 245
gi|164604|gb|AAA31096.1|/1-385  199 LWDQGLVSQDLFSVYLSSNDDS--GSVVLLGGIDS 231

gi|18904|emb|CAA39602.1|/1-508  246 KHYVGEHTYVPVTQKGYWQFDMGDVLVGGKSTGFC 280
gi|164604|gb|AAA31096.1|/1-385  232 SYYTGSLNWVPVSVEGYWQITLDSITMDG-ETIAC 265

gi|18904|emb|CAA39602.1|/1-508  281 AGGCAAIADSGTSLLAGPTAIITEINEKIGAAGVV 315
gi|164604|gb|AAA31096.1|/1-385  266 SGGCQAIVDTGTSLLTGPTSAIANIQSDIGAS--- 297

gi|18904|emb|CAA39602.1|/1-508  316 SQECKTIVSQYGQQILDLLLAETQPKKICSQVGLC 350
gi|164604|gb|AAA31096.1|/1-385      -----------------------------------

gi|18904|emb|CAA39602.1|/1-508  351 TFDGTRGVSAGIRSVVDDEPVKSNGLRADPMCSAC 385
gi|164604|gb|AAA31096.1|/1-385      -----------------------------------

gi|18904|emb|CAA39602.1|/1-508  386 EMAVVWMQNQLAQNKTQDLILDYVNQLCNRLPSPM 420
gi|164604|gb|AAA31096.1|/1-385  298 -------------------------------ENSY 301

gi|18904|emb|CAA39602.1|/1-508  421 GESAVDCGSLGSMPDIEFTIGGKKFALKPEEYILK 455
gi|164604|gb|AAA31096.1|/1-385  302 GEMVISCSSIDSLPDIVFTINGVQYPLSPSAYILQ 336

gi|18904|emb|CAA39602.1|/1-508  456 VGEGAAAQCISGFTAMDIPPPRGPLWILGDVFMGP 490
gi|164604|gb|AAA31096.1|/1-385  337 DDD----SCTSGFEGMDVPTSSGELWILGDVFIRQ 367

gi|18904|emb|CAA39602.1|/1-508  491 YHTVFDYGKLRIGFAKAA                  508
gi|164604|gb|AAA31096.1|/1-385  368 YYTVFDRANNKVGLAPVA                  385
```

Sequence 2 ID: gi|164604|gb|AAA31096.1| Residue: ASP (91)

```
gi|18904|emb|CAA39602.1|/1-508    1 MGTRGLALALLAAVLLLQTVLPAASEAEGLVRIA 34
gi|164604|gb|AAA31096.1|/1-385    1 M--KWL-L-LLS---LV--VL---SEC--LVKVP 20

gi|18904|emb|CAA39602.1|/1-508   35 L-KKRPIDRNSRVATG-LSGGEEQPLL-SGA-NP 64
gi|164604|gb|AAA31096.1|/1-385   21 LVRKKSL-RQNLIKNGKL---KD--FLKTHKHNP 48

gi|18904|emb|CAA39602.1|/1-508   65 L-R--SEEE---GDIVALKNYMNAQYFGEIGVGT 92
gi|164604|gb|AAA31096.1|/1-385   49 ASKYFPEAAALIGD-EPLENYLDTEYFGTIGIGT 81

gi|18904|emb|CAA39602.1|/1-508   93 PPQKFTVIFDTGSSNLWVPSAKCYFSIACYLHSR 126
gi|164604|gb|AAA31096.1|/1-385   82 PAQDFTVIFDTGSSNLWVPSVYCS-SLACSDHNQ 114

gi|18904|emb|CAA39602.1|/1-508  127 YKAGASSTYKKNGKPAAIQYGTGSIAGYFSEDSV 160
gi|164604|gb|AAA31096.1|/1-385  115 FNPDDSSTFEATSQELSITYGTGSMTGILGYDTV 148

