Sequence Alignment: A General Overview

COMP 571

Luay Nakhleh, Rice University

Life through Evolution

- All living organisms are related to each other through evolution
- This means: any pair of organisms, no matter how different, have a common ancestor sometime in the past, from which they evolved
- Evolution involves
  - inheritance: passing of characteristics from parent to offspring
  - variation: differentiation between parent and offspring
  - selection: favoring some organisms over others

Sequence Variations Due to Mutations

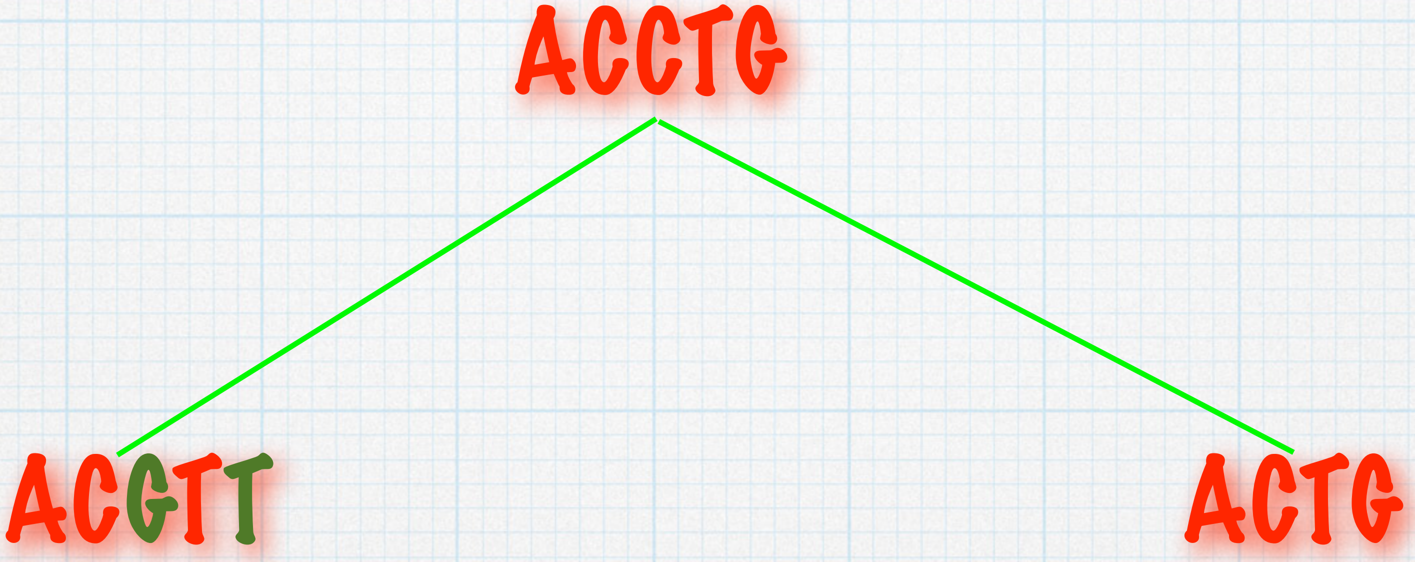
- Mutations and selection over millions of years can result in considerable divergence between present-day sequences derived from the same ancestral sequence.
- The base pair composition of the sequences can change due to point mutation (substitutions), and the sequence lengths can vary due to insertions/deletions

