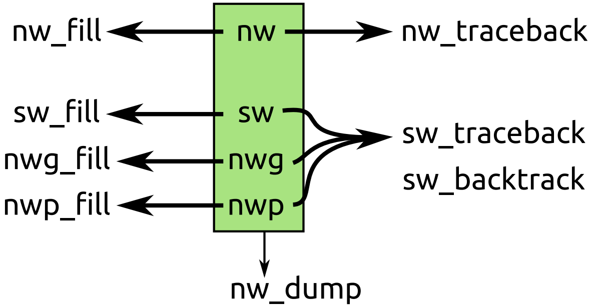
Practical Bioinformatics

Mark Voorhies

5/28/2015

Functions in homework solution (dp2.py)



makeIdent 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
	0							
G								
A								
G								
C								
G								
G								
A								

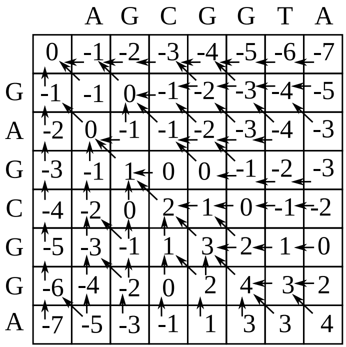
```
def nw_fill(seq1, seq2, s, e):
    # Initialize dp matrix
    #    first dimension  = seq1 positions
    #    second dimension = seq2 positions
    #    m[i][j] = best score for subalignment
    #              of seq1[:i],seq2[:j]

    m = [[0]]
    # Initialize pointer matrix, a two dimensional
    #   matrix of lists of (row, column) pointers
    p = [[None]]
```

Needleman-Wunsch with g=0

$$\begin{array}{c|c|c|c|c|c|c|c|c|}
 & & A & G & C & G & G & T & A \\
\hline
 & 0 & \leftarrow -1 & \leftarrow -2 & \leftarrow -3 & \leftarrow -4 & \leftarrow -5 & \leftarrow -6 & \leftarrow -7 \\
\hline
G & & & & & & & & \\
\hline
A & & & & & & & & \\
\hline
G & & & & & & & & \\
\hline
C & & & & & & & & \\
\hline
G & & & & & & & & \\
\hline
G & & & & & & & & \\
\hline
A & & & & & & & & \\
\hline
\end{array}$$

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

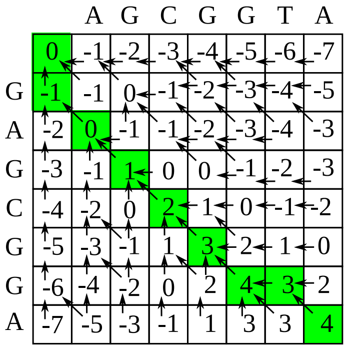
Needleman-Wunsch with g=0



		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))
```

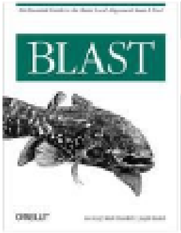
Needleman-Wunsch with g=0

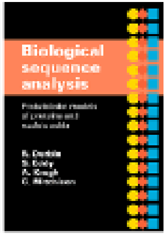


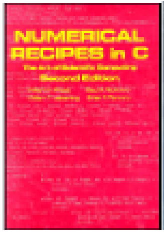
		A	G	C	G	G	T	A
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	-1	0	-1	-2	-3	-4	-5
A	-2	0	-1	-1	-2	-3	-4	-3
G	-3	-1	1	0	0	-1	-2	-3
C	-4	-2	0	2	1	0	-1	-2
G	-5	-3	-1	1	3	2	1	0
G	-6	-4	-2	0	2	4	3	2
A	-7	-5	-3	-1	1	3	3	4

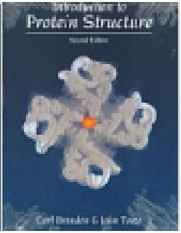
```python
# Start at bottom right corner
curpos = (len(seq1),len(seq2))
aligned1 = ""
aligned2 = ""

