Identifying and Aligning Homologs

Fran Lewitter

Bioinformatics and Research Computing @ Whitehead Institute

6/6/12

Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor.

Doolittle RF, Hunkapiller MW, Hood LE, Devare SG, Robbins KC, Aaronson SA, Antoniades HN. *Science* 221:275-277, 1983.

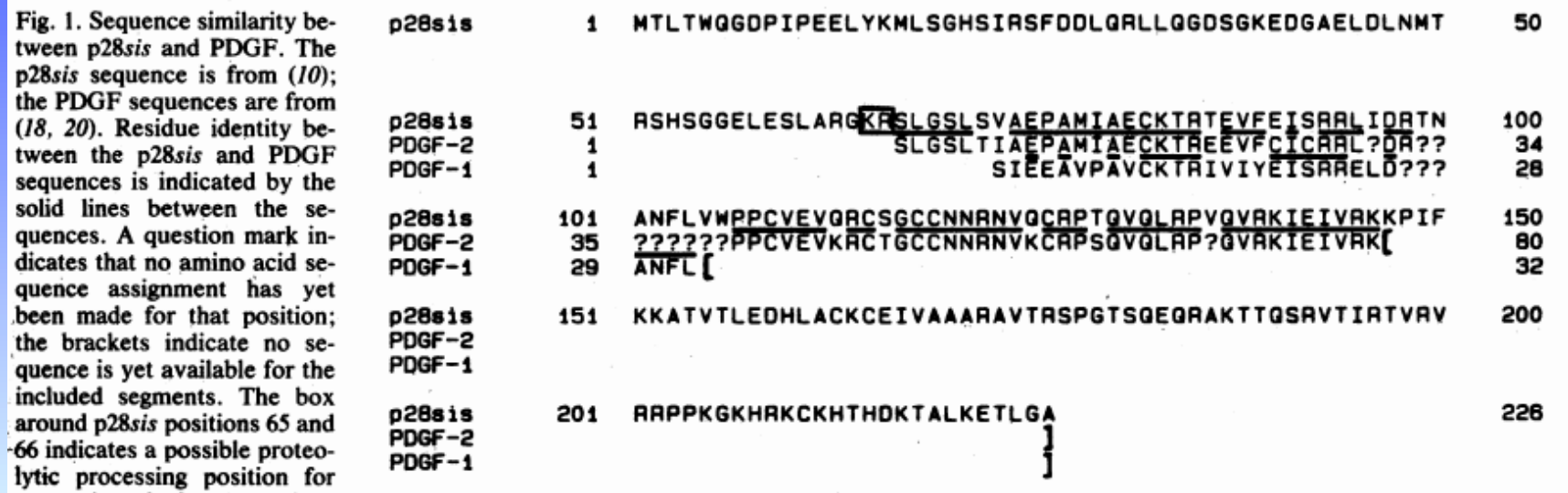
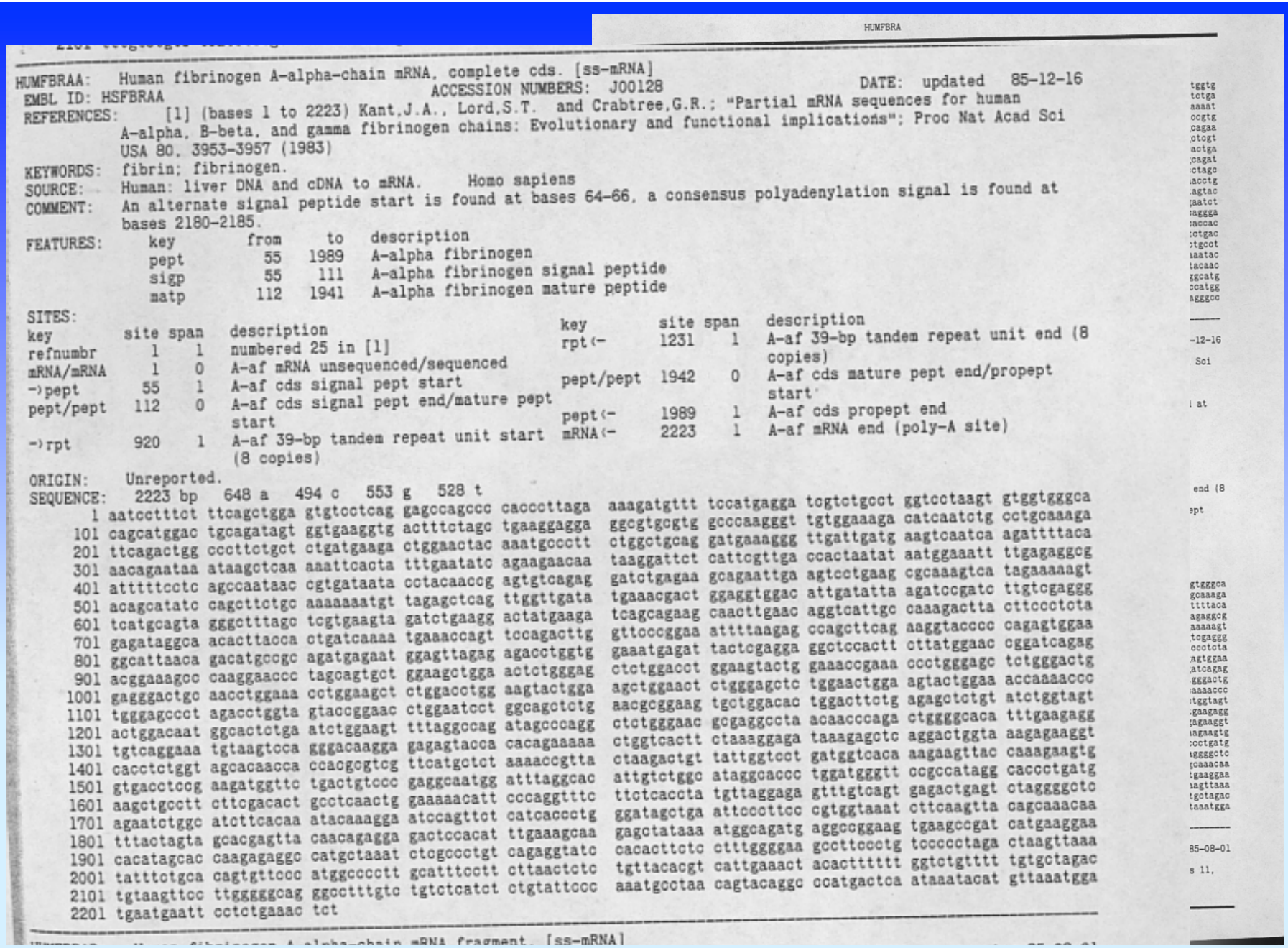


Fig. 1. Sequence similarity between p28*sis* and PDGF. The p28*sis* sequence is from (*10*); the PDGF sequences are from (*18, 20*). Residue identity between the p28*sis* and PDGF sequences is indicated by the solid lines between the sequences. A question mark indicates that no amino acid sequence assignment has yet been made for that position; the brackets indicate no sequence is yet available for the included segments. The box around p28*sis* positions 65 and 66 indicates a possible proteolytic processing position for generation of a fragment of p28*sis* corresponding to PDGF-2. Single letter abbreviations for the amino acid residues are as follows: A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine.

Identifying and Aligning Homologs (Whitehead Institute)

What I hope you'll learn

- What we can learn from sequence alignments
- Fundamentals of alignments
- Tools for building alignments



```
HUMFBRA

HUMFBRAA:   Human fibrinogen A-alpha-chain mRNA, complete cds. [ss-mRNA]
EMBL ID: HSFBRAA                          ACCESSION NUMBERS:  J00128          DATE:  updated   85-12-16
REFERENCES:     [1] (bases 1 to 2223) Kant,J.A., Lord,S.T.  and Crabtree,G.R.; "Partial mRNA sequences for human
          A-alpha, B-beta, and gamma fibrinogen chains: Evolutionary and functional implications"; Proc Nat Acad Sci
          USA 80, 3953-3957 (1983)
KEYWORDS:   fibrin; fibrinogen.
SOURCE:     Human: liver DNA and cDNA to mRNA.     Homo sapiens
COMMENT:    An alternate signal peptide start is found at bases 64-66, a consensus polyadenylation signal is found at
          bases 2180-2185.
FEATURES:       key         from     to   description
                pept          55   1989   A-alpha fibrinogen
                sigp          55    111   A-alpha fibrinogen signal peptide
                matp         112   1941   A-alpha fibrinogen mature peptide

SITES:
key         site span   description                                  key         site span   description
refnumbr       1    1   numbered 25 in [1]                          rpt(-       1231    1   A-af 39-bp tandem repeat unit end (8
mRNA/mRNA      1    0   A-af mRNA unsequenced/sequenced                                      copies)
-)pept        55    1   A-af cds signal pept start                  pept/pept   1942    0   A-af cds mature pept end/propept
pept/pept    112    0   A-af cds signal pept end/mature pept                                 start
                        start                                       pept(-      1989    1   A-af cds propept end
-)rpt         920    1   A-af 39-bp tandem repeat unit start         mRNA(-      2223    1   A-af mRNA end (poly-A site)
                        (8 copies)

ORIGIN:     Unreported.
SEQUENCE:   2223 bp   648 a   494 c   553 g   528 t
       1 aatcctttct ttcagctgga gtgtcctcag gagccagccc caccettaga  aaagatgttt tccatgagga tcgtctgcct ggtcctaagt gtggtgggca
     101 cagcatggac tgcagatagt ggtgaaggtg actttctagc tgaaggagga  ggcgtgcgtg gcccaagggt tgtggaaaga catcaatctg cctgcaaaga
     201 ttcagactgg cccttctgct ctgatgaaga ctggaactac aaatgccctt  ctggctgcag gatgaaaggg ttgattgatg aagtcaatca agattttaca
     301 aacagaataa ataagctcaa aaattcacta tttgaatatc agaagaacaa  taaggattct cattcgttga ccactaatat aatggaaatt ttgagaggcg
     401 atttttcctc agccaataac cgtgataata cctacaaccg agtgtcagag  gatctgagaa gcagaattga agtcctgaag cgcaaagtca tagaaaaagt
     501 acagcatatc cagcttctgc aaaaaaatgt tagagctcag ttggttgata  tgaaacgact ggaggtggac attgatatta agatccgatc ttgtcgaggg
     601 tcatgcagta gggctttagc tcgtgaagta gatctgaagg actatgaaga  tcagcagaag caacttgaac aggtcattgc caaagactta cttccctcta
     701 gagataggca acacttacca ctgatcaaaa tgaaaccagt tccagacttg  gttcccggaa attttaagag ccagcttcag aaggtacccc cagagtggaa
     801 ggcattaaca gacatgccgc agatgagaat ggagttagag agacctggtg  gaaatgagat tactcgagga ggctccactt cttatggaac cggatcagag
     901 acggaaagcc caaggaaccc tagcagtgct ggaagctgga actctgggag  ctctggacct ggaagtactg gaaaccgaaa ccctgggagc tctgggactg
    1001 gagggactgc aacctggaaa cctggaagct ctggacctgg aagtactgga  agctggaact ctgggagctc tggaactgga agtactggaa accaaaaccc
    1101 tgggagccct agacctggta gtaccggaac ctggaatcct ggcagctctg  aacgcggaag tgctggacac tggacttctg agagctctgt atctggtagt
    1201 actggacaat ggcactctga atctggaagt tttaggccag atagcccagg  ctctgggaac gcgaggccta acaacccaga ctggggcaca tttgaagagg
    1301 tgtcaggaaa tgtaagtcca gggacaagga gagagtacca cacagaaaaa  ctggtcactt ctaaaggaga taaagagctc aggactggta aagagaaggt
    1401 cacctctggt agcacaacca ccacgcgtcg ttcatgctct aaaaccgtta  ctaagactgt tattggtcct gatggtcaca aagaagttac caaagaagtg
    1501 gtgacctccg aagatggttc tgactgtccc gaggcaatgg atttaggcac  attgtctggc ataggcaccc tggatgggtt ccgccatagg caccctgatg
    1601 aagctgcctt cttcgacact gcctcaactg gaaaaacatt cccaggtttc  ttctcaccta tgttaggaga gtttgtcagt gagactgagt ctaggggctc
    1701 agaatctggc atcttcacaa atacaaagga atccagttct catcaccctg  ggatagctga attcccttcc cgtggtaaat cttcaagtta cagcaaacaa
    1801 tttactagta gcacgagtta caacagagga gactccacat ttgaaagcaa  gagctataaa atggcagatg aggccggaag tgaagccgat catgaaggaa
    1901 cacatagcac caagagaggc catgctaaat ctcgccctgt cagaggtatc  cacacttctc ctttggggaa gccttcccctg tccccctaga ctaagttaaa
    2001 tatttctgca cagtgttccc atggcccctt gcatttcctt cttaactctc  tgttacacgt cattgaaact acacttttt ggtctgtttt tgtgctagac
    2101 tgtaagttcc ttgggggcag ggcctttgtc tgtctcatct ctgtattccc  aaatgcctaa cagtacaggc ccatgactca ataaatacat gttaaatgga
    2201 tgaatgaatt cctctgaaac tct

HUMFBRA   Human fibrinogen A-alpha-chain mRNA fragment. [ss-mRNA]
```

Topics to Cover

- Introduction
  - Why do alignments?
  - Definitions
  - Scoring alignments
- Pairwise Alignment methods
- Multiple sequence alignments
- Pre-computed alignment resources

Why do alignments

- Use sequence similarity to infer homology and/or structural similarity between 2 or more genes/ proteins
- Identify more conserved regions of a protein, potentially identifying regions of most functional importance
- Compare and contrast homologs (perhaps into groups) based on shared positions or regions
- Infer evolutionary distance from sequence dissimilarity

Evolutionary Basis of Sequence Alignment

- ***Similarity*** - observable quantity, such as percent identity
- ***Homology*** - conclusion drawn from data that two genes share a common evolutionary history; no metric is associated with this
  - ***Paralog*** – genes related by duplication
  - ***Ortholog*** – genes related by speciation

More Definitions

- An ***alignment*** is a mutual arrangement of two sequences, which exhibits where the two sequences are similar, and where they differ.
- An ***optimal alignment*** is one that exhibits the most correspondences and the least differences. It is the alignment with the highest score. May or may not be biologically meaningful.

