Introduction to Alignments

Lec˙06˙slides

Outline for today

1. More of the big picture
2. Sequence Alignment
   - Principles of Sequence Alignment
   - Scoring Alignments
   - Substitution Matrices
   - Inserting Gaps
   - Types of Alignment
3. Summary
4. Reading for next class

Outline

1 More of the big picture

2 Sequence Alignment
Principles of Sequence Alignment
Scoring Alignments
Substitution Matrices
Inserting Gaps
Types of Alignment

3 Summary

4 Reading for next class

Classic example of paralogs and pseudogenes

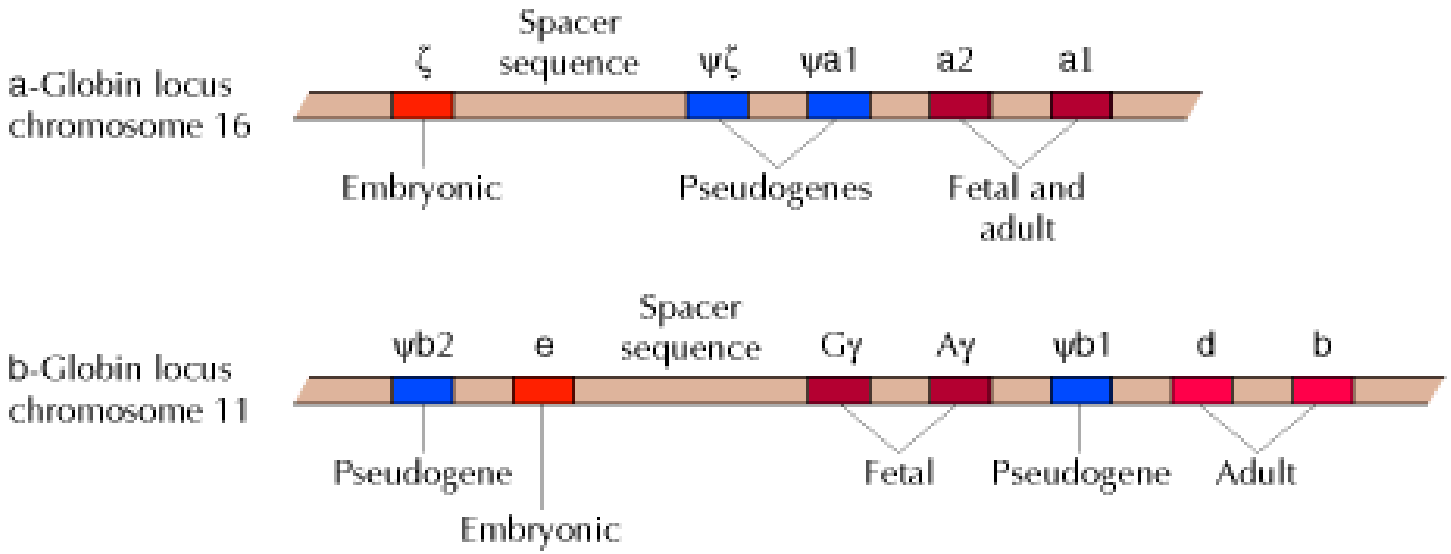


Figure 4.5. Globin gene families Members of the human α- and β-globin gene families are clustered on chromosomes 16 and 11, respectively. Each family contains genes that are specifically expressed in embryonic, fetal, and adult tissues, in addition to nonfunctional gene copies (pseudogenes).

The Cell - A Molecular Approach. 2nd ed. Cooper, Geoffrey M.

Review: Key idea from last lecture

- BLAST looks for sequence similarity
  - Example: `AAACCG` is similar to `TAACCG`

- Homology + evolution creates sequence similarity

83% identity

```
Human catggccctgtggatgcgcctcctgcccctgctggcgctgctggccc
      |||||||||||||||||| |||||||||||||||| |||||   ||
Mouse catggccctgtggatgcgcttcctgcccctgctggccctgctcttcc
```

Therefore, we look for sequence similarity as evidence of homology

Therefore, we use BLAST to find evidence of homology.

Key idea of bioinformatics: Alignment → Similarity score → Homology

Line of reasoning:

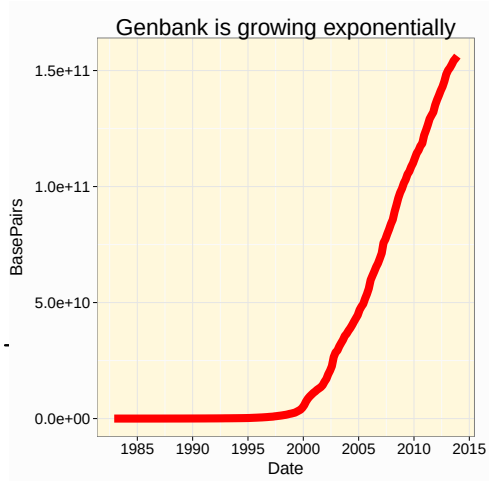
1. We have a DNA or protein sequence
2. We align it against another DNA or protein sequence
3. We measure the similarity or alignment score
4. If the score is strong, we have evidence for homology.

But wait:

1. What's the difference between sequence similarity and homology?
2. If two sequences are similar, are the genes homologous?
3. How can we use this relationship? Why is this so important?

What BLAST is really about

1. I have a DNA or protein sequence
2. Because I'm a biologist,
   I want to find homologs
3. GenBank & Refseq & others have
   lots of sequence data $\rightarrow \rightarrow \rightarrow \rightarrow \rightarrow$
4. Align (BLAST) my sequence against each sequence in database
5. High scoring alignments may be homologues.



Outline

1 More of the big picture

2 Sequence Alignment
- Principles of Sequence Alignment
- Scoring Alignments
- Substitution Matrices
- Inserting Gaps
- Types of Alignment

3 Summary

4 Reading for next class

Outline

1 More of the big picture

2 Sequence Alignment
Principles of Sequence Alignment
Scoring Alignments
Substitution Matrices
Inserting Gaps
Types of Alignment

