Gene Sequence Analysis

Lecture 1: **Homology Searching**

23/05/2024

Phuc-Loi Luu, PhD

Adapted from Dr. Morgan Langille and Dr. G P S Raghava


22 22 X X X X X X Y Y Y
17 17 17 18 18 18 19 19 19 20 20 20 21 21
13 13 14 14 14 14 15 15 15 15 16 16 16 16
10 10 10 10 11 11 11 11 11 12 12 12 12 12 13 13
7 7 7 7 8 8 8 8 8 8 9 9 9 9 9 10
4 5 5 5 5 5 5 5 6 6 6 6 6 7 7
2 2 3 3 3 3 3 3 3 4 4 4 4 4 4
1 1 1 1 1 1 2 2 2 2 2 2 2

Outline

- What is homology, orthologs, paralogs?
- Local vs global alignment
- E-values, bit scores, "coverage", identity vs similarity
- Different blast flavours (blastn, blastp, tblastn, etc.)
- Blast (Web)

What is homology?

- Homology refers to shared ancestry
- Two sequences are homologous if they are derived from a common ancestral sequence
- One sequence by itself is not informative
  - it must be analyzed by comparative methods against existing sequence databases to develop hypothesis concerning relatives and function.

What is homology?

Mouse *Pax6* gene:

GTATCCAACGGTTGTGTGAGTAAAATTCTGGGCAGGTATTACGAGACTGGCTCCATCAGA

Fly *eyeless* gene:

Genetic similarity to mouse: 76.66%
Protein similarity to mouse: 100%

GTATCAAATGGATGTGTGAGCAAAATTCTCGGGAGGTATTATGAAACAGGAAGCATACGA

Shark eye control gene:

Genetic similarity to mouse: 85%
Protein similarity to mouse: 100%

GTGTCCAACGGTTGTGTCAGTAAAATCCTGGGCAGATACTATGAAACAGGATCCATCAGA

Squid eye control gene:

Genetic similarity to mouse: 78.33%
Protein similarity to mouse: 100%

GTCTCCAACGGCTGCGTTAGCAAGATTCTCGGACGGTACTATGAGACGGGCTCCATAAGA

Flatworm eye control gene:

Genetic similarity to mouse: 71.66%
Protein similarity to mouse: 100%

GTGTCTAATGGTTGTGTTAGTAAAATACTTTGCCGATATTATGGAACAGGTTCTATTAAA

https://evolution.berkeley.edu/evolibrary/article/1_0_0/eyes_10

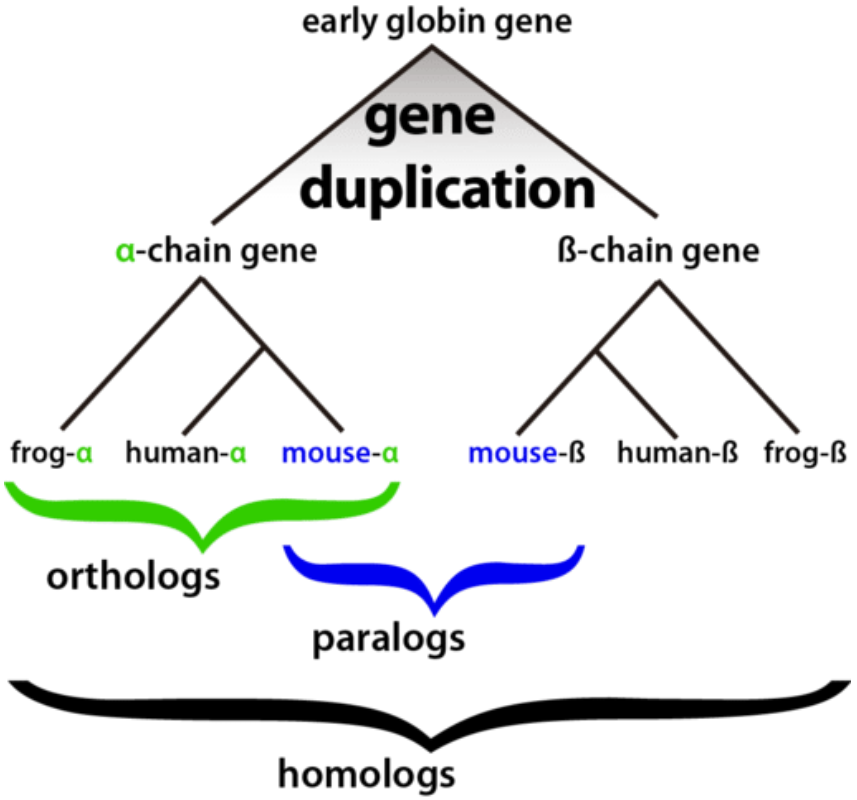
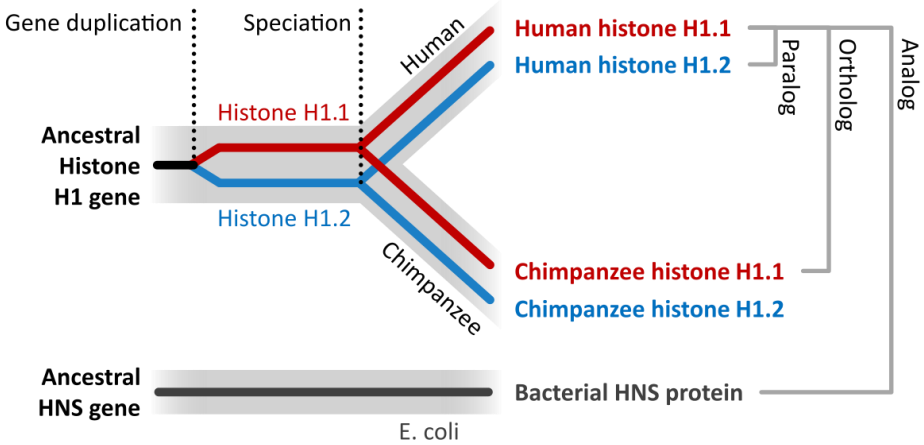