Alignment Concepts

- ***Global alignment*** - Needleman-Wunsch (1970) maximizes the number of matches between the sequences along the entire length of the sequences.
- ***Local alignment*** - Smith-Waterman (1981) produces the highest scoring regional match between two sequences.
- ***Insertion and Deletions*** (indels)
- ***Affine gap costs*** - a scoring system for gaps within alignments that charges a penalty for the existence of a gap and an additional per-residue penalty proportional to the gap's length

Global vs Local Alignment

GLOBAL

```
L G P S S K Q T G K G S — S R I W D N
L N — I T K S A G K G A I M R L G D A
```

LOCAL

```
- - - - - - - T G K G - - - - - - - -
- - - - - - - A G K G - - - - - - - -
```

From Mount, *Bioinformatics, 2004, pg 71*

Possible Alignments

```
A:  T C A G A C G A G T G
B:  T C G G A G C T G

I.    T C A G A C G A G T G
      T C G G A - - G C T G

II.   T C A G A C G A G T G
      T C G G A - G C - T G

III.  T C A G A C G A G T G
      T C G G A - G - C T G
```

Identifying and Aligning Homologs (Whitehead Institute)

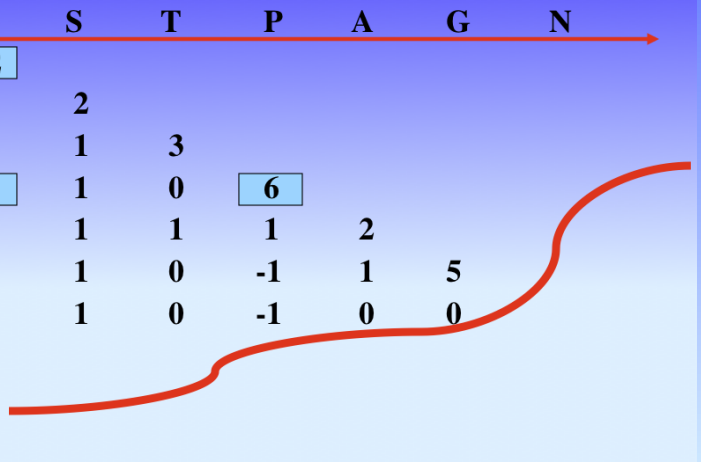
Nucleotide vs Protein

- If comparing protein coding genes, use protein sequences because of less noise
- If protein sequences are very similar, it might be more instructive to use DNA sequences
- If interested in DNA alignment of coding sequences, first do a protein alignment and use it as a template for aligning DNA sequences

AA Scoring Matrices

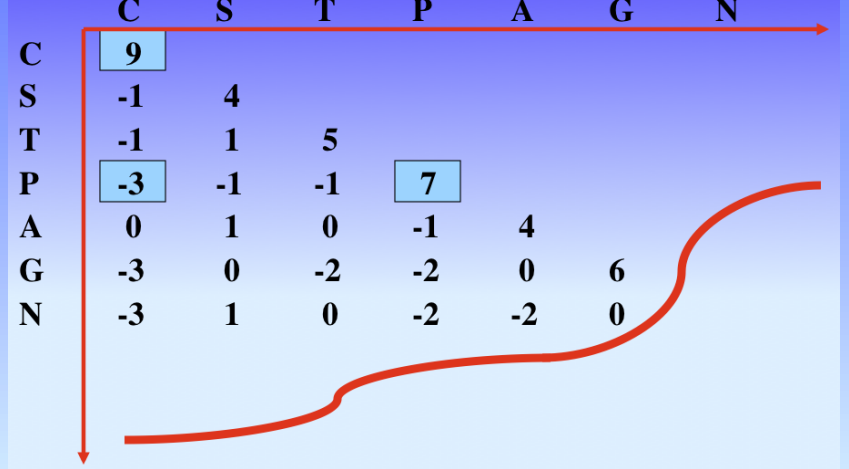
Part of PAM 250 Matrix

	C	S	T	P	A	G	N
C	12						
S	0	2					
T	-2	1	3				
P	-3	1	0	6			
A	-2	1	1	1	2		
G	-3	1	0	-1	1	5	
N	-4	1	0	-1	0	0	



Part of BLOSUM 62 Matrix

	C	S	T	P	A	G	N
C	9						
S	-1	4					
T	-1	1	5				
P	-3	-1	-1	7			
A	0	1	0	-1	4		
G	-3	0	-2	-2	0	6	
N	-3	1	0	-2	-2	0	



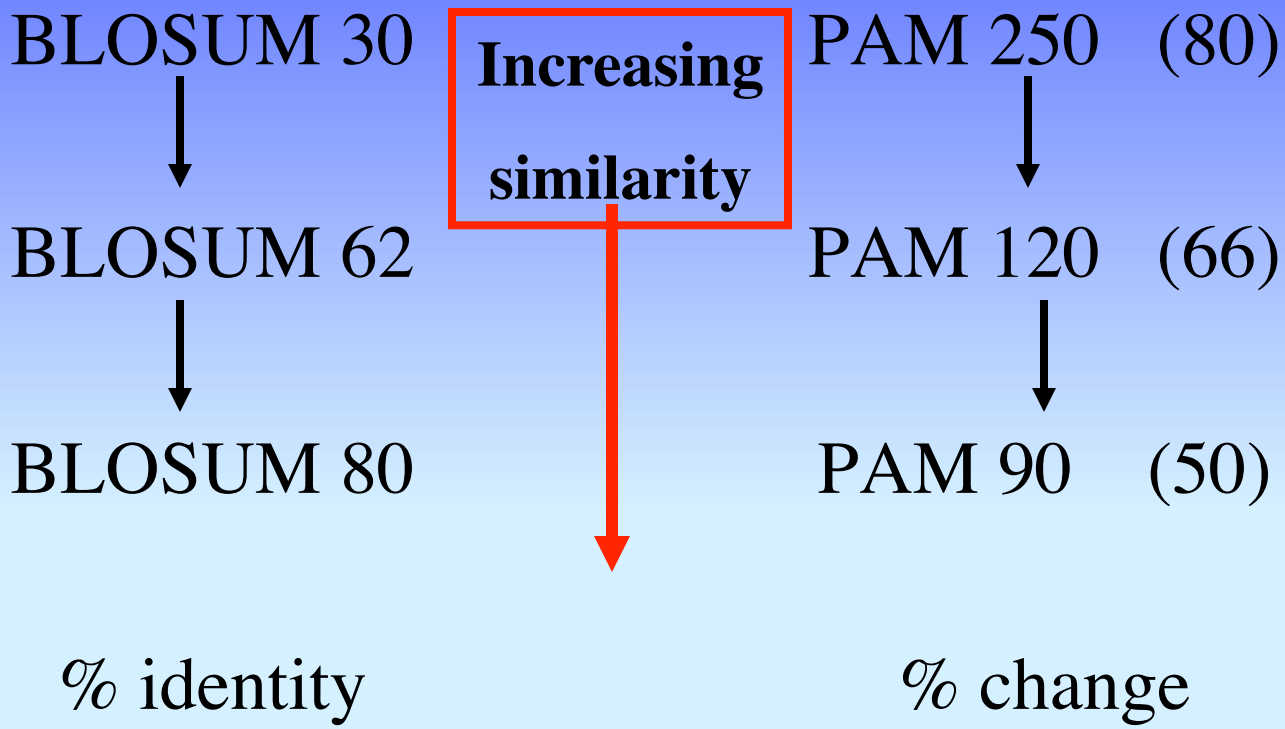
- ***PAM*** **-** point accepted mutation based on *global* alignment [evolutionary model]

- ***BLOSUM*** - block substitutions based on *local* alignments [similarity among conserved sequences]

Log-odds= $\frac{\text{pair in homologous proteins}}{\text{pair in unrelated proteins by chance}}$

Log-odds= $\frac{\text{obs freq of aa substitutions}}{\text{freq expected by chance}}$

Substitution Matrices



Scoring for BLAST Alignments

```
Score = 94.0 bits (230), Expect = 6e-19
Identities = 45/101 (44%), Positives = 54/101 (52%), Gaps = 7/101 (6%)

Query: 204 YTGPFCDV----DTKASCYDGRGLSYRGLARTTLSGAPCQPWASEATYRNVTAEQ---AR 256
           Y+  FC      +  + CY G G +YRG    T SGA C PW S      V   Q   A+
Sbjct: 198 YSSEFCSTPACSEGNSDCYFGNGSAYRGTHSLTESGASCLPWNSMILIGKVYTAQNPSAQ 257

Query: 257 NWGLGGHAFCRNPDNDIRPWCFVLNRDRLSWEYCDLAQCQT 297
             GLG H +CRNPD D +PWC VL   RL+WEYCD+  C T
Sbjct: 258 ALGLGKHNYCRNPDGDAKPWCHVLKNRRLTWEYCDVPSCST 298
```

Based on BLOSUM62

```
Position  1: Y - Y =   7
Position  2: T - S =   1
Position  3: G - S =   0
Position  4: P - E =  -1
        . . .
Position  9: - - P = -11
Position 10: - - A =  -1
        . . .
             Sum     230
```

What's significant?

- High confidence - >40% identity for long alignments (Rost, 1999 found that sequence alignments unambiguously distinguish between protein pairs of similar and non-similar structure when the pairwise sequence identity >40%)
- "Twilight zone" – blurry - 20-35% identity
- "Midnight zone" - <20% identity

Topics to Cover

- Introduction
- Pairwise Alignment methods
  - Dot plot analysis
  - Exhaustive methods; Dynamic programming algorithm (Smith-Waterman (Local), Needleman-Wunsch (Global))
  - Heuristic methods; Approximate methods; word or k-tuple (FASTA, BLAST, BLAT)
- Multiple sequence alignments
- Pre-computed alignment resources

Comparing two sequences

- DOTLET – Dot Plot (http://myhits.isb-sib.ch/cgi-bin/dotlet)
- NCBI

  BLAST 2 Sequences (www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi)
- EBI – (http://www.ebi.ac.uk/Tools/psa/)

  GLOBAL

  needle (EMBOSS) - Needleman -Wunsch

  stretcher (EMBOSS) – modification of N-W

  LOCAL

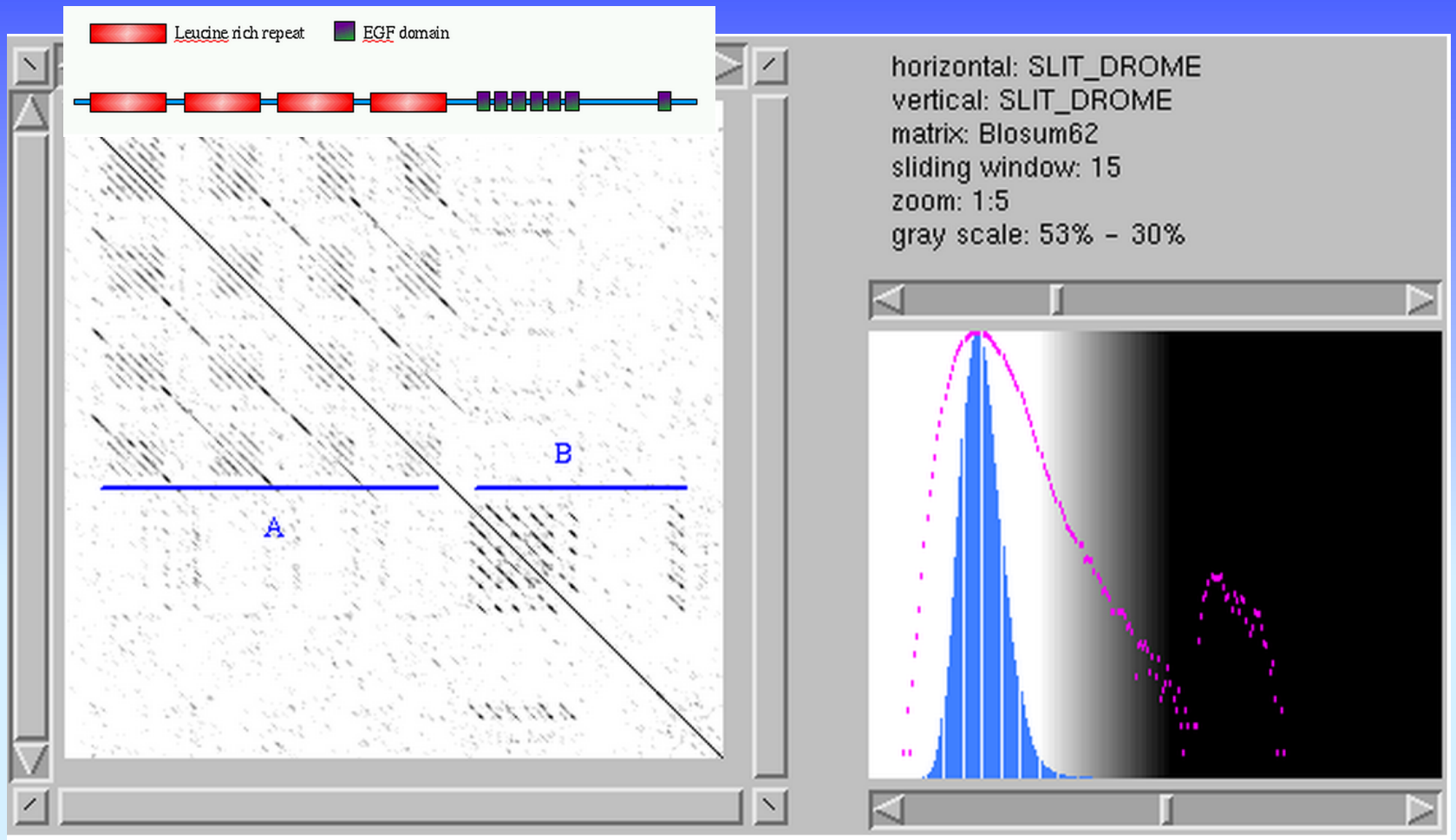
  water (EMBOSS) - Smith-Waterman

  matcher (EMBOSS) - uses algorithm based on LALIGN

Dot Plot

- Graphical way of looking at alignment of 2 sequences
- Look at structure of a sequence by doing a self comparison
- Method first described by Gibbs and McIntyre (1970)
- Can find direct or inverted repeats in sequences

