

Practical Bioinformatics

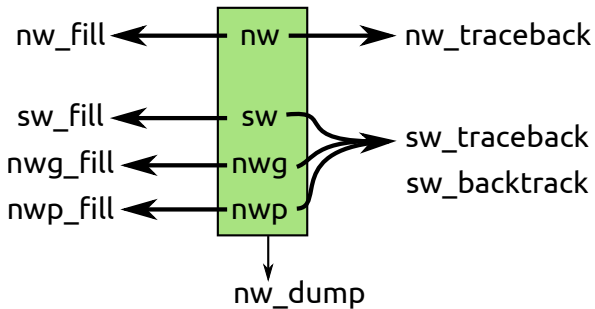
Mark Voorhies

5/9/2017

Habits are things you get for free, without requiring any special work.

–Cory Doctorow Advice to Writers, 4/5/2012

Functions in homework solution (dp2.py)



makelident

gapped_score

Needleman-Wunsch with $g=0$

	A	G	C	G	G	T	A	
G								
A								
G								
C								
G								
G								
A								

Needleman-Wunsch with $g=0$

	A	G	C	G	G	T	A
G	0						
A							
G							
C							
G							
G							
A							

```
def nw_fill(seq1, seq2, s, e):  
    # Initialize dp matrix  
    #     first dimension = seq1 positions  
    #     second dimension = seq2 positions  
    #     m[i][j] = best score for subalignment  
    #               of seq1[:i], seq2[:j]  
  
    m = [[0]]  
    # Initialize pointer matrix, a two dimensional  
    # matrix of lists of (row, column) pointers  
    p = [[None]]
```

Needleman-Wunsch with $g=0$

	A	G	C	G	G	T	A	
	0	-1	-2	-3	-4	-5	-6	-7
G								
A								
G								
C								
G								
G								
A								

```
# Fill first row as leading gaps  
for j in range(len(seq2)):  
    m[-1].append(m[0][j]+e)  
    p[-1].append([(0, j)])
```

Needleman-Wunsch with $g=0$

	A	G	C	G	G	T	A	
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	-1	0	-1	-2	-3	-4	-5
A	-2	0	-1	-1	-2	-3	-4	-3
G	-3	-1	1	0	0	-1	-2	-3
C	-4	-2	0	2	1	0	-1	-2
G	-5	-3	-1	1	3	2	1	0
G	-6	-4	-2	0	2	4	3	2
A	-7	-5	-3	-1	1	3	3	4

```
for i in range(len(seq1)):
    # First column is leading gaps
    m.append([m[i][0]+e])
    p.append([(i,0)])
    for j in range(len(seq2)):
        # Score for aligning seq1[i] with seq2[j]
        match = m[i][j]+s[seq1[i]][seq2[j]]
        # Score for aligning seq1[i] with a gap
        hgap = m[i+1][j]+e
        # Score for aligning seq2[i] with a gap
        vgap = m[i][j+1]+e

        best = max(match, vgap, hgap)
        m[-1].append(best)
        p[-1].append([])

        if (match == best):
            p[-1][-1].append((i, j))
        if (hgap == best):
            p[-1][-1].append((i+1, j))
        if (vgap == best):
            p[-1][-1].append((i, j+1))
```

Needleman-Wunsch with $g=0$

	A	G	C	G	G	T	A	
G	0	-1	-2	-3	-4	-5	-6	-7
A	-1	-1	0	-1	-2	-3	-4	-5
G	-2	0	-1	-1	-2	-3	-4	-3
C	-3	-1	1	0	0	-1	-2	-3
G	-4	-2	0	2	1	0	-1	-2
G	-5	-3	-1	1	3	2	1	0
G	-6	-4	-2	0	2	4	3	2
A	-7	-5	-3	-1	1	3	3	4

```
# Start at bottom right corner
curpos = (len(seq1), len(seq2))
aligned1 = ""
aligned2 = ""
```

```
exitFlag = False
for i in range(len(seq1)+len(seq2)):
    plist = p[curpos[0]][curpos[1]]
    if(plist is None):
        exitFlag = True
        break
    nextpos = plist[0]
    # Check for vgap
    if(nextpos[0] == curpos[0]):
        aligned1 = "-" + aligned1
    else:
        aligned1 = seq1[nextpos[0]] + aligned1
    # Check for hgap
    if(nextpos[1] == curpos[1]):
        aligned2 = "-" + aligned2
    else:
        aligned2 = seq2[nextpos[1]] + aligned2
    curpos = nextpos
```

```
if(exitFlag == False):
    print "WARNING: Unexpected exit from traceback"
```