Photo courtesy of: Popo H. Liao (Own work), via Wikimedia Commons

Types of homologs

- Orthologs
  - Think same gene in different organism
  - Often thought to have similar function
- Paralogs
  - Think gene duplication
  - Less likely to have similar function



https://en.wikipedia.org/wiki/Sequence_homology

What is similarity?

- Similarity is a measure of the likeness between sequences.
- Gene searching tools calculate the similarity between sequences and rank more similar sequences higher.
- Sequences can NOT be partially homologous
  - WRONG: Gene X is 80% homologous to Gene Y
- Sequences can be partially similar
  - CORRECT: Gene X has 80% identity to Gene Y

Identity vs Similarity

- Identity is a percentage measurement that states how many characters in the sequence are identical

- Similarity can also be used as a metric which means how many characters are “positive scoring”

Identity vs Similarity

Similarity in sequence alignment is the resemblance between two sequences when compared. This fact is dependent on the identity of sequences. Similarity depicts the extent to which the residues are aligned. Hence, similar sequences contain similar properties. In bioinformatics, similarity is a tool to assess the likeness between two proteins.

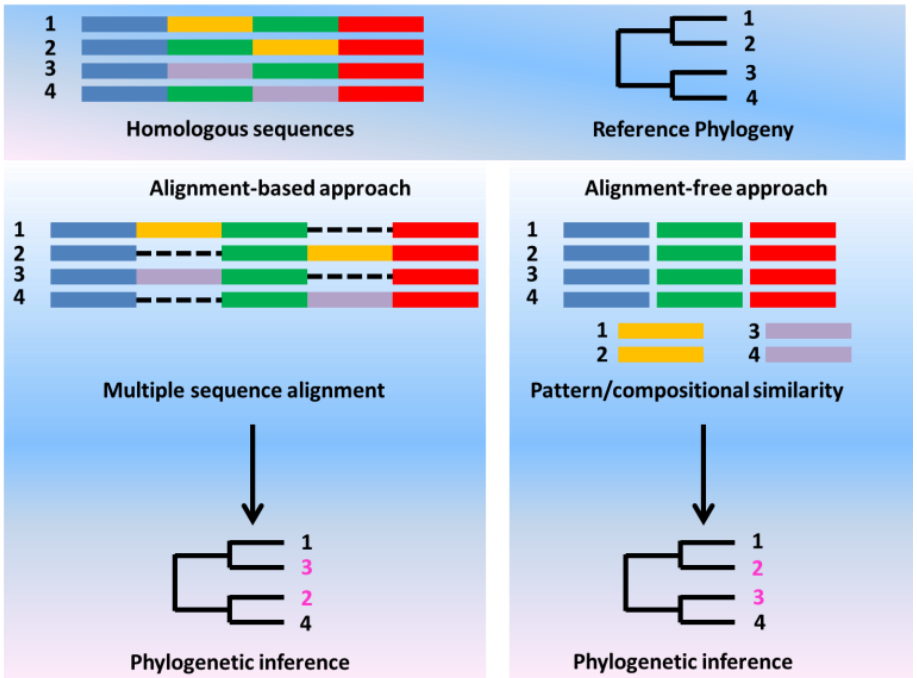
Identity in sequence alignment is the number of characters that match exactly between two different sequences. Hence, gaps do not count when assessing identity. The measurement is considered to be relational to the shorter sequence among the two sequences. It significantly implies that it has the effect where the sequence identity is not transitive. If X=Y and Y=Z, then X is not necessarily equal to Z. This is deduced in terms of the identity distance measure.