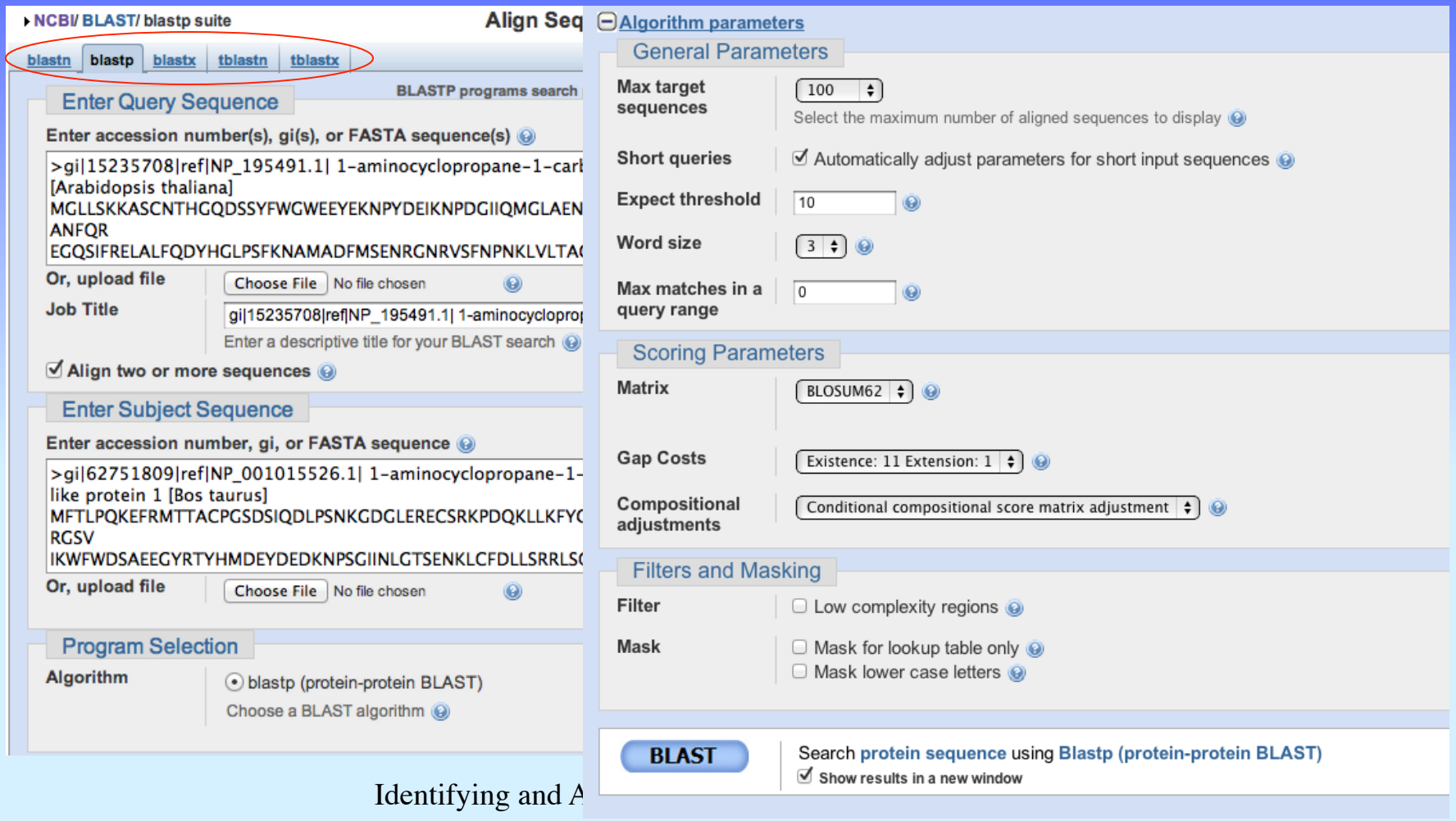
Dot Matrix Comparison



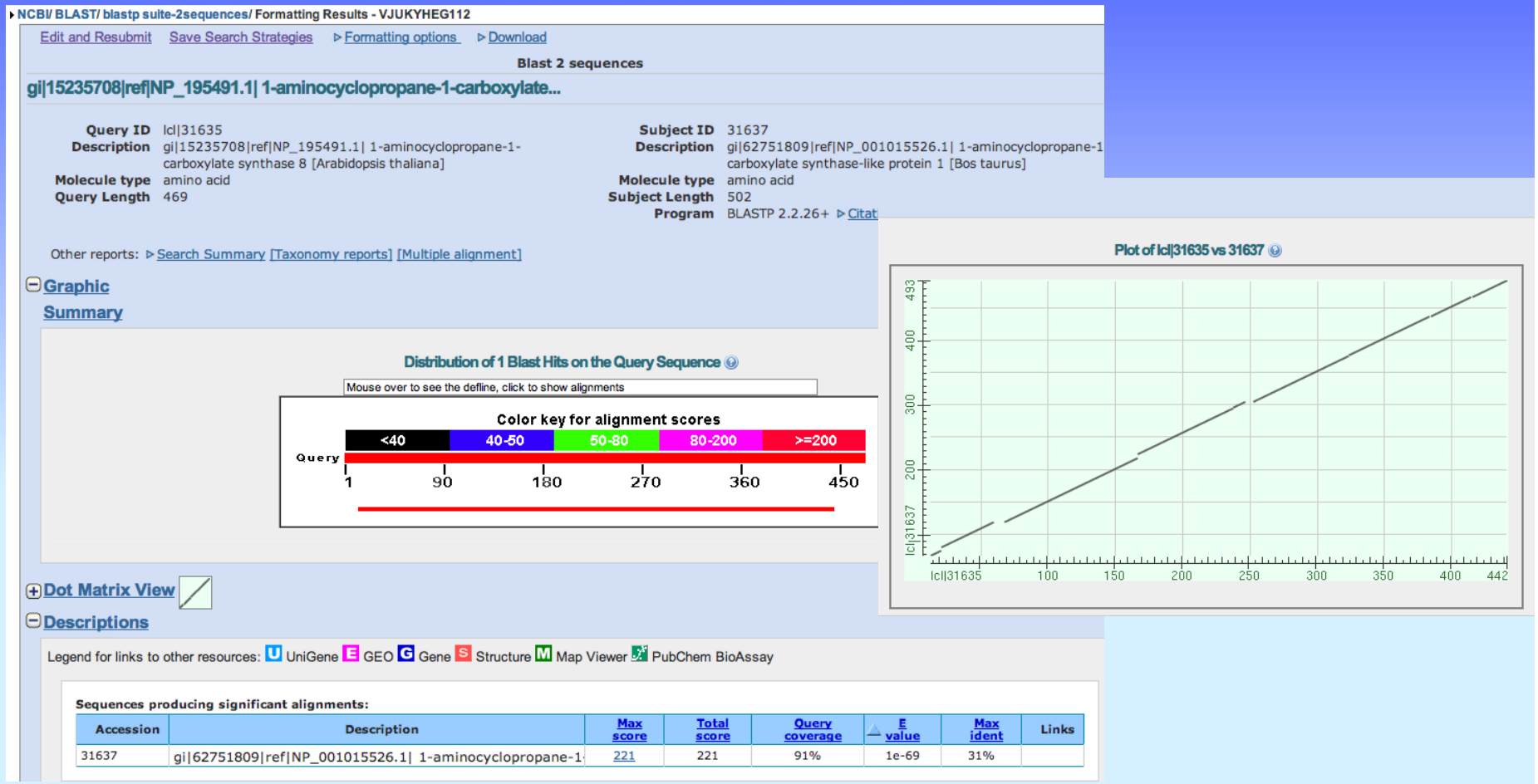
http://myhits.isb-sib.ch/cgi-bin/dotlet

Identifying and Aligning Homologs (Whitehead Institute)

NCBI - Blast2seq

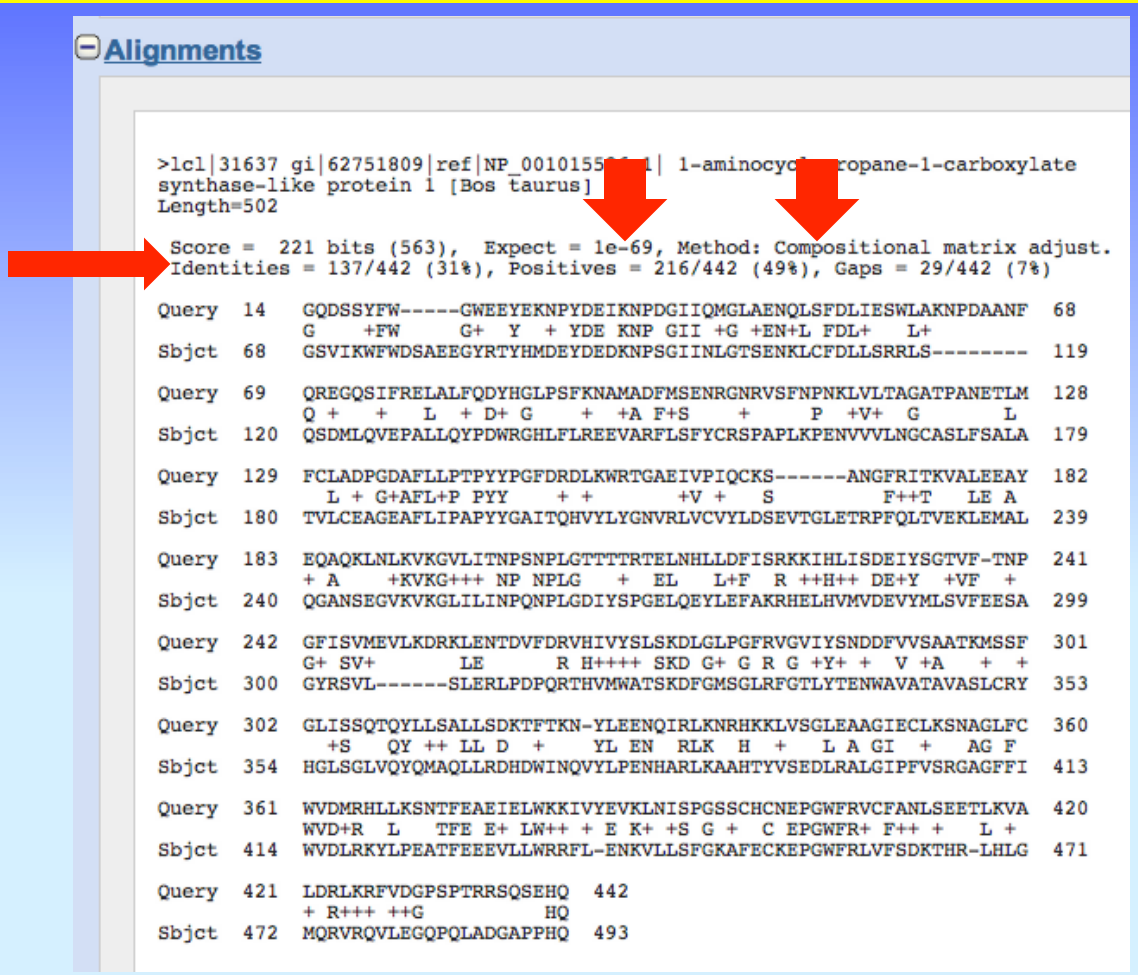


NCBI - Blast2seq



Identifying and Aligning Homologs (Whitehead Institute)