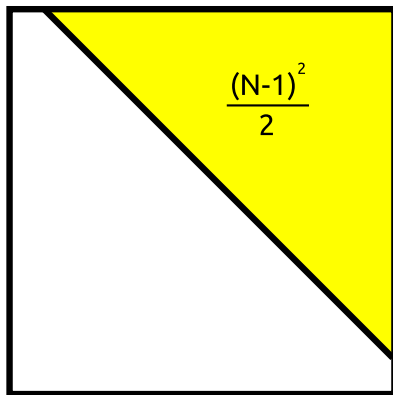

Evolution implies a self-consistent model



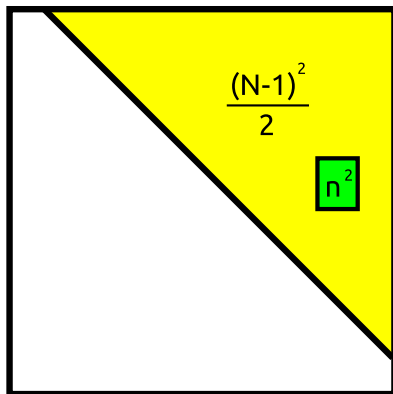
Distances
(Pairwise relationships)

Topology
(Evolutionary history)

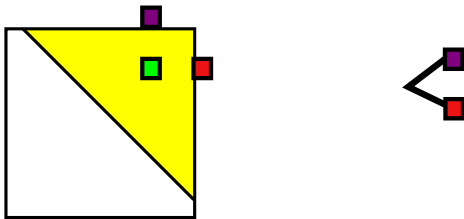
Measure all pairwise distances by dynamic programming



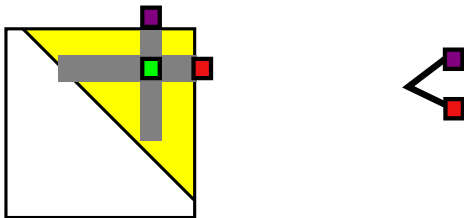
Measure all pairwise distances by dynamic programming



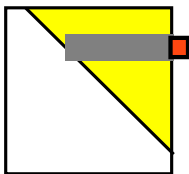
Generate a guide tree by UPGMA



Generate a guide tree by UPGMA



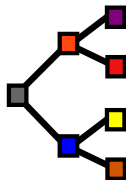
Generate a guide tree by UPGMA



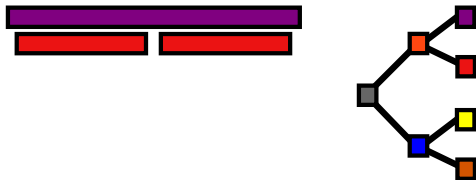
Generate a guide tree by UPGMA



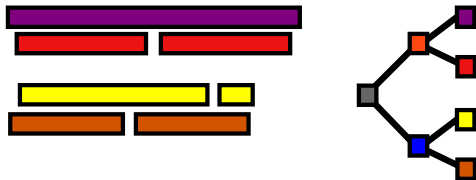
Generate a guide tree by UPGMA



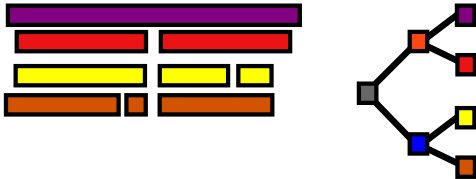
Progressive alignment following the guide tree



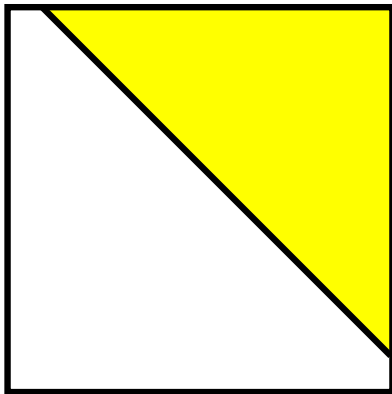
Progressive alignment following the guide tree