3 Summary

4 Reading for next class

Two DNA sequences optimally aligned

FASTA input:

```
>Rat
CAAGTA

>Cow
CAATTA
```

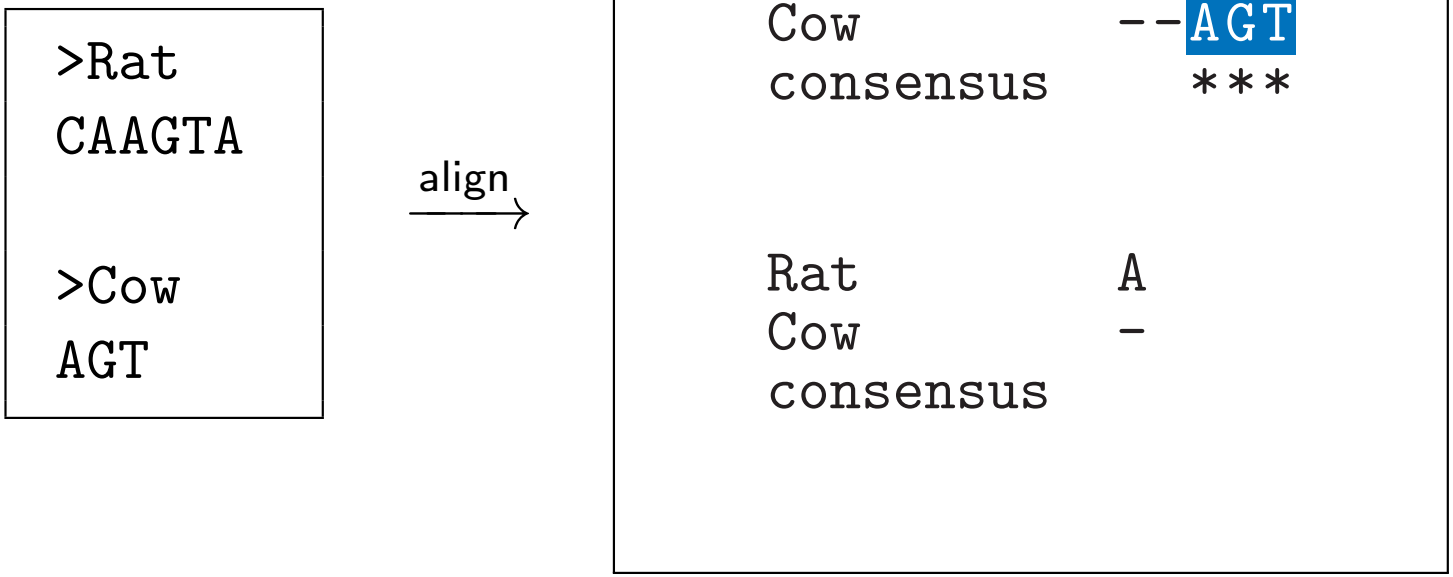
align →

Alignment output:

```
Rat         CAAGT
Cow         CAATT
consensus   ***  *

Rat         A
Cow         A
consensus   *
```

How would you do this by hand?

Slightly more complicated example



How would you do this by hand?

How would you score this alignment's quality?

Gaps hugely complicate finding optimal alignment


```
>Rat
THISSEQUENCE

>Cow
THISISASEQUENCE
```

align →

```
Rat         THIS---SEQUENCE
Cow         THISISASEQUENCE
consensus   ****   ********
```

How would you do this by hand?

How would you score this alignment's quality?

Outline

1 More of the big picture

2 Sequence Alignment
Principles of Sequence Alignment
Scoring Alignments
Substitution Matrices
Inserting Gaps
Types of Alignment

3 Summary

4 Reading for next class

We can score alignment quality


```
>Rat
CAAGTA

>Cow
AGT
```

align →

Good alignment

$\frac{3}{6} = 50\%$ identity

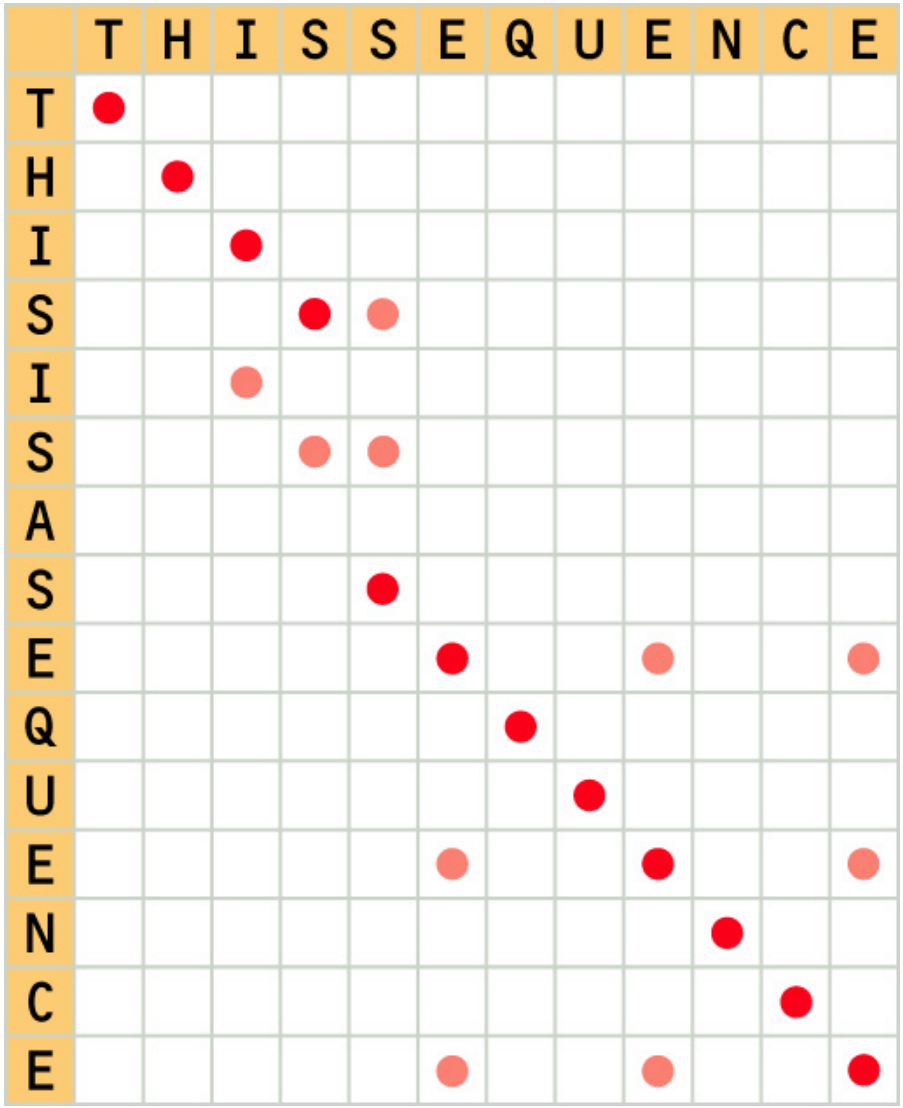
```
Rat          CAAGT
Cow          --AGT
consensus      ***

Rat          A
Cow          -
consensus
```

Bad alignment

$\frac{1}{6} = 17\%$ identity

```
Rat          CAAGT
Cow          -AGT-
consensus     *

Rat          A
Cow          -
consensus
```

