Syllabus

- Sep 30 Introduction & Genome Assembly
- Oct 2 Sequence Comparison
- Oct 7 Gene Modeling
- Oct 9 Gene Function Identification – Read intro to HMM on blackboard
- Oct 14 OCTOBER BREAK
- Oct 16 Comparative Genomics
- Oct 21 Homology, alignments and motifs
- Oct 25 Pathways and PPI
- Oct 28 Structural Genomics / Protein Structure Prediction
- Nov 4 Protein Modeling
- Nov 8 EXAM

- Gribskov@purdue.edu – Lilly G-233

Genome Annotation

Genomics

- Please sign your work by using 4 "~" (tildes). for me the four tildes become 'gribskov 09 October 2008'.
- If you change the title (e.g., smo9 project) this will create a new page with that name. That's OK, but update the link below in the *Sequence files* section
- See the user manual for help with creating and editing pages (yes you can)

- Tools

Pending tasks

- ~~Post sequences(Gribskov)~~ - uploaded smo1-smo8, gribskov 09 October 2008
- ~~Create Accounts (Gribskov)~~ - added all accounts, gribskov 09 October 2008
- choose your partner and sequence (class)
- go to your project page, add your names as "annotators", and start working

Sequence files

Each is approximately 100kb. Sequences are from *Selaginella moellendorfii.* To indicate your choice of project, add the names of each team after the project by having each partner edit this page and add four tildes(~). The wiki will automatically replce the tildes with your name and the current date (3 would get just your name). Click on the *project page* link and start adding material to your page.

- smo1 project page jbrazelt 20 October 2008 csilva 20 October 2008
- smo2 project page
- smo3 project page - gandino 09 October 2008 shorter 20 October 2008
- smo4 project page
- smo5 project page -pkongsil 15 October 2008nyookong 17 Oct 2008
- smo6 project page - lin104 10 October 2008 wang4 10 October 2008
- smo7 project page
- smo8 project page - afahey 20 Oct 2008mandreat 20 October 2008

Missing
dmukherj
sowmya

example

- smo9 project page - gribskov 09 October 2008 gribskov 16 October 2008

Suggestions

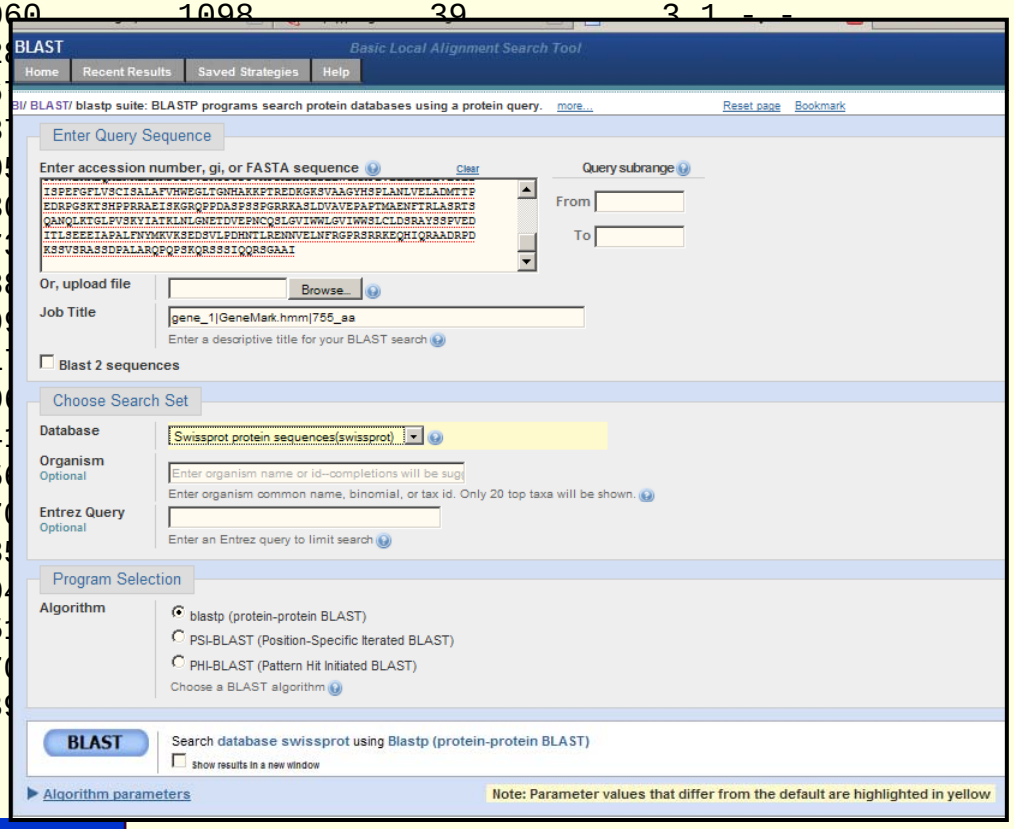
- See the user manual for help with creating and editing pages
- To get a sense of how to use the wiki
  - Add something to your personal page (see my page at upper left)
  - Use your personal page as a to-do list
  - Restrict access to your to-do list

Genome Annotation

Genomics

Gene1 - Genemark

```
1  20   -  Terminal        553        696        144          3 1 - -
 1  19   -  Internal       1060       1098         39          3 1 - -
 1  18   -  Internal       12
 1  17   -  Internal       16
 1  16   -  Internal       18
 1  15   -  Internal       20
 1  14   -  Internal       23
 1  13   -  Internal       27
 1  12   -  Internal       28
 1  11   -  Internal       29
 1  10   -  Internal       31
 1   9   -  Internal       40
 1   8   -  Internal       44
 1   7   -  Internal       45
 1   6   -  Internal       47
 1   5   -  Internal       48
 1   4   -  Internal       50
 1   3   -  Internal       55
 1   2   -  Internal       57
 1   1   -  Initial        58
```