DNA Sequence Evolution

DNA Sequence Evolution

ACCTG

DNA Sequence Evolution

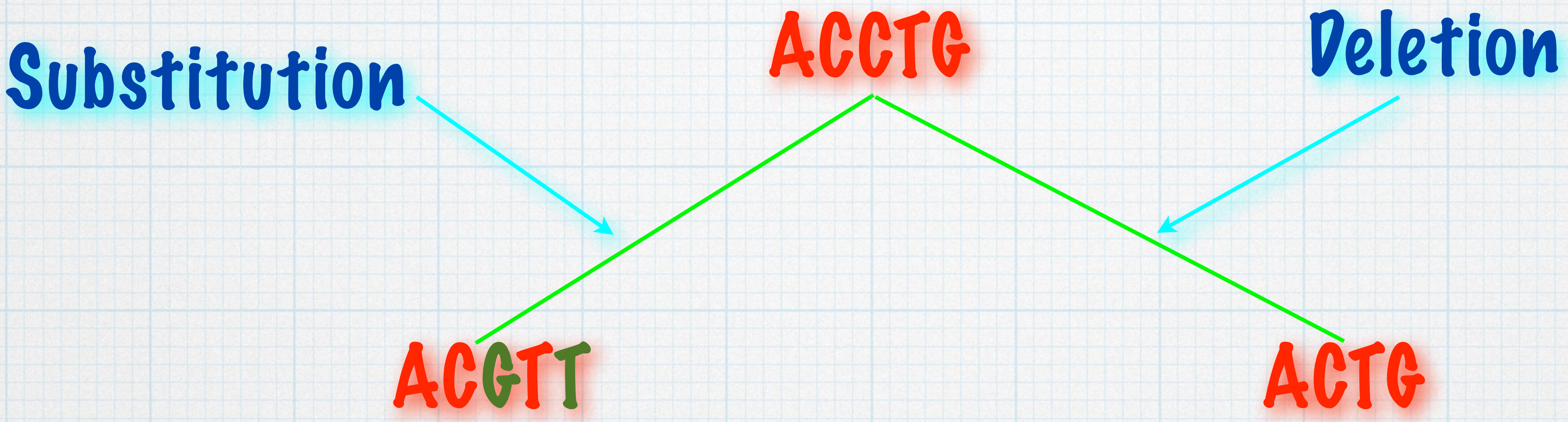


ACCTG

ACGTT

ACTG

DNA Sequence Evolution



ACCTG

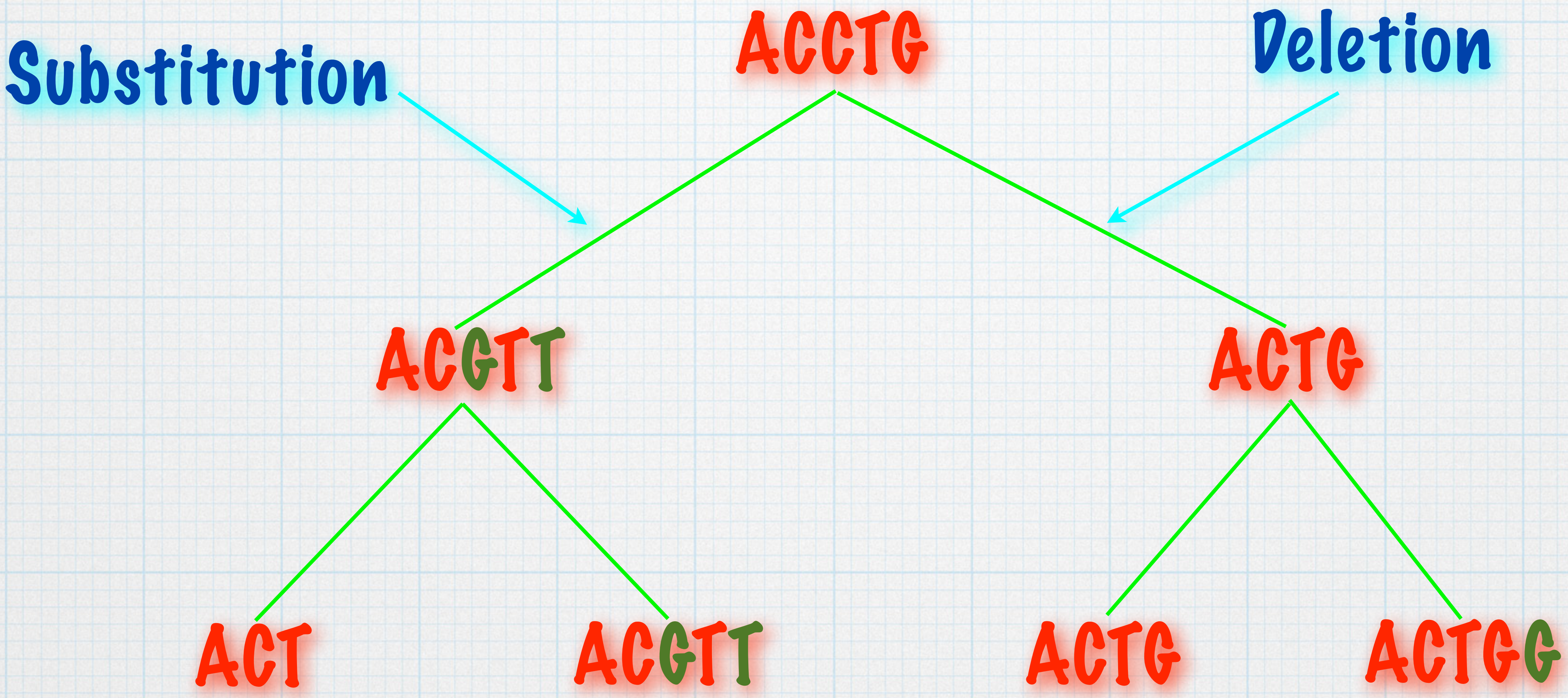
Substitution

Deletion

ACGTT

ACTG

DNA Sequence Evolution



Substitution

ACCTG

Deletion

ACGTT

ACTG

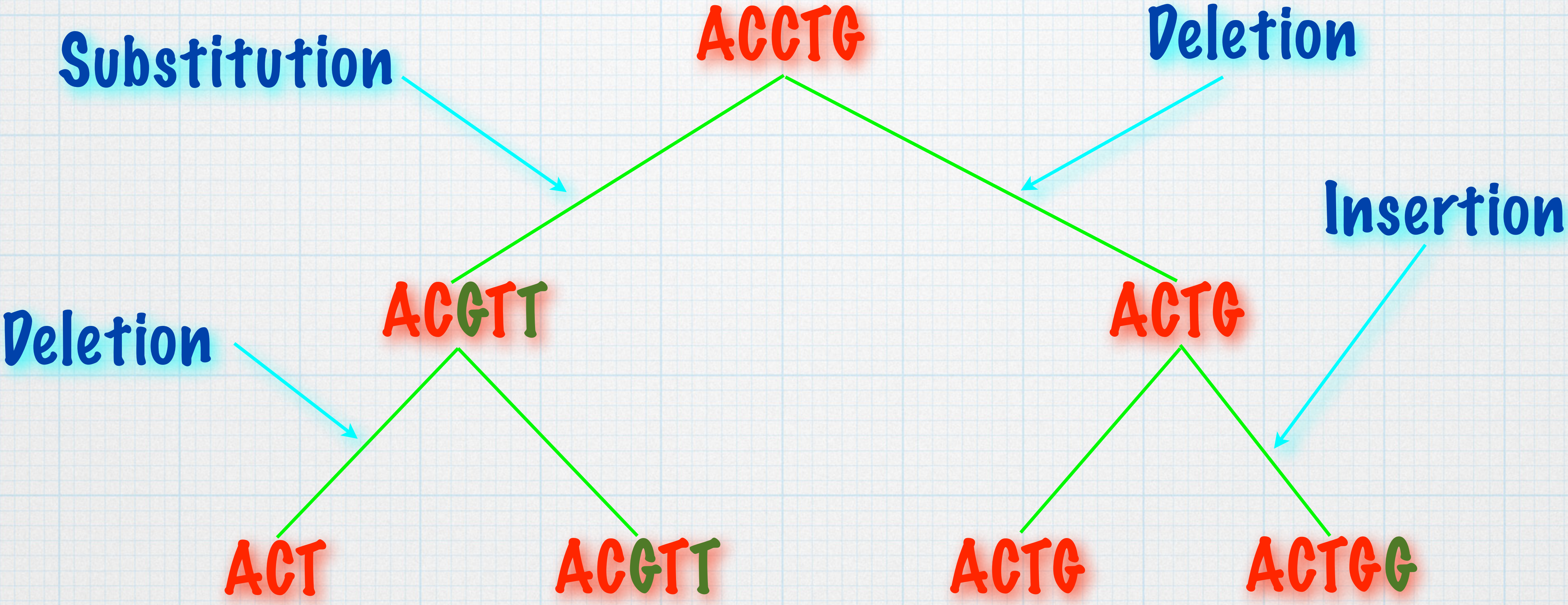
ACT

ACGTT

ACTG

ACTGG

DNA Sequence Evolution



ACCTG

Substitution

Deletion

ACGTT

ACTG

Deletion

Insertion

ACT

ACGTT

ACTG

ACTGG

Principles of Sequence Alignment

1. Alignment is the task of locating "equivalent" regions of two or more sequences to maximize their similarity

Mismatches

T	H	A	T	S	E	Q	U	E	N	C	E
T	H	I	S	S	E	Q	U	E	N	C	E

T	H	I	S	I	S	A	-	S	E	Q	U	E	N	C	E
T	H	-	-	-	-	A	T	S	E	Q	U	E	N	C	E

gap (indels: insertions/deletions)