X = AAGGCTT, Y = AAGGC and Z = AAGGCAT.

Identity between X and Y is 100% {5 identical nucleotides / min[length(X),length(Y)]}.

Identity between Y and Z is also 100%.

But identity between X and Z is only 85% {(6 identical nucleotides / 7)}.



Selected annotation from match **P03372**

Accession	Entry name	Status	Protein names	Organism	Length	Gene names
P03372	ESR1_HUMAN	★	**Estrogen receptor**	Homo sapiens (Human)	595	**ESR1** ESR NR3A1

Alignment 1 against **P03372**

Score	218	**E-value**	4.0×10^{-21}
Identity	48.0%	**Positives**	68.0%
Query length	76	**Match length**	595
Position	**P03372** matches from 183 to 252 (70AA), in the query sequence from 2 to 71 (70AA)		

```
2     RICGVCGDRATGFHFNAMTCEGCKGFFRRSMKRKALFTCPFNGDCRITKDNRRHCQACRL    61 Query
      R C VC D A+G+H+   +CEGCK FF+RS++    + CP    C I K+ R+ CQACRL
183   RYCAVCNDYASGYHYGVWSCEGCKAFFKRSIQGHNDYMCPATNQCTIDKNRRKSCQACRL   242 P03372

62    KRCVDIGMMK    71 Query
      ++C ++GMMK
243   RKCYEVGMMK   252 P03372
```


https://www.differencebetween.com/difference-between-similarity-and-identity-in-sequence-alignment

Identity vs Similarity

Identity

The extent to which two (nucleotide or amino acid) sequences are **invariant** (identical).

Similarity

The extent to which (nucleotide or amino acid) sequences are **related**. The extent of similarity between two sequences can be based on percent sequence identity and/or conservation. In BLAST similarity refers to a positive matrix score. This is quite flexible (see later examples of DNA polymerases) – similar across the whole sequence *or* similarity restricted to domains !

Homology

Similarity attributed to **descent from a common ancestor**.

Assessing Sequence Similarity

```
Rbn                                   KETAAAKFERQHMD
Lsz     KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNT

Rbn     SST  SAASSSNYCNQMMKSRNLTKDRCKEMNTFVHESLA
Lsz    QATNRNTDGSTDYGILQINSRWWCNDGRTP      GSRN

Rbn    DVQAVCSQKNVACKNGQTNCYQSYSTMSITDCRETGSSKY
Lsz    LCNIPCSALLSSDITASVNC  AKKIVSDGDGMNAWVAWR

Rbn    PNACYKTTQANKHITVACEGNPYVPHFDASV
Lsz    NRCKGTDVQA   WIRGCRL
```

is this alignment significant?

DNA scoring systems

```
Sequence 1  ACTACCAGTTCATTTGATACTTCTCAAA
                     | |      |        ||
Sequence 2      TACCATTACCGTGTTAACTGAAAGGACTTAAAGACT
```