gi|18904|emb|CAA39602.1|/1-508  161 TVGDLVVKDQEF-IEATKEPGITFLV-AKFDGIL 192
gi|164604|gb|AAA31096.1|/1-385  149 QVGGISDTNQIFGLSET-EPG-SFLYYAPFDGIL 180

gi|18904|emb|CAA39602.1|/1-508  193 GLGFKEISV-GKAVPVWYKMIEQGLVS-DPVFSF 224
gi|164604|gb|AAA31096.1|/1-385  181 GLAYPSISASG-ATPVFDNLWDQGLVSQD-LFSV 212

gi|18904|emb|CAA39602.1|/1-508  225 WLNRHVDEGEGGEIIFGGMDPKHYVGEHTYVPVT 258
gi|164604|gb|AAA31096.1|/1-385  213 YLSSN-DDS-GSVVLLGGIDSSYYTGSLNWVPVS 244

gi|18904|emb|CAA39602.1|/1-508  259 QKGYWQFDMGD-VLVGGKSTGFCAGGCAAIADSG 291
gi|164604|gb|AAA31096.1|/1-385  245 VEGYWQITL-DSITMDGETIA-CSGGCQAIVDTG 276

gi|18904|emb|CAA39602.1|/1-508  292 TSLLAGPTAIITEINEKIGAAGVVSQECKTIVSQ 325
gi|164604|gb|AAA31096.1|/1-385  277 TSLLTGPTSAIA--N--I-------Q------S- 292

gi|18904|emb|CAA39602.1|/1-508  326 YGQQILDLLLAETQPKKICSQVGLCTFDGTRGVS 359
gi|164604|gb|AAA31096.1|/1-385  293 ------D----------I----G----------- 295

gi|18904|emb|CAA39602.1|/1-508  360 AGIRSVVDDEPVKSNGLRADPMCSACEMAVVWMQ 393
gi|164604|gb|AAA31096.1|/1-385  296 A---S----E--NSYG----------EM-V---- 305

gi|18904|emb|CAA39602.1|/1-508  394 NQLAQNKTQDLILDYVNQLCNRLPSPMGESAVDC 427
gi|164604|gb|AAA31096.1|/1-385  306 -----------I----S--C----S----S-ID- 312

gi|18904|emb|CAA39602.1|/1-508  428 GSLGSMPDIEFTIGGKKFALKPEEYILKVGEGAA 461
gi|164604|gb|AAA31096.1|/1-385  313 -SL---PDIVFTINGVQYPLSPSAYILQ-DDDS- 340

gi|18904|emb|CAA39602.1|/1-508  462 AQCISGFTAMDIPPPRGPLWILGDVFMGPYHTVF 495
gi|164604|gb|AAA31096.1|/1-385  341 --CTSGFEGMDVPTSSGELWILGDVFIRQYYTVF 372
```

Sequence 2 ID: gi|164604|gb|AAA31096.1|

Comparing files on *NIX

```
# List all differences between two text files
# (empty output for identity)
diff HvSs.gap0.0.both.aln HvSs.gap0.0.mult.aln
# Report only whether the files differ
# (empty output for identity)
diff -q HvSs.gap0.0.both.aln HvSs.gap0.0.mult.aln
```

(*NIX = *BSD, OS X, Solaris, Linux, Windows with Cygwin, ...)

Homework

- Use a text or sequence editor to create a spliced variant of HvPhytepsin that can be aligned to the full HsSaposinC sequence
- Find the GenBank entries for HvPhytepsin and SsPepsinogen (tip: use the identifiers from the FASTA files) and find the corresponding transcript sequences.
  - How easy is it to align the proteins vs. the transcripts?
  - Can you tell if you are getting equivalent results from the two alignments?
- Try repeating this exercise for a pair of sequences where genomic sequence is available; *e.g.*, *A. nidulans* VosA (ABQ18268.1), and *H. capsulatum* Ryp2 (ACB59236.1).