Bioinformatics I

Pairwise alignment of DNA & protein (using matrices)

Dr Manaf A Guma

University Of Anbar- college of Applied sciences-Hit

Department of applied chemistry

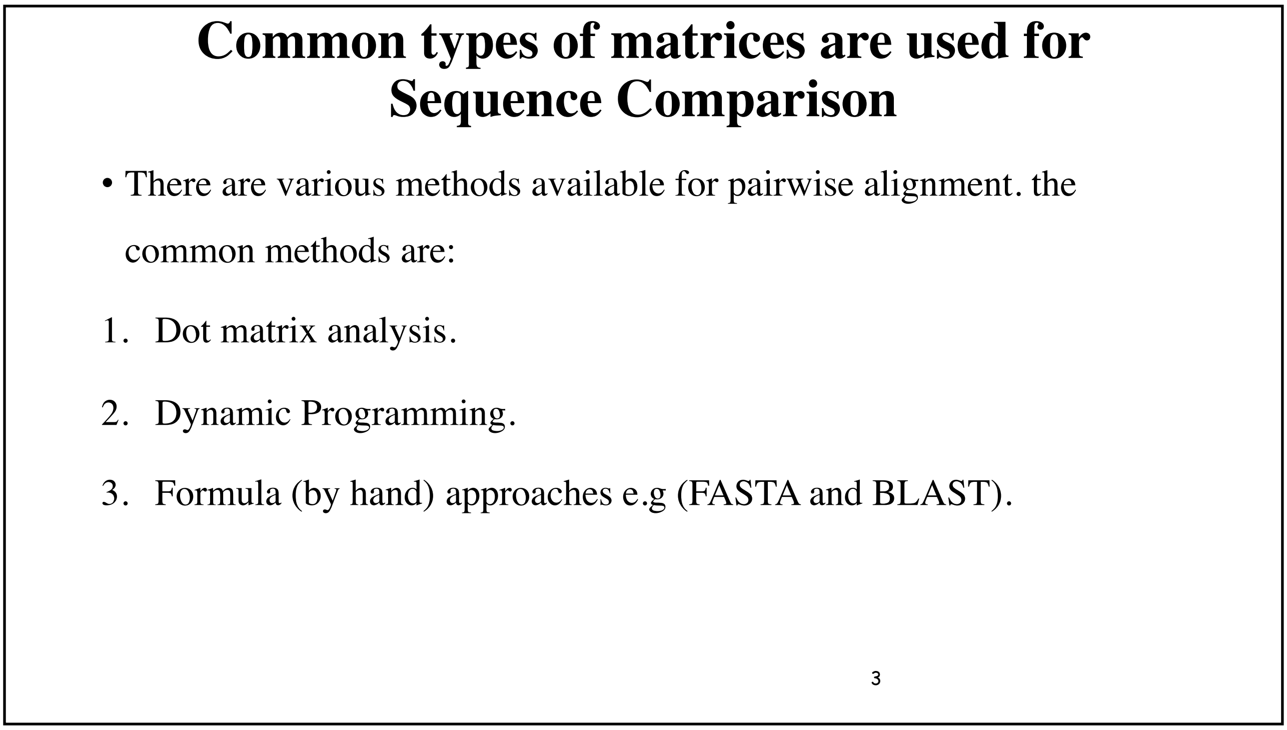
1

What are the purposes of pairwise alignment comparison?