	A	C	G	T
A	**1**	0	0	0
C	0	**1**	0	0
G	0	0	**1**	0
T	0	0	0	**1**

Match:	5	x	1	=	5
Mismatch:	19	x	0	=	0
Score:					5

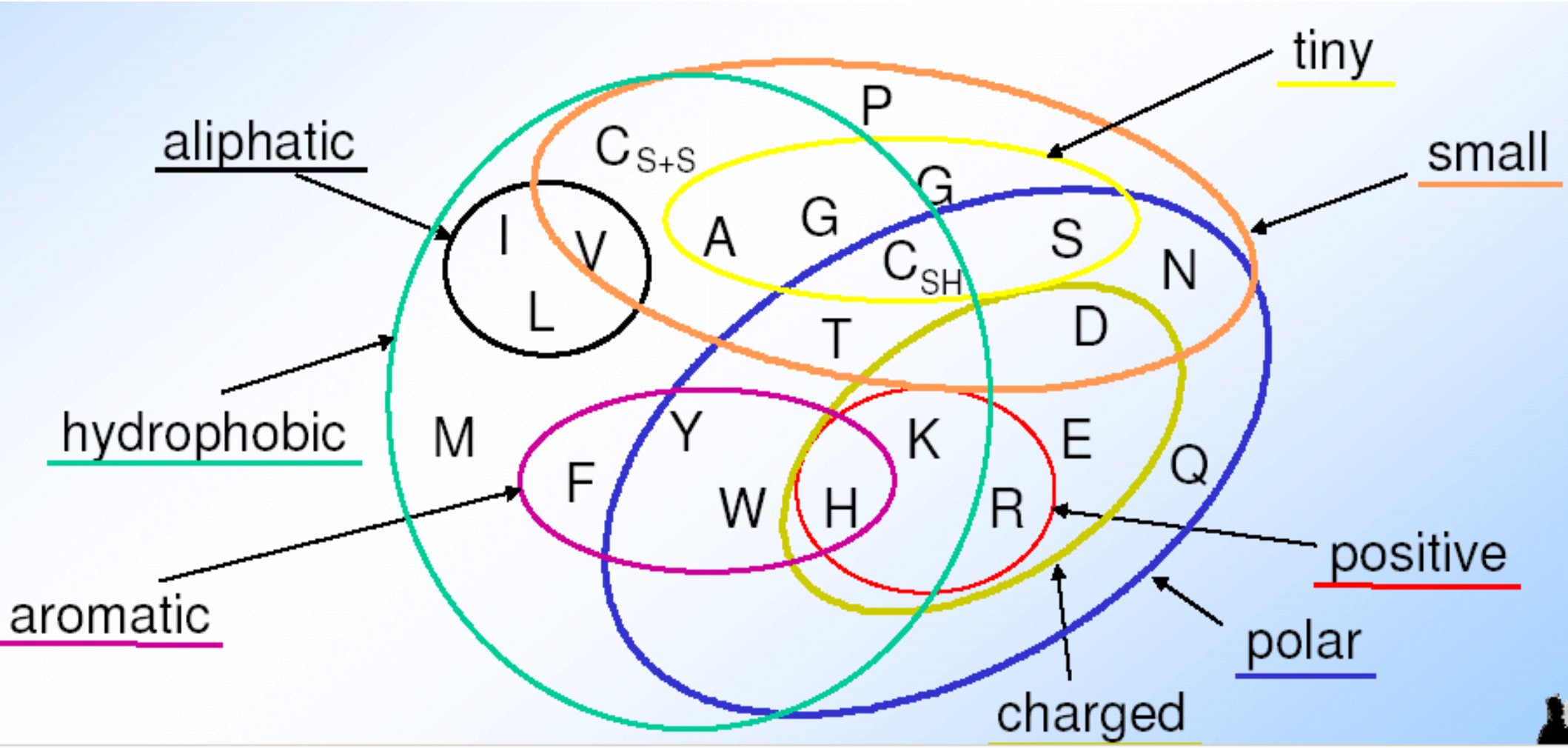
The Scoring Schemes or Weight Matrices

Genetic Code Scoring

- Fitch 1966 based on Nucleotide Base change required (0,1,2,3)
- Required to interconvert the codons for the two amino acids
- Rarely used nowadays

Complication: **„inexact“ is not binary (1|0) but something *relative***

Amino acids have different physical and biochemical properties that are/are not important for function and thus influence their probability to be replaced in evolution



The Scoring Schemes or Weight Matrices

Chemical Similarity Scoring

- ❖Similarity based on Physio-chemical properties
- ❖MacLachlan 1972, Based on size, shape, charge and polar
- ❖Score 0 for opposite (e.g. E & F) and 6 for identical character