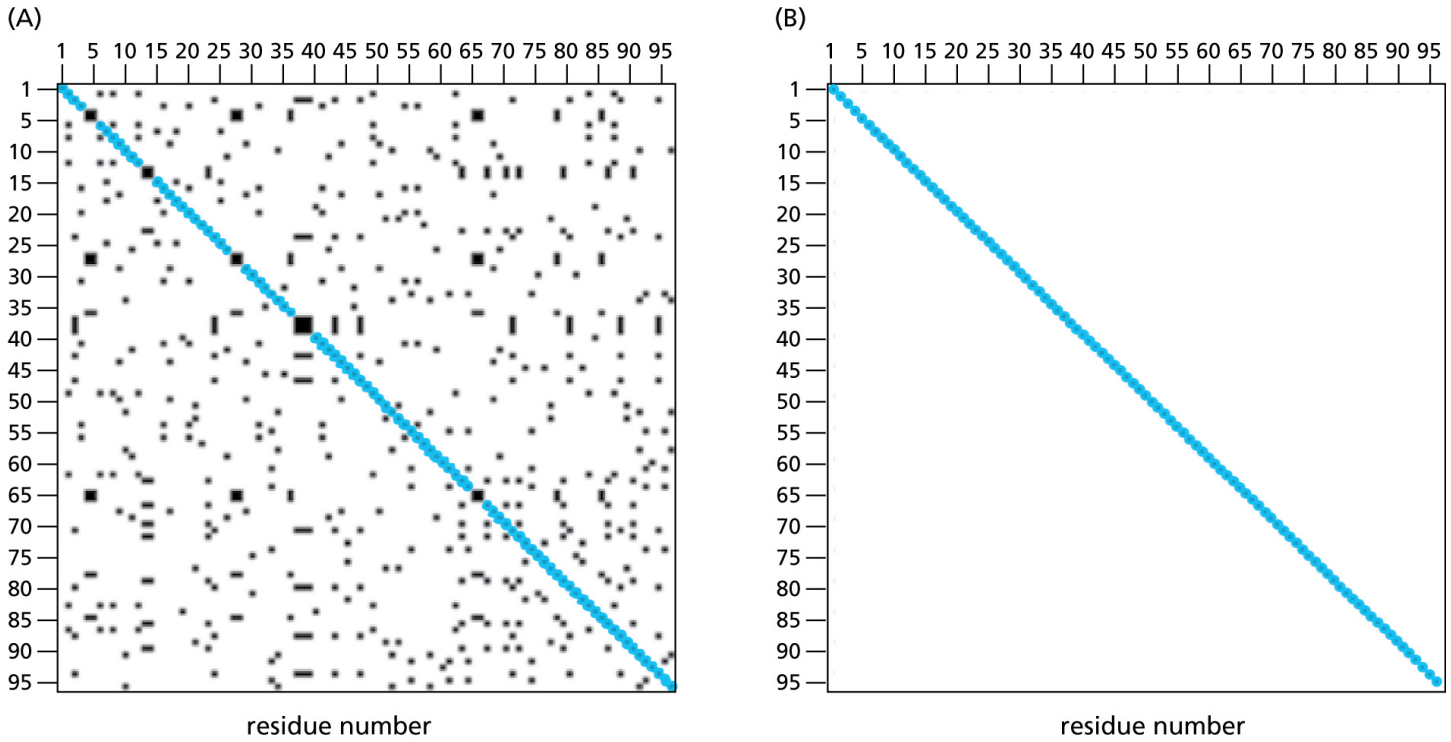
Dotplots align two sequences



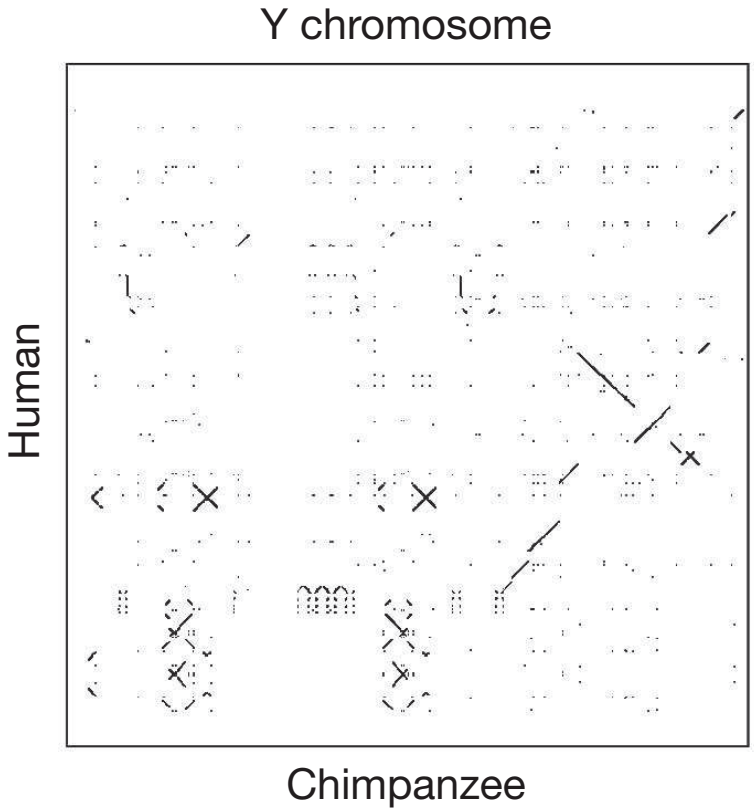
Red: Sequence match

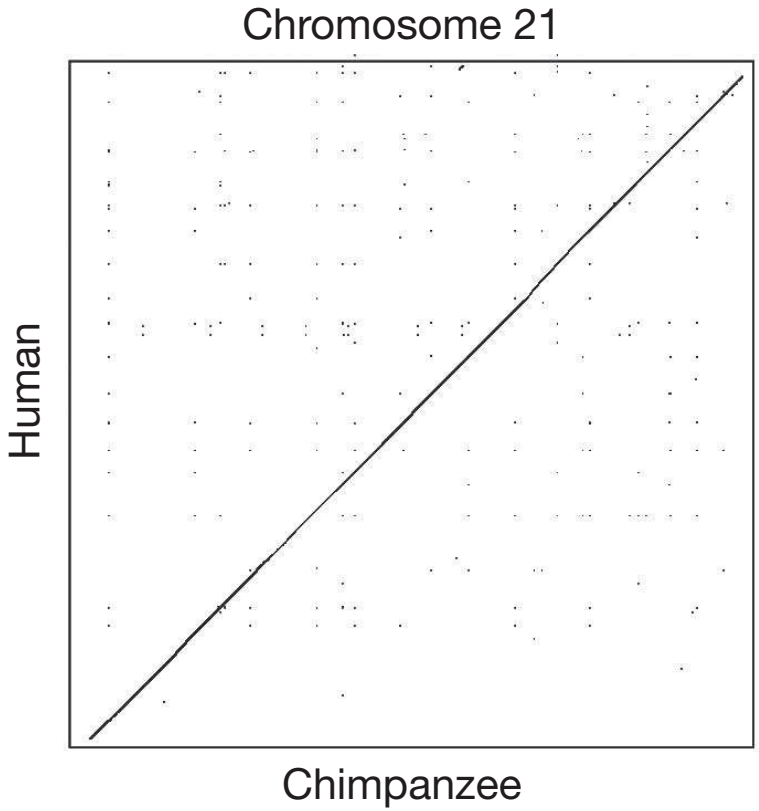
Pink: accidental match

Remove noise with window and minimum identity



Chimp vs human dotplots, 2 different chromosomes

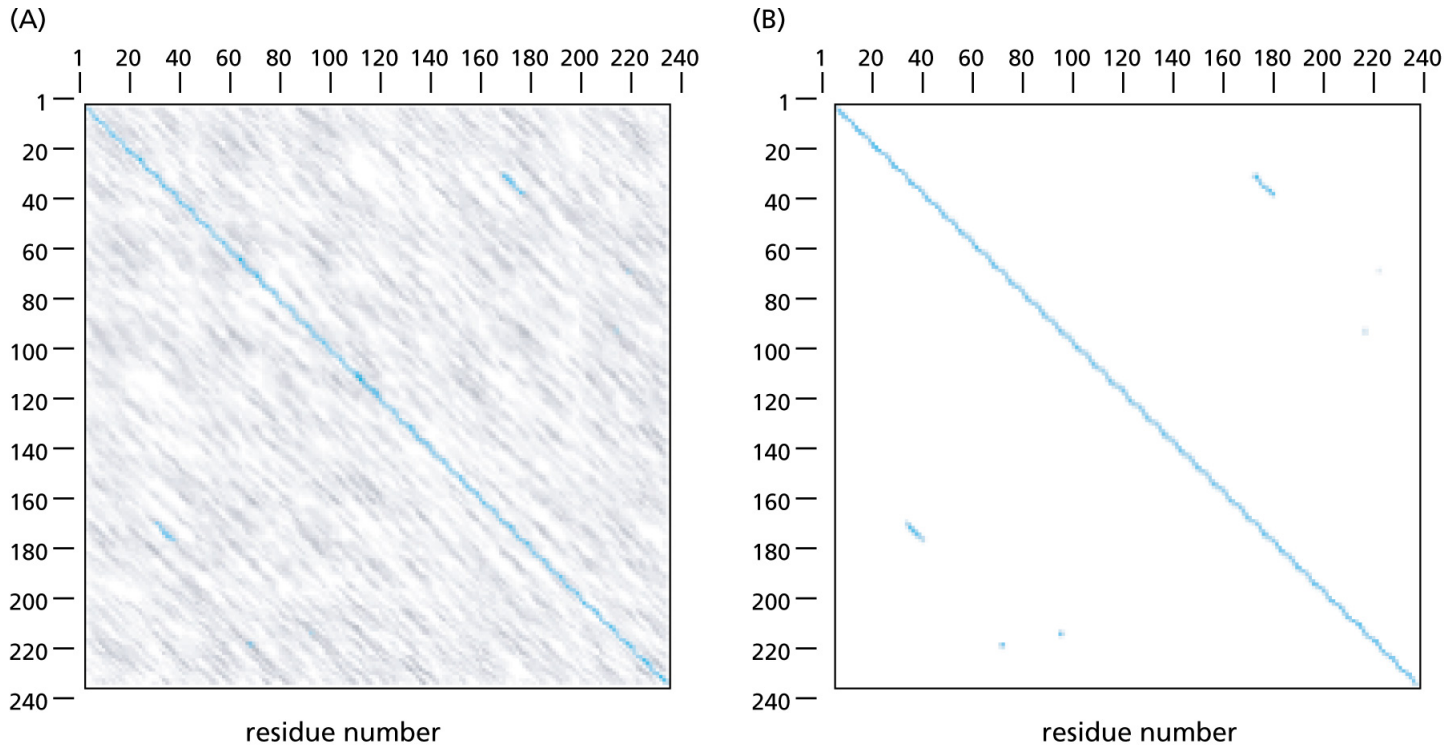




What's going on here?

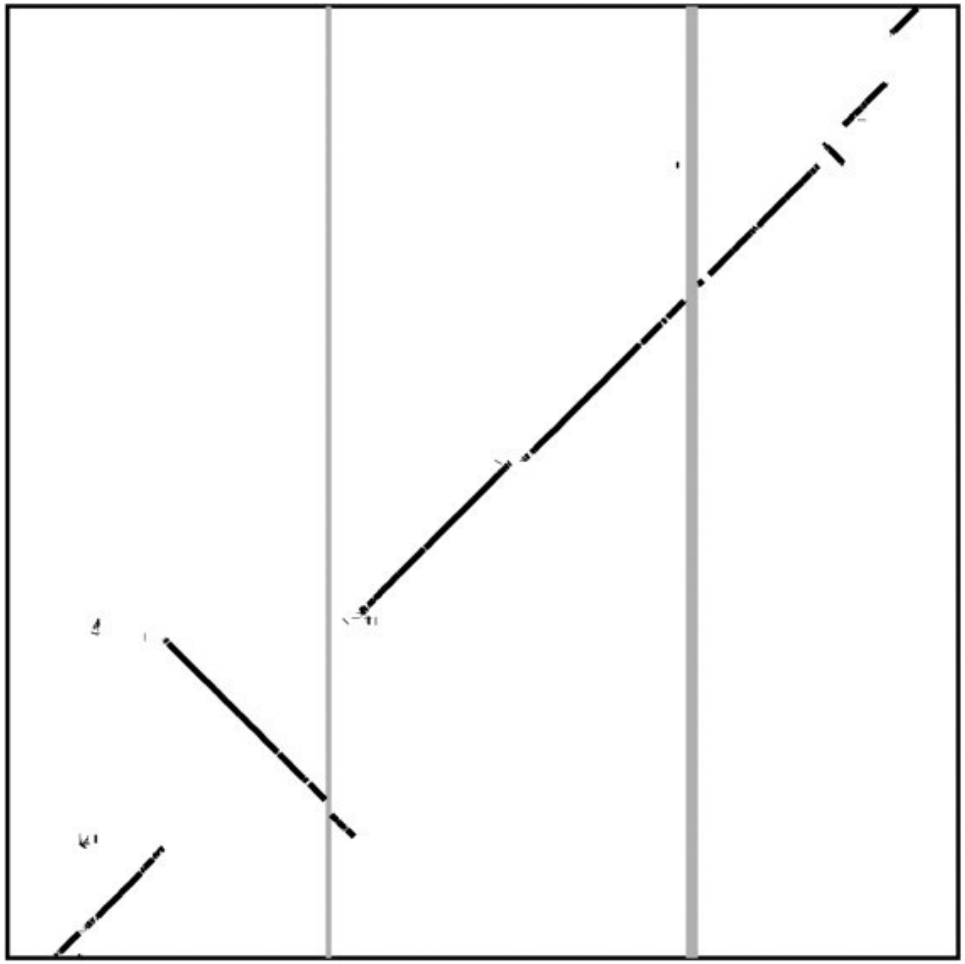
Hughes, et al., Nature, Vol 463, 28 Jan 2010

Dotplots reveal repeats