Principles of Sequence Alignment

2. Alignment can reveal homology between sequences
    1. **Similarity** is descriptive term that tells about the degree of match between the two sequences
    2. **Homology** tells that the two sequences evolved from a common ancestral sequence
    3. Sequence similarity does not not always imply a common function
    4. Conserved function does not always imply similarity at the sequence level
    5. Convergent evolution: sequences are highly similar, but are not homologous

Principles of Sequence Alignment

3. It is easier to detect homology when comparing protein sequences than when comparing nucleic acid sequences

    1. The probability of a “match by chance” is much higher in DNA sequences than in protein sequences.

    2. The genetic code is redundant: identical amino acids can be encoded by different codons.

    3. The complex 3D structure of a protein, and hence its function, is determined by the amino acid sequence. Hence, conserving function leads to fewer changes in the amino acids than in the nucleotide sequence.

Scoring Alignments

- Given two sequences, the number of possible alignments is exponential
- Finding the "correct" alignment involves defining a **scoring scheme** and finding an alignment with optimal score

Scoring Alignments: The Main Principles

- Alignments of related sequences should give good scores compared with alignments of randomly chosen sequences
- The correct alignment of two related sequences should ideally be the one that gives the best score
- (In practice, the correct alignment does not necessarily have the best score, since no "perfect" scoring scheme has been devised)

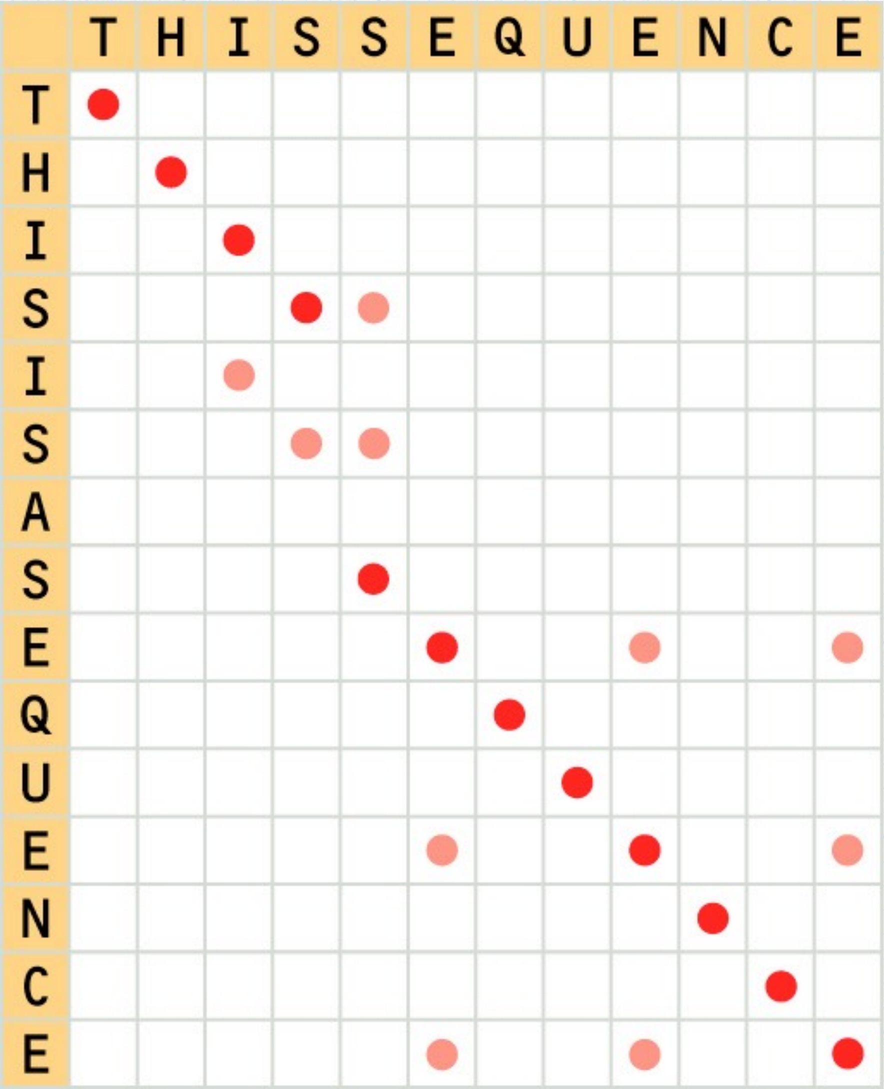
Percent Identity as a Measure for Quantifying Sequence Similarity

- **Identity** is the number of identical bases or amino acids matched between two aligned sequences

- **Percent identity** is obtained by dividing this number by the total length of the aligned sequences and multiplying by 100

- Sequence similarity based on identity is usually visualized using the **dot-plot** representation

Dot-plot



	T	H	I	S	S	E	Q	U	E	N	C	E
T	●											
H		●										
I			●									
S				●	○							
I			○									
S				○	○							
A												
S					●							
E						●			○			○
Q							●					
U								●				
E						○			●			○
N										●		
C											●	
E						○			○			●