- The purposes of pairwise **alignment** comparison are (using the matrices or the manual methods):

1. To find the score of the identity between two sequences.
2. To find whether two (or more) genes or proteins are evolutionarily related to each other.
3. To find structurally or functionally similar regions within proteins

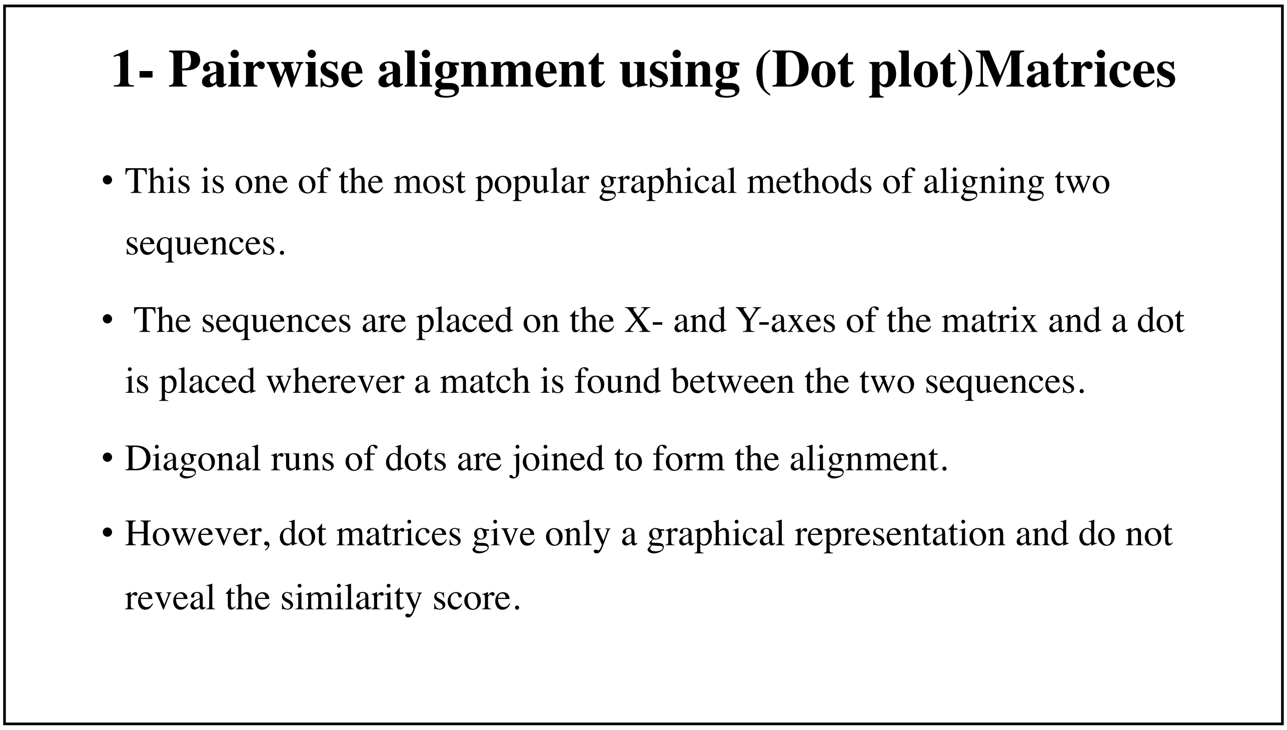
2



Common types of matrices are used for Sequence Comparison

- There are various methods available for pairwise alignment. the common methods are:

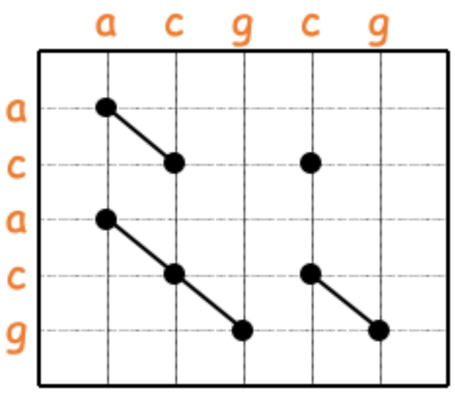
1. Dot matrix analysis.
2. Dynamic Programming.
3. Formula (by hand) approaches e.g (FASTA and BLAST).