NCBI - Blast2seq



Alignments

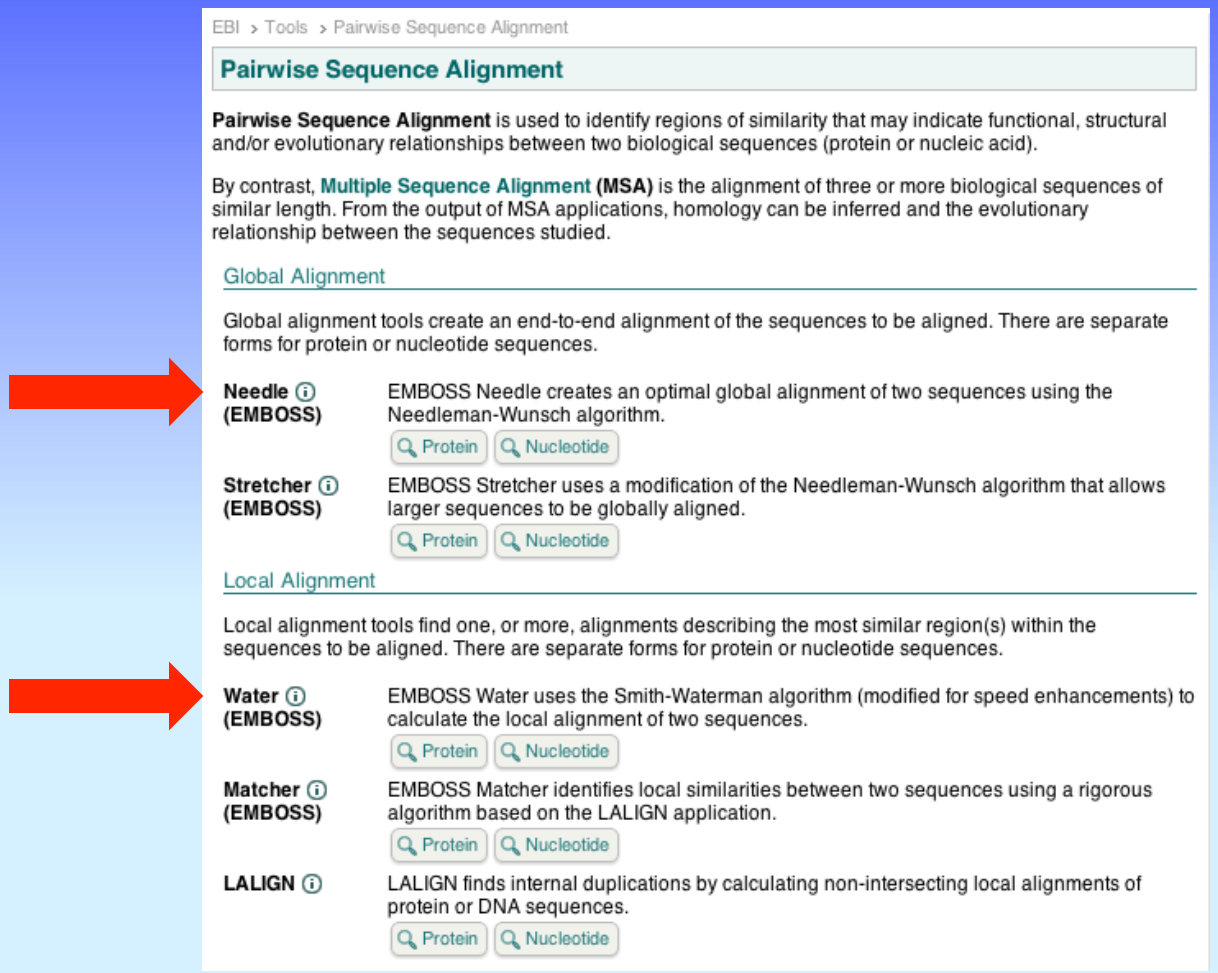
>lcl|31637 gi|62751809|ref|NP_001015[illegible]1| 1-aminocyc[illegible]ropane-1-carboxylate synthase-like protein 1 [Bos taurus]
Length=502

Score = 221 bits (563), Expect = 1e-69, Method: Compositional matrix adjust.
Identities = 137/442 (31%), Positives = 216/442 (49%), Gaps = 29/442 (7%)

```
Query  14   GQDSSYFW-----GWEEYEKNPYDEIKNPDGIIQMGLAENQLSFDLIESWLAKNPDAANF  68
            G    +FW     G+  Y  + YDE KNP GII +G +EN+L FDL+   L+
Sbjct  68   GSVIKWFWDSAEEGYRTYHMDEYDEDKNPSGIINLGTSENKLCFDLLSRRLS--------  119

Query  69   QREGQSIFRELALFQDYHGLPSFKNAMADFMSENRGNRVSFNPNKLVLTAGATPANETLM  128
            Q +   +   L  + D+ G    +  +A F+S    +     P  +V+  G       L
Sbjct  120  QSDMLQVEPALLQYPDWRGHLFLREEVARFLSFYCRSPAPLKPENVVVLNGCASLFSALA  179

Query  129  FCLADPGDAFLLPTPYYPGFDRDLKWRTGAEIVPIQCKS------ANGFRITKVALEEAY  182
              L + G+AFL+P PYY    + +       +V +   S          F++T   LE A
Sbjct  180  TVLCEAGEAFLIPAPYYGAITQHVYLYGNVRLVCVYLDSEVTGLETRPFQLTVEKLEMAL  239

Query  183  EQAQKLNLKVKGVLITNPSNPLGTTTTRTELNHLLDFISRKKIHLISDEIYSGTVF-TNP  241
            + A    +KVKG+++ NP NPLG   +  EL   L+F  R ++H++ DE+Y  +VF  +
Sbjct  240  QGANSEGVKVKGLILINPQNPLGDIYSPGELQEYLEFAKRHELHVMVDEVYMLSVFEESA  299

Query  242  GFISVMEVLKDRKLENTDVFDRVHIVYSLSKDLGLPGFRVGVIYSNDDFVVSAATKMSSF  301
            G+ SV+       LE      R H++++ SKD G+ G R G +Y+ +  V +A   +  +
Sbjct  300  GYRSVL------SLERLPDPQRTHVMWATSKDFGMSGLRFGTLYTENWAVATAVASLCRY  353

Query  302  GLISSQTQYLLSALLSDKTFTKN-YLEENQIRLKNRHKKLVSGLEAAGIECLKSNAGLFC  360
              +S   QY ++ LL D  +    YL EN  RLK  H  +   L A GI  +   AG F
Sbjct  354  HGLSGLVQYQMAQLLRDHDWINQVYLPENHARLKAAHTYVSEDLRALGIPFVSRGAGFFI  413

Query  361  WVDMRHLLKSNTFEAEIELWKKIVYEVKLNISPGSSCHCNEPGWFRVCFANLSEETLKVA  420
            WVD+R  L   TFE E+ LW++ + E K+ +S G +  C EPGWFR+ F++ +   L +
Sbjct  414  WVDLRKYLPEATFEEEVLLWRRFL-ENKVLLSFGKAFECKEPGWFRLVFSDKTHR-LHLG  471

Query  421  LDRLKRFVDGPSPTRRSQSEHQ  442
            + R+++ ++G           HQ
Sbjct  472  MQRVRQVLEGQPQLADGAPPHQ  493
```

Identifying and Aligning Homologs (Whitehead Institute)

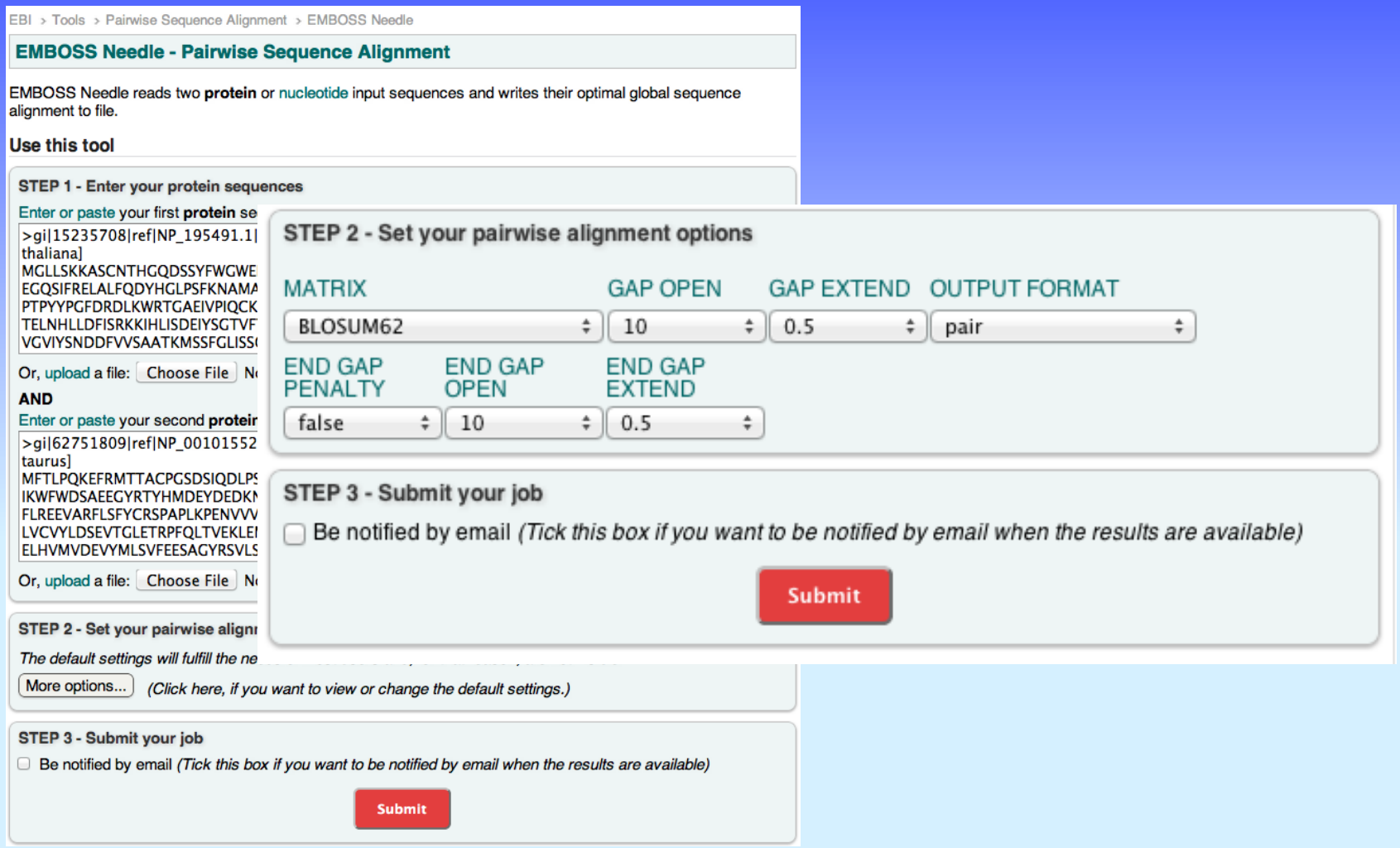
Pairwise Alignment Tools at EBI



http://www.ebi.ac.uk/Tools/psa/

Identifying and Aligning Homologs (Whitehead Institute)

Needle (global)



Identifying and Aligning Homologs (Whitehead Institute)

Needle (global)

EBI › Tools › Pairwise Sequence Alignment › EMBOSS Needle

EMBOSS Needle - Alignment

Alignment | Submission Details | Submit Another Job

Alignment

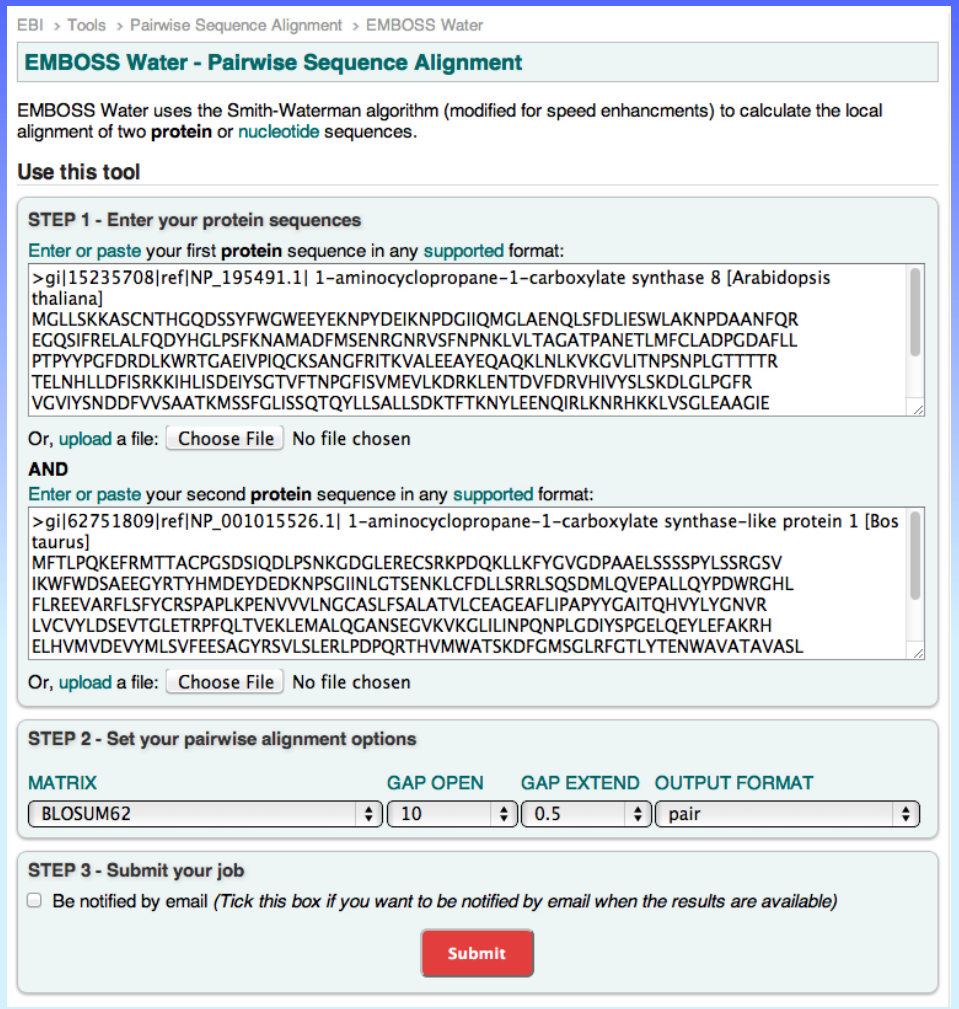
View Alignment File

```
########################################
# Program: needle
# Rundate: Sun 20 May 2012 23:24:21
# Commandline: needle
#    -auto
#    -stdout
#    -asequence emboss_needle-I20120520-232408-0601-60957810-o
#    -bsequence emboss_needle-I20120520-232408-0601-60957810-o
#    -datafile EBLOSUM62
#    -gapopen 10.0
#    -gapextend 0.5
#    -endopen 10.0
#    -endextend 0.5
#    -aformat3 pair
#    -sprotein1
#    -sprotein2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: NP_195491.1
# 2: NP_001015526.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 545
# Identity:      140/545 (25.7%)
# Similarity:    223/545 (40.9%)
# Gaps:          119/545 (21.8%)
# Score: 575.5
#
#
#=======================================
```

```
NP_195491.1         1 ------------------------------------------------      0

NP_001015526.       1 MFTLPQKEFRMTTACPGSDSIQDLPSNKGDGLERECSRKPDQKLLKFYGV     50

NP_195491.1         1 ----MGLLSKKASCNTHGQDSSYFW-----GWEEYEKNPYDEIKNPDGII     41
                        ..|.|.....::.|....:||     |:..|..:.|||.|||.|||
NP_001015526.      51 GDPAAELSSSSPYLSSRGSVIKWFWDSAEEGYRTYHMDEYDEDKNPSGII    100

NP_195491.1        42 QMGLAENQLSFDLIESWLAKNPDAANFQREGQSIFRELALFQDYHGLPSF     91
                      .:|.:||:|.|||:...|:         |.:...:...|..:.|:.|....
NP_001015526.     101 NLGTSENKLCFDLLSRRLS--------QSDMLQVEPALLQYPDWRGHLFL    142

NP_195491.1        92 KNAMADFMSENRGNRVSFNPNKLVLTAGATPANETLMFCLADPGDAFLLP     141
                      :..:|.|:|....:....|..:|:..|.......|...|.:.|:|||:|
NP_001015526.     143 REEVARFLSFYCRSPAPLKPENVVVLNGCASLFSALATVLCEAGEAFLIP    192

NP_195491.1       142 TPYYPGFDRDLKWRTGAEIVPIQCKS------ANGFRITKVALEEAYEQA    185
                      .|||....:.:.......:|.:...|      ...|::|...||.|.:.|
NP_001015526.     193 APYYGAITQHVYLYGNVRLVCVYLDSEVTGLETRPFQLTVEKLEMALQGA    242

NP_195491.1       186 QKLNLKVKGVLITNPSNPLGTTTTRTELNHLLDFISRKKIHLISDEIYSG    235
                      ....:||||:::.||.||||...:..||...|:|..|.::|::.||:|..
NP_001015526.     243 NSEGVKVKGLILINPQNPLGDIYSPGELQEYLEFAKRHELHVMVDEVYML    292

NP_195491.1       236 TVF-TNPGFISVMEVLKDRKLENTDVFDRVHIVYSLSKDLGLPGFRVGVI    284
                      :|| .:.|:.||:      .||.....|.|::::.|||.|:.|.|.|.:
NP_001015526.     293 SVFEESAGYRSVL------SLERLPDPQRTHVMWATSKDFGMSGLRFGTL    336

NP_195491.1       285 YSNDDFVVSAATKMSSFGLISSQTQYLLSALLSDKTFTKN-YLEENQIRL    333
                      |:.:..|.:|...:..:..:|...||.::.||.|..:... ||.||..||
NP_001015526.     337 YTENWAVATAVASLCRYHGLSGLVQYQMAQLLRDHDWINQVYLPENHARL    386

NP_195491.1       334 KNRHKKLVSGLEAAGIECLKSNAGLFCWVDMRHLLKSNTFEAEIELWKKI    383
                      |..|..:...|.|.||..:...||.|.|||:|..|...|||.|:.||::.
NP_001015526.     387 KAAHTYVSEDLRALGIPFVSRGAGFFIWVDLRKYLPEATFEEEVLLWRRF    436

NP_195491.1       384 VYEVKLNISPGSSCHCNEPGWFRVCFANLSEETLKVALDRLKRFVDG---    430
                      : |.|:.:|.|.:.:.|.||||||:.|::.:..  |.:.:.|:::.::|
NP_001015526.     437 L-ENKVLLSFGKAFECKEPGWFRLVFSDKTHR-LHLGMQRVRQVLEGQPQ    484

NP_195491.1       431 ------PSPTRRSQSEHQRLKNLRKMKVSNWVFRLSFHDREPEER    469
                            |...:..|..|:
NP_001015526.     485 LADGAPPHQIQEPQGPHR----------------------------    502
```