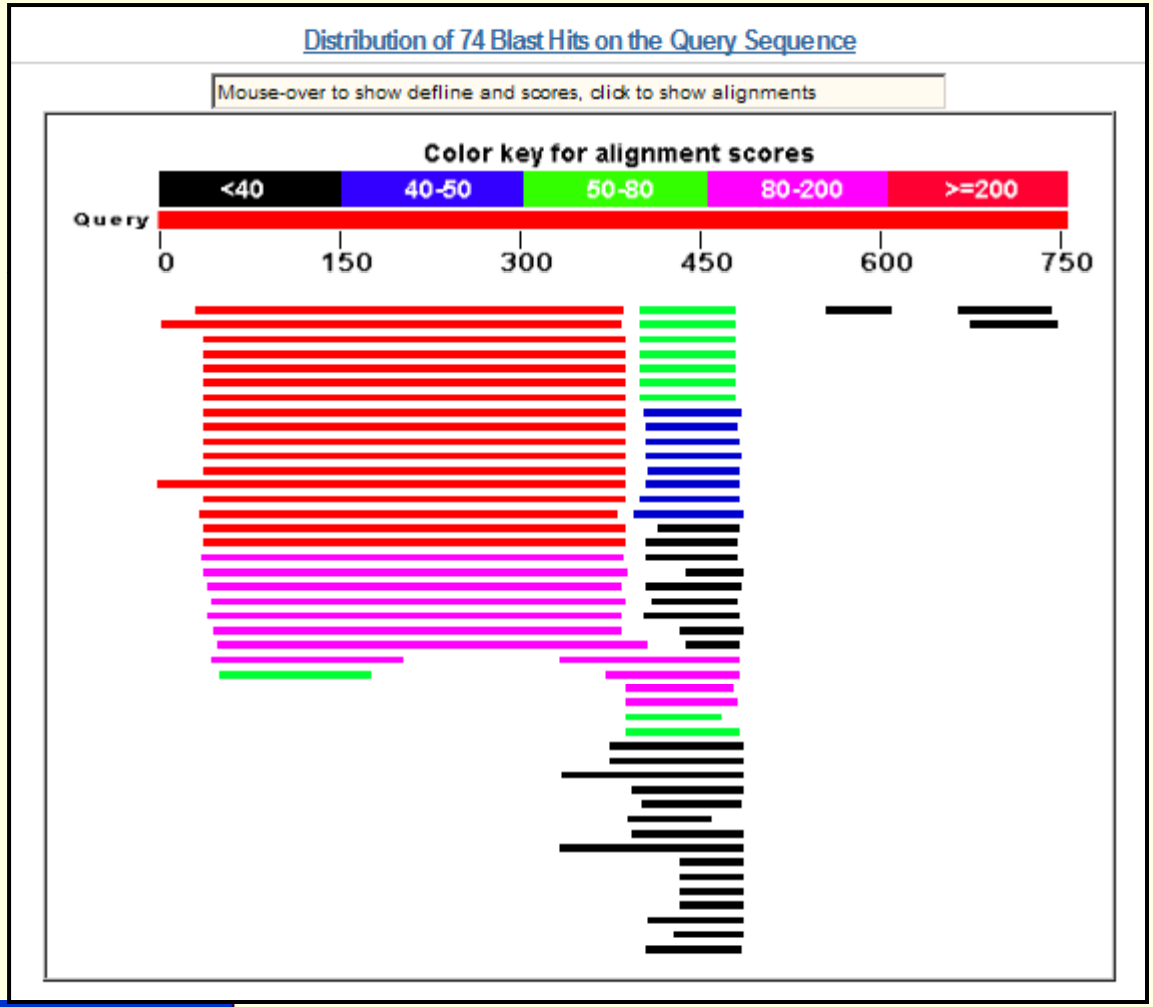


Gribskov 6.3

Genome Annotation

Genomics

BLAST vs Swiss-Prot



Gribskov 6.4

Genome Annotation

BLAST Search vs SwissProt

```
>sp|Q67ZU1|LIP2_ARATH Triacylglycerol lipase 2 precursor
Length=418

 GENE_ID: 831268 MPL1 | MPL1 (MYZUS PERSICAE-INDUCED LIPASE 1); catalytic
[Arabidopsis thaliana]

 Score =  314 bits (805),  Expect = 1e-84, Method: Compositional matrix adjust.
 Identities = 151/364 (41%), Positives = 223/364 (61%), Gaps = 8/364 (2%)

Query  33   RFAEDSFCSTLVLVHGYPCQEFKVTTPDGYILRVHRIPHGVAGVSSPSP---KPVFLQHG  89
            R A    C++ V + GY C+E  V T DGYIL + RIP G AG  +       +PV +QHG
Sbjct  45   RTAAGGICASSVHIFGYKCEEHDVVTQDGYILNMQRIPEGRAGAVAGDGGKRQPVLIQHG  104

Query  90   VLQGGDDWVFYPPRNSLGFVLADEGFDVWIGNLRGTHWSRQHVSYSSGDKAYWDWTWDEH  149
            +L  G  W+  P   +L  +LAD+GFDVW+GN RGT +SR+H   +   +A+W+WTWDE
Sbjct  105  ILVDGMSWLLNPADQNLPLILADQGFDVWMGNTRGTRFSRRHKYLNPSQRAFWNWTWDEL  164

Query  150  AQYDLPAMLNLVHENTGSELYYVGHSQGTLIALAAFSESKLMNVVRAAVLLSPIAYLKGM  209
              YDLPAM + +H  TG +++Y+GHS GTLI  A+FSE  L++ VR+A +LSP+AYL  M
Sbjct  165  VSYDLPAMFDHIHGLTGQKIHYLGHSLGTLIGFASFSEKGLVDQVRSAAMLSPVAYLSHM  224

Query  210  TSTLSRLAALLYMDQIYDDIGLSQFTLESGIGAYLLRNLC-SLDPRCADLLVLVTGRNCC  268
            T+ +  +AA  ++ +    +G  +F  +SG+    ++ +C      C DL+ ++TG+NCC
Sbjct  225  TTVIGDIAAKTFLAEATSILGWPEFNPKSGLVGDFIKAICLKAGIDCYDLVSVITGKNCC  284

Query  269  FNASLTSYYRQFEPQGSSTKNLVHLAQMVRTGLFAKFDYGSSLGNIRAYSQVVPPTYEPA  328
             NAS    +   EPQ +STKN++HLAQ VR     K++YGSS  NI+ Y Q +PP Y  +
Sbjct  285  LNASTIDLFLANEPQSTSTKNMIHLAQTVRDKELRKYNYGSSDRNIKHYGQAIPPAYNIS  344

Query  329  NIPKSFPVFLVYGGKDTLSTAQGVQELAKRL----VCTQQTLFLPNYAHADFVVGTRARQ  384
             IP   P+F  YGG D+L+  + V+ L  +      +      F+ +YAHADF++G  A+
Sbjct  345  AIPHELPLFFSYGGLDSLADVKDVEFLLDQFKYHDIDKMNVQFVKDYAHADFIMGVTAKD  404

Query  385  DVFD  388
             V++
Sbjct  405  VVYN  408
```

Genome Annotation

Genomics

BLAST vs SwissProt

```
>sp|Q54GN2.1|KTHY_DICDI  RecName: Full=Thymidylate kinase; AltName: Full=dTMP kinase
Length=222

 Score = 97.1 bits (240),  Expect = 3e-19, Method: Composition-based stats.