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"
```

RTFM

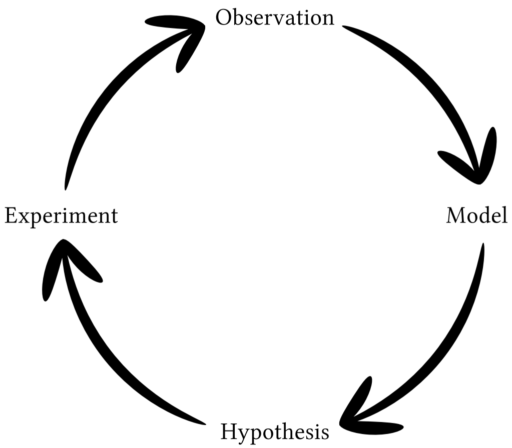




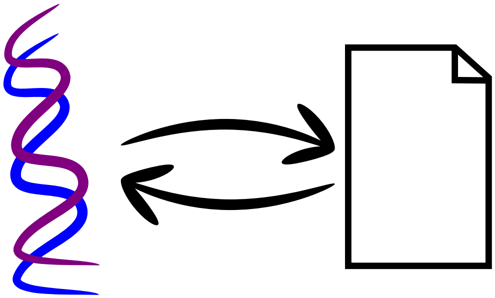




- For tools:
  - Read the manual
  - Read the paper
- Good general references:
  - The O'Reilly BLAST book
  - Durbin, Eddy, Krogh, and Mitcheson (HMMs)
  - Numerical Recipes
  - Branden & Tooze

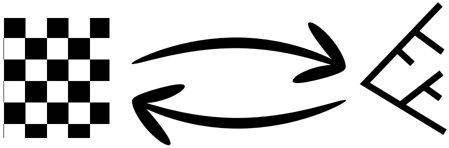

Observation
Model
Hypothesis
Experiment

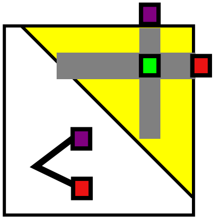
Every object should have an isomorphism to a file

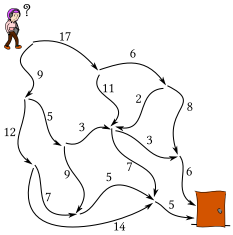


- Export, audit, edit, and import *independent* of a given program.
- Standard file formats for portability.
- Don't be afraid to look inside and hack on *your* data files.

A few techniques can solve many problems

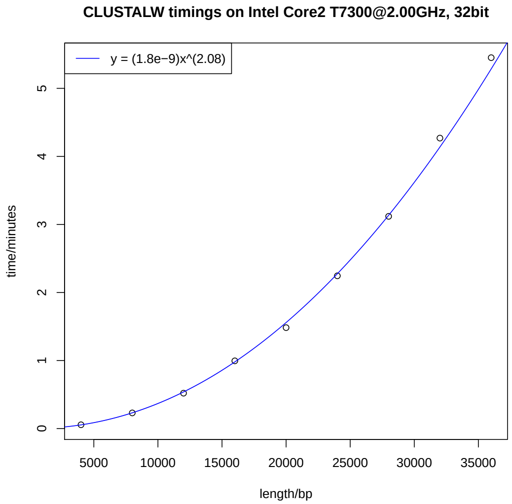




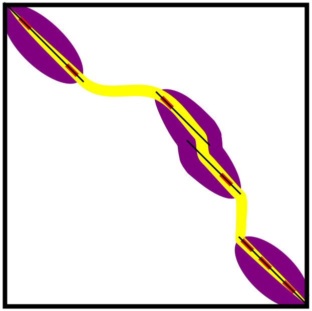


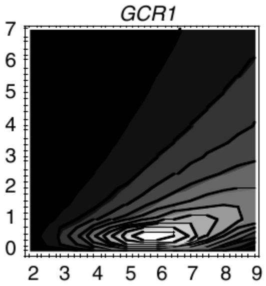
Iteration, clustering, dynamic programming, ...

Run times are predictable and measurable



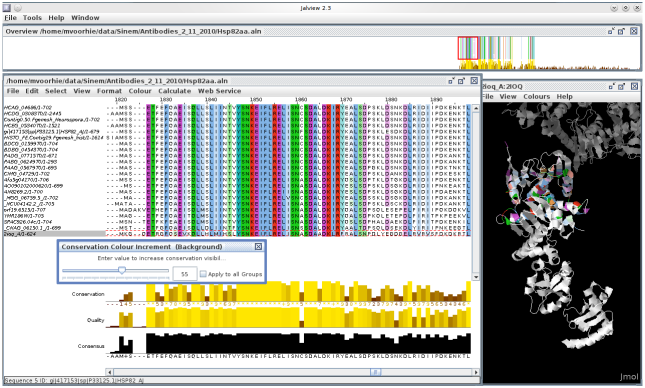
Heuristics and stochastic sampling for hard problems





BLAST, HMMer3, MrBayes, ...

Evolution is a rich source of information


Jalview 2.3
File Tools Help Window
Overview /home/mvoorhie/data/Sinem/Antibodies_2_11_2010/Hsp82aa.aln
/home/mvoorhie/data/Sinem/Antibodies_2_11_2010/Hsp82aa.aln
File Edit Select View Format Colour Calculate Web Service
2ioq_A:2IOQ
File View Colours Help
Jmol