What's going on here biologically?

What's going on here?



http://www.pnas.org/content/101/7/1916/F1.large.jpg

Dot plotter demos

Traditional dot plotters:

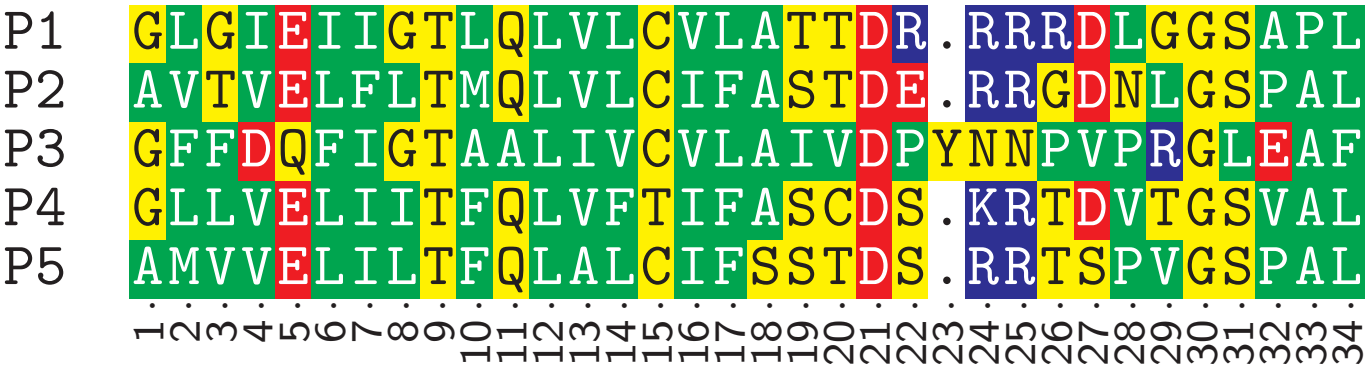
- `http://myhits.isb-sib.ch/cgi-bin/dotlet`
- `http://www.vivo.colostate.edu/molkit/dnadot/`
- `http://bioinfo.lifl.fr/yass/yass.php`

Try BLAST 2 pairwise sequences at NCBI:

- `http://blast.ncbi.nlm.nih.gov/Blast.cgi`
  Click: **Align two or more sequences**