 Identities = 57/151 (37%), Positives = 84/151 (55%), Gaps = 21/151 (13%)

Query  336  VFLVYGGKDTLSTAQGVQELAKRLVCTQQTLFLPNYAHADFVVGTRARQDVFDPLWPART  395
            +F+++ G D +  +  VQ L   +   Q+   LP          T++ +      +P RT
Sbjct  20   LFILFEGVDRVGKSTQVQSLTNHISNVQK---LP----------TKSLR------FPDRT  60

Query  396  SRLRCLL--FLGRGAELDDRAVHILFSPNRWEKRLQMENKLMARTTLVVDRYSYSGVAFS  453
            + +  ++  +L     +DDRA+H+LFS NRWE R  +   L   T +VVDRYSYSGVA+S
Sbjct  61   TPIGQIINQYLQNATNMDDRALHLLFSSNRWEARDSILELLNNGTNIVVDRYSYSGVAYS  120

Query  454  AAKGLDLEWCKAPEVELLAADLVLYLDISPE  484
            AAKG+D +WC A E   L   DL+ YL +S E
Sbjct  121  AAKGIDFDWCYACEKGLPKPDLIFYLSMSSE  151


>sp|P36590|KTHY_SCHPO  Thymidylate kinase (dTMP kinase)
Length=210

 GENE ID: 2538707 tmp | thymidylate kinase Tmp1 [Schizosaccharomyces pombe]
(10 or fewer PubMed links)

 Score = 88.2 bits (217),  Expect = 2e-16, Method: Compositional matrix adjust.
 Identities = 42/112 (37%), Positives = 71/112 (63%), Gaps = 4/112 (3%)

Query  374  ADFVVGTRARQDVFDPLWPARTSRL--RCLLFLGRGAELDDRAVHILFSPNRWEKRLQME  431
             D ++    + ++F   +P RT+ +   +   +L    +L+D+ +H+LFS NRWE    +
Sbjct  28   VDKLISQHEKAELFK--FPDRTTAIGKKIDDYLKESVQLNDQVIHLLFSANRWETIQYIY  85

Query  432  NKLMARTTLVVDRYSYSGVAFSAAKGLDLEWCKAPEVELLAADLVLYLDISP  483
             ++    T ++DRY++SG+AFSAAKGLD EWCK+P+  L   DLV++L++ P
Sbjct  86   EQINKGVTCILDRYAFSGIAFSAAKGLDWEWCKSPDRGLPRPDLVIFLNVDP  137


>sp|P00572|KTHY_YEAST  Thymidylate kinase (dTMP kinase)
Length=216

 GENE ID: 853520 CDC8 | Cdc8p [Saccharomyces cerevisiae] (Over 10 PubMed links)

 Score = 84.7 bits (208),  Expect = 2e-15, Method: Composition-based stats.
 Identities = 42/95 (44%), Positives = 62/95 (65%), Gaps = 6/95 (6%)

Query  391  WPARTSRLRCLL---FLGRGAELDDRAVHILFSPNRWEKRLQMENKLMARTTLVVDRYSY  447
            +P R++R+  L+          +L D+A+H+LFS NRWE   +++  L+    +V+DRY Y
Sbjct  38   FPERSTRIGGLINEYLTDDSFQLSDQAIHLLFSANRWEIVDKIKKDLLEGKNIVMDRYVY  97

Query  448  SGVAFSAAK---GLDLEWCKAPEVELLAADLVLYL  479
            SGVA+SAAK   G+DL+WC   P+V LL   DL L+L
Sbjct  98   SGVAYSAAKGTNGMDLDWCLQPDVGLLKPDLTLFL  132
```