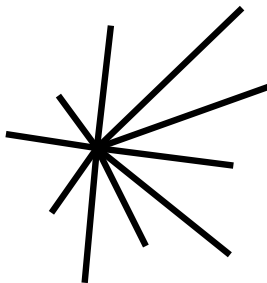
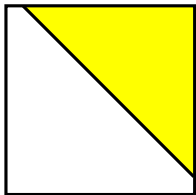
Progressive alignment following the guide tree



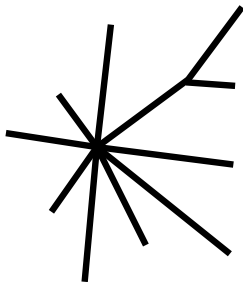
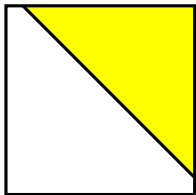
Measure distances directly from the alignment



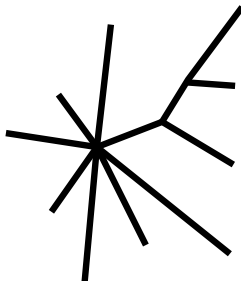
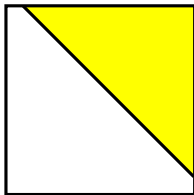
Generate neighbor-joining tree from new distances



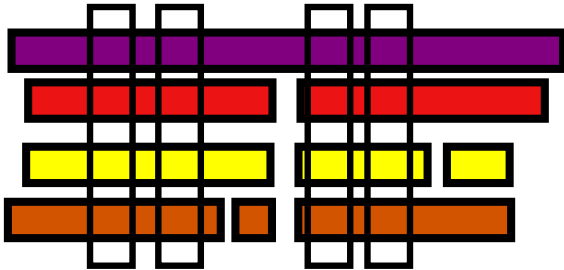
Generate neighbor-joining tree from new distances



Generate neighbor-joining tree from new distances



Generate bootstrap values from subsets of the alignment



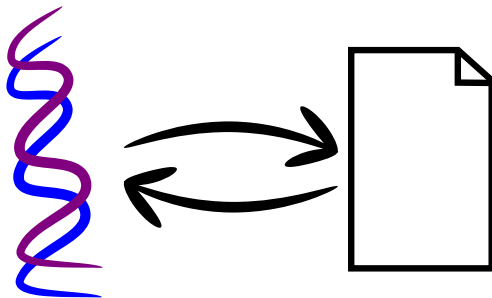
Viewing the alignment and tree in JALVIEW

The screenshot displays the JALVIEW software interface. The main window shows a multiple sequence alignment of protein sequences from various sources, including *HCAG_046862.702*, *HCDO_03083702-2445*, *Contig0_30_Fgenseh_Neurospora_1-702*, *HCIG_03340702-1221*, *gi417153|sp|P33125.1|HSPB2_A1-679*, *INSTD_FE_Contig19_Fgenseh_insl/2-1614*, *BCDO_011599702-704*, *BOBO_04548702-704*, *PADO_07715702-671*, *PADO_06249702-290*, *PADO_05679702-495*, *CIMG_047292-702*, *Af3g0421702-706*, *HC0901020006201-699*, *AN809.2/2-700*, *JMG_06759_5_2-702*, *JNC00142.2_1-705*, *gk19_65252-707*, *YMR186W2-705*, *SPAC206.04c2-704*, *CIMG_06159_1_1-699*, and *Dirg_A1-624*. The alignment is color-coded by conservation, with a 'Conservation Colour Increment (Background)' dialog box open, showing a slider set to 55 and the 'Apply to All Groups' option checked. Below the alignment, there are three bar charts: 'Conservation' (yellow), 'Quality' (yellow), and 'Consensus' (black). The 'Consensus' bar shows the sequence: `--AAM+G--...-ETFEFOAEISQLLSLINTVYSNKEIFLRELINSQDALDKIYEALSDPSKLDNSKDLRIDIPDKNKTL`. The sequence ID is `gi417153|sp|P33125.1|HSPB2_A1`. On the right side, a 3D ribbon diagram of a protein structure is shown, with a 'Jmol' label at the bottom right. The ribbon is colored by conservation, matching the alignment. The Jmol window title is 'Jmol A:2IQ' and it has a menu bar with 'File View Colours Help'.