Outline

Score can include amino acid similarity

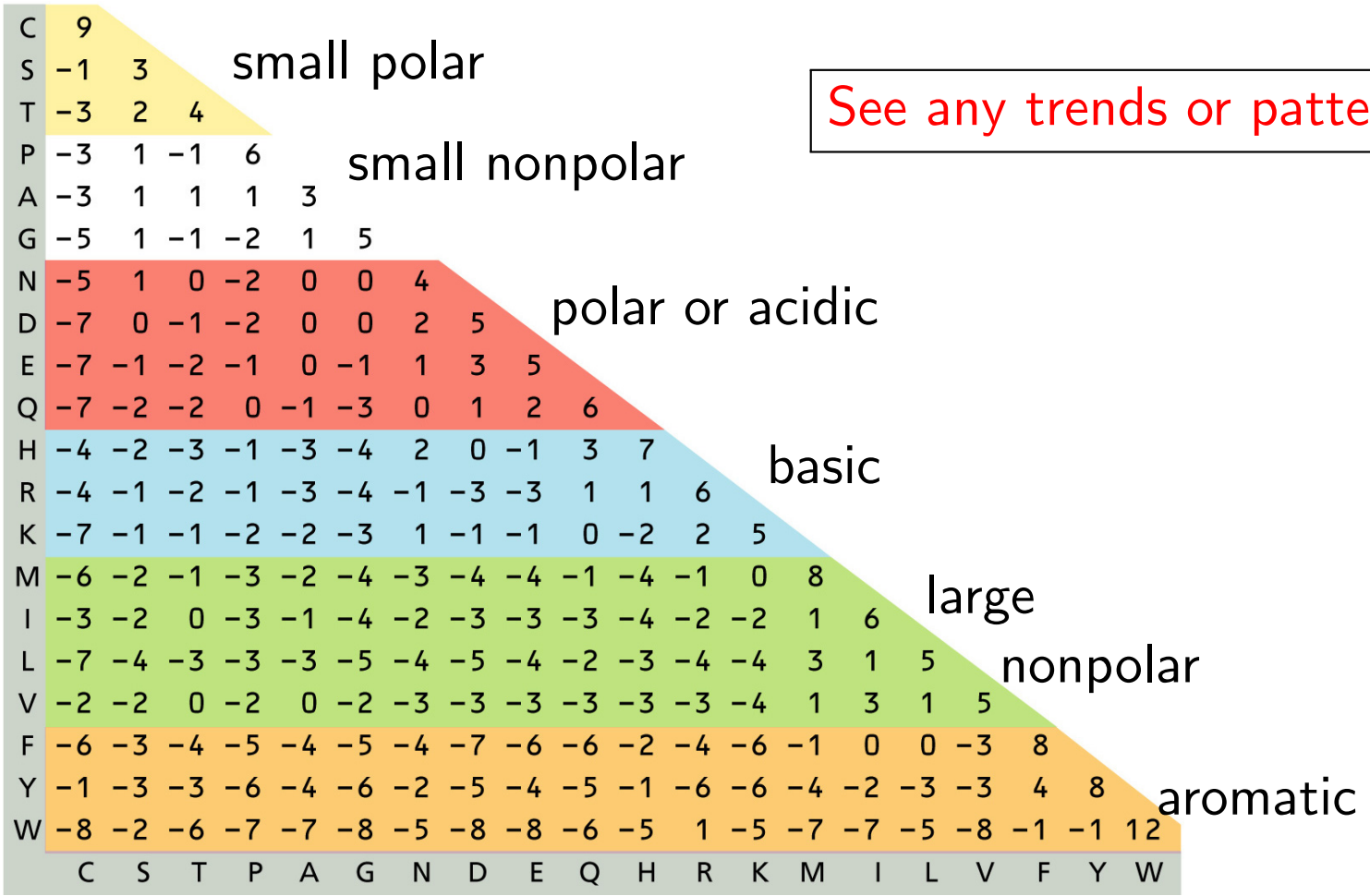


X acidic (−)
X basic (+)
X polar uncharged
X hydrophobic nonpolar

What biological process happened here?

Hint: Look down column 2.

PAM120 substitution matrix

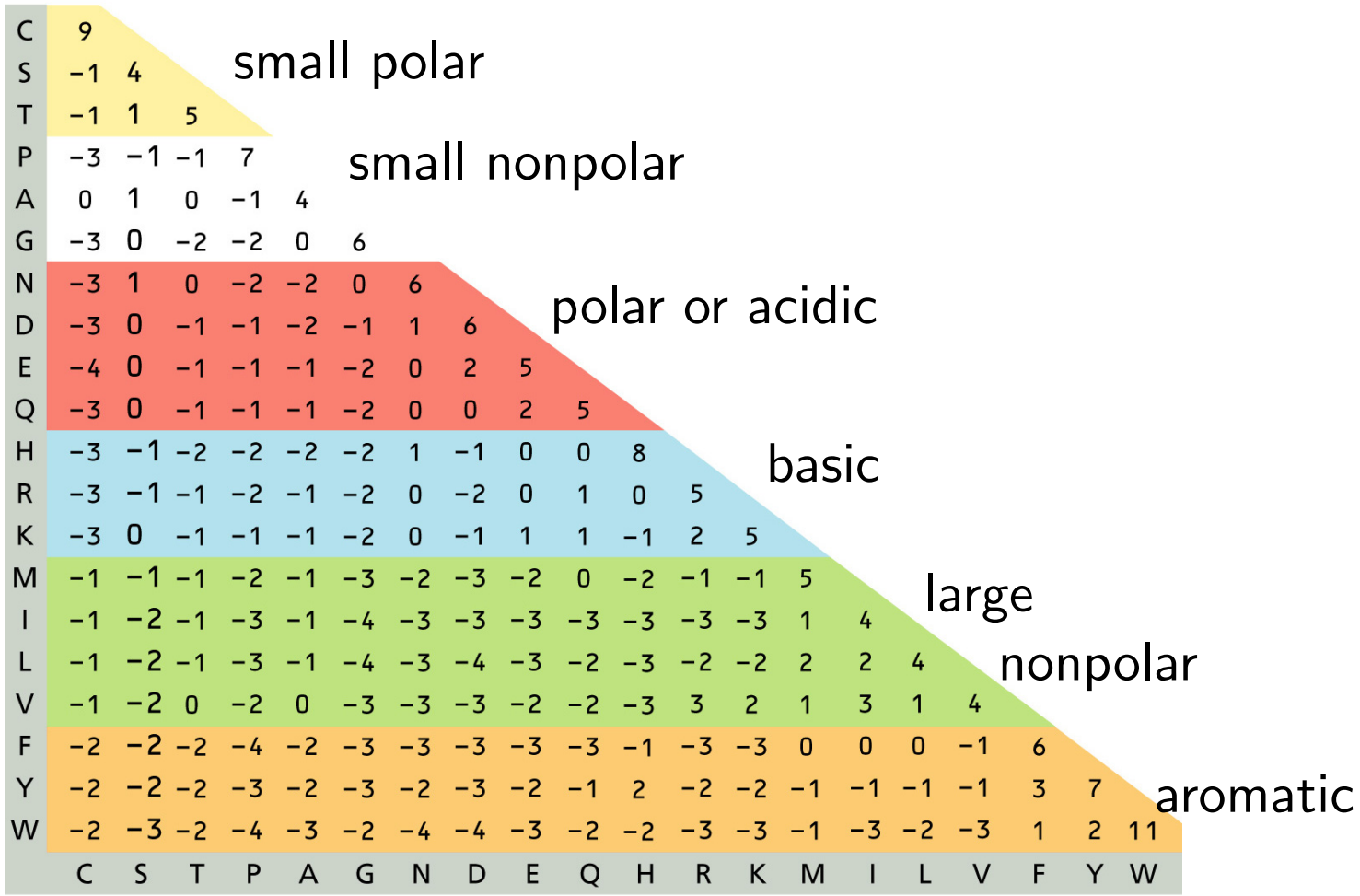


See any trends or patterns?