Gribskov 6.6

FGeneSH – BLAST vs SwissProt

```
>sp|Q67ZU1|LIP2_ARATH Triacylglycerol lipase 2 precursor
Length=418

 GENE_ID: 831268 MPL1 | MPL1 (MYZUS PERSICAE-INDUCED LIPASE 1); catalytic
[Arabidopsis thaliana]

 Score =  251 bits (641),  Expect = 7e-66, Method: Compositional matrix adjust.
 Identities = 131/364 (35%), Positives = 195/364 (53%), Gaps = 55/364 (15%)

Query  35   RFAEDSFCSTLVLVHGYPCQEFKVTTPDGYILRVHRIPHGVAGVSSPSP---KPVFLQHG  91
            R A    C++ V + GY C+E  V T DGYIL + RIP G AG  +       +PV +QHG
Sbjct  45   RTAAGGICASSVHIFGYKCEEHDVVTQDGYILNMQRIPEGRAGAVAGDGGKRQPVLIQHG  104

Query  92   VL-------------------------------------------------QAYWDWTWDEH  104
            +L                                                 +A+W+WTWDE
Sbjct  105  ILVDGMSWLLNPADQNLPLILADQGFDVWMGNTRGTRFSRRHKYLNPSQRAFWNWTWDEL  164

Query  105  AQYDLPAMLNLVHENTGSELYYVGHSQGTLIALAAFSESKLMNVVRAAVLLSPIAYLKGM  164
              YDLPAM + +H  TG +++Y+GHS GTLI  A+FSE  L++ VR+A +LSP+AYL  M
Sbjct  165  VSYDLPAMFDHIHGLTGQKIHYLGHSLGTLIGFASFSEKGLVDQVRSAAMLSPVAYLSHM  224

Query  165  TSTLSRLAALLYMDQIYDDIGLSQFTLESGIGAYLLRNLC-SLDPRCADLLVLVTGRNCC  223
            T+ +  +AA  ++ +    +G  +F  +SG+    ++ +C      C DL+ ++TG+NCC
Sbjct  225  TTVIGDIAAKTFLAEATSILGWPEFNPKSGLVGDFIKAICLKAGIDCYDLVSVITGKNCC  284

Query  224  FNASLTSYYRQFEPQGSSTKNLVHLAQMVRTGLFAKFDYGSSLGNIRAYSQVVPPTYEPA  283
             NAS    +   EPQ +STKN++HLAQ VR     K++YGSS  NI+ Y Q +PP Y  +
Sbjct  285  LNASTIDLFLANEPQSTSTKNMIHLAQTVRDKELRKYNYGSSDRNIKHYGQAIPPAYNIS  344

Query  284  NIPKSFPVFLVYGGKDTLSTAQGVQELAKRL----VCTQQTLFLPNYAHADFVVGTRARQ  339
             IP   P+F  YGG D+L+  + V+ L  +      +       F+ +YAHADF++G  A+
Sbjct  345  AIPHELPLFFSYGGLDSLADVKDVEFLLDQFKYHDIDKMNVQFVKDYAHADFIMGVTAKD  404

Query  340  DVFD  343
             V++
Sbjct  405  VVYN  408
```

Sequence Alignment

Alignment

- Provides a one-to-one picture of the residues or bases in the sequences that correspond

```
 1 VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF.... 46
   .||.::.  | ..||||:|. .:.|.|.| |:| : |.| . |..|
 1 GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLK 50
             .          .          .          .
47 ..DLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVD 94
     |  .:|.::| || .| .||.. : . :. ...:.:|.: || | ::.
51 SEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIP 100
```

- Computational problem is putting in gaps

Sequence Alignment

Genomics

Dynamic Programming Alignment

- Calculation of alignment score using affine gap model

```
BOVGH  G G T G C C A C -  T C C C A - - - - - C T G
         : : : : : : :    : :   :             : : :
A02321 A G T G C C A C C  C C C A A T G C C G C T G
      -3+4+4+4+4+4+4+4-12-3+4+4-3+4  -12-4x4 +4+4+4
```

matches	52
mismatches	-9
gaps	-40
Overall Score	3

	A	C	G	T
A	4			
C	-3	4		
G	-3	-3	4	
T	-3	-3	-3	4

```
Gap opening   = -12
Gap extension = -4
```

Gribskov 6.9

Sequence Alignment

Genomics

Dynamic Programming Alignment

- Much faster than brute force but too slow for searching
- Dynamic Programming is a "Rigorous" method
  - **Rigorous: no approximations, all numbers and positions of gaps**
  - **EVERY possible alignment is considered**
  - **Requires time proportional to product of sequence lengths, O(nm)**
- Dynamic Programming yields an Optimal Alignment:
  - **Scores for matches/mismatches**
  - **Penalties for gaps**
  - **Score = (Matches x Match_Score) + (Mismatches x Mismatch_Score) + Gap_Score**
- Dynamic Programming uses an Affine Gap Model:
  - **Lower Penalty to EXTEND a Gap than to CREATE or Open a new Gap**
  - **Gap_Score = Gap_Open + Gap_extend x Exten_Length**
  - **Generally this is a negative score, i.e., a Penalty**