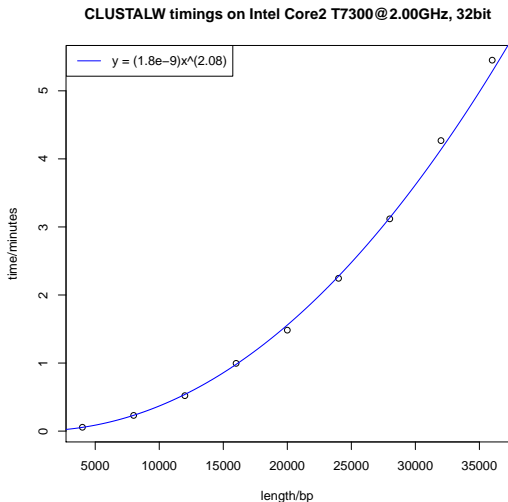
- For tools:
 - Read the manual
 - Read the paper
- Good general references:
 - The O'Reilly BLAST book
 - Durbin, Eddy, Krogh, and Mitcheson (HMMs)
 - Numerical Recipes
 - Branden & Tooze

Every object should have an isomorphism to a file

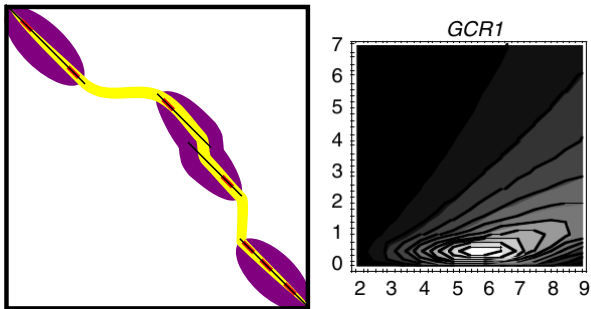


- Export, audit, edit, and import *independent* of a given program.
- Standard file formats for portability.
- Don't be afraid to look inside and hack on *your* data files.

Run times are predictable and measurable

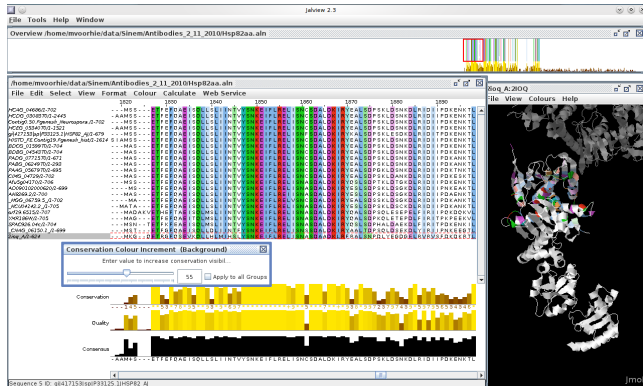


Heuristics and stochastic sampling for hard problems



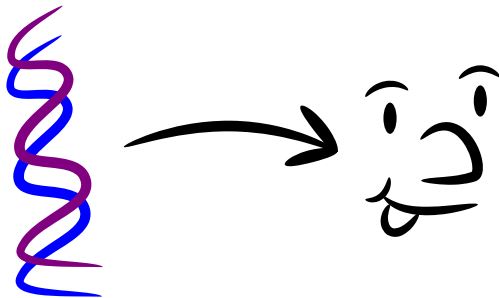
BLAST, HMMer3, simulated annealing, ...

Evolution is a rich source of information

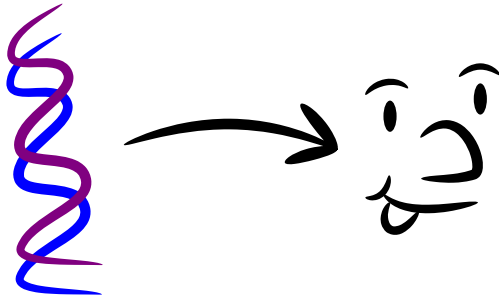


- Infer homology from sequence similarity
- More sequences provide more information

Phenotype is more diverse than Genotype

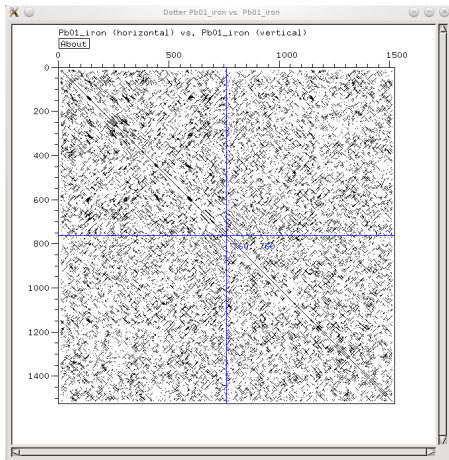


Phenotype is more diverse than Genotype



- Make sure you know what you are measuring
- Nucleic acid sequences are especially easy to address
- Many phenotypes can be analyzed by common numerical methods