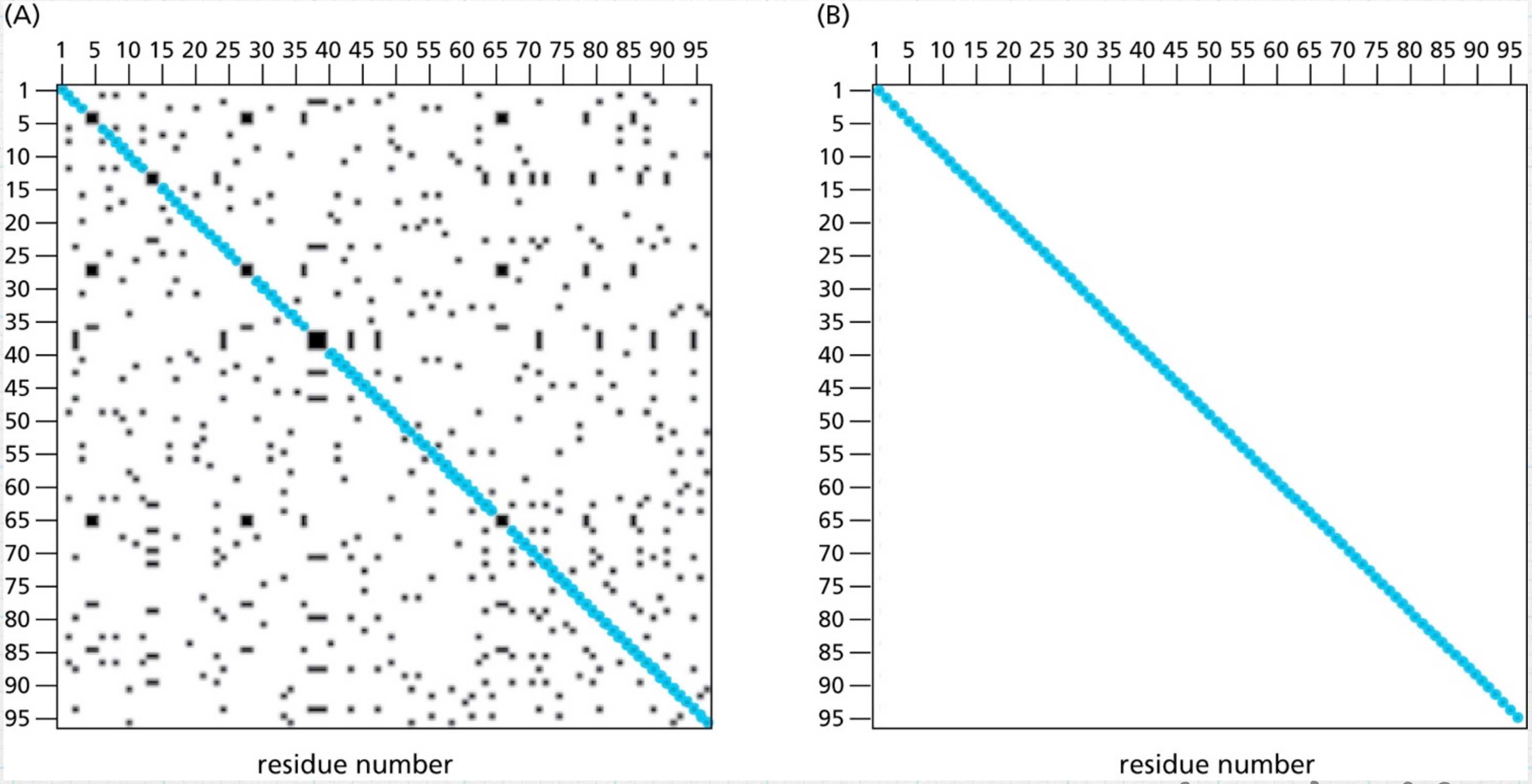
Red dots represent identities that are due to true matching of identical residue-pairs and pink dots represent identities that are due to noise (matching of random identical residue-pairs)

Dot-plot

- Dot-plots suffer from background noise
- To overcome this problem it is necessary to apply a filter
- The most-commonly used filtering method uses a sliding window and requires that the comparison achieves some minimum identity score summed over that window before being considered

Dot-plot

Dot-plot of an SH2 sequence compared to itself



window length = 1

window length = 10
min. ident. score = 3

Scoring Alignments

- Genuine matches do not have to be identical
  - Mutations that replace one amino acid with another having similar physiochemical properties are more likely to have been accepted during evolution
  - So, pairs of amino acids with similar properties will often represent genuine matches rather than matches occurring randomly
- To take this into account, percent identity is replaced by **percent similarity**

Scoring Alignments

- There is a minimum percent identity that can be accepted as significant
- It has been found that, in general, sequence pairs with identity at or greater than 30% over their whole length are pairs of structurally similar proteins
- In the region between 20% and 30% identity, referred to as the **twilight zone**, evolutionary relatedness may exist but cannot be reliably assumed in the absence of other evidence

Substitution Matrices and Gaps

- Scoring schemes have to represent two salient features of an alignment

1. They must reflect the degree of similarity of each pair of residues (the likelihood that both are derived from the same residue in the presumed common ancestral sequence). This is achieved through the use of an appropriate **substitution matrix**