Gribskov 6.10

Sequence Alignment

Dynamic Programming Alignment

- If we know that a partial alignment is optimal, then we have only three choices for the next step:

```
Given two Sequences:     A G T G C       and        A G G C

and that we have                A G T
this partial                          G C
Alignment shown:                      G C
                              A G
```

Then we have the following three choices for the next step:

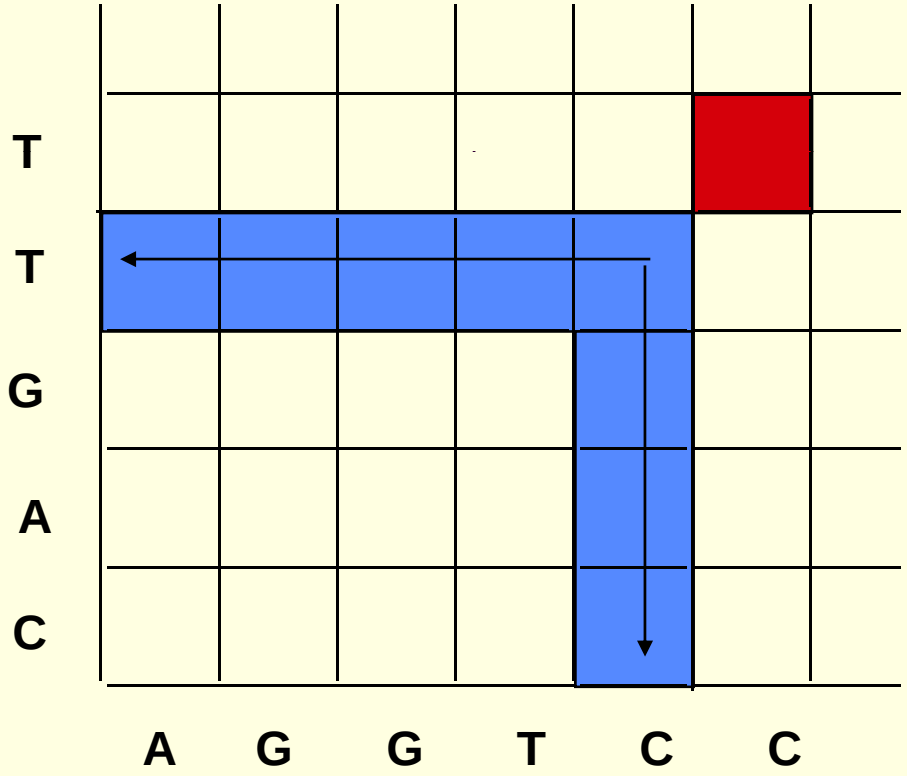
```
1) A G              2) A G              3) A G T
      T G C               T G C                 . G C
      G G C               . G C                 G G C
    A                   A G                     A
```