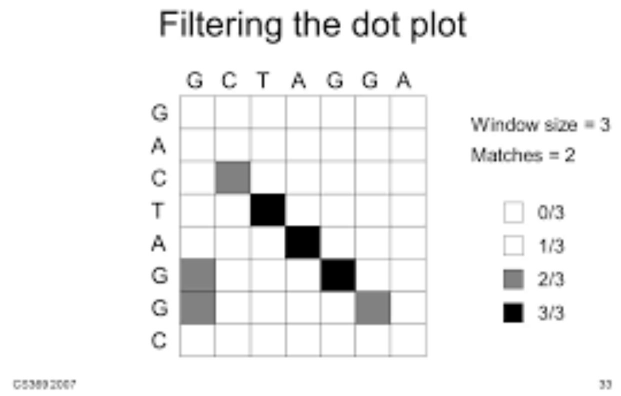


1- Pairwise alignment using (Dot plot)Matrices

- This is one of the most popular graphical methods of aligning two sequences.
- The sequences are placed on the X- and Y-axes of the matrix and a dot is placed wherever a match is found between the two sequences.
- Diagonal runs of dots are joined to form the alignment.
- However, dot matrices give only a graphical representation and do not reveal the similarity score.

How do ‘’Dot matrices” look like?




Filtering the dot plot

	G	C	T	A	G	G	A
G							
A							
C		▨					
T			■				
A				■			
G	▨				■		
G	▨					▨	
C							

Window size = 3
Matches = 2

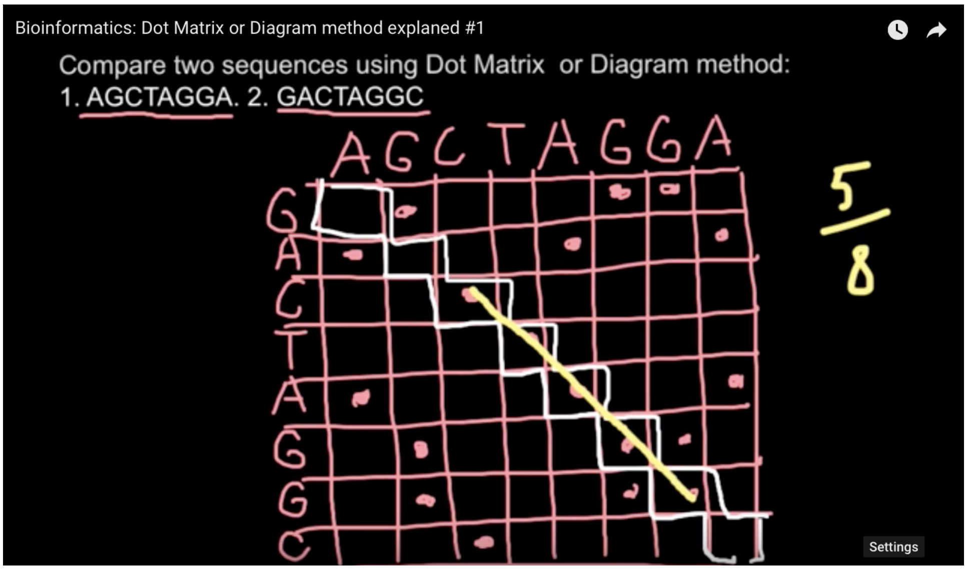
□ 0/3
□ 1/3
▨ 2/3
■ 3/3

CS389 2007

33

5

Dot matrix comparison by finding how many coincidences for alignment

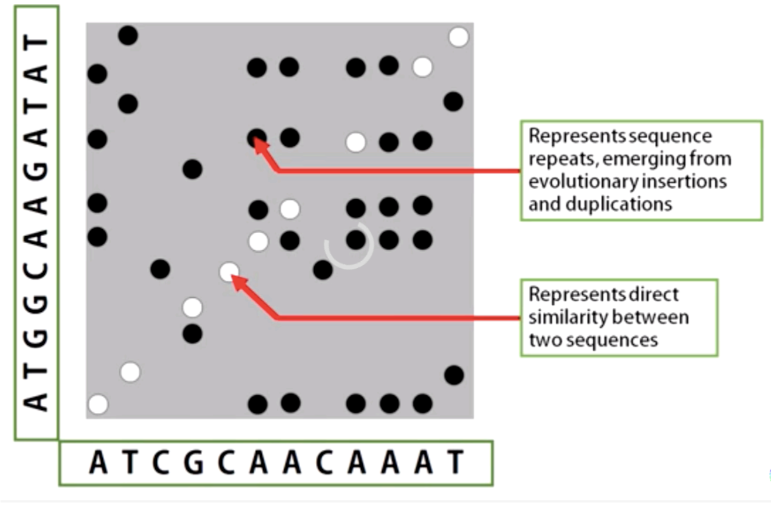

Bioinformatics: Dot Matrix or Diagram method explaned #1

Compare two sequences using Dot Matrix or Diagram method:
1. AGCTAGGA. 2. GACTAGGC