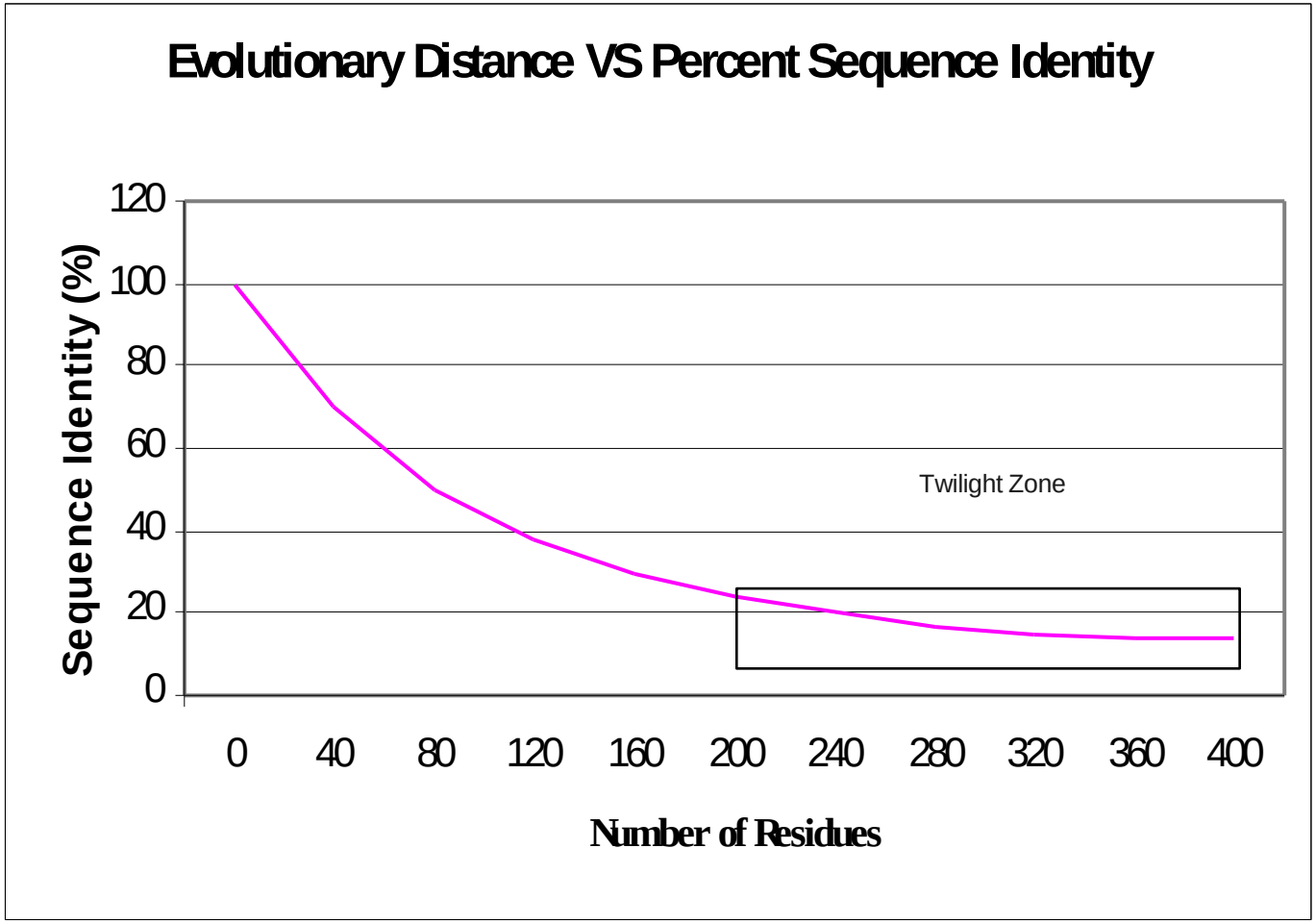
The Scoring Schemes or Weight Matrices

Observed Substitutions or PAM matrices

- Based on Observed Substitutions
- Chicken and Egg problem
- Dayhoff group in 1977 align sequence manually
- Observed Substitutions or point mutation frequency
- MATRICES are PAM30, PAM250, PAM100 etc

```
AILDCTGRTG……
ALLDCTGR--……
SLIDCSAR-G……
AILNCTL-RG……
```