```
                                          1820      1830      1840      1850      1860      1870      1880      1890
HCAG_04686/1-702                          ---MSS---ETFEFQAEISQLLSLIINTVYSNKEIFLRELISNCSDALDKIRYEALSDPSKLDSNKDLRIDIIPDKENKTL
HCDG_03083T0/1-2445                       -AAMSS---ETFEFQAEISQLLSLIINTVYSNKEIFLRELISNCSDALDKIRYEALSDPSKLDSNKDLRIDIIPDKENKTL
Contig0.50.Fgenesh_Neurospora./1-702      ---MSS---ETFEFQAEISQLLSLIINTVYSNKEIFLRELISNCSDALDKIRYEALSDPSKLDSNKDLRIDIIPDKENKTL
HCEG_05340T0/1-1521                       -AAMSS---ETFEFQAEISQLLSLIINTVYSNKEIFLRELISNCSDALDKIRYEALSDPSKLDSNKDLRIDIIPDKENKTL
gi|417153|sp|P33125.1|HSP82_AJ/1-679      ---MSS---ETFEFQAEISQLLSLIINTVYSNKEIFLRELISNFSDALDKIRYKALSDPSKLESDKDLRIDITPDKENKTL
HISTO_FE.Contig19.Fgenesh_hist/1-1614     SIAMSS---ETFEFQAEISQLLSLIINTVYSNKEIFLRELISNCSDALDKIRYEALSDPSKLDSNKDLRIDIIPDKENKTL
BDCG_01599T0/1-704                        ---MAS---ETFEFQAEISQLLSLIINTVYSNKEIFLRELISNCSDALDKIRYEALSDPSKLDSNKDLRIDIIPDKENKTL
BDBG_04543T0/1-704                        ---MAS---ETFEFQAEISQLLSLIINTVYSNKEIFLRELISNCSDALDKIRYEALSDPSKLDSNKDLRIDIIPDKENKTL
PADG_07715T0/1-671                        ---MAS---ETFEFQAEISQLLSLIINTVYSNKEIFLRELISNCSDALDKIRYEALSDPSKLDSNKDLRIDIIPDKANKTL
PABG_06249T0/1-293                        ---MAS---ETFEFQAEISQLLSLIINTVYSNKEIFLRELISNCSDALDKIRYEALSDPSKLDSNKDLRIDIIPDKANKTL
PAAG_05679T0/1-695                        ---MAS---ETFEFQAEISQLLSLIINTVYSNKEIFLRELISNCSDALDKIRYEALSDPGKLDSNKDLRIDIIPDKTNKTL
CIMG_04729/1-702                          ---MAS---ETFEFQAEISQLLSLIINTVYSNKEIFLRELISNCSDALDKIRYEALSDPSKLDANKDLRIDIIPDKESKTL
Afu5g04170/1-706                          ---MSS---ETFEFQAEISQLLSLIINTVYSNKEIFLRELISNASDALDKIRYQSLSDPTKLDTGKDLRIDIIPDKENKTL
AO090102000620/1-699                      ----MS---ETFEFQAEISQLLSLIINTVYSNKEIFLRELISNASDALDKIRYESLSDPSKLDSGKDLRIDIIPNKEAKTL
AN8269.2/1-700                            ---MAS---ETFEFQAEISQLLSLIINTVYSNKEIFLREIISNASDALDKIRYESLSDPSKLDSGKDLRIDIIPDAENKTL
_MGG_06759.5_/1-702                       ----MA---ETFEFQAEISQLLSLIINTVYSNKEIFLRELVSNCSDALDKIRYEALSDSSKLDSGKDLRIDIIPDKANKTL
_NCU04142.2_/1-705                        --MATA---ETFEFQAEISQLLSLIINTVYSNKEIFLRELISNASDALDKIRYESLSDPSKLDSCKDLRIDIIPDKANKTL
orf19.6515/1-707                          ---MADAKVETHEFTAEISQLMSLIINTVYSNKEIFLRELISNASDALDKIRYQALSDPSQLESEPELFIRIIPQKDQKVL
YMR186W/1-705                             ---MAG---ETFEFQAEITQLMSLIINTVYSNKEIFLRELISNASDALDKIRYQALSDPKQLETEPDLFIRITPKPEEKVL
SPAC926.04c/1-704                         ---MSN--TETFKFEAEISQLMSLIINTVYSNKEIFLRELISNASDALDKIRYQSLSDPHALDAEKDLFIRITPDKENKIL
_CNAG_06150.1_/1-699                      ---MST---ETFGFQAEISQLLDLIINTFYSNKEIFLRELISNSSDALDKIRYAALTDPSQLDSEKDLYIRIIPNKEEGTL
2ioq_A/1-624                              ---MKG--QETRGFQSEVKQLLHLMIHSLYSNKEIFLRELISNASDAADKLRFRALSNPDLYEGDGELRVRVSFDKDKRTL