2. They must asses the validity of **inserted gaps**

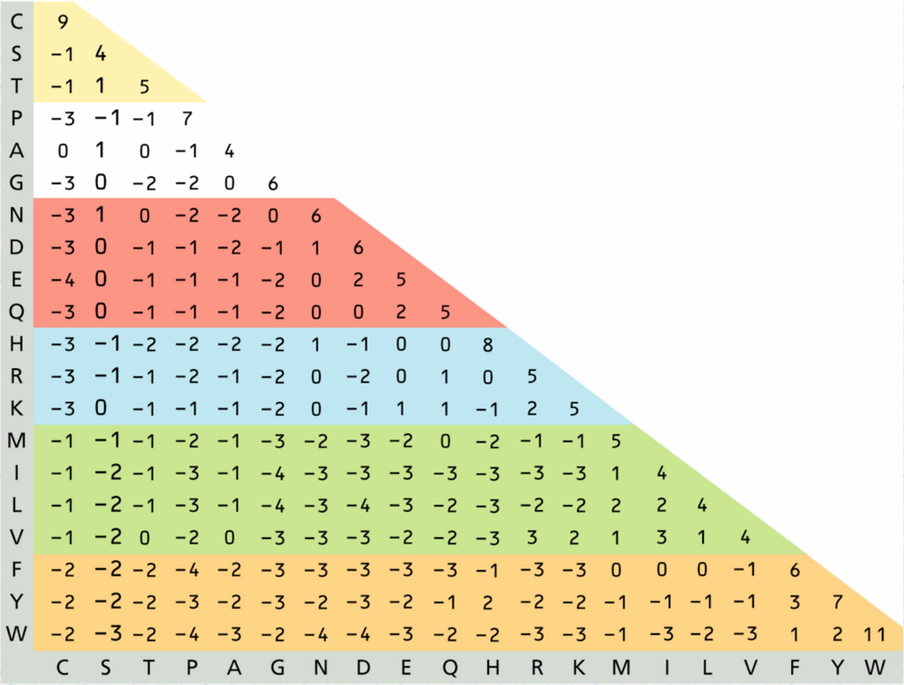
Substitution Matrices

- Mostly apply to amino acids
- The score of a protein sequence alignment is assigned to each aligned pair of amino acids by reference to a substitution matrix, which defines values for all possible pairs of residues (a 20x20 matrix)
- Several such matrices have been devised, and which one to use depends on the "evolutionary distance" of the sequences being aligned

Substitution Matrices

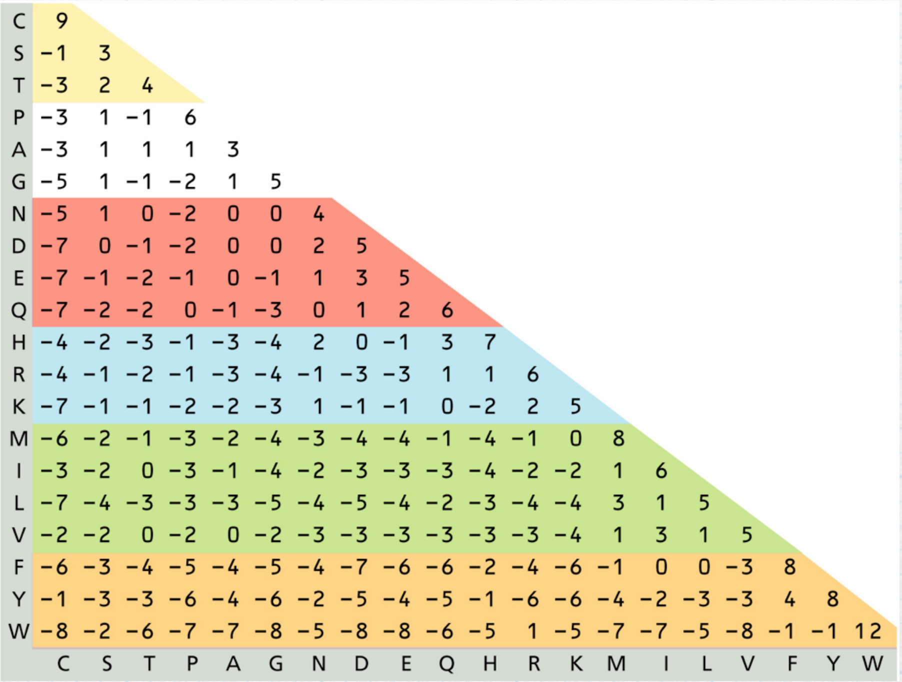
- Commonly used substitution matrices include **PAM** (Point Accepted Mutation) and **BLOSUM** (Blocks of Amino Acid Substitution Matrix)
- PAM substitution matrices were derived based on substitution frequencies in sets of closely related protein sequences
- BLOSUM substitution matrices were derived based on mutation data in highly conserved local regions of sequences

Substitution Matrices



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
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

BLOSUM62



C	9																			
S	-1	3																		
T	-3	2	4																	
P	-3	1	-1	6																
A	-3	1	1	1	3															
G	-5	1	-1	-2	1	5														
N	-5	1	0	-2	0	0	4													
D	-7	0	-1	-2	0	0	2	5												
E	-7	-1	-2	-1	0	-1	1	3	5											
Q	-7	-2	-2	0	-1	-3	0	1	2	6										
H	-4	-2	-3	-1	-3	-4	2	0	-1	3	7									
R	-4	-1	-2	-1	-3	-4	-1	-3	-3	1	1	6								
K	-7	-1	-1	-2	-2	-3	1	-1	-1	0	-2	2	5							
M	-6	-2	-1	-3	-2	-4	-3	-4	-4	-1	-4	-1	0	8						
I	-3	-2	0	-3	-1	-4	-2	-3	-3	-3	-4	-2	-2	1	6					
L	-7	-4	-3	-3	-3	-5	-4	-5	-4	-2	-3	-4	-4	3	1	5				
V	-2	-2	0	-2	0	-2	-3	-3	-3	-3	-3	-3	-4	1	3	1	5			
F	-6	-3	-4	-5	-4	-5	-4	-7	-6	-6	-2	-4	-6	-1	0	0	-3	8		
Y	-1	-3	-3	-6	-4	-6	-2	-5	-4	-5	-1	-6	-6	-4	-2	-3	-3	4	8	
W	-8	-2	-6	-7	-7	-8	-5	-8	-8	-6	-5	1	-5	-7	-7	-5	-8	-1	-1	12
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

PAM120

Substitution Matrices

- There are other matrices: JTT (Jones et al., 1992), VT (Muller and Vingron, 2000), STR (uses information about protein structure), SLIM and PHAT (for membrane proteins), ...
- It is very hard to determine which matrix to use
- When aligning distantly related sequences, PAM250 and BLOSUM-50 are preferable, whereas when aligning closely related sequences PAM120 and BLOSUM-80 may be better