Twilight Zone



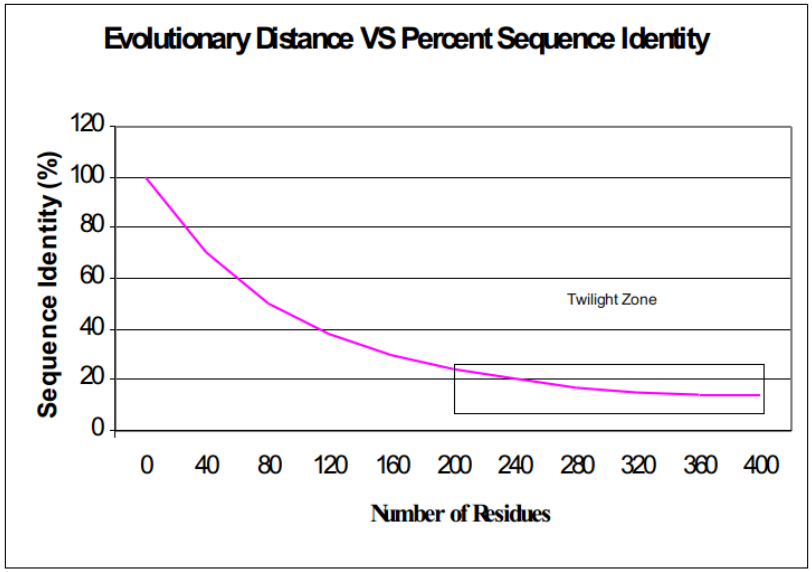
Some Simple Suggestions

- **If two sequence are > 100 residues and > 25% identical, they are likely related**
- **If two sequences are 15-25% identical they may be related, but more tests are needed**
- **If two sequences are < 15% identical they are probably not related**

Importance of Similarity

Rule-of-thumb:
If your sequences are more than **100 amino acids** long (or 100 nucleotides long) you can considered them as homologues if **25%** of the **aa** are identical (**70%** of **nucleotide** for DNA). Below this value you enter the twilight zone.

Twilight zone = protein sequence similarity between ~0-20% identity: is not statistically significant, i.e. could have arisen by chance.



Global vs Local

- Alignments can be global or local (this is algorithm specific)
  - A global alignment is an optimal alignment that includes all characters from each sequence (Multiple Sequence Alignment)
  - A local alignment is an optimal alignment that includes only the most similar local region or regions (e.g BLAST).

Local Alignment

```
Target Sequence
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| |||||||||||||||
Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Global Alignment

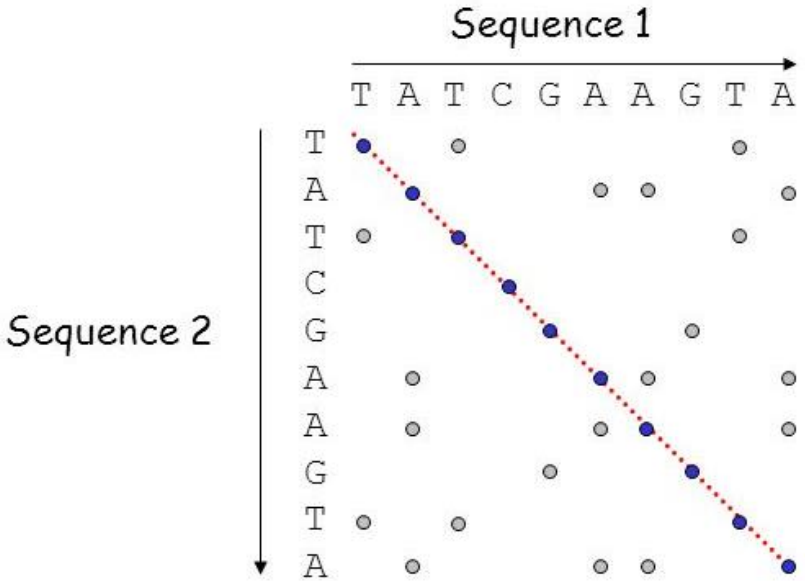
```
Target Sequence
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
Query Sequence
```


https://www.majordifferences.com/2016/05/difference-between-global-and-local.html

Dot Plots

Dotplot gives an overview of all possible alignments

The ideal case: two identical sequences



Every word in one sequence is aligned with each word in the second sequence

The dotplot generates a diagonal

But there are more matches