Conservation                              ---145---**53*78*95**93*9*697**********+9**5***7**+*938*99723797439*597956534346*
Quality
Consensus                                 -AAM+S---ETFEFQAEISQLLSLIINTVYSNKEIFLRELISNCSDALDKIRYEALSDPSKLDSNKDLRIDIIPDKENKTL
```

Conservation Colour Increment (Background)

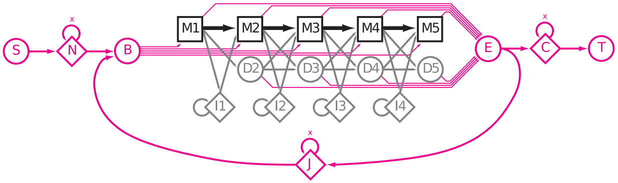
Enter value to increase conservation visibil...

55 ☐ Apply to all Groups

Sequence 5 ID: gi|417153|sp|P33125.1|HSP82_AJ

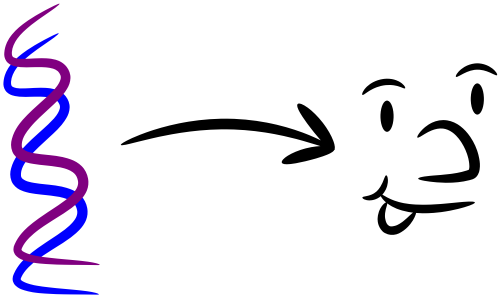
- Infer homology from sequence similarity
- More sequences provide more information

HMMs capture position and gap information

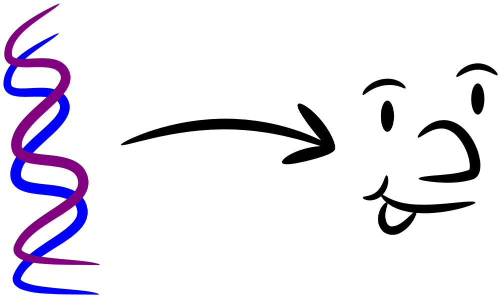


(Image from Sean Eddy, PLoS Comp. Biol. 4:e1000069)

Phenotype is more diverse than Genotype

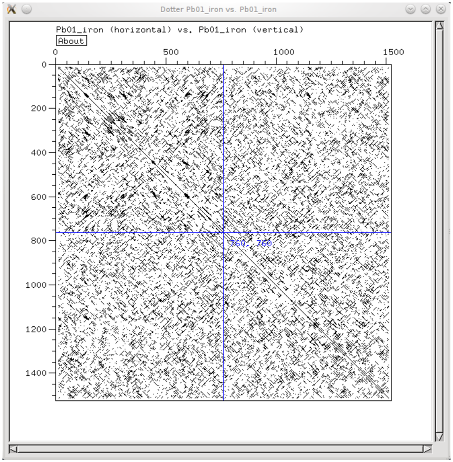


Phenotype is more diverse than Genotype

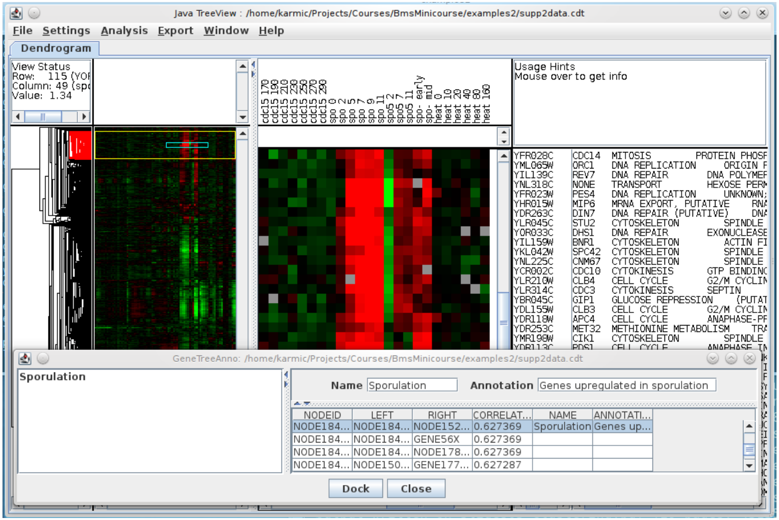


- Make sure you know what you are measuring
- Nucleic acid sequences are especially easy to address
- Many phenotypes can be analyzed by common numerical methods

Start from an unbiased view


Dotter Pb01_iron vs. Pb01_iron
Pb01_iron (horizontal) vs. Pb01_iron (vertical)
About
0
500
1000
1500
0
200
400
600
800
1000
1200
1400
760, 760

Tools should support aggregation and annotation


Java TreeView : /home/karmic/Projects/Courses/BmsMinicourse/examples2/supp2data.cdt
File Settings Analysis Export Window Help
Dendrogram
cdc15 170
cdc15 190
cdc15 210
cdc15 230
cdc15 250
cdc15 270
cdc15 290
spo 0
spo 2
spo 5
spo 7
spo 9
spo 11
spo5 2
spo5 7
spo5 11
spo- early
spo- mid
heat 0
heat 10
heat 20
heat 40
heat 80
heat 160
GeneTreeAnno: /home/karmic/Projects/Courses/BmsMinicourse/examples2/supp2data.cdt
Dock
Close