Identifying and Aligning Homologs (Whitehead Institute)

Water (local)



Identifying and Aligning Homologs (Whitehead Institute)

Water (local)

EBI › Tools › Pairwise Sequence Alignment › EMBOSS Water

EMBOSS Water - Alignment

Alignment | Submission Details | Submit Another Job

Alignment

View Alignment File

```
########################################
# Program: water
# Rundate: Sun 20 May 2012 23:28:18
# Commandline: water
#    -auto
#    -stdout
#    -asequence emboss_water-I20120520-232815-0355-55247721-oy.asequence
#    -bsequence emboss_water-I20120520-232815-0355-55247721-oy.bsequence
#    -datafile EBLOSUM62
#    -gapopen 10.0
#    -gapextend 0.5
#    -aformat3 pair
#    -sprotein1
#    -sprotein2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: NP_195491.1
# 2: NP_001015526.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 453
# Identity:     139/453 (30.7%)
# Similarity:   220/453 (48.6%)
# Gaps:          29/453 ( 6.4%)
# Score: 578.5
#
#
#=======================================
```

```
NP_195491.1         3 LLSKKASCNTHGQDSSYFW-----GWEEYEKNPYDEIKNPDGIIQMGLAE     47
                      |.|.....::.|....:||      |:..|..:.|||.|||.|||.:|.:|
NP_001015526.      57 LSSSSPYLSSRGSVIKWFWDSAEEGYRTYHMDEYDEDKNPSGIINLGTSE    106

NP_195491.1        48 NQLSFDLIESWLAKNPDAANFQREGQSIFRELALFQDYHGLPSFKNAMAD     97
                      |:|.|||:...|:          |.:...:...|..:.|:.|....:..:|.
NP_001015526.     107 NKLCFDLLSRRLS--------QSDMLQVEPALLQYPDWRGHLFLREEVAR    148

NP_195491.1        98 FMSENRGNRVSFNPNKLVLTAGATPANETLMFCLADPGDAFLLPTPYYPG    147
                      |:|....:.....|..:|:..|.......|...|.:.|:|||:|.|||..
NP_001015526.     149 FLSFYCRSPAPLKPENVVVLNGCASLFSALATVLCEAGEAFLIPAPYYGA    198

NP_195491.1       148 FDRDLKWRTGAEIVPIQCKS------ANGFRITKVALEEAYEQAQKLNLK    191
                      ..:.:.......:|.:...|      ...|::|...||.|.:.|....:|
NP_001015526.     199 ITQHVYLYGNVRLVCVYLDSEVTGLETRPFQLTVEKLEMALQGANSEGVK    248

NP_195491.1       192 VKGVLITNPSNPLGTTTTRTELNHLLDFISRKKIHLISDEIYSGTVF-TN    240
                      |||:::.||.||||...:..||...|:|..|.::|::.||:|..:|| .:
NP_001015526.     249 VKGLILINPQNPLGDIYSPGELQEYLEFAKRHELHVMVDEVYMLSVFEES    298

NP_195491.1       241 PGFISVMEVLKDRKLENTDVFDRVHIVYSLSKDLGLPGFRVGVIYSNDDF    290
                      .|:.||:      .||......|.|::::.|||.|:.|.|.|.|.:|:.:..
NP_001015526.     299 AGYRSVL------SLERLPDPQRTHVMWATSKDFGMSGLRFGTLYTENWA    342

NP_195491.1       291 VVSAATKMSSFGLISSQTQYLLSALLSDKTFTKN-YLEENQIRLKNRHKK    339
                      |.:|...:..:..:|...||.::.||.|..:... ||.||..|||..|..
NP_001015526.     343 VATAVASLCRYHGLSGLVQYQMAQLLRDHDWINQVYLPENHARLKAAHTY    392

NP_195491.1       340 LVSGLEAAGIECLKSNAGLFCWVDMRHLLKSNTFEAEIELWKKIVYEVKL    389
                      :...|.|.||..:...||.|.|||:|..|...|||.|:.||::.: |.|:
NP_001015526.     393 VSEDLRALGIPFVSRGAGFFIWVDLRKYLPEATFEEEVLLWRRFL-ENKV    441

NP_195491.1       390 NISPGSSCHCNEPGWFRVCFANLSEETLKVALDRLKRFVDGPSPTRRSQS    439
                      .:|.|.:..|.||||||:.|::.:.. |.:.:.|:::.::|.........
NP_001015526.     442 LLSFGKAPPCKEPGWFRLVFSDKTHR-LHLGMQRVRQVLEGQPQLADGAP    490

NP_195491.1       440 EHQ    442
                      .||
NP_001015526.     491 PHQ    493
```

Identifying and Aligning Homologs (Whitehead Institute)

Topics to Cover

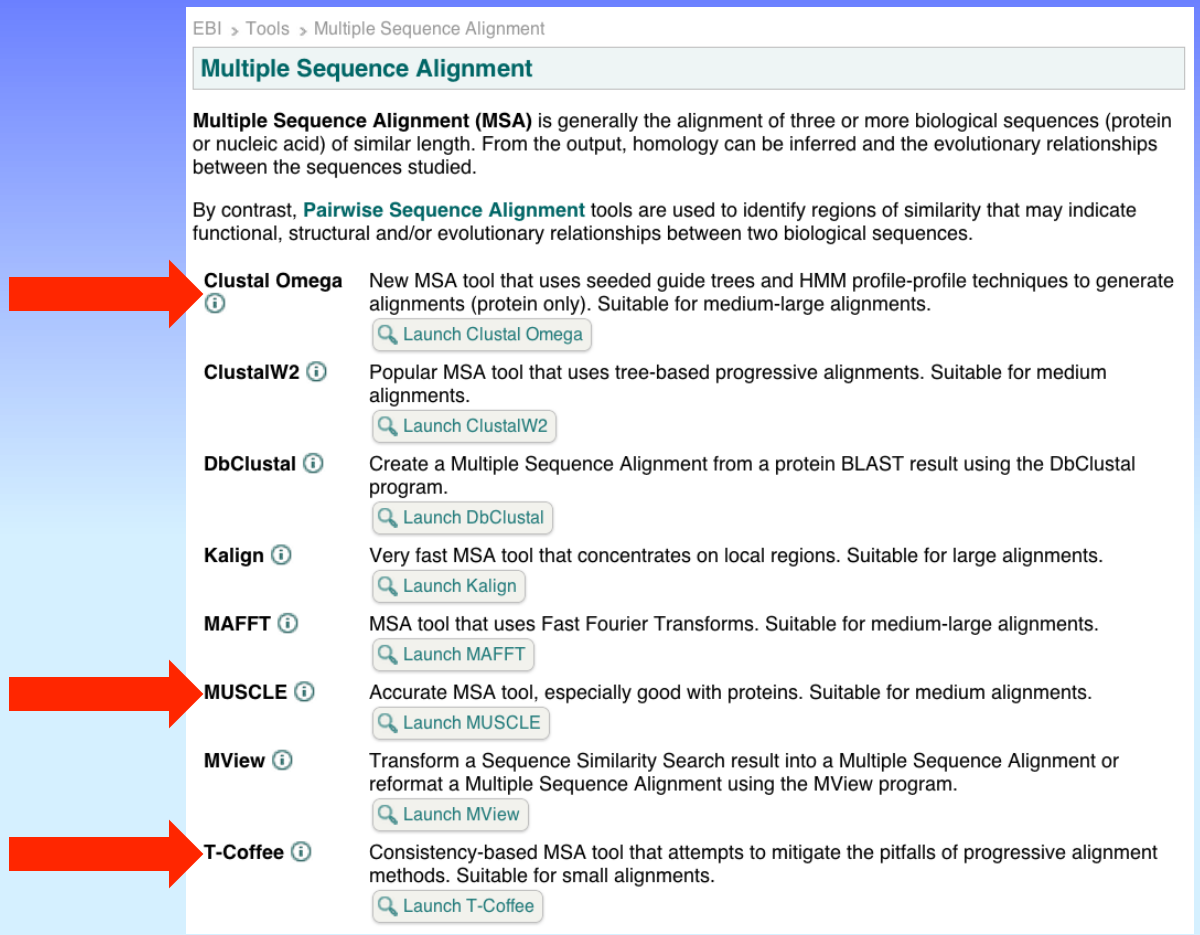
- Introduction
- Pairwise Alignment methods
- Multiple sequence alignments
- Pre-computed alignment resources

Multiple Sequence Alignment

- Additional sequences can help resolve ambiguities found in a pairwise comparison
- Remove uninformative sequences
- Dynamic programming techniques require prohibitively large computer resources
- Tree or hierarchical methods (successive pairwise alignments), Consistency-based methods, Template-based methods