PAM = Point Accepted Mutation

- Margaret Dayhoff, 1978 – foundational work but still used, still OK
- Aligned closely-related proteins, counted amino acid substitutions at each position
- Family of matrices:
  - PAM1 = 1/100 AA changed
    = short evolutionary span
    = fairly intolerant of evolution
  - PAM250 = 250/100 AA have changed
    = distant evolution span
    = more tolerant of distant evolution

BLOSUM–62 substitution matrix



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

BLOSUM = BLOcks SUbstitution

- Henikoff & Henikoff 1992, 1996, very widely used
- Larger sample size than PAM
- Looked at distantly–related *blocks*
- Larger numbers: more distantly–related proteins: BLOSUM45 captures greater evolutionary distance than BLOSUM90

Recommended substitution matrices for BLAST

Query length	Substitution matrix	Gap costs
< 35	PAM-30	(9,1)
$35 - 50$	PAM-70	(10,1)
$50 - 85$	BLOSUM-80	(10,1)
> 85	BLOSUM-62	(10,1)

http://www.ncbi.nlm.nih.gov/blast/html/sub_matrix.html

Outline

1 More of the big picture

2 Sequence Alignment
Principles of Sequence Alignment
Scoring Alignments
Substitution Matrices
Inserting Gaps
Types of Alignment

3 Summary

4 Reading for next class

Indels = Insertions or Deletions


```
>Rat
THISSEQUENCE

>Cow
THISISASEQUENCE
```

align →

```
Rat         THIS---SEQUENCE
Cow         THISISASEQUENCE
consensus   ****   ********
```

What happened here biologically?

Hint: Two obvious possibilities. . .

High gap insertion penalty versus low penalty

(A)

```
Bovine PI-3Kinase p110a       LNWENPDIMSELLFQNNEIIFKNGDDLRQDMLTLQIIRIMENIWQNQGLDLRMLPYGCLSIGDCVGLIEVVRNSHTIMQIQCKGGLKGAL
cAMP-dependent protein kinase --WENPAQNTAHLDQFERIKTLGTGSFGRVMLVKHMETGNHYAMKILDKQKVVKLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDNSNLY

Bovine PI-3Kinase p110a       QFNSHTLHQWLKDKNKGEIYDAAIDLFTRSCAGYCVATFILGIGDRHNSNIMVKDDGQLFHIDFGHFLDHKKKKFGYKRERVPFVLTQDF
cAMP-dependent protein kinase MVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCGTPEYLAP

Bovine PI-3Kinase p110a       LIVISKGAQECTKTREFERFQEMCYKAYLAIRQHANLFINLFSMMLGSGMPELQSFDDIAYIRKTLALDKTEQEALEYFMKQMNDAHHGG
cAMP-dependent protein kinase EIILSKGYNKAVDWWALGVLIYEMAAGYPPFFADQPIQIYEKIVSGKVRFPSHFSSDLKDLLRNLLQVDLTKRFGNLKNGVNDIKNHKWF

Bovine PI-3Kinase p110a       WTTKMDWIFHTIKQHALN-----------------------------------
cAMP-dependent protein kinase ATTDWIAIYQRKVEAPFIPKFKGPGDTSNFDDYEEEEIRVXINEKCGKEFSEF
```

(B)

```
Bovine PI-3Kinase p110a       LNWENPDIMSELLFQNNEIIFKNGDDLRQDMLTLQIIRIMENIWQNQGLDLRMLPYGCLSIGDCVGLIEVVRNSHTIMQIQCKGGLKGAL
cAMP-dependent protein kinase ?-WENPAQNTAHLDQFERIKTLGTGSFGRVMLVKHM--ETGNHYAMKILDKQKV-VKLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDN-

Bovine PI-3Kinase p110a       QFNSHTLHQWLKDKNKGEIYDAAIDLFTRSCAGYCVATFILGIGDRHNSNIMVKD-DGQLFHIDFGHFLDHKKKKFGYKRERVPFVL--T
cAMP-dependent protein kinase -SNLYMVMEYVPGGEMFSHLRR-IGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCGT

Bovine PI-3Kinase p110a       QDFL---IVISKGAQECTKTREFERF-QEMC--YKAYLAIRQHANLFINLFSMMLGSGMPELQSFDDIAYIRKTLALDKTEQEALEYFMK
cAMP-dependent protein kinase PEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFA-DQPIQIYEKIVSGKVRF--PSHFSSDLKDLLRNLLQVDLTKR--FGNLKN

Bovine PI-3Kinase p110a       QMNDAHHGGWTTKMDWI-----------------------FHTIKQHAL----N----------
cAMP-dependent protein kinase GVNDIKNHKWFATTDWIAIYQRKVEAPFIPKFKGPGDTSNFDDYEEEEIRVXINEKCGKEFSEF
```

We can include gaps in the similarity score

Example protein alignment scoring system:

1. Lookup BLOSUM-62 score at each position.
2. Add these scores.
3. Subtract 10 for each indel position.

A better protein alignment scoring system:

1. Lookup BLOSUM-62 score at each position.
2. Add these scores
3. Subtract 10 for each new string of gaps
4. Subtract 2 for each position in a gap.

Hard question:
Why does the second scoring system better reflect evolution?

Outline

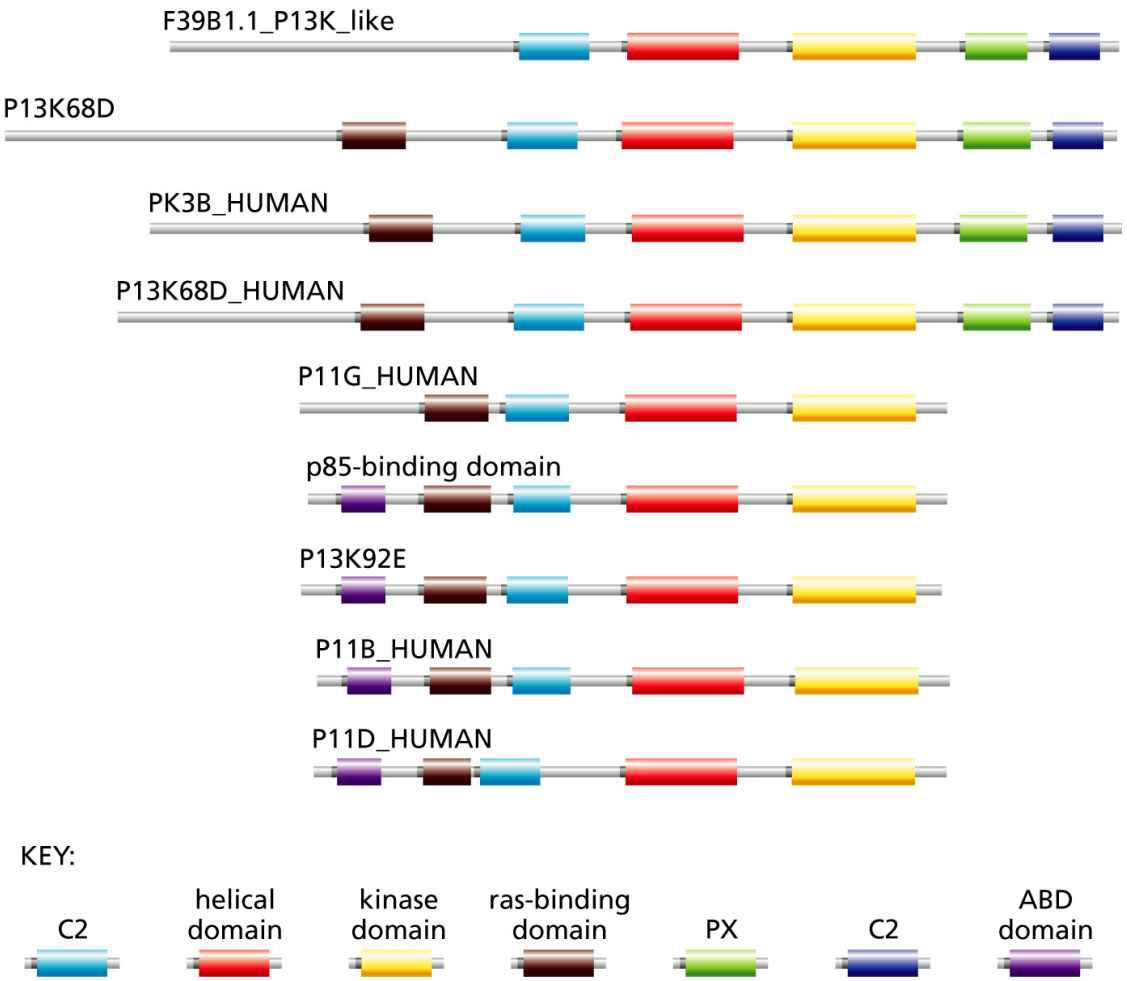
1 More of the big picture

2 Sequence Alignment
Principles of Sequence Alignment
Scoring Alignments
Substitution Matrices
Inserting Gaps
Types of Alignment