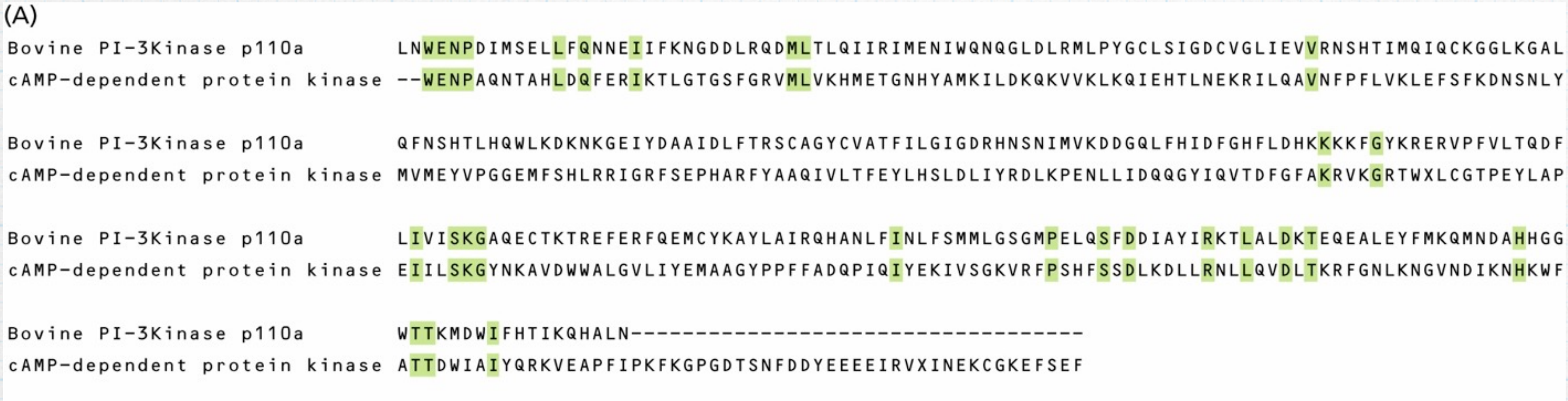
Handling Gaps

- Homologous sequences are often of different lengths as the result of **insertions** and **deletions** (**indels**) that have occurred in the sequences as they diverged from the ancestral sequences
- Their alignment is generally dealt with by inserting gaps in the sequences to achieve as correct a match as possible
- Gaps must be introduced judiciously
- To place limits on the introduction of gaps, alignment programs use a **gap penalty**: each time a gap is introduced, the penalty is subtracted from the score

Handling Gaps

- Structural analysis has shown that fewer indels occur in sequences of structural importance, and that insertions tend to be several residues long rather than just a single residue long
- This informs what gap penalty model to define
- Different gap penalty models may result in different alignments

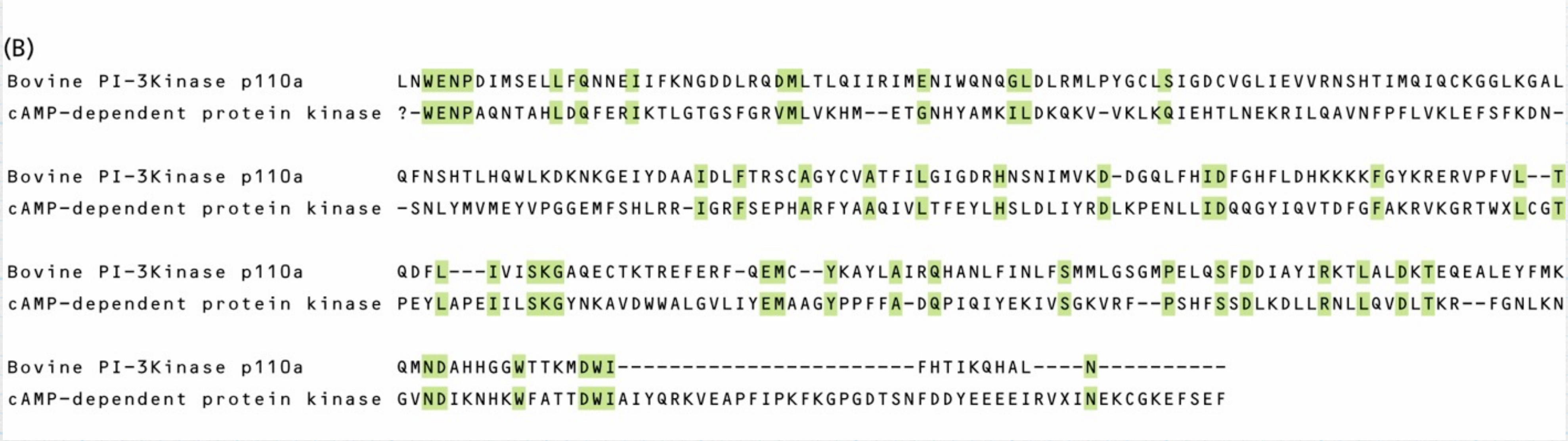
Handling Gaps



(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
```

Very high gap penalty results in gaps only at beginning and end, and 10% sequence identity



(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
```

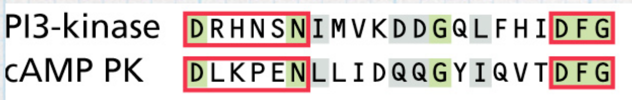
Very low gap penalty results in many more gaps, and 18% sequence identity

Types of Alignment

- **Global alignment**: Aligning the whole sequences
  - Appropriate when aligning two very closely related sequencs
- **Local alignment**: Aligning certain regions in the sequences
  - Appropriate for aligning multi-domain protein sequences
- It is important to use the "appropriate" type

