CHAPTER 4: PRODUCING AND ANALYZING SEQUENCE ALIGNMENTS

Dr. Garrett Dancik

Motivation

- You have recently sequenced a gene and its CDS begins with
  - GGCGGAGCCAGGCCGGCCTAGAGTCACTTCTCC
- You have isolated a protein and its amino acid sequence is
  - MGKEIPTDAPWEAQHADKWDKMTMKELIDKICWTKTA
- Questions:
  - What does this protein do?
  - What are the important functional regions?
  - Do other organisms have similar genes or proteins?
- To answer these questions we can find similar sequences, identified through sequence alignments, using tools such as BLAST

Sequence alignment

- Two sequences should be aligned in such a way that maximizes their *similarity*
  - If they derive from a common ancestor, characters (bases or amino acids) derived from the same ancestral base should be aligned
  - Shared domains in proteins (and important regions in nucleotide sequences) should align, even if the sequences are not similar overall
- Alignment should take into account biological mutations and other events
  - Point mutations
  - Insertions or deletions (indels)
  - Gene duplications and pseudogenes (a gene copy that does not produce a functional protein)
    - The human genome has up to 20,000 pseudogenes!

Sequence alignment example

- Consider the alignment of two hypothetical protein sequences:

  THISSEQUENCE and THATSEQUENCE

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

Sequence alignment example (different lengths)

- Now consider the alignment of two hypothetical protein sequences:

THATSEQUENCE and THISISASEQUENCE,

where the amino acids I, S, and A were inserted into one of the original sequences

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

- When aligning both sequences from the beginning
  - similarity which is obvious to us is lost
  - false matches are created because of differences in length

Sequence alignment example (different lengths)

- The solution is to introduce a **gap**, which corresponds to an insertion or a deletion and is usually indicated by a dash (-) in an alignment

```
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
```

- There are always multiple possible alignments, and the best alignment is not always obvious
- The alignment must be selected using a quantitative scoring measure

Sequence homology

- **Similarity** is a descriptive term indicating that two or more sequences have a certain degree of identity or likeness
- **Homologous sequences** (or homologues) are sequences that are descended from a common ancestor
- Homologous genes will accumulate different mutations (**divergent evolution**) during the course of evolution and their sequences are often not identical.
- **Convergent evolution** is when sequences with high similarity are not homologous
- Alignments cannot distinguish between homology and convergent evolution

Homology is more easily detected from protein sequences

- Number of possible characters in nucleotides vs. proteins?
- Matches in nucleotide sequences are more likely due to chance than matches in protein sequences
- The genetic code is redundant
  - Identical amino acid sequences can be encoded by different nucleotide sequences
  - Nucleotide sequences are more likely to change over time
- Structure and function of a protein is determined by its amino acid sequence (although this is determined by the nucleotide sequence)