6

6

Another example



Represents sequence repeats, emerging from evolutionary insertions and duplications

Represents direct similarity between two sequences

7

Give an interpretation for the matrices.

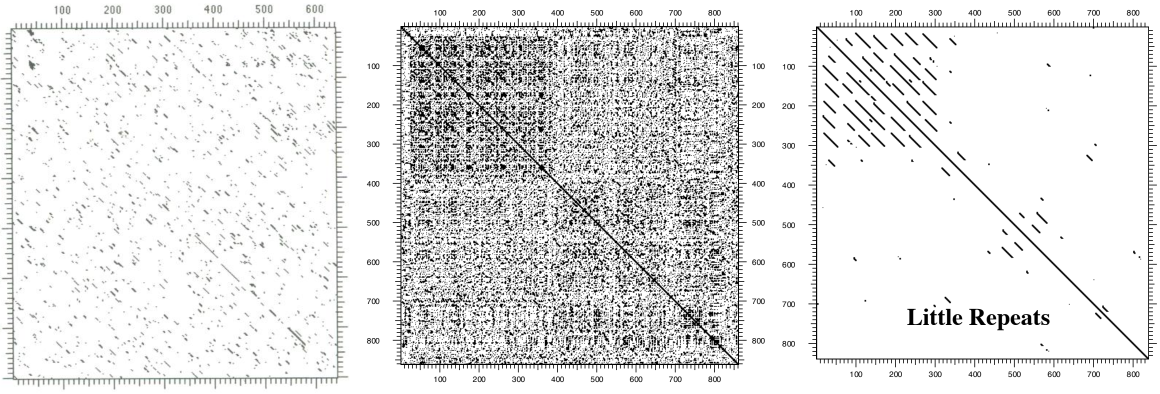
1. Regions of similarity appear as diagonal runs of dots.
2. Reverse diagonals (perpendicular to diagonal) indicate inversions
3. Reverse diagonals crossing diagonals (Xs) indicate palindromes.
4. Link can separate diagonals to form **alignment** with *gaps;* each amino acid. or base can only be used once (Can't double back)

8

8

What are the artifact of Dot matrices? By Filtering?

- Dot matrices for long sequences can be noisy due to insignificant matches.



9

What are the uses of dot matrices for?

1. Aligning two proteins or two nucleic acid sequences.
2. Finding amino acid repeats within a protein by comparing a protein sequence to itself.
3. Repeats appear as a set of diagonal runs stacked vertically and/or horizontally.