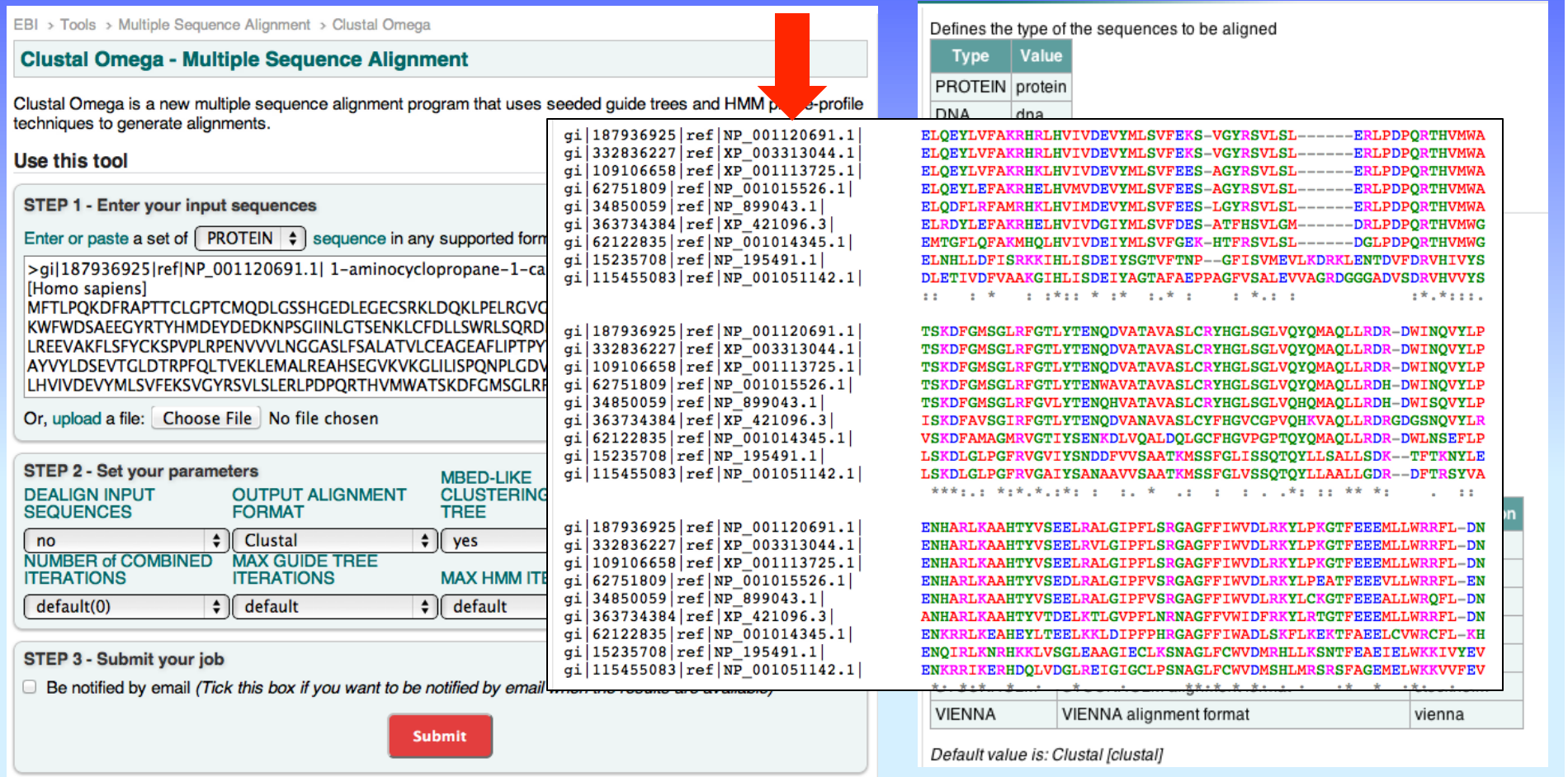
Multiple Sequence Alignment

http://www.ebi.ac.uk/Tools/msa/



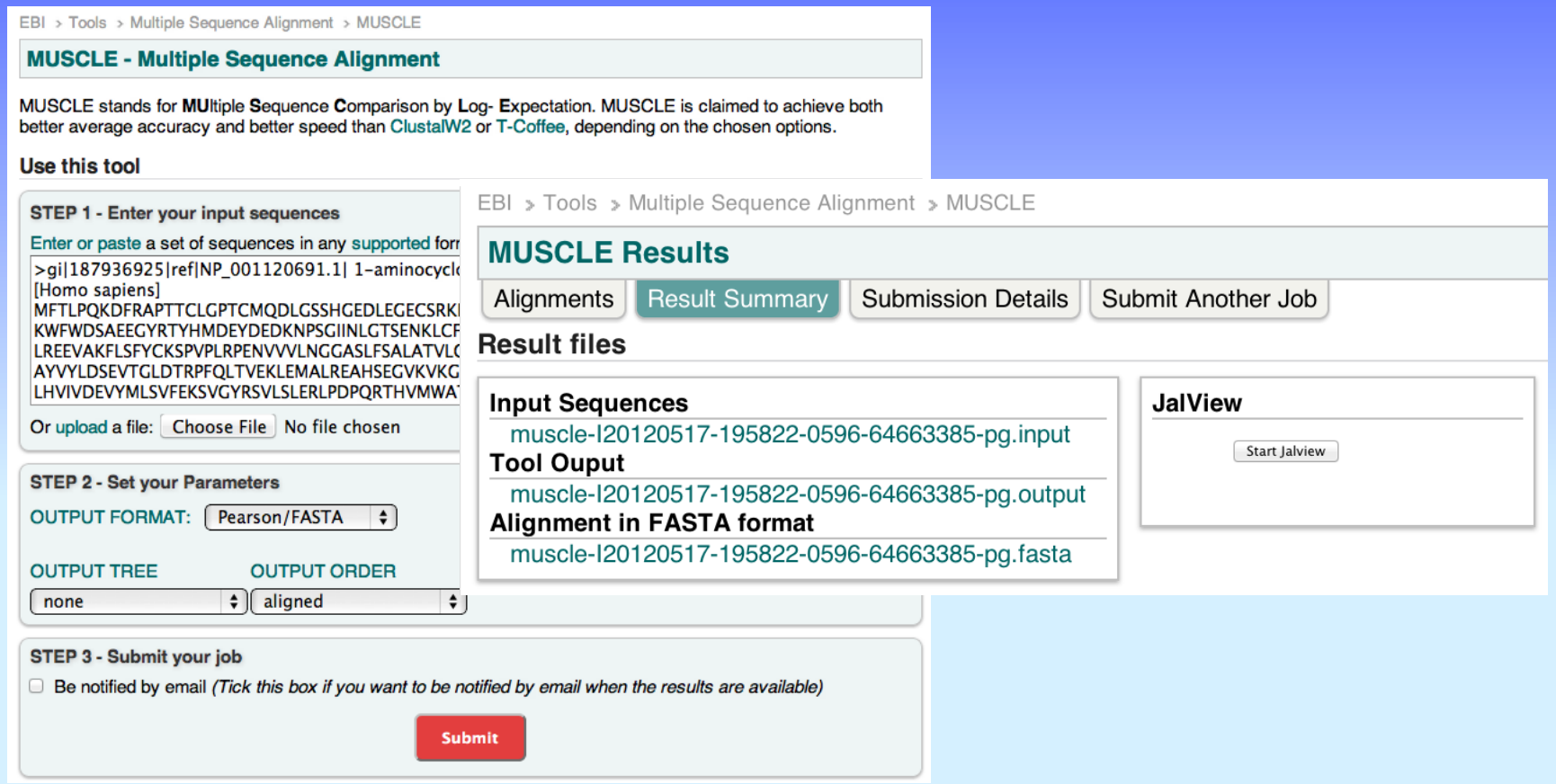
Identifying and Aligning Homologs (Whitehead Institute)

Clustal Omega



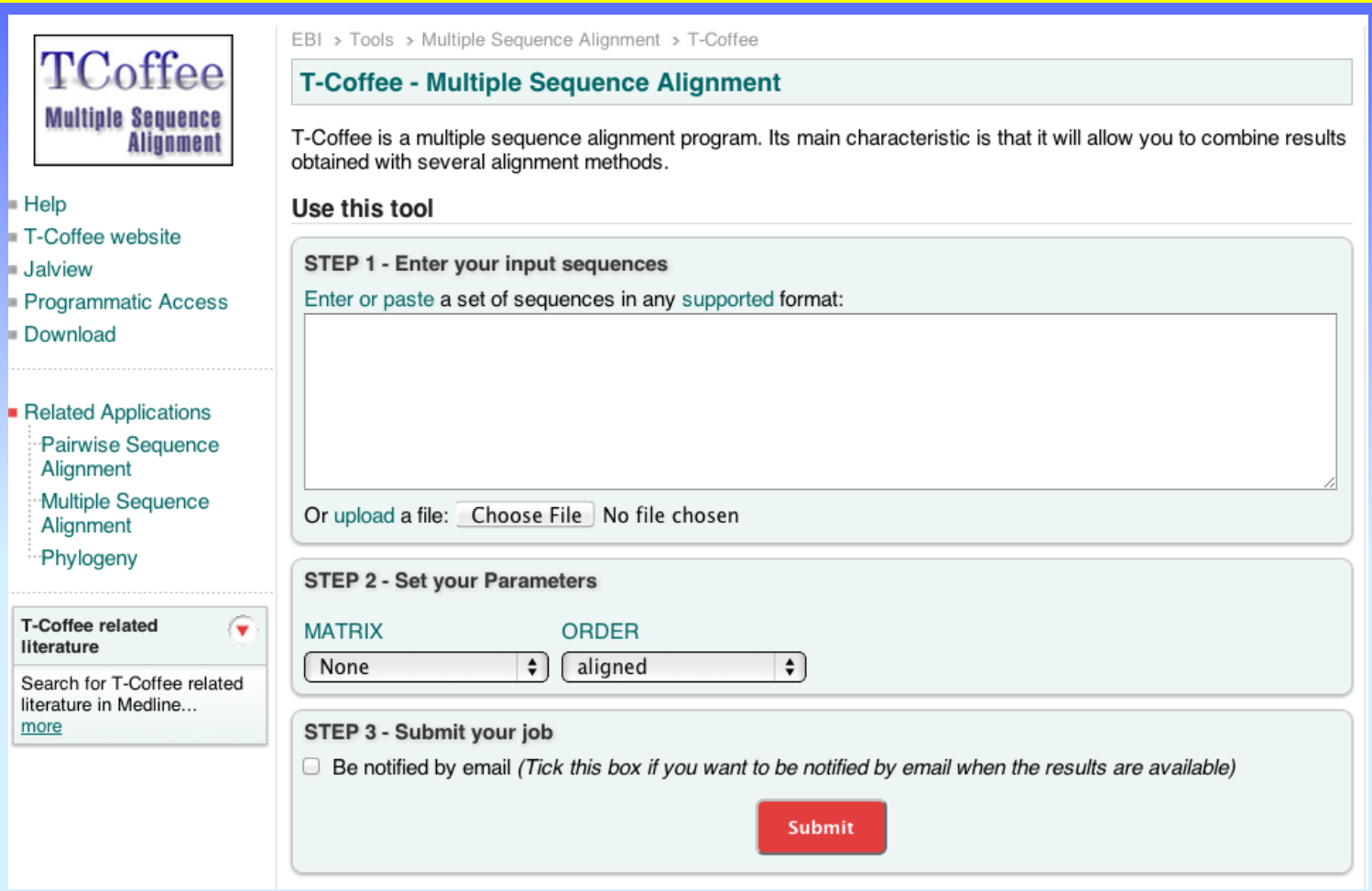
Identifying and Aligning Homologs (Whitehead Institute)

Muscle



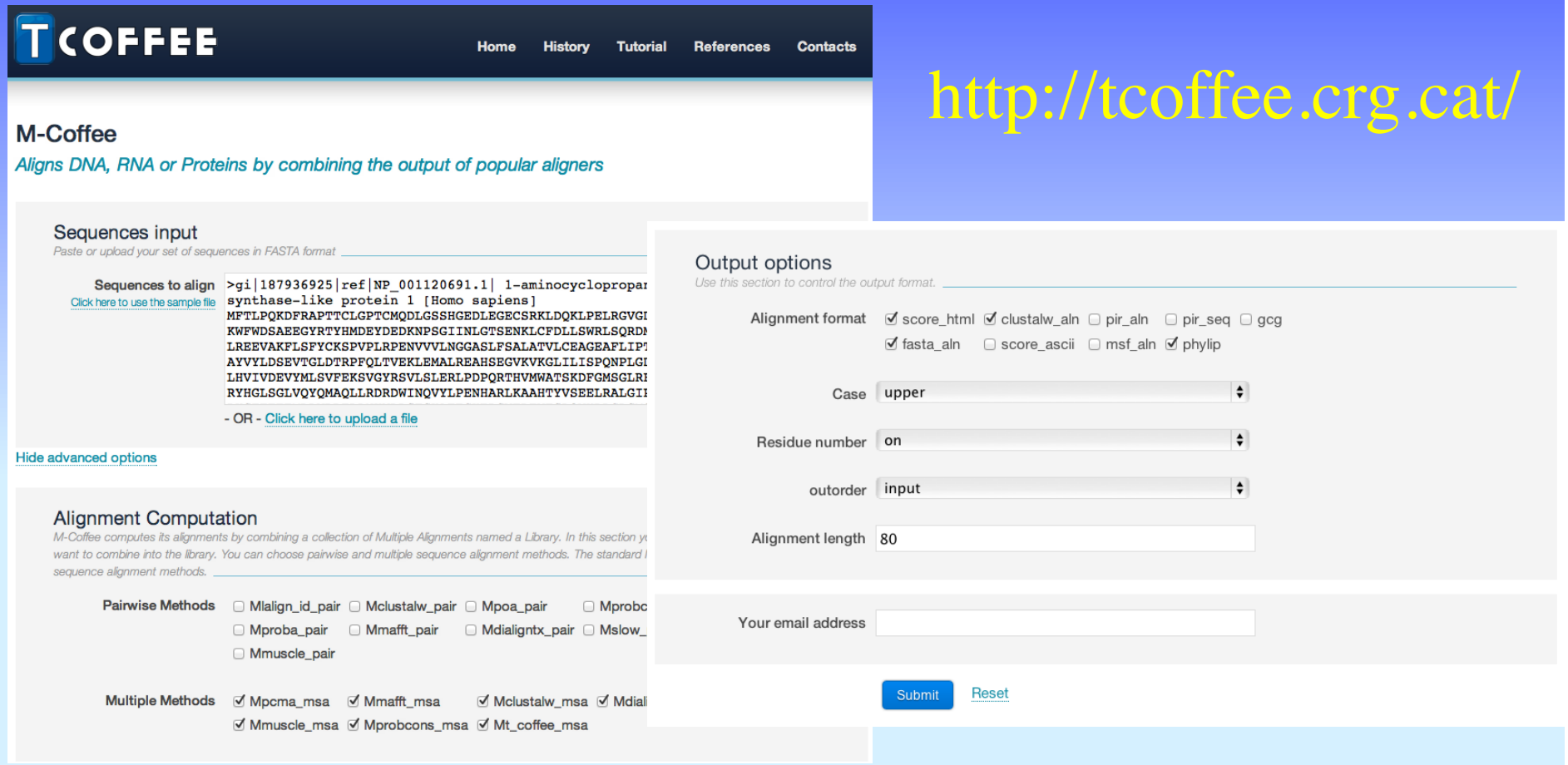
Identifying and Aligning Homologs (Whitehead Institute)