Scoring alignments

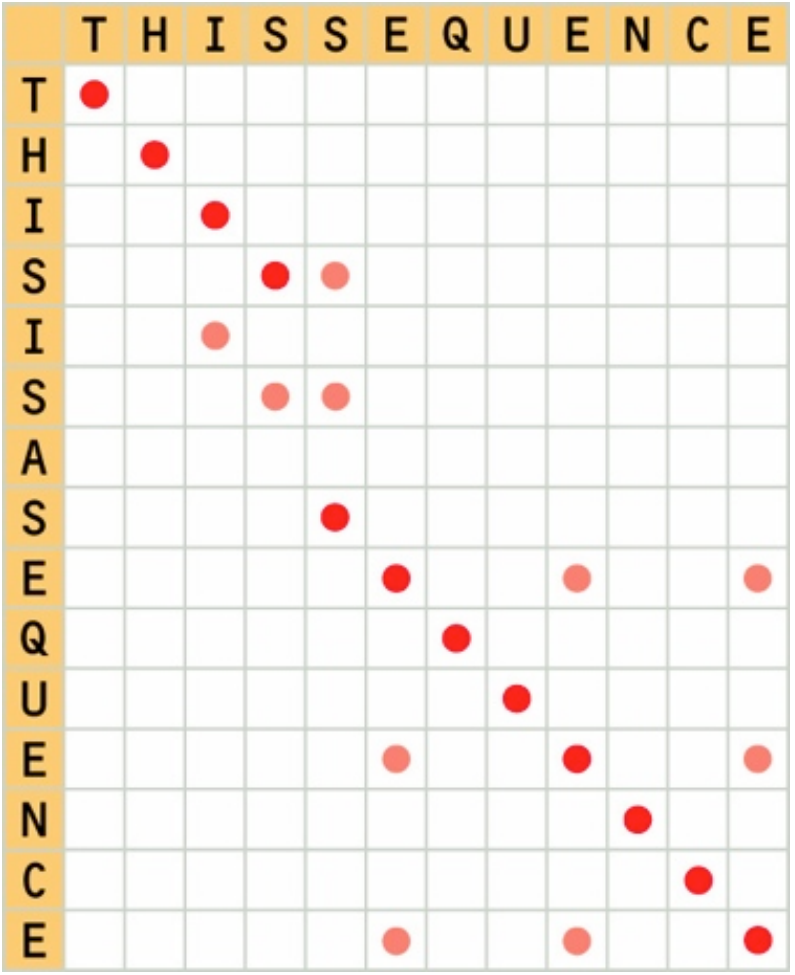
- Since multiple alignments are always possible, the best possible alignment is determined based on an alignment **score**
  - The **optimal alignment** is the alignment with the best score (multiple optimal alignments are possible)
  - **Suboptimal alignments** have slightly less scores than the best one
- The **percentage** or **percent identity** of an alignment is equal to the number of identical matches in an alignment divided by the length of the alignment (including gaps)

```
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
```