Sequence Alignment

Genomics

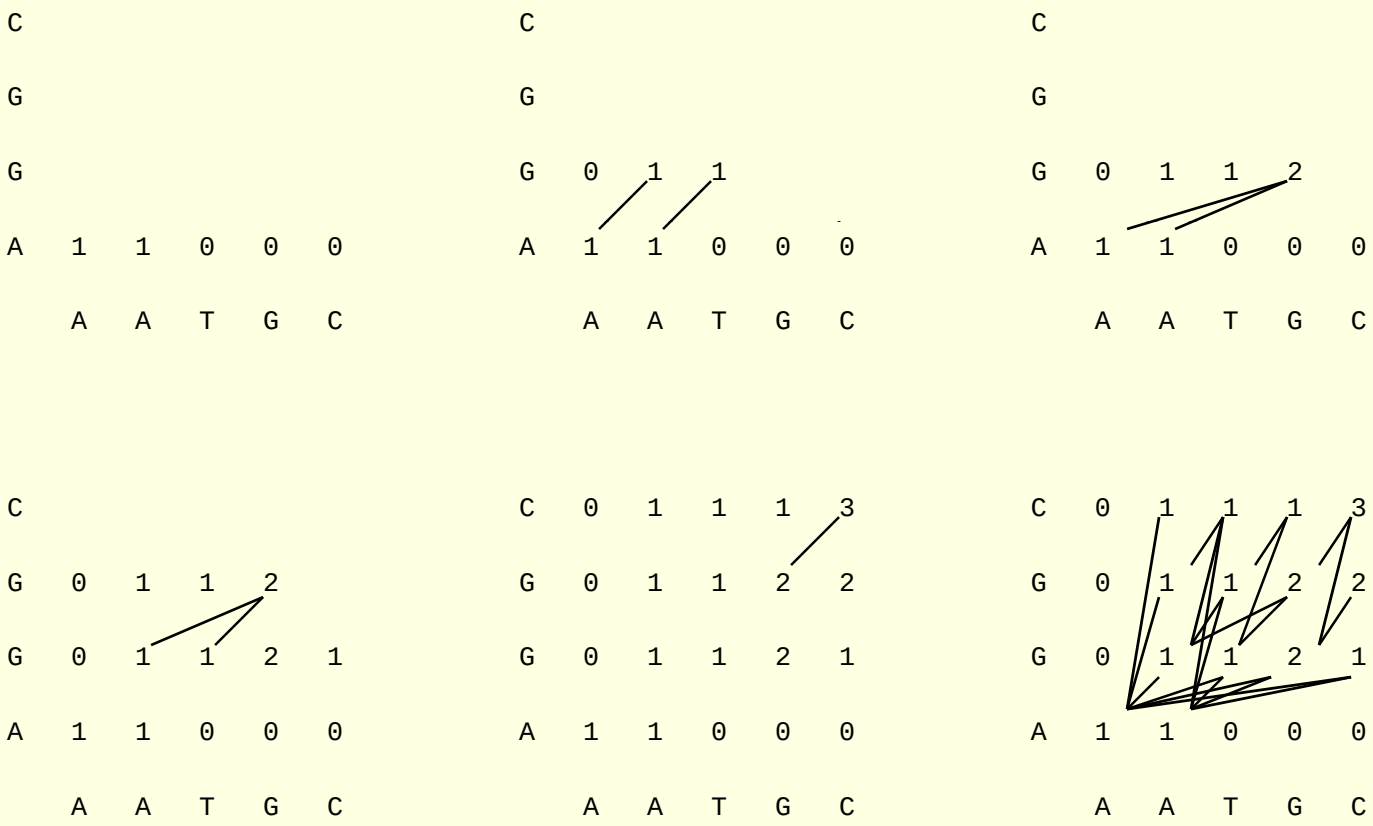
Dynamic Programming Alignment

- Forward fill of score matrix



Gribskov 6.12

Dynamic Programming Alignment



Gribskov 6.13

Sequence Alignment

Dynamic Programming Alignment

- Calculations: Posn Score + Max of Previous Column / Previous Row

```
C
G  0  1  1  2
G  0  1  1  2  1
A  1  1  0  0  0
   A  A  T  G  C
```

```
C
G  0  1  1  2  ?
G  0  1  1  2  1
A  1  1  0  0  0
   A  A  T  G  C
```

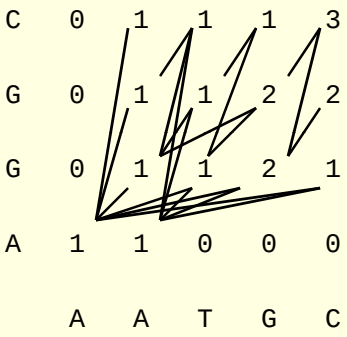
```
C
G  0  1  1  2  2
G  0  1  1  2  1
A  1  1  0  0  0
   A  A  T  G  C
```

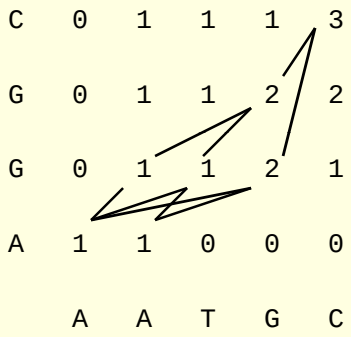
```
C  0  1  ?
G  0  1  1  2
G  0  1  1  2  1
A  1  1  0  0  0
   A  A  T  G  C
```

```
C  0  1  1  1  ?
G  0  1  1  2  2
G  0  1  1  2  1
A  1  1  0  0  0
   A  A  T  G  C
```

```
C  0  1  1  1  3 = 1 + 2
G  0  1  1  2  2
G  0  1  1  2  1
A  1  1  0  0  0
   A  A  T  G  C
```

Dynamic Programming Alignment





5 equivalent alignments, EVERY ONE IS OPTIMAL

AG.GC	**A.GGC**	**.AGGC**	**.A.GGC**	**A..GGC**
AATGC	**AATGC**	**AATGC**	**AATG.C**	**AATG.C**

Sequence Alignment

Scoring Alignments and Searches

- DNA – uses identity matching in most cases
- Protein – uses empirically derived scoring tables that capture complex chemical similarity of amino acid residues
  - **PAM system**
  - **BLOSUM system**
- Both can be tuned to be better for certain pairs of sequences

Scoring Systems - Log-odds matrices

- A log-odds scoring system evaluates the relative probabilities of a match representing true homology versus the chance that a match occurs at random, i.e. the relative probability of two models

$$s_{ij} = ln(q_{ij} / p_i \; p_j)$$

- Normally, one multiplies probabilities - since these are log probabilities you get the total probability by adding them up
- When added up over a matching segment, you get the probability that the segment represents homology relative to the probability that it represents a random match, i.e. how much more likely than chance is it that the matching segment represents homology

Sequence Alignment

Scoring

- Dayhoff's matrix
  - **Most influential and enduring, work began in late 60's**
  - **1978 version unchallenged until early 90's**
- BLOSUM
  - **May be slightly better than PAM**
- Others
  - **Gonnet**
  - **Physical properties**

Sequence Alignment

Scoring - Dayhoff / PAM

- Model: There is some universal underlying pattern of residues that can mutationally substitute
- Positions are independent
- The only thing that affects mutation is what the residue is now
- Look at closely related sequences so that alignments are unambiguous
  - **About 1 change / 100 residues**
  - **1 PAM (Percent Accepted Mutations)**
- 1572 changes in 71 families

Sequence Alignment

Scoring - Dayhoff/PAM

- 1 PAM represents a single step of evolution
- Extrapolate model to long evolutionary distances as a series of identical steps
- PAM250 (MDM78) is 250 PAM distance
- 120 PAM may be a bit better
- Log-odds form
  - **log[Pmodel(i,j) / Pexpect(i,j)]**
- Good statistical discriminators
- Sum over alignment gives log-probability ratio that alignment is better than chance (random model)
- Problems
  - **Not all positions are the same**
  - **Evolutionary rates vary greatly within a sequence**
  - **Most mutable positions were inadvertently selected**