3 Summary

4 Reading for next class

Multi-domain proteins and alignment



Local versus global alignment

- **Local alignment**
  - Best-known algorithm: BLAST family
  - Looks for any regions of similarity
  - Ignores regions on dissimilarity
  - Useful for fishing expedition
  - Great for finding protein domains

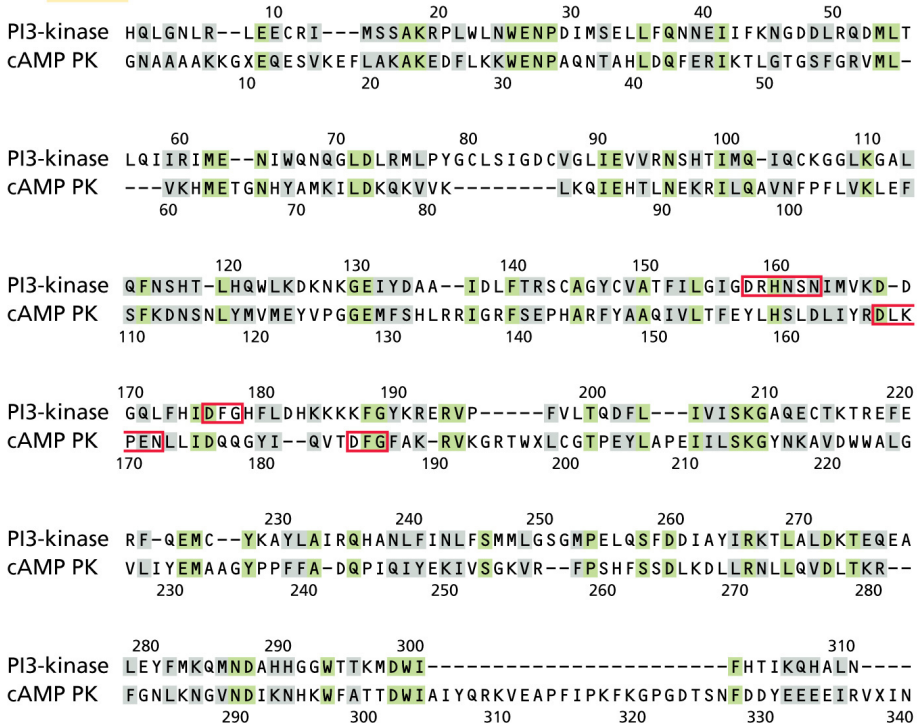
- **Global alignment**
  - Best-known algorithm: CLUSTALW
  - Aligns entire sequence
  - Can produce poor results in multi-domain proteins
  - May need manual adjustment after alignment
  - Very useful with protein families, e.g., hemoglobin

Example of local versus global alignment

(A) local (BLAST)

```
PI3-kinase  DRHNSNIMVKDDGQLFHIDFG
cAMP PK     DLKPENLLIDQQGYIQVTDFG
```

(B) global (CLUSTALW)



Critical kinase residues:

- DLKPEN
- DFG

Local alignment found them

Global alignment didn't find them

When would we want local alignment?

When would we want global alignment?

Pairwise alignment versus multiple alignment

Pairwise alignment: Aligns two sequences (BLAST)

```
AQP1nuc.SEQ   GCTGTCA-TGTATATCATCGCC
AQP2nuc.SEQ   GCTGCCT-TCTATGTGGCTGCC
```

Multiple alignment: Aligns $\geq$ two sequences (CLUSTALW)

```
AQP1nuc.SEQ   GCTGTCA-TGTATATCATCGCC
AQP2nuc.SEQ   GCTGCCT-TCTATGTGGCTGCC
AQP3nuc.SEQ   GCTGCCCATCTACACACTGGCA
AQP4nuc.SEQ   TCTGTCT-TCTACATCACTGCG
AQP5nuc.SEQ   GCTGTCT-TCTACGTGGCAGCC
```

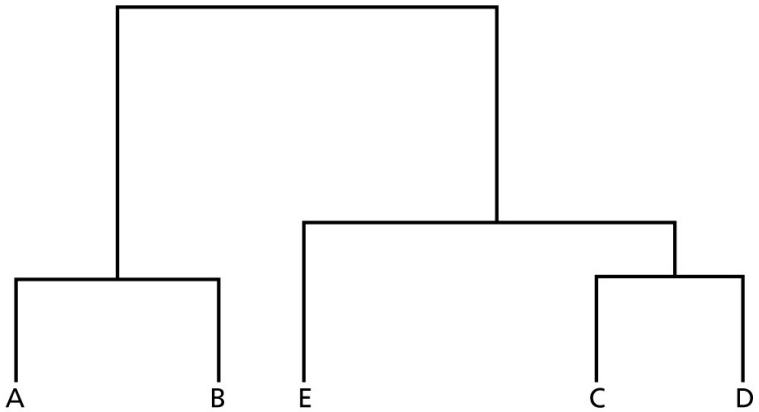
Multiple sequence alignment can discover conserved and critical residues

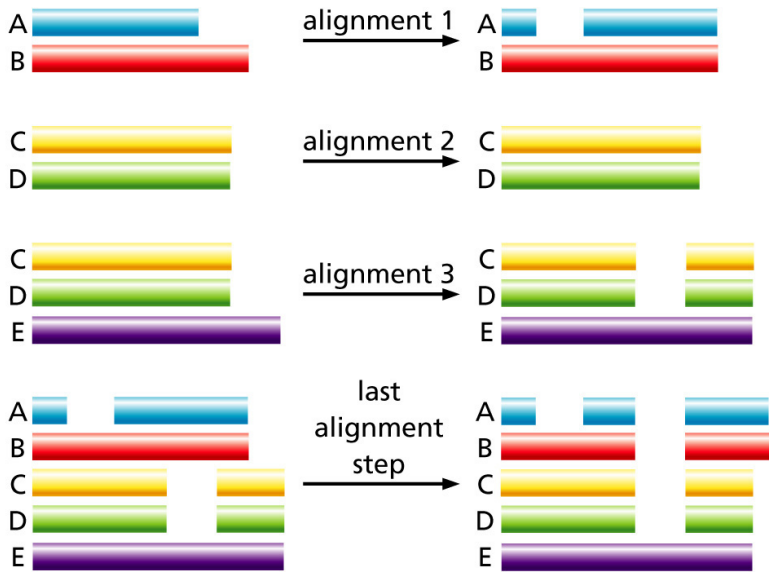
```
(A)  p110α          TFILGIGDRHNSNIMVKDDG-QLFHIDFGHFLDHKKKKFGYKRERVPFVLT--QDFLIVI 142
     cAMP-kinase    QIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCGTPEYLAPE 179

(B)  p110β          SYVLGIG----------DRHSDNINVKKTGQLFHIDFGHILGNFKSKFGIKRERVPFILT 136
     p110δ          TYVLGIG----------DRHSDNIMIRESGQLFHIDFGHFLGNFKTKFGINRERVPFILT 136
     p110α          TFILGIG----------DRHNSNIMVKDDGQLFHIDFGHFLDHKKKKFGYKRERVPFVLT 135
     p110γ          TFVLGIG----------DRHNDNIMITETGNLFHIDFGHILGNYKSFLGINKERVPFVLT 135
     p110_dicti     TYVLGIG----------DRHNDNLMVTKGGRLFHIDFGHFLGNYKKKFGFKRERAPFVFT 135
     cAMP-kinase    QIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCG--TPEYLA 177
```

What's going on here biologically?