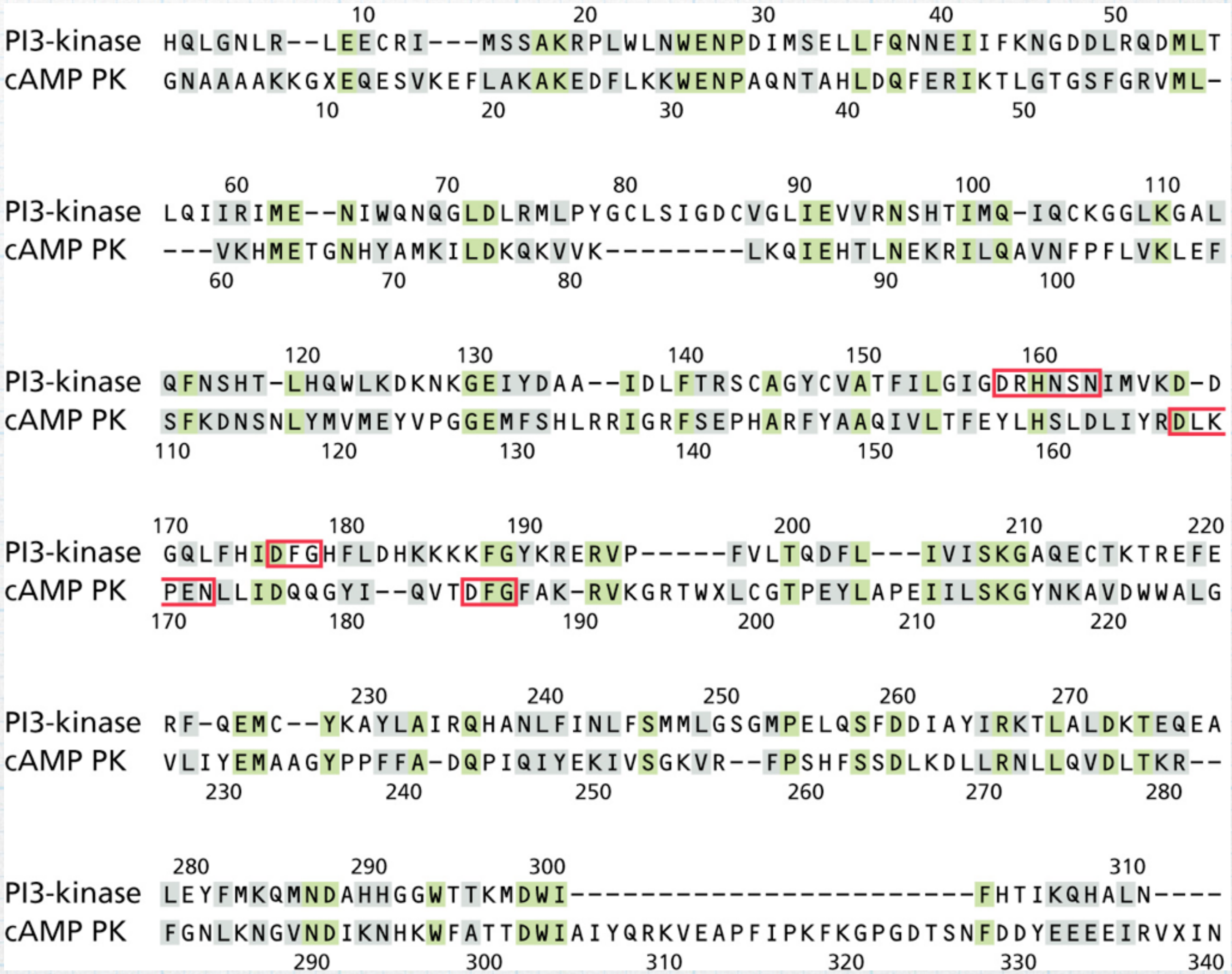
Local vs. Global Alignment

Local



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

Global

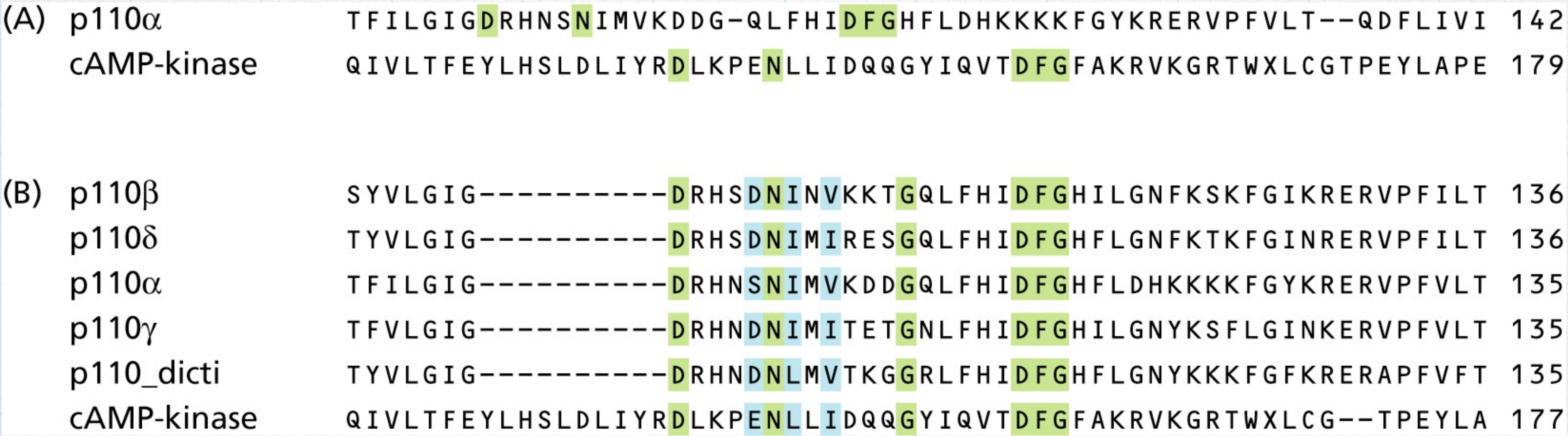


```
                       10          20        30        40        50
PI3-kinase  HQLGNLR--LEECRI---MSSAKRPLWLNWENPDIMSELLFQNNEIIFKNGDDLRQDMLT
cAMP PK     GNAAAAKKGXEQESVKEFLAKAKEDFLKKWENPAQNTAHLDQFERIKTLGTGSFGRVML-
                     10       20        30        40        50

                60          70        80        90       100        110
PI3-kinase  LQIIRIME--NIWQNQGLDLRMLPYGCLSIGDCVGLIEVVRNSHTIMQ-IQCKGGLKGAL
cAMP PK     ---VKHMETGNHYAMKILDKQKVVK--------LKQIEHTLNEKRILQAVNFPFLVKLEF
               60        70       80                60        100

                   120       130         140       150       160
PI3-kinase  QFNSHT-LHQWLKDKNKGEIYDAA--IDLFTRSCAGYCVATFILGIGDRHNSNIMVKD-D
cAMP PK     SFKDNSNLYMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLK
            110      120       130       140       150       160

            170      180       190            200           210      220
PI3-kinase  GQLFHIDFGHFLDHKKKKFGYKRERVP-----FVLTQDFL---IVISKGAQECTKTREFE
cAMP PK     PENLLIDQQGYI--QVTDFGFAK-RVKGRTWXLCGTPEYLAPEIILSKGYNKAVDWWALG
            170      180           190       200      210       220

                       230       240       250       260      270
PI3-kinase  RF-QEMC--YKAYLAIRQHANLFINLFSMMLGSGMPELQSFDDIAYIRKTLALDKTEQEA
cAMP PK     VLIYEMAAGYPPFFA-DQPIQIYEKIVSGKVR--FPSHFSSDLKDLLRNLLQVDLTKR--
              230        240       250         260       270        280

             280       290       300                             310
PI3-kinase  LEYFMKQMNDAHHGGWTTKMDWI-----------------------FHTIKQHALN----
cAMP PK     FGNLKNGVNDIKNHKWFATTDWIAIYQRKVEAPFIPKFKGPGDTSNFDDYEEEEIRVXIN
                   290       300       310        320      330       340
```