- The above alignment is optimal and has a percent identity of 11/16 = 68.75%

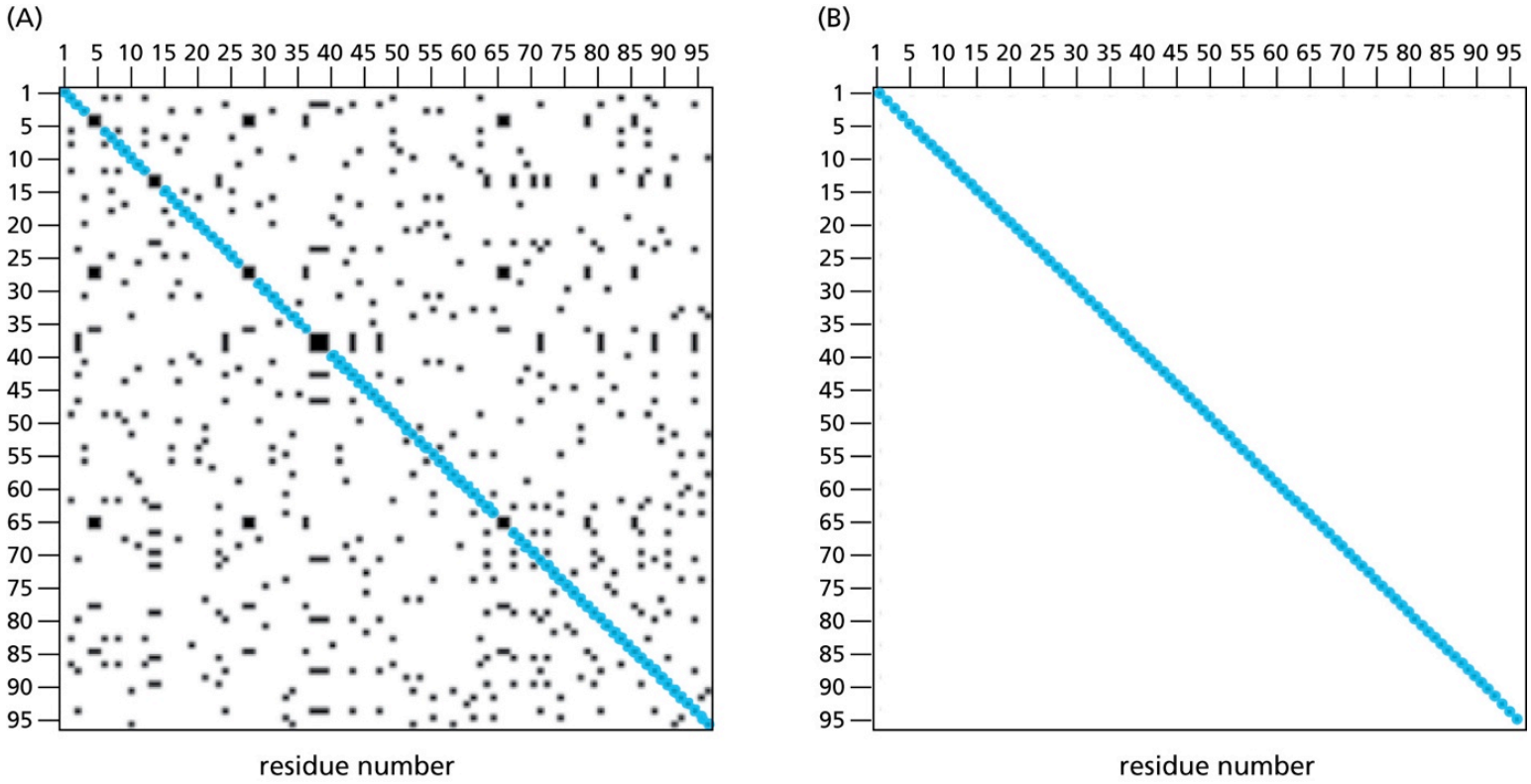
Dot-plots

- A **dot-plot** is a display of the alignment of two sequences that visualizes sequence similarity graphically
- A dot indicates identity between characters of each sequence
- Interruptions along the diagonal indicate a gap
- In addition to visualizing overall similarity, dot-plots can indicate intrasequence repeats