Sequence Alignment

Scoring - BLOSUM (BLOcks SUbstitution Matrix)

- Based on PROSITE signatures
  - **Signatures are short expressions like C-X-X-C-X-X-X-C**
- locally align each feature to get "blocks"
- blocks contain sequences at all different evolutionary distances and may be highly biased (e.g. many identical sequences)

Scoring - BLOSUM

- Dealing with bias and distance
  - **Cluster all sequences with less than X% identities**

    Clustered sequences count as 1 sequence

    if X is 100% it simply removes identical sequences

    if X < 100% it reduces the weight on closely related sequences

    The bigger the X in BLOSUMX, the more suitable the scoring table is for very similar sequences
- Calculate substitution frequencies and log-odds matrix
- This gives a BLOSUM X table
  - **BLOSUM 62 - sequences greater than 62% identical are clustered**
  - **BLOSUM 80 - sequences greater than 80% identical are clustered**

Sequence Alignment

BLOSUM

```
TCMN_STRGA ( 331) IADLGGGDGWFLAQILRRHPHATGLLMDLPRVA  74
TCMO_STRGA ( 173) FVDLGGARGNLAAHLHRAHPHLRATCFDLPEME  81
ZRP4_MAIZE ( 204) LVDVGGGIGAAAQAISKAFPHVKCSVLDLAHVV  68
CHMT_POPTM ( 204) LVDVGGGTGAVVNTIVSKYPSIKGINFDLPHVI  41
COMT_EUCGU ( 205) VVDVGGGTGAVLSMIVAKYPSMKGINFDLPHVI  42
COMT_MEDSA ( 204) LVDVGGGTGAVINTIVSKYPTIKGINFDLPHVI  47
CRTF_RHOSH ( 205) LMDVGGGTGAFLAAVGRAYPLMELMLFDLPVVA  59
OMTA_ASPPA ( 250) VVDVGGGRGHLSRRVSQKHPHLRFIVQDLPAVI  47
```

METHYLTRANSFERASE BI

Unweighted (BLOSUMn) count of transitions for column 1

$c_{FF} = 0$ $c_{FI} = 1$ $c_{FL} = 4$ $c_{FV} = 2$

$c_{II} = 0$ $c_{IL} = 4$ $c_{IV} = 2$

$c_{LL} = 6$ $c_{LV} = 8$

$c_{VV} = 1$

BLOSUM

Unweighted (BLOSUMn) count of transitions for column 1

$c_{FF} = 0$	$c_{FI} = 1$	$c_{FL} = 4$	$c_{FV} = 2$
	$c_{II} = 0$	$c_{IL} = 4$	$c_{IV} = 2$
		$c_{LL} = 6$	$c_{LV} = 8$
			$c_{VV} = 1$

N = 28 transitions, $\mathbf{f_{ij} = c_{ij} / N}$

$f_{FF} = 0.00$	$f_{FI} = 0.04$	$f_{FL} = 0.14$	$f_{FV} = 0.07$
	$f_{II} = 0.00$	$f_{IL} = 0.14$	$f_{IV} = 0.07$
		$f_{LL} = 0.21$	$f_{LV} = 0.29$
			$f_{VV} = 0.04$

Log-0dds $\mathbf{s_{ij} = f_{ij} / p_i p_j}$

Background frequencies, $\mathbf{p_i}$***, from database***

- $p_F = 0.0397$
- $p_I = 0.0529$
- $p_L = 0.0917$
- $p_V = 0.0649$

Sequence Alignment

BLOSUM

```
TCMN_STRGA ( 331) IADLGGGDGWFLAQILRRHPHATGLLMDLPRVA  74
TCMO_STRGA ( 173) FVDLGGARGNLAAHLHRAHPHLRATCFDLPEME  81
ZRP4_MAIZE ( 204) LVDVGGGIGAAAQAISKAFPHVKCSVLDLAHVV  68

COMT_EUCGU ( 205) VVDVGGGTGAVLSMIVAKYPSMKGINFDLPHVI  42
CHMT_POPTM ( 204) LVDVGGGTGAVVNTIVSKYPSIKGINFDLPHVI  41
COMT_MEDSA ( 204) LVDVGGGTGAVINTIVSKYPTIKGINFDLPHVI  47

CRTF_RHOSH ( 205) LMDVGGGTGAFLAAVGRAYPLMELMLFDLPVVA  59
OMTA_ASPPA ( 250) VVDVGGGRGHLSRRVSQKHPHLRFIVQDLPAVI  47
```

} 1 sequence

COMT_EUCGU, CHMT_POPTM, and COMT_MEDSA are all >80 identical

- Sequences are not independent because they are closely related

Weighted (BLOSUM 80) count of transitions for column 1

$c_{FF} = 0$ $c_{FI} = 1$ $c_{FL} = 2.67$ $c_{FV} = 1.33$

$c_{II} = 0$ $c_{IL} = 2.67$ $c_{IV} = 1.33$

$c_{LL} = 2.33$ $c_{LV} = 3.33$

$c_{VV} = 0.33$

Sequence Alignment

Scoring Systems - BLOSUM

Weighted (BLOSUM 80) count of transitions for column 1
N = 15 transitions

$f_{FF} = 0$	$f_{FI} = 0.07$	$f_{FL} = 0.18$	$f_{FV} = 0.09$
	$f_{II} = 0$	$f_{IL} = 0.18$	$f_{IV} = 0.09$
		$f_{LL} = 0.16$	$f_{LV} = 0.22$
			$f_{VV} = 0.02$