```
View Status
Row:    115 (YOR
Column: 49 (spo
Value:  1.34
```

```
Usage Hints
Mouse over to get info
```

```
YFR028C  CDC14  MITOSIS                 PROTEIN PHOSP
YML065W  ORC1   DNA REPLICATION            ORIGIN R
YIL139C  REV7   DNA REPAIR              DNA POLYMER
YNL318C  NONE   TRANSPORT               HEXOSE PERM
YFR023W  PES4   DNA REPLICATION            UNKNOWN;
YHR015W  MIP6   MRNA EXPORT, PUTATIVE      RNA
YDR263C  DIN7   DNA REPAIR (PUTATIVE)      DNA
YLR045C  STU2   CYTOSKELETON               SPINDLE
YOR033C  DHS1   DNA REPAIR              EXONUCLEASE
YIL159W  BNR1   CYTOSKELETON               ACTIN FI
YKL042W  SPC42  CYTOSKELETON               SPINDLE
YNL225C  CNM67  CYTOSKELETON               SPINDLE
YCR002C  CDC10  CYTOKINESIS             GTP BINDING
YLR210W  CLB4   CELL CYCLE              G2/M CYCLIN
YLR314C  CDC3   CYTOKINESIS             SEPTIN
YBR045C  GIP1   GLUCOSE REPRESSION        (PUTAT
YDL155W  CLB3   CELL CYCLE              G2/M CYCLIN
YDR118W  APC4   CELL CYCLE              ANAPHASE-PR
YDR253C  MET32  METHIONINE METABOLISM      TRA
YMR198W  CIK1   CYTOSKELETON               SPINDLE
YDR113C  PDS1   CELL CYCLE              ANAPHASE IN
```

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		

Groovy Packages

- numpy
- matplotlib
- scipy
- networkx
- h5py
- rpy
- pandas
- Pycluster (and BioPython)
- MySQLdb
- scikit-image
- scikit-learn

Science is a Conversation

- Follow computational methods as they evolve (Web of Science, PubMed RSS, arXiv, Haldane's Sieve...)

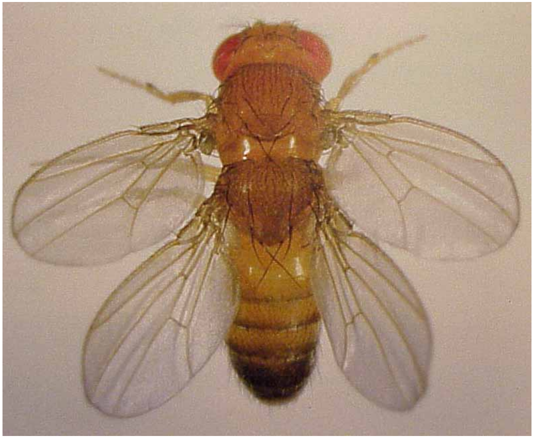
Science is a Conversation

- Follow computational methods as they evolve (Web of Science, PubMed RSS, arXiv, Haldane's Sieve...)
- As a reviewer, insist on availability of source code

Science is a Conversation

- Follow computational methods as they evolve (Web of Science, PubMed RSS, arXiv, Haldane's Sieve...)
- 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