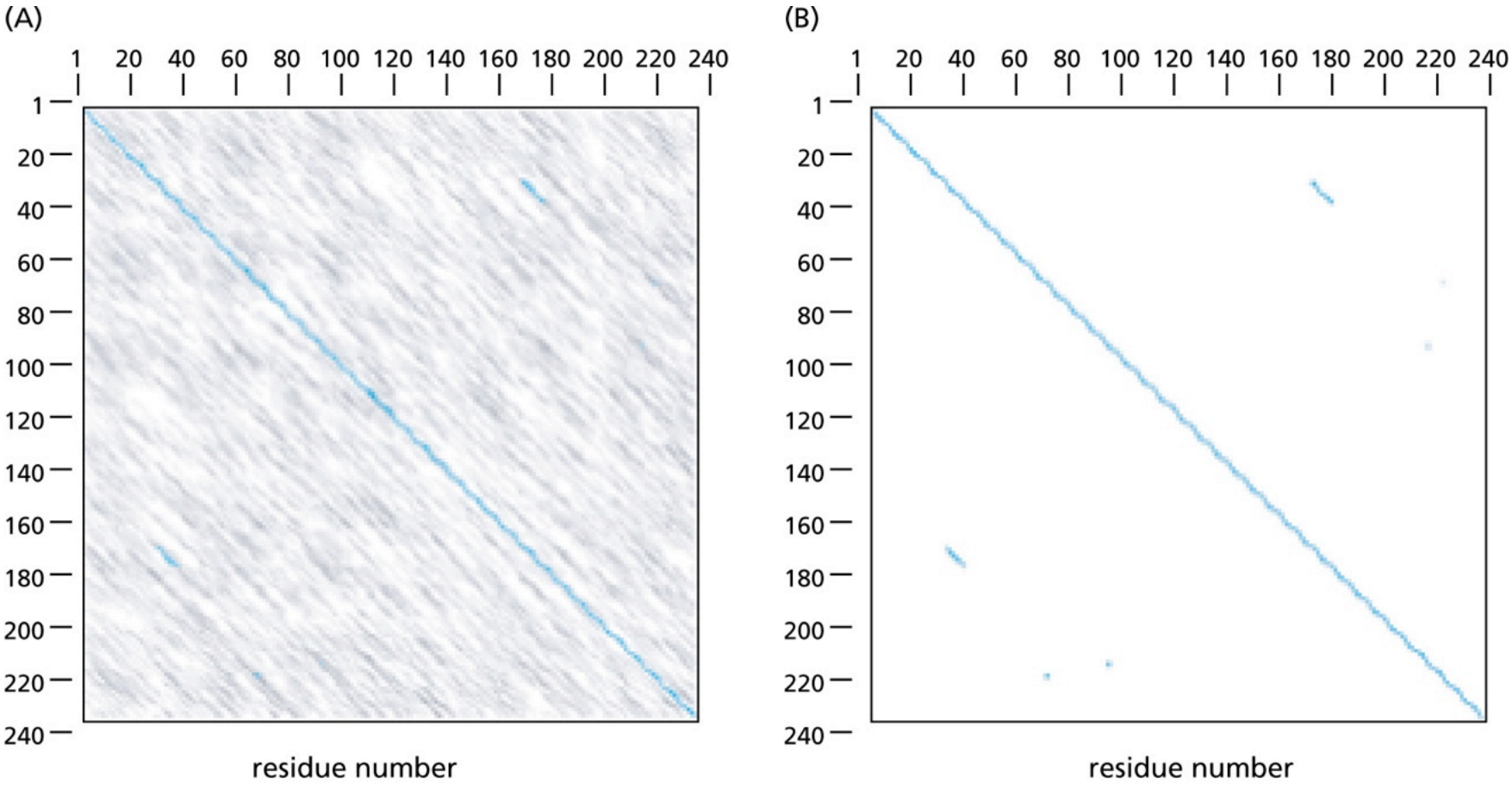
Dot-plots and background noise

A. Dot-plot of an SH2 domain with itself
B. The same dot-plot but with background noise removed, based on a window of 10 residues and a minimum identity score within each window of 3



Dot-plots showing BRCA2 repeat domain

Background is removed using a window of 30 and a minimum score of 5



Similarity versus identity

- Genuine matches do not have to be identical
- Certain non-identical amino acids may have
  - Similar physical and chemical properties
  - May be more likely to be present at the same region than others in related sequences
- Percent similarity is calculated in the same way as percent identity but both identical and similar matches are considered

```
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
```

- Isoleucine (I) and alanine (A) are hydrophobic; serine (S) and threonine (T) are polar
- Percent similarity is 12/15 = 80%

Substitution matrices

- For protein sequences, the score for each aligned pair of amino acids is determined by a **substitution matrix**, which has values for all possible pairs of residues.
- Example using BLOSUM-62 matrix:

Seq1:	T	H	I	S	S	E	Q	U	E	N	C	E
Seq2:	T	H	A	T	S	E	Q	U	E	N	C	E
Score:	5	8	-1	1	4	5	5	0	5	6	9	5