10

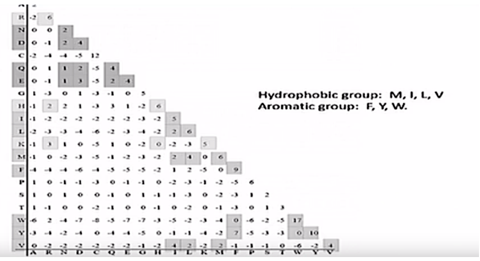
2- PAM matrices

- **P**oint **a**ccepted **m**utation **matrix** known as a PAM.
- It is also called **P**ercent **A**ccepted **M**utation.
- Dayhoff and colleagues defined the PAM1 matrix as that which produces 1 accepted point mutation per 100 amino acid residues.
- **PAM matrix** is designed to compare two sequences which are a specific number of PAM units apart.
- Only mutations are allowed.
- https://www.youtube.com/watch?v=UCtP5-KtB94, https://www.youtube.com/watch?v=F8WdDfpQqCM

11

PAM matrices are calculated by BLAST websites

- PAM matrices are also used as a scoring matrix when comparing DNA sequences or protein sequences to judge the quality of the alignment.
- This form of scoring system is utilized by a wide range of alignment software including BLAST.
- PAM250 corresponds to 20% amino acid identity, represents 250 mutations per 100 residues.
- If you times (multiply) PAM1 by itself 250 times you will get substitution matrix like this:

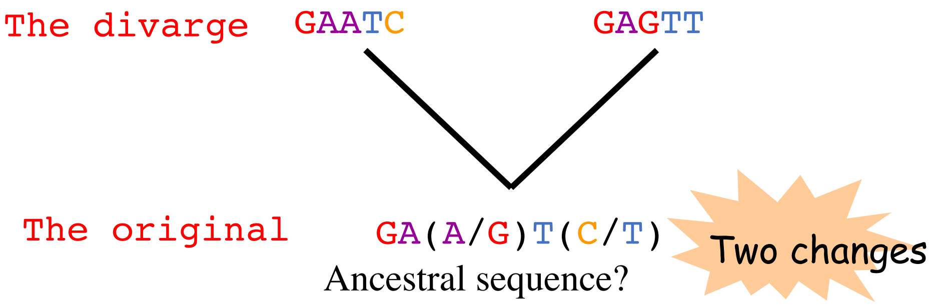


	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

Hydrophobic group: M, I, L, V
Aromatic group: F, Y, W.

12

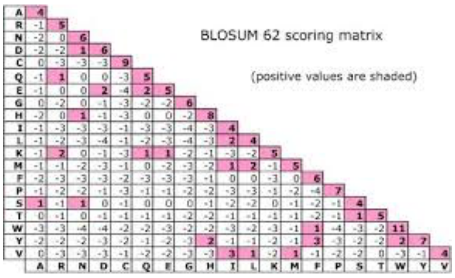
What are PAM matrices based on? PAM matrices are based on a simple evolutionary model



- So, only mutations are allowed
- Sites evolve independently

3- BLOSUM Matrices

1. **Blo**cks **Su**bstitution **M**atrix.
2. It is a matrix that calculates scores for each position which are obtained frequencies of substitutions in blocks of local alignments of protein sequences.
3. For example BLOSUM62 is derived from sequence alignments with no more than 62% identity.



BLOSUM 62 scoring matrix

(positive values are shaded)

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

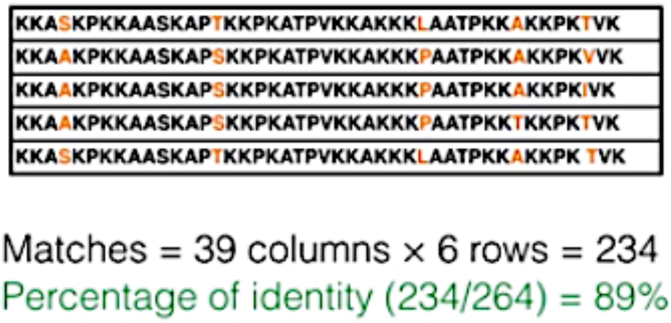
What BLOSUM is based on?