Start from an unbiased view



Tools should support aggregation and annotation

Java TreeView : /home/karmic/Projects/Courses/BmsMinicourse/examples2/supp2data.cdt

File Settings Analysis Export Window Help

Dendrogram

View Status
Row: 115 (YLR028C)
Column: 49 (sp0_10)
Value: 1.34

Usage Hints
Mouse over to get info

cdcl5_100
cdcl5_100
cdcl5_200
cdcl5_200
cdcl5_250
cdcl5_270
cdcl5_290
spo_0
spo_5
spo_7
spo_9
spo_11
spo5_2
spo5_11
spo-early
spo-mid
heat_0
heat_10
heat_20
heat_40
heat_60
heat_100

YFR028C CDC14 MITOSIS PROTEIN PHOS
YML069W ORC1 DNA REPLICATION ORIGIN F
YIL139C REV7 DNA REPAIR DNA POLYMEF
YNL318C NONE TRANSPORT HEXOSE PERM
YFR023W PES4 DNA REPLICATION UNKNOWN:
YHR015W MIP6 MRNA EXPORT, PUTATIVE RNA
YDR263C DLM7 DNA REPAIR (PUTATIVE) DNA
YLR045C STU2 CYTOSKELETON SPINDLE
YOR033C DHS1 DNA REPAIR EXONUCLEASE
YIL159W BNR1 CYTOSKELETON ACTIN FI
YKL042W SPC42 CYTOSKELETON SPINDLE
YML225C CNM67 CYTOSKELETON SPINDLE
YCR092C CDC10 CYTOKINESIS GTP BINDING
YLR210W CLB4 CELL CYCLE G2/M CYCLIN
YLR314C CDC3 CYTOKINESIS SEPTIN
YBR045C GIP1 GLUCOSE REPRESSION (PUTA
YDL159W CLB3 CELL CYCLE G2/M CYCLIN
YDR118W APC4 CELL CYCLE ANAPHASE-PF
YDR253C MET32 METHIONINE METABOLISM TRP
YMR190W CLK1 CYTOSKELETON SPINDLE
YDR113C PDS1 CELL CYCLE ANAPHASE-T

GeneTreeAnno: /home/karmic/Projects/Courses/BmsMinicourse/examples2/supp2data.cdt

Sporulation

Name Sporulation Annotation Genes upregulated in sporulation

NODEID	LEFT	RIGHT	CORRELAT...	NAME	ANNOTATI...
NODE184...	NODE184...	NODE152...	0.627369	Sporulation	Genes up...
NODE184...	NODE184...	GENE56X	0.627369		
NODE184...	NODE184...	NODE178...	0.627369		
NODE184...	NODE150...	GENE177...	0.627287		

Dock Close

- Follow computational methods as they evolve (Web of Science; RSS: PubMed, GEO, arXiv, ...)

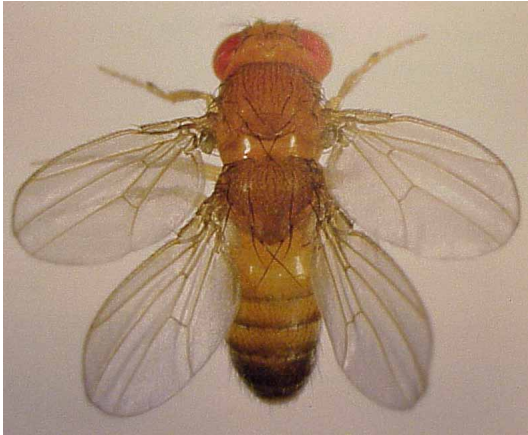
Science is a Conversation

- Follow computational methods as they evolve (Web of Science; RSS: PubMed, GEO, arXiv, ...)
- As a reviewer, insist on availability of source code

Science is a Conversation

- Follow computational methods as they evolve (Web of Science; RSS: PubMed, GEO, arXiv, ...)
- As a reviewer, insist on availability of source code
- Draw on your classmates' expertise

We understand systems by breaking them



Source: Peter A. Lawrence via <http://www.bio.davidson.edu/courses/molbio/ubx/ubx.html>