T-Coffee



Identifying and Aligning Homologs (Whitehead Institute)

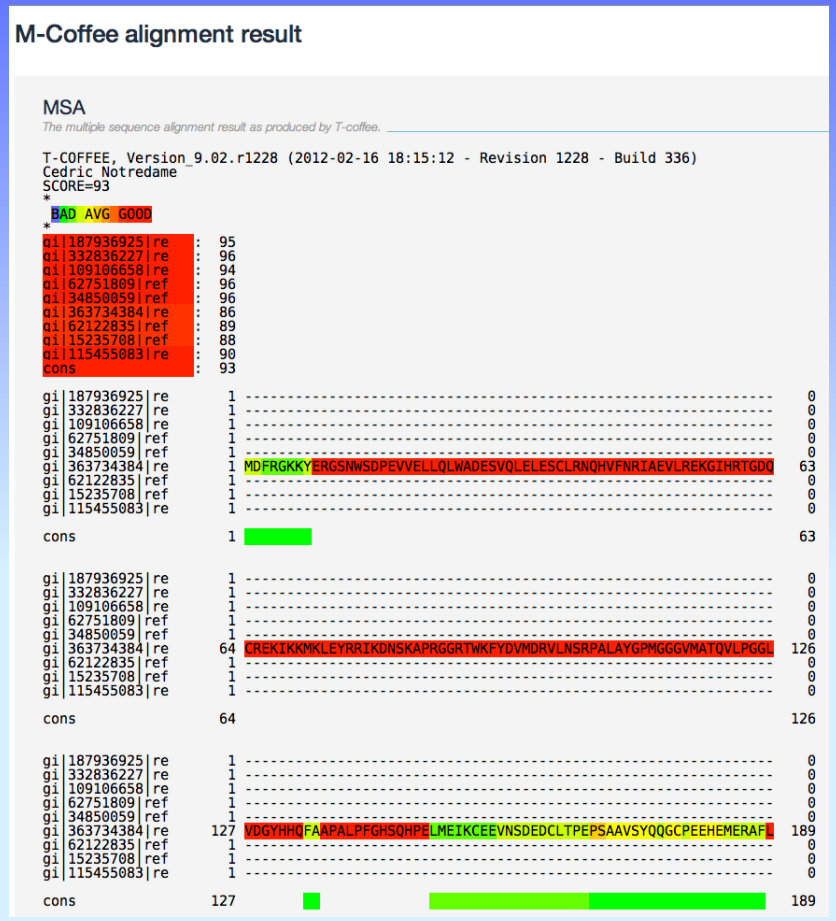
M-Coffee



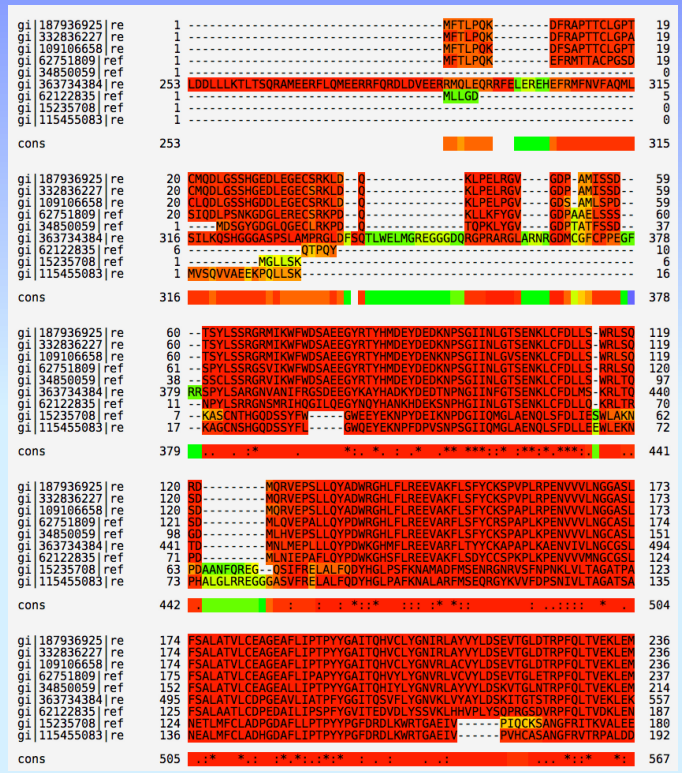
http://tcoffee.crg.cat/

Identifying and Aligning Homologs (Whitehead Institute)

M-Coffee



http://tcoffee.crg.cat/



Identifying and Aligning Homologs (Whitehead Institute)

Other Considerations

Web vs Command Line

- More options to change parameters
- Process lots of alignments in one command

Which web page to use?

- Alignment home page vs uniform interface

Alternatives

- Use subsequences or subset of sequences

Realign by hand

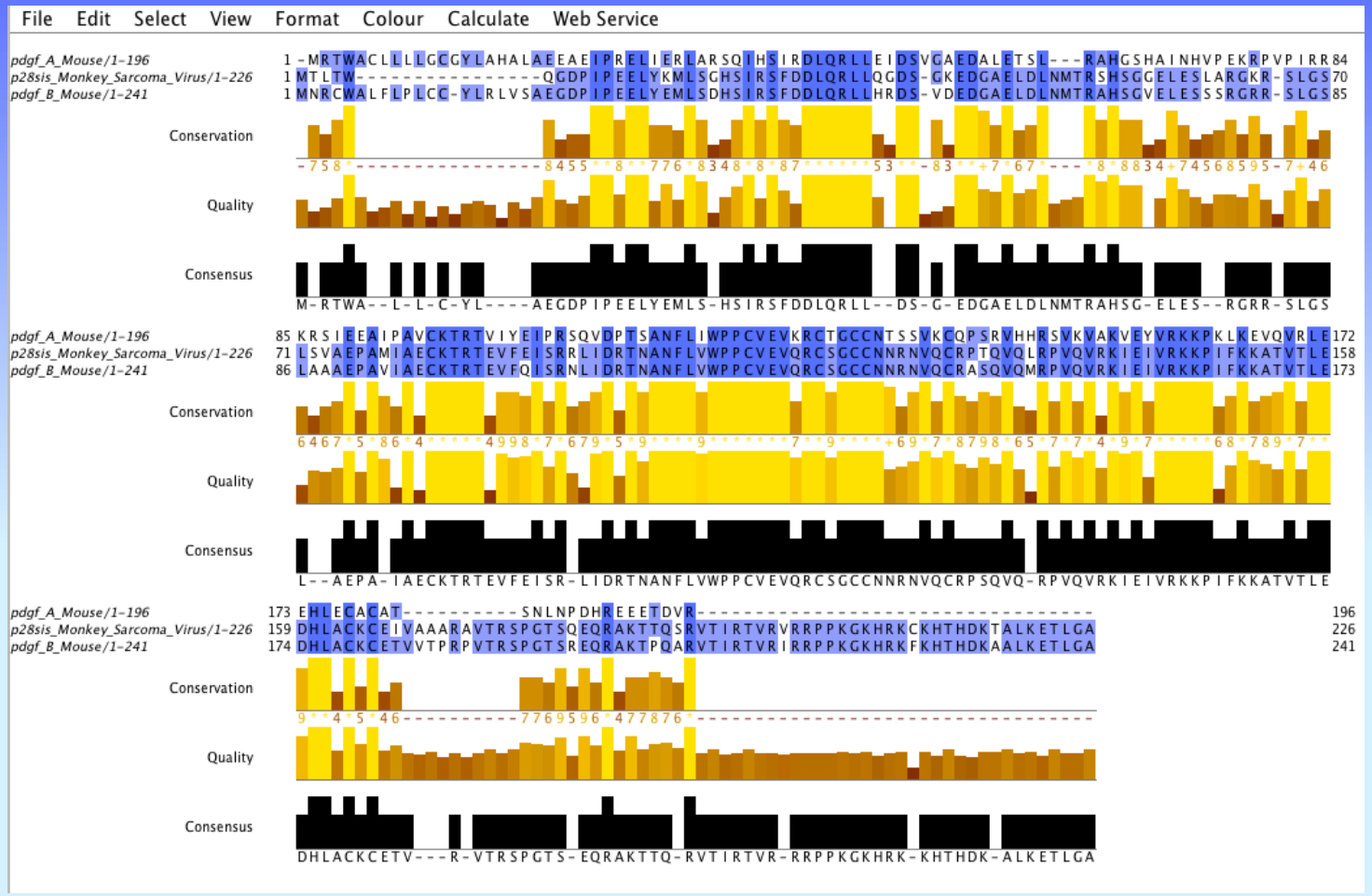
Doing Lots of Alignments

Our favorite method is to use the T-COFFEE suite (more specifically, M-Coffee) to run multiple alignment methods and then create a consensus alignment, a sort of a meta-alignment. This can be done with a single command like

```
t_coffee my_proteins.fa
        -method=t_coffee_msa,mafft_msa,probcons_msa,muscle_msa
        -output=fasta_aln
```

The final consensus alignment will appear in the file my_proteins.fasta_aln, which can them be viewed in JalView.

Jalview (http://www.jalview.org)

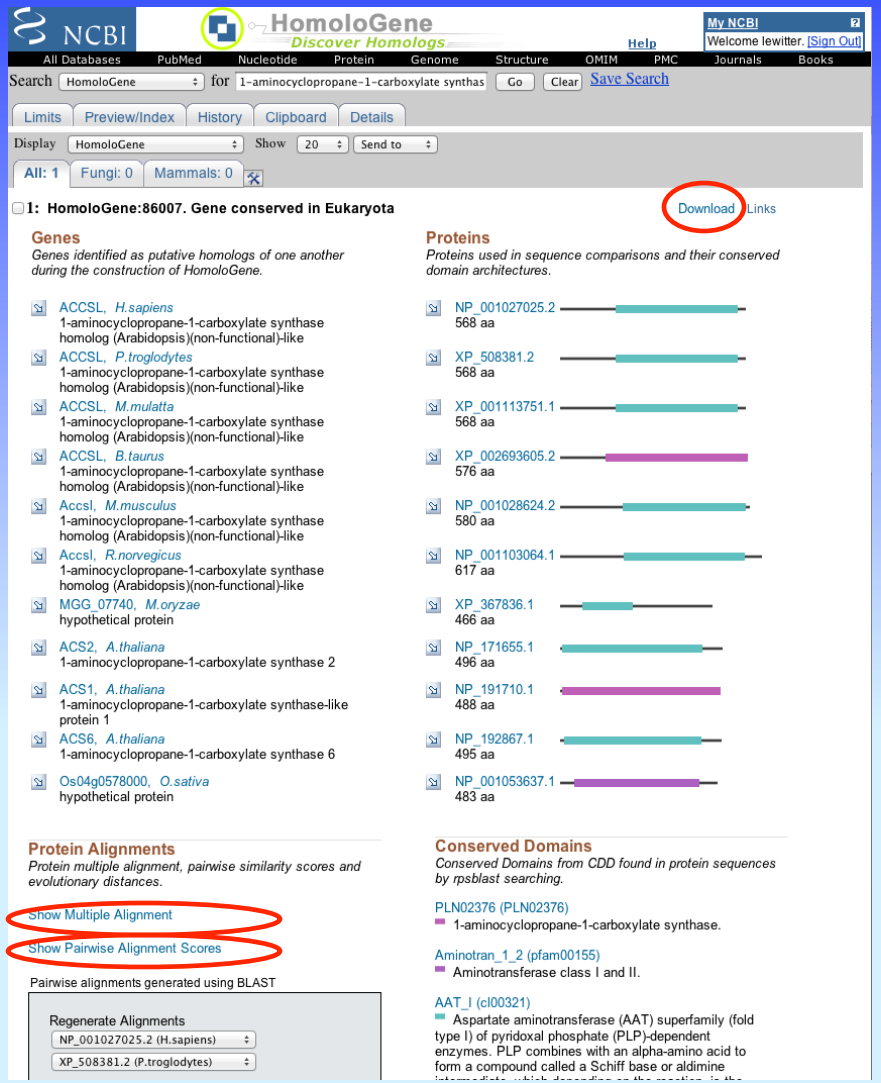


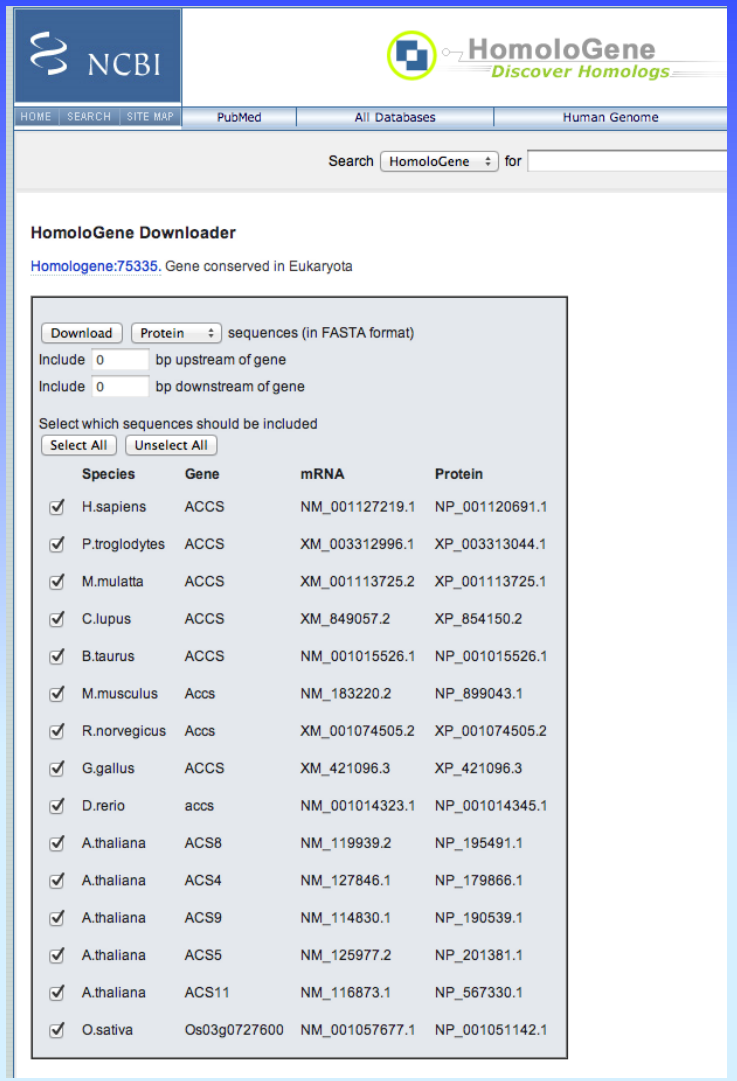
Identifying and Aligning Homologs (Whitehead Institute)