CLUSTALW operates in several steps

1. Compute alignment scores between all sequence pairs
2. Take highest scoring pair, create a consensus from the pair
3. Take next highest scoring pair, create a consensus sequence
4. Repeat until all sequences are merged into a single consensus.





Alignment programs produce different results

(A) structural/functional alignment from BAliBase

```
1csy  SHEKMPWFHGKISREESEQIVLIGSKTNGKFLIRARD--NNGSYALCLLHEGKVLHYRIDKDKTGKLSIPEGK-KFDTLWQLVEHYSYKA------DGLLRVL-TVPCQK
1gri  EMKPHPWFFGKIPRAKAEEML-SKQRHDGAFLIRESES-APGDFSLSVKFGNDVQHFKVLRDGAGKYFL-WVV-KFNSLNELVDYHRSTS-VSRNQQIFLRDIEQVPQQ-
1aya  ---MRRWFHPNITGVEAENLLLTRG-VDGSFLARPSKS-NPGDFTLSVRRNGAVTHIKIQN--TGDYYDLYGGEKFATLAELVQYYMEHHGQLKEKNGDVIEL-KYPLN-
2pna  -LQDAEWYWGDISREEVNEKLRDT--ADGTFLVRDASTKMHGDYTLTLRKGGNNKLIKIFH-RDGKYGFSDPL-TFNSVVELINHYRNES-LAQYNPKLDVKL-LYPVS-
1bfi  HHDEKTWNVGSSNRNKAENLLRGK--RDGTFLVRESS--KQGCYACSVVVDGEVKHCVINKTATG-YGFAEPYNLYSSLKELVLHYQHTS-LVQHNDSLNVTL-AYPVYA
```

(B) DIALIGN multiple sequence alignment

```
1csy  SHEKMPWFHGKISREESEQIVLIGSKT-NGKFLIRAR-DN--NGSYALCLLHEGKVLHYRIDKDKTGKLSIPEGKK-FDTLWQLVEHYSYKA-------DGLLRVLT-VPCQK
1gri  EMKPHPWFFGKIPRAKAEEML--SKQRHDGAFLIRESESA--PGDFSLSVKFGNDVQHFKVLRDGAGKYFLWVV-K-FNSLNELVDYHRST--SVSRNQQIFLRDIEQVPQQ-
1aya  M---RRWFHPNITGVEAENLLLTRGV--DGSFLARPSKSN--PGDFTLSVRRNGAVTHIKIQNTGDYYDLYG-GEK-FATLAELVQYYMEHHGQLKEKNGDV-IELK-YPLN-
2pna  LQDAE-WYWGDISREEVNEKL--RDTA-DGTFLVRDA-STKMHGDYTLTLRKGGNNKLIKIFHRDGKYGFSD-PLT-FNSVVELINHYRNE--SLAQYNPKLDVKLL-YPVS-
1bfi  HHDEKTWNVGSSNRNKAENLL--RGKR-DGTFLVRES-SK--QGCYACSVVVDGEVKHCVINKTATGYGFAE-PYNLYSSLKELVLHYQHT--SLVQHNDSLNVTLA-YPVYA
```

(C) ClustalW multiple sequence alignment

```
1csy  SHEKMPWFHGKISREESEQIVLIGSKTNGKFLIRARDN--NGSYALCLLHEGKVLHYRIDKDKTGKLSIPEGKKFD-TLWQLVEHYSYK------ADGLLRVLTVPCQK
1gri  EMKPHPWFFGKIPRAKAEE-MLSKQRHDGAFLIRESES-APGDFSLSVKFGNDVQHFKVLRDGAGKY-FLWVVKFN-SLNELVDYHRSTS-VSRNQQIFLRDIEQVPQQ
1aya  ---MRRWFHPNITGVEAEN-LLLTRGVDGSFLARPSKS-NPGDFTLSVRRNGAVTHIKIQNT-GDYYDLYGGEKFA-TLAELVQYYMEHHGQLKEKNGDVIELKYPLN-
2pna  -LQDAEWYWGDISREEVN--EKLRDTADGTFLVRDASTKMHGDYTLTLRKGGNNKLIKIFHR-DGKYGFSDPLTFN-SVVELINHYRNES-LAQYNPKLDVKLLYPVS-
1bfi  HHDEKTWNVGSSNRNKAE--NLLRGKRDGTFLVRESSK--QGCYACSVVVDGEVKHCVINKT-ATGYGFAEPYNLYSSLKELVLHYQHTS-LVQHNDSLNVTLAYPVYA
```

(D) divide-and-conquer multiple sequence alignment

```
1csy  SHEKMPWFHGKISREESEQIVLIGSKTNGKFLIRA-RDNN-GSYALCLLHEGKVLHYRIDKDKTGKLSIPEGKK-FDTLWQLVEHY-SY----KADGLLRV-L-TVPCQK
1gri  EMKPHPWFFGKIPRAKAEEMLS-KQRHDGAFLIRE-SESAPGDFSLSVKFGNDVQHFKVLRDGAGK-YFLWVVK-FNSLNELVDYH-RSTSVSRNQQIFLRDIEQVPQQ-
1aya  ---MRRWFHPNITGVEAENLLL-TRGVDGSFLARP-SKSNPGDFTLSVRRNGAVTHIKIQNTGDYY-DLYGGEK-FATLAELVQYYMEHHGQLKEKNGDVIEL-KYPLN-
2pna  -LQDAEWYWGDISREEVNEKL--RDTADGTFLVRDASTKMHGDYTLTLRKGGNNKLIKIFHRDGKY-GFSDPLT-FNSVVELINHY-RNESLAQYNPKLDVKL-LYPVS-
1bfi  HHDEKTWNVGSSNRNKAENLL--RGKRDGTFLVRE-SSKQ-GCYACSVVVDGEVKHCVINKTATGY-GFAEPYNLYSSLKELVLHY-QHTSLVQHNDSLNVTL-AYPVYA
```

If the inputs are the same, how can this happen?

What's going on here?

Why can't programs find the best alignment?

Two kinds of computational problems:

1. Tractable:
   - Runs in a "reasonable" amount of time
   - Can find best solution
2. Intractable:
   - Requires huge amount of computer time to find best solution → would take longer than age of universe.
   - Famous examples:
     - Traveling salesman problem
     - Multiple sequence alignment

Q: So what do we do?

A: Try for an *approximate answer* in a reasonable amount of time
→ heuristic solution → most of bioinformatics.
What are the practical implications?

Outline

1 More of the big picture

2 Sequence Alignment
- Principles of Sequence Alignment
- Scoring Alignments
- Substitution Matrices
- Inserting Gaps
- Types of Alignment

3 Summary

4 Reading for next class

Outline for today

1. More of the big picture
2. Sequence Alignment
   - Principles of Sequence Alignment
   - Scoring Alignments
   - Substitution Matrices
   - Inserting Gaps
   - Types of Alignment
3. Summary
4. Reading for next class

Outline

1 More of the big picture

2 Sequence Alignment
- Principles of Sequence Alignment
- Scoring Alignments
- Substitution Matrices
- Inserting Gaps
- Types of Alignment

3 Summary

4 Reading for next class

Chapter 3

Section "Pairwise Alignment and Limits of Detection"
to
End of Chapter

Pages	Notes
94–112	Read

Chapter 4

Section "Introduction"
to
Section "Stand-alone BLAST"

Pages	Notes
121–135	Read