BLOSUM n, N = 28 transitions

$f_{FF} = 0.00$	$f_{FI} = 0.04$	$f_{FL} = 0.14$	$f_{FV} = 0.07$
	$f_{II} = 0.00$	$f_{IL} = 0.14$	$f_{IV} = 0.07$
		$f_{LL} = 0.21$	$f_{LV} = 0.29$
			$f_{VV} = 0.04$

Scoring Systems - PAM vs BLOSUM

- PAM
  - **Based on explicit evolutionary model**
  - **Represents a specific evolutionary distance**
  - **Ranges from identical to completely random**
- BLOSUM
  - **Based on empirical frequencies**
  - **Always a blend of distances as seen in the database/PROSITE**
  - **Narrower range than PAM matrix**

Sequence Alignment

Genomics

Scoring - PAM vs BLOSUM

- *Upper matrix difference from PAM 160*
- *Lower matrix BLOSUM 62*

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
	0	-1	1	0	2	1	1	2	1	2	0	0	2	4	1	5	1	2	-2	5	C
		2	0	-2	0	-1	0	0	0	1	0	0	0	1	0	1	-1	1	1	-1	S
C	9		2	-1	-1	-1	0	0	0	0	0	0	-1	0	-1	1	0	1	1	3	T
S	-1	4		2	-2	-1	-1	0	0	-1	-1	-1	1	1	0	-1	0	0	2	1	P
T	-1	1	5		2	-1	-2	-2	-1	0	0	1	1	0	0	1	0	1	1	2	A
P	-3	-1	-1	7		2	0	-1	-2	0	1	1	0	0	-1	0	-1	1	2	4	G
A	0	1	0	-1	4		3	-1	-1	0	0	1	-1	0	-1	0	-1	0	0	0	N
G	-3	0	-2	-2	0	6		2	-1	-1	-1	0	-1	0	0	0	0	2	1	3	D
N	-3	1	0	-2	-2	0	6		1	0	0	2	2	1	-1	0	0	2	2	4	E
D	-3	0	-1	-1	-2	-1	1	6		0	-2	0	1	1	-1	0	0	1	3	3	Q
E	-4	0	-1	-1	-1	-2	0	2	5		2	-1	0	1	0	-1	0	1	2	2	H
Q	-3	0	-1	-1	-1	-2	0	0	2	5		-1	-1	0	-1	1	0	1	3	-4	R
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8		1	-2	-1	1	1	2	3	1	K
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5		-2	-1	-1	0	1	2	4	M
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5		-1	1	0	0	1	3	I
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5		-1	0	-1	1	2	L
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4		0	1	2	4	V
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4		-1	-2	1	F
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4		-1	2	Y
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		-1	W
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Gribskov 6.28

Scoring Systems - Log-odds matrices

- A log-odds scoring system evaluates the relative probabilities of a match representing true homology versus the chance that a match occurs at random, i.e. the relative probability of two models

$$s_{ij} = ln(q_{ij} / p_i \; p_j)$$

- Normally, one multiplies probabilities - since these are log probabilities you get the total probability by adding them up
- When added up over a matching segment, you get the probability that the segment represents homology relative to the probability that it represents a random match, i.e. how much more likely than chance is it that the matching segment represents homology

Target frequencies

- Karlin and Altschul showed that for MSPs (Maximum Sequence Pairs), amino acids a_i and a_j will be aligned with frequency approaching

$$q_{ij} = p_i \, p_j \, e^{-\lambda s}$$

where p_i and p_j are the expected probabilities of observing the amino acid residues and s is the match score

- A given scoring matrix will try to align the residues according to the above equation, so q_{ij} are a characteristic set of target frequencies for the scoring matrix S

- The correct scoring system is the one in which the target frequencies are the same as the frequencies of the actual aligned residues

Scoring Systems - Log-odds matrices

- by rearranging the target frequency equation from the previous slide, equation we get:

$$s_{ij} = [\, \ln (q_{ij} / p_i\, p_j\,) \,] / \lambda$$

- All scoring systems can therefore be looked on as log-odds matrices with an implied set of target frequencies!
- Since multiplying a log-odds scoring system by a constant won't change the relative score for local alignments, λ can be looked on as a scaling factor that we can choose to suit our convenience.
- One convenient choice for λ is *ln2*, so that the scores can be thought of as representing bits

Scoring Systems - Information

- A bit of information is the amount of information needed to distinguish between 2 possibilities, i.e. one yes-no question.
- Taking λ as ln2 , the Karlin-Altschul equation becomes

$$p = KNe^{-\lambda S} \qquad p = KN\,2^{-S}$$

- Rearranging gives the score required to find a given number of MSPs with score S

$$S = log2\,(K/p) + log_2 N$$

- K is generally about 0.1 so the first term basically disappears
- The amount of information required to distinguish an MSP from chance therefore depends entirely on N, the size of the comparison
- N is the product of the lengths for two sequences, or the size of the database times the sequence length for a search

Scoring Systems - Information

- How much information do you need to find something interesting?
- An MSP of about 16 bits is required for significance in a pairwise comparison of two 250 long sequences

  log2 (2502) = 15.93 bits
- For a 250 residue protein sequence and the NCBI nr database,

  log2 (145,000,000 x 250) = 35.0 bits
- For a 1000 base long DNA sequence and the NCBI nr database,

  log2 (2,360,000,000 x 750) = 40.7 bits