Topics to Cover

- Introduction
- Pairwise Alignment methods
- Multiple sequence alignments
- Pre-computed alignment resources
  - Homologene (NCBI)
  - Ensembl (EBI)

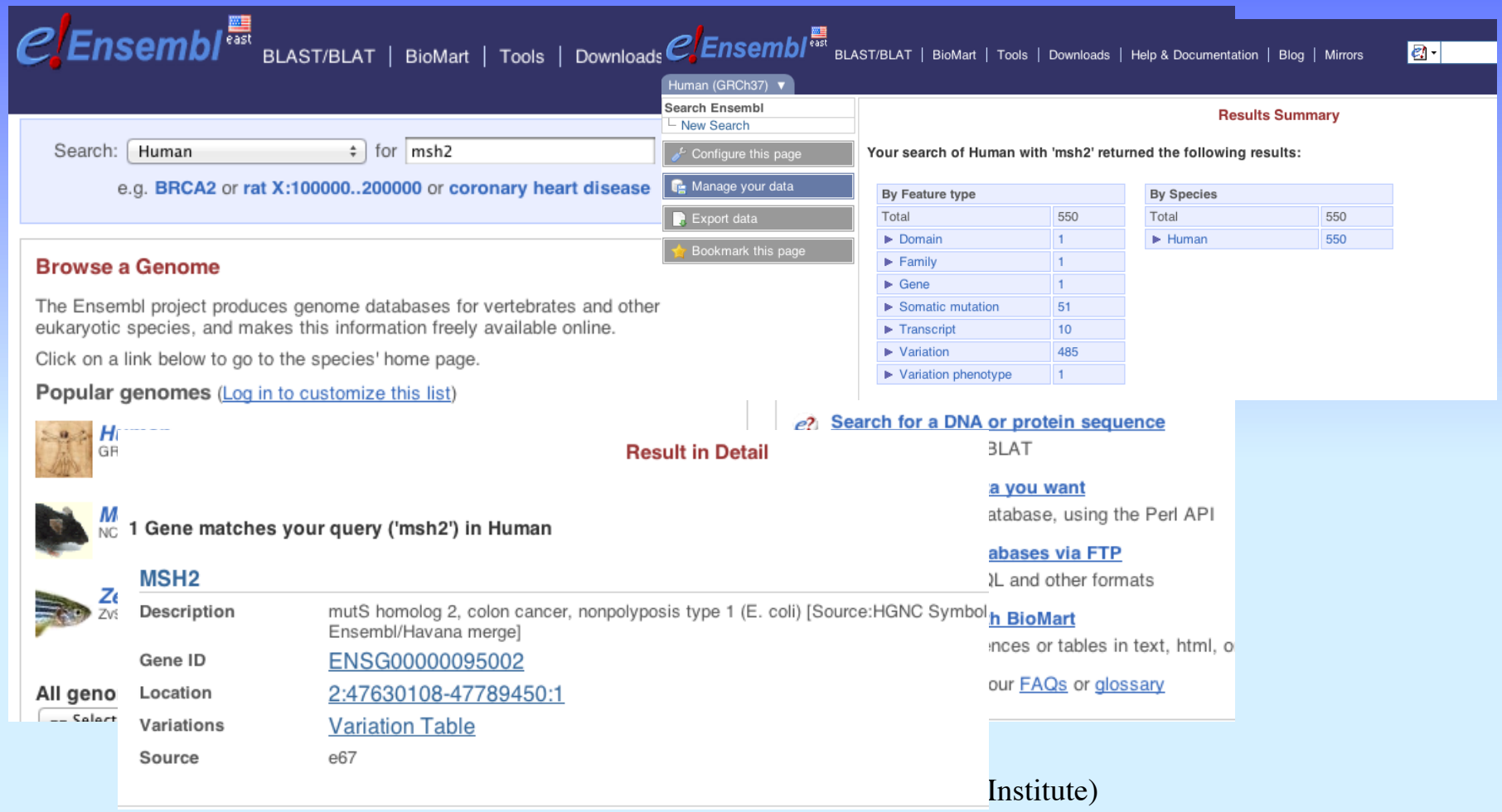
HOMOLOGENE: http://www.ncbi.nlm.nih.gov/homologene





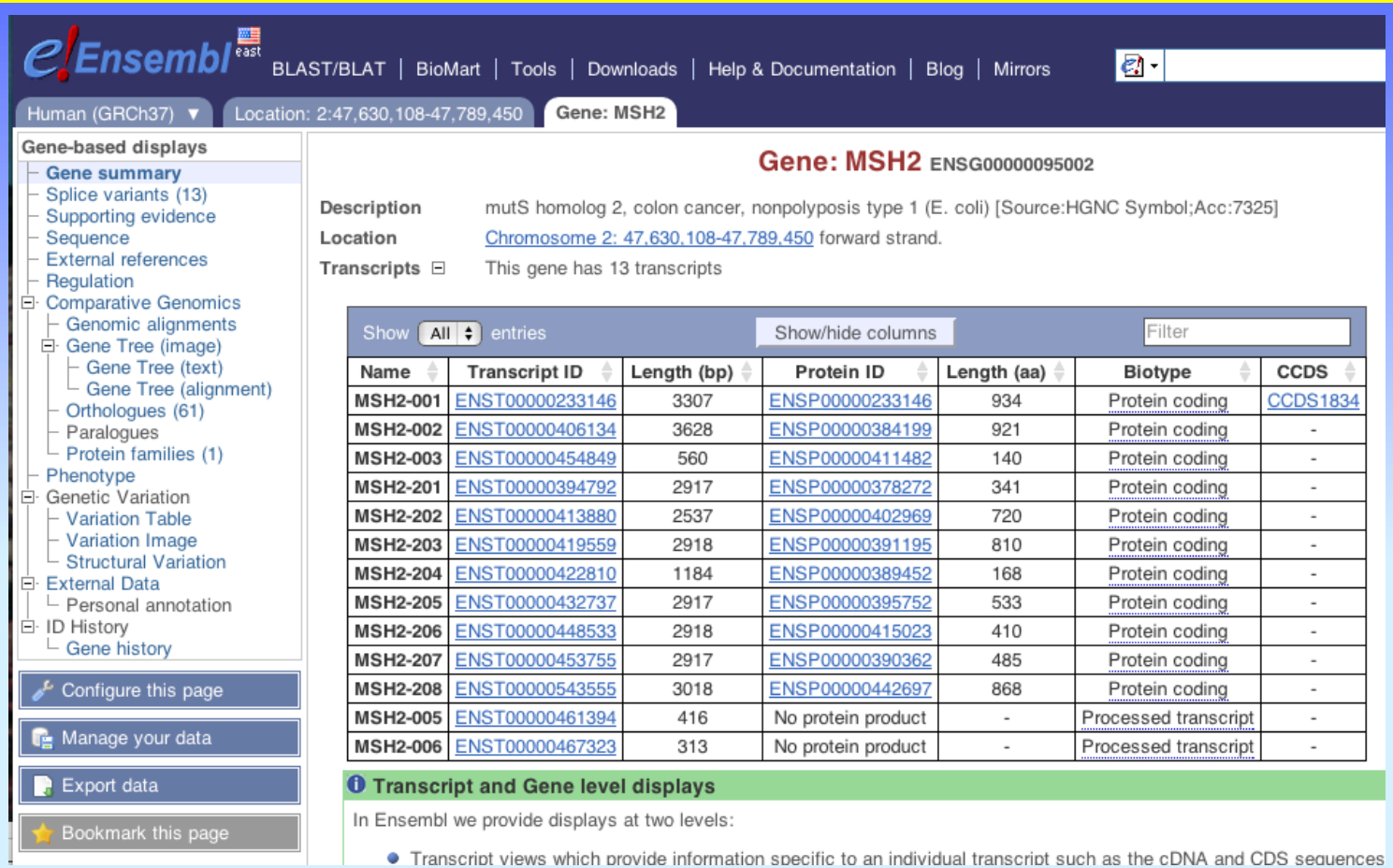
Identifying and Aligning Homologs (Whitehead Institute)

Ensembl (www.ensembl.org/)



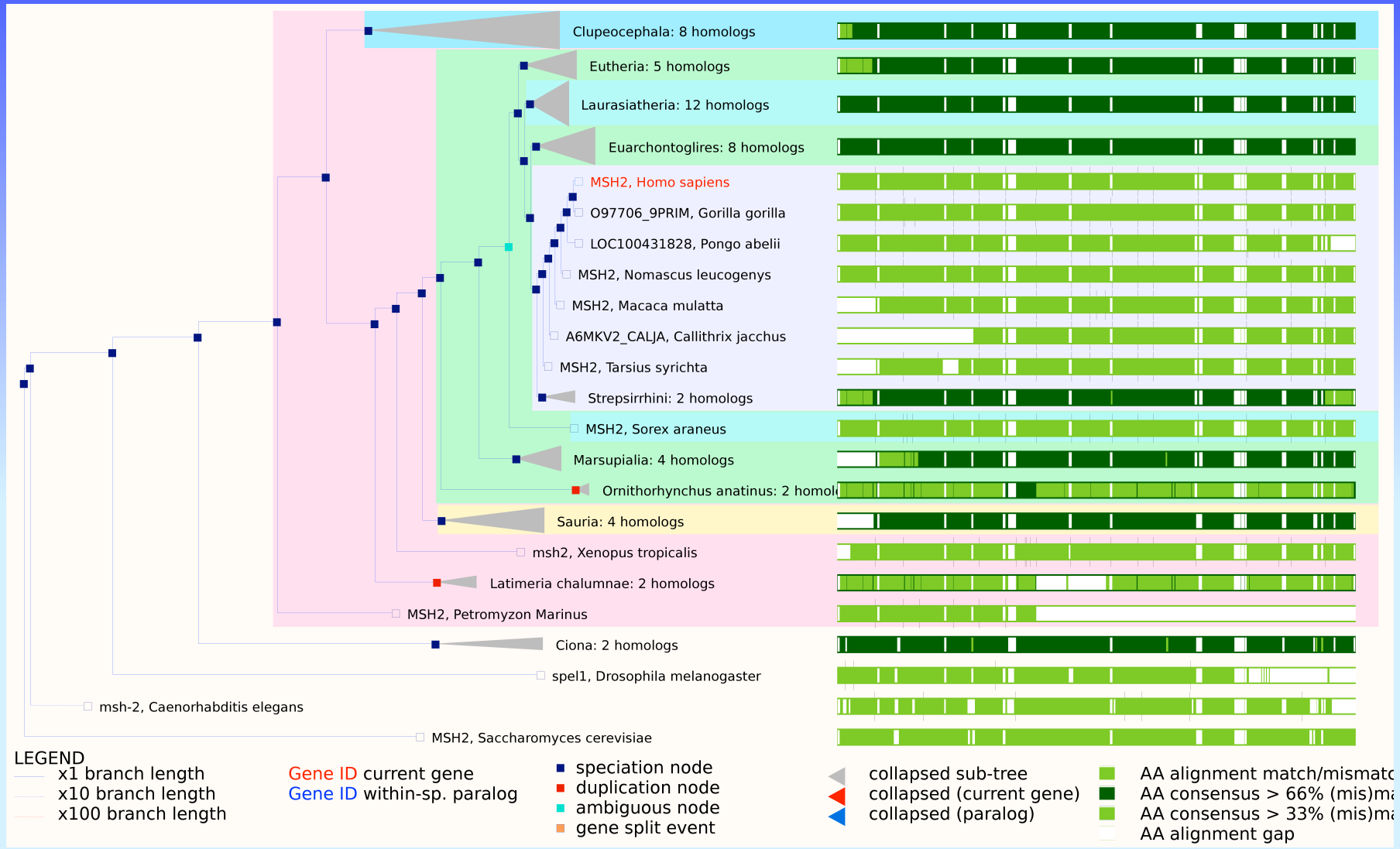
Institute)

Ensembl



Identifying and Aligning Homologs (Whitehead Institute)

Genome Research 2008: PMID 19029536



What I hope you learned

- What we can learn from sequence alignments
- Fundamentals of alignments
- Tools for building alignments

References

1. Web links within handout plus
   - Local EMBOSS - http://iona.wi.mit.edu/bio/tools/emboss/
   - Local NCBI - http://tak.wi.mit.edu/blast/
2. SOPs - https://gir.wi.mit.edu/trac/wiki/barc/SOPs
3. What's on tak - https://tak.wi.mit.edu/trac/wiki/Packages
4. Help pages for each application
   - Click, click, click
5. Genome Technology - http://jura.wi.mit.edu/bio/education/GT/
6. Marketa Zvelebil and Jeremy O. Baum – Understanding Bioinformatics, Garland Science 2008

Journal Articles

- Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005;39:309-38. PMID: 16285863
- Ensembl- *NAR* 2012, PMID: 22086963
- Homologene – *NAR* 2012, PMID: 22140104
- Clustal Omega –*Mol Syst Biol*, PMID: 21988835
- T-Coffee – *NAR* 2011, PMID: 21558174
- Challenges for MSA in hi-thruput era – *Bioinformatics* 2009, PMID: 19648142
- MSA - *Curr Opin Struct Biol*. 2006, PMID: 16679011