- It is based on comparisons of **blocks of sequences** derived from the Blocks database.
- The block length is 60 amino acids. (without any gaps or frequencies).
- Blocks database refers to the alignment not to the individual sequence.
- BOLSUM matrices tell the % of matching.
- It can be 100% even if there is a substitution.
- It tells how much the sequence is conserved!

15

What does block mean in BLOUSM method?

- It means creating a block of 2 seq or multiple sequences which refers to a best alignment in order to recognize the mutations, gaps and penalties in each row.
- The block presents the same length of sequences ' about 60 letter of amino acids or nucleotide'.



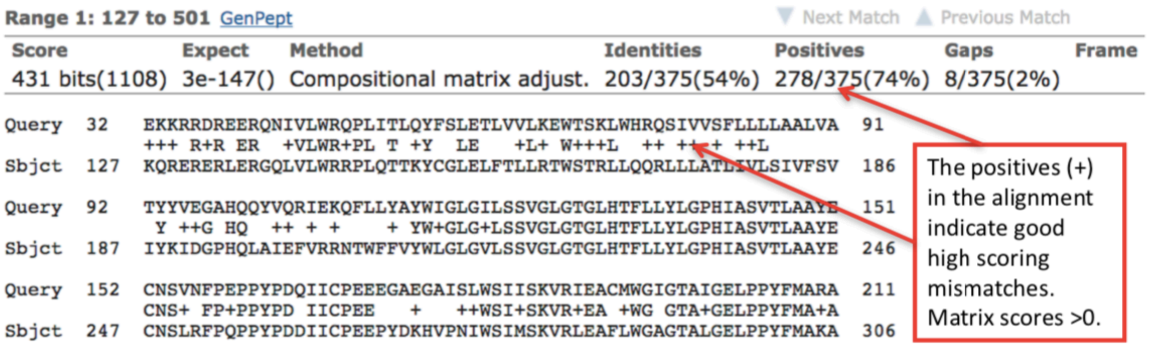
```
KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVK
KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKVVK
KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKIVK
KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKTKKPKTVK
KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPK TVK
```

Matches = 39 columns × 6 rows = 234

Percentage of identity (234/264) = 89%

16

BLOSUM in BLAST



```
Range 1: 127 to 501 GenPept                                  ▼ Next Match  ▲ Previous Match
Score            Expect     Method                        Identities     Positives      Gaps       Frame
431 bits(1108)  3e-147()  Compositional matrix adjust.  203/375(54%)  278/375(74%)  8/375(2%)

Query  32   EKKRRDREERQNIVLWRQPLITLQYFSLETLVVLKEWTSKLWHRQSIVVSFLLLLAALVA  91
            +++ R+R ER  +VLWR+PL T +Y  LE   +L+ W+++L  ++ ++ + ++L
Sbjct  127  KQRERERLERGQLVLWRRPLQTTKYCGLELFTLLRTWSTRLLQQRLLLATLIVLSIVFSV  186

Query  92   TYYVEGAHQQYVQRIEKQFLLYAYWIGLGILSSVGLGTGLHTFLLYLGPHIASVTLAAYE  151
             Y ++G HQ  ++ + +    + YW+GLG+LSSVGLGTGLHTFLLYLGPHIASVTLAAYE
Sbjct  187  IYKIDGPHQLAIEFVRRNTWFFVYWLGLGVLSSVGLGTGLHTFLLYLGPHIASVTLAAYE  246

Query  152  CNSVNFPEPPYPDQIICPEEEGAEGAISLWSIISKVRIEACMWGIGTAIGELPPYFMARA  211
            CNS+ FP+PPYPD IICPEE   +   ++WSI+SKVR+EA +WG GTA+GELPPYFMA+A
Sbjct  247  CNSLRFPQPPYPDDIICPEEPYDKHVPNIWSIMSKVRLEAFLWGAGTALGELPPYFMAKA  306
```

The positives (+) in the alignment indicate good high scoring mismatches. Matrix scores >0.

Mis-matching !!!!!

17