Types of Alignment

- **Pairwise alignment**: Aligning a pair of sequences
  - Computationally "easy"
- **Multiple alignment**: Aligning more than two sequences
  - Computationally "hard"
  - Advantageous when sequences of low similarity are being aligned

Pairwise vs. Multiple Alignment



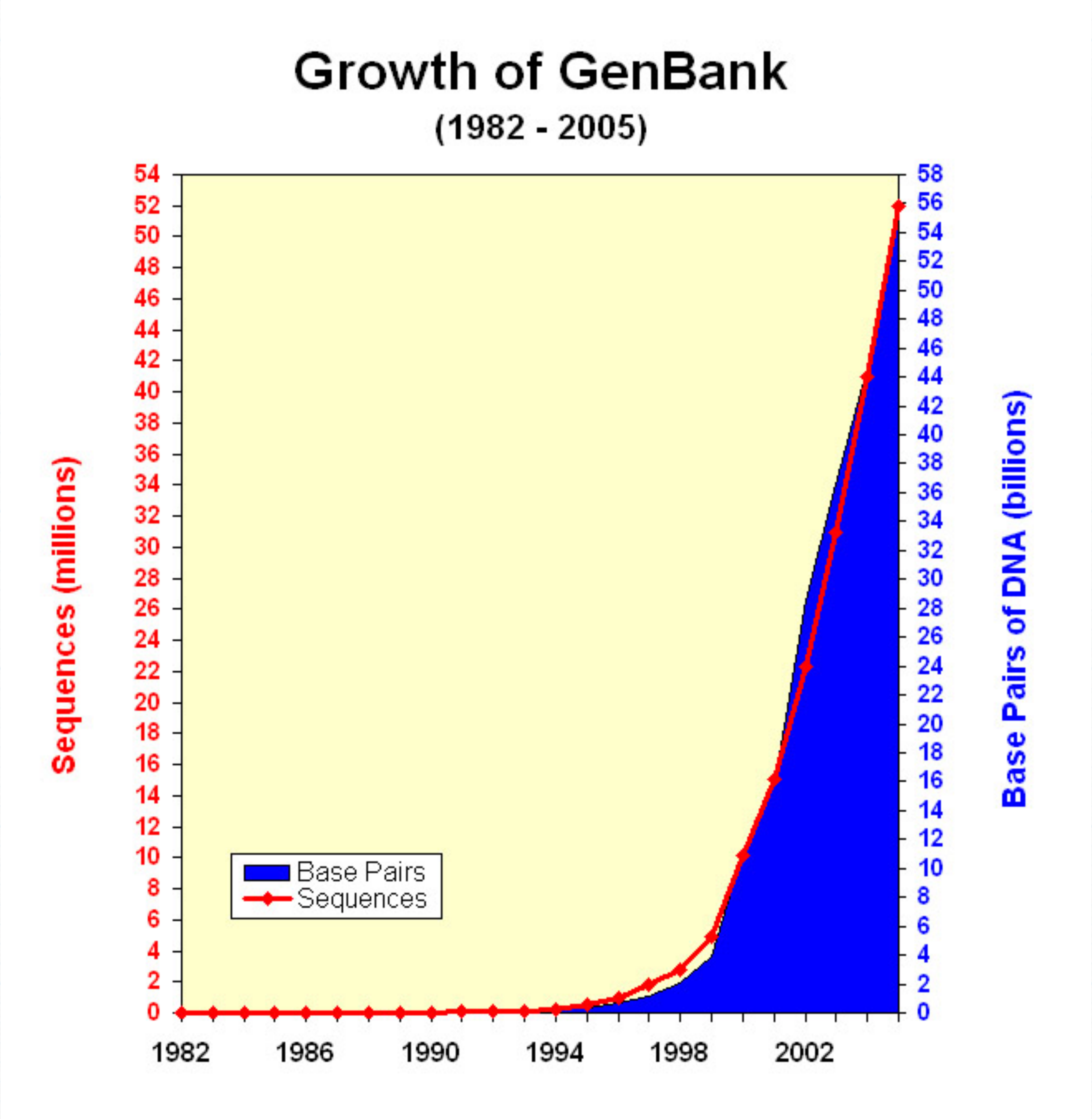
```
(A) p110α          TFILGIGDRHNSNIMVKDDG-QLFHIDFGHFLDHKKKKFGYKRERVPFVLT--QDFLIVI 142
    cAMP-kinase    QIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCGTPEYLAPE 179


(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
```

The pairwise alignment does not align the important active-site residues
The multiple alignment does align the important active-site residues

Searching Databases

- Searching databases has become an integral part of molecular biology
- Searching sequence databases typically entails finding sequences in the database that are "similar" to a query sequence
- Such a search amounts to aligning the query sequence to sequences in the database and returning ones with "good" alignment score
- Given the sizes of available biological databases, heuristics are employed instead of the exact, yet expensive, alignment algorithms

Growth of GenBank



Source: http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

Database Search

- Database search needs to be both sensitive, in order to detect distantly related homologs and avoid false-negative searches, and also specific, in order to reject unrelated sequences with high similarity (false positives)
- The two most-commonly used heuristics for sequence database search are FASTA and BLAST

- In the next set of slides we will cover
  - the theory behind scoring schemes
  - algorithmic techniques for computing optimal alignments
  - significance of alignments