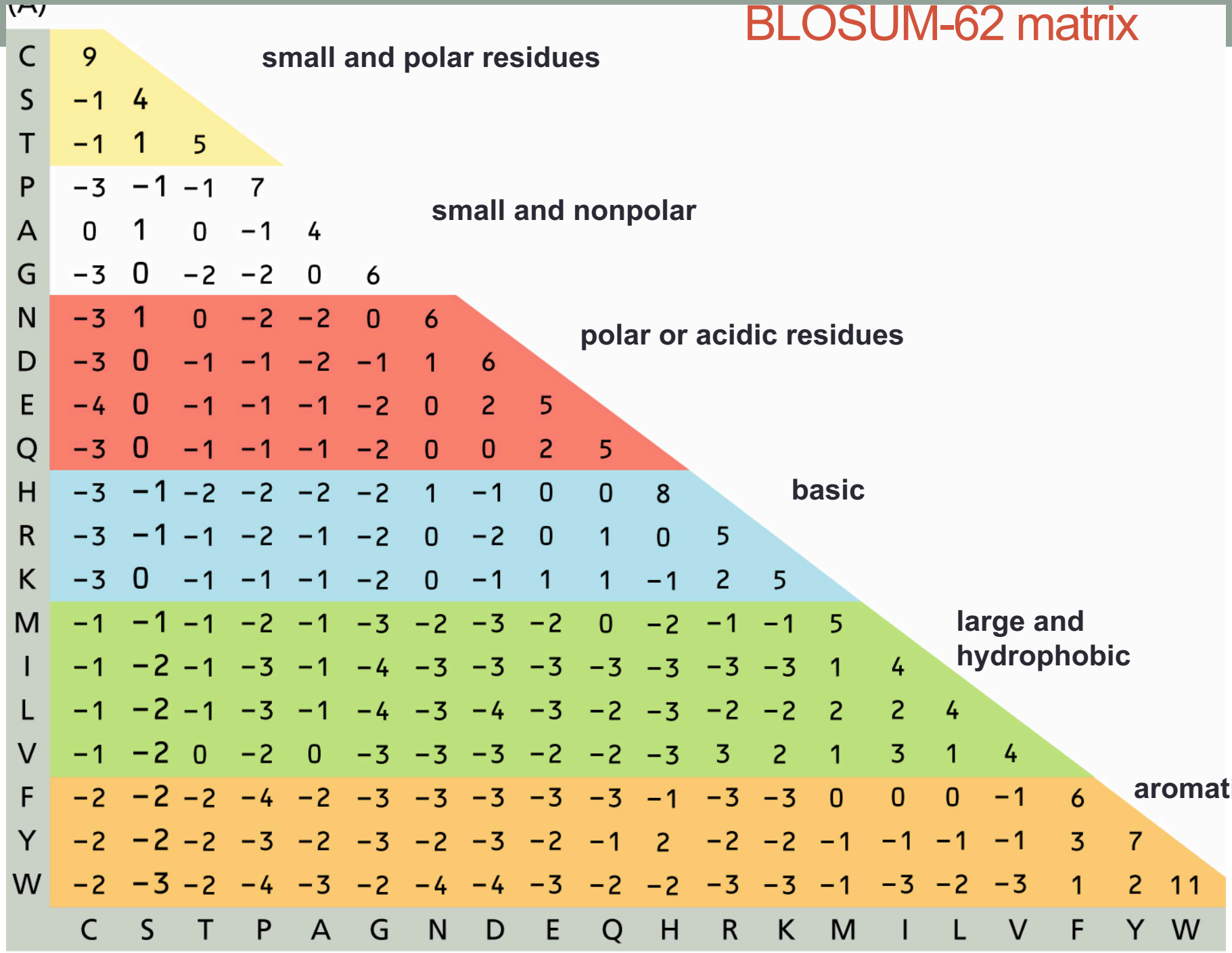
This alignment has an overall score (S) of 52

Note: there is no amino acid U, so the score for U paired with U is taken to be 0

Substitution matrices

- BLOSUM matrices
  - BLOck SUbstitution Matrix
  - Based on local alignments to detect conserved short regions
  - Sequences grouped based on percent identity, where the percent identify threshold for grouping determines the specific BLOSUM matrix
    - BLOSUM-62 is based on grouping aligned sequences with no more than 62% identity
  - Substitution frequencies are then calculated
  - Positive scores indicate conservative (more likely) substitutions
  - Negative scores indicate non-conservative (less likely) substitutions
  - All BLOSUM matrices are based on observed alignments

BLOSUM-62 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

Substitution matrices

- Point Accepted Mutation (PAM) matrices
  - Based on amino acid frequencies in alignment of similar and homologous protein sequences
  - Probabilities were calculated for whether a given amino acid mutates to any other over a given period of time
  - The logarithm of this probability gives the substitution score
  - Based on number of changes from each amino acid and total number of occurrences
  - There are multiple PAM matrices and the PAM # corresponds to the number of accepted point mutations per 100 residues.
  - All PAM matrices are based on PAM1; others are inferred.
  - For example, the PAM250 contains scores based on an expected evolutionary distance corresponding to 250 point accepted mutations for every 100 amino acid residues

PAM vs. BLOSUM Substitution matrices

- Choice depends on evolutionary distance
- For closely related sequences
  - Use higher BLOSUM number and lower PAM number
- For distantly related sequences
  - Use lower BLOSUM number or higher PAM number

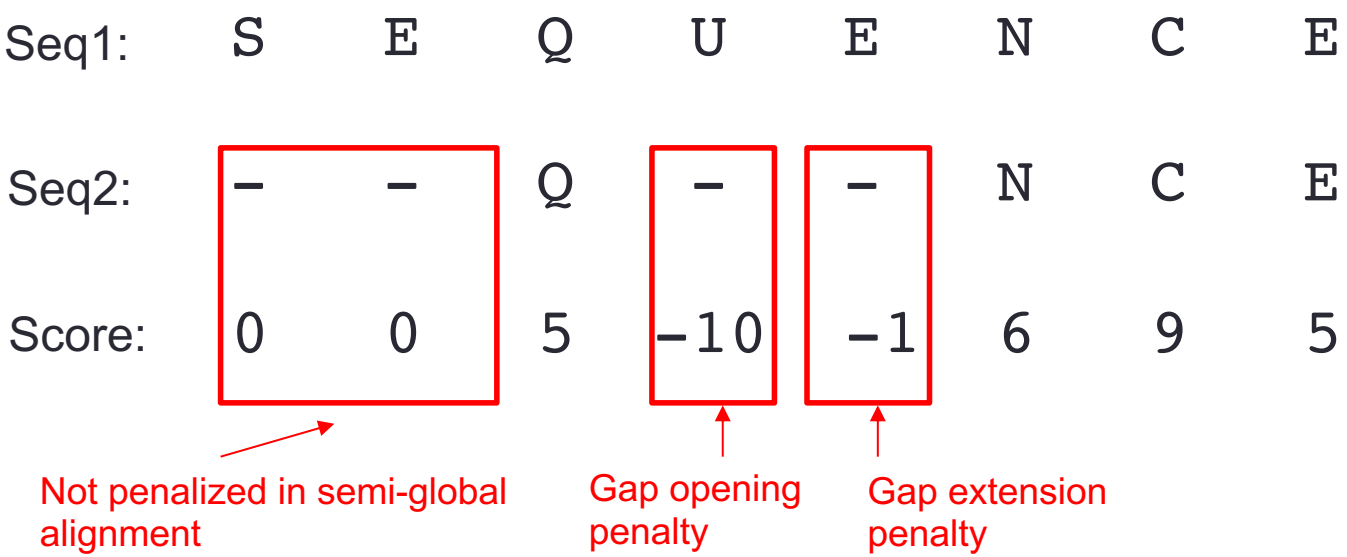
Inserting Gaps

- A **gap** in a sequence alignment indicates an insertion or deletion in the sequence
- When a gap is introduced, a **gap opening penalty** is added to the score
  - Insertions and deletions are not likely to occur in regions of structural importance
- Insertions tend to be several residues long
  - A smaller **gap extension** penalty is added each time a gap is extended
- Gaps cannot be aligned with each other

Gap Penalties

- A "gap" (composed of a sequence of gap characters in the alignment, e.g., - - -) has a penalty composed of a **gap opening penalty** for the initial character and a **gap extension penalty** for each subsequent character. Typically gaps are not penalized if they occur at the beginning or end of the alignment (this is known as a semi-global alignment)

- Here we use a gap opening penalty of 10 and a gap extension penalty of 1

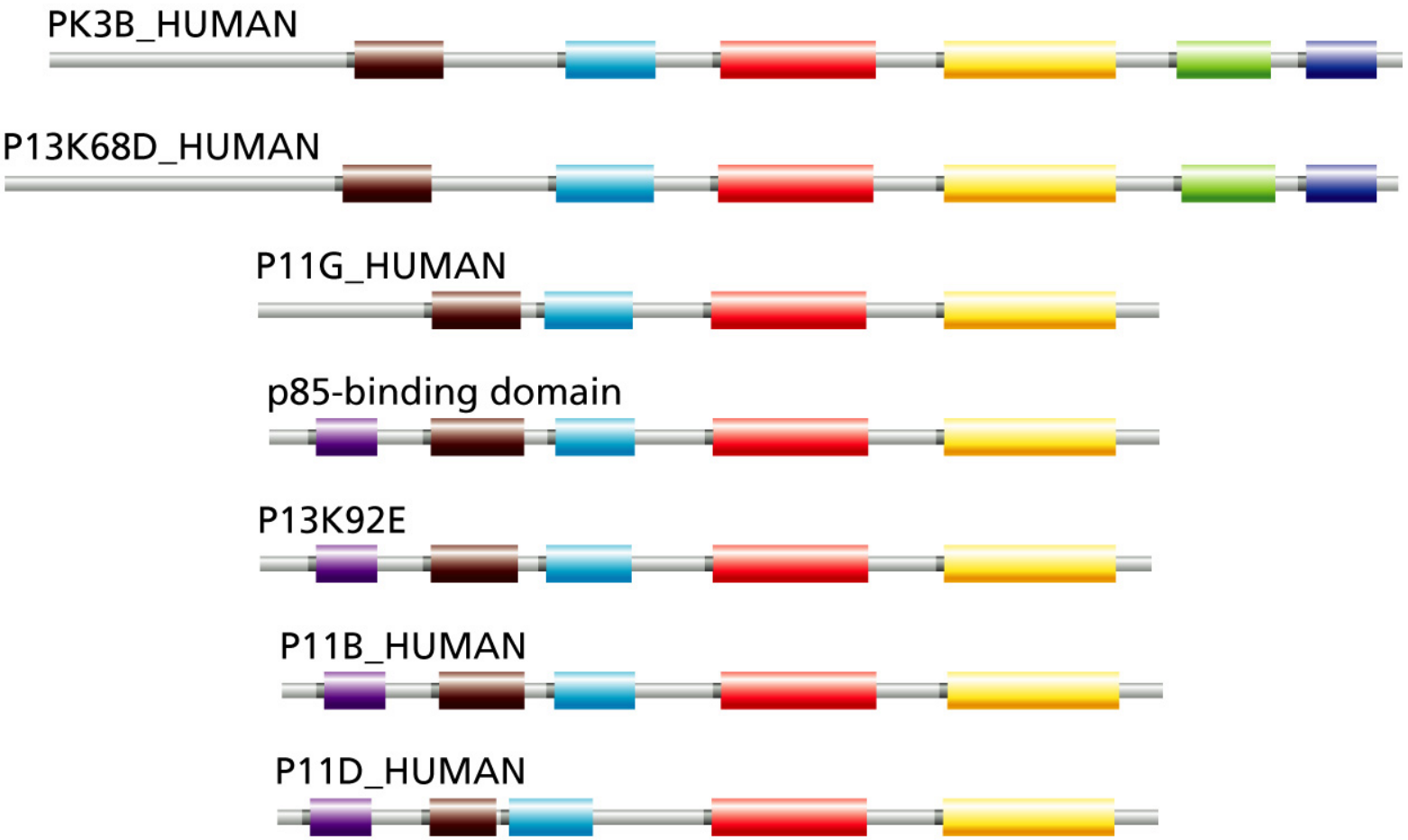


This semi-global alignment has an overall score (S) of 14

Types of alignments

- A (**semi**) **global alignment** aligns two sequences across their entire lengths
  - Appropriate for homologous sequences
- A **local alignment** detects shared regions (e.g., domains) which may be missed in global alignments
- A **pairwise alignment** is the alignment of two sequences
- A **multiple alignment** is the simultaneous alignment of more than two sequences

Many proteins have multiple domains



Local and global alignments

(A) local

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

(B) 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                  90                  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
```

Alignment algorithms (preview)

- Needleman-Wunsch (1970) and variations:
  - for aligning two sequences
  - uses dynamic programming to "consider" all possible alignments (10^{600} for two sequences of length 1000!)
- FASTA: uses a heuristic method for efficient searches (though not guaranteed to find the optimal solution)
  - Creates dictionary of *k*-tuples for the query sequence which is checked against sequences in the database
  - A local alignment algorithm is used to complete the alignment
- BLAST (Basic Local Alignment Search Tool): also fast and uses a heuristic
  - Finds short matches (which do not have to be exact)
  - Then uses local alignment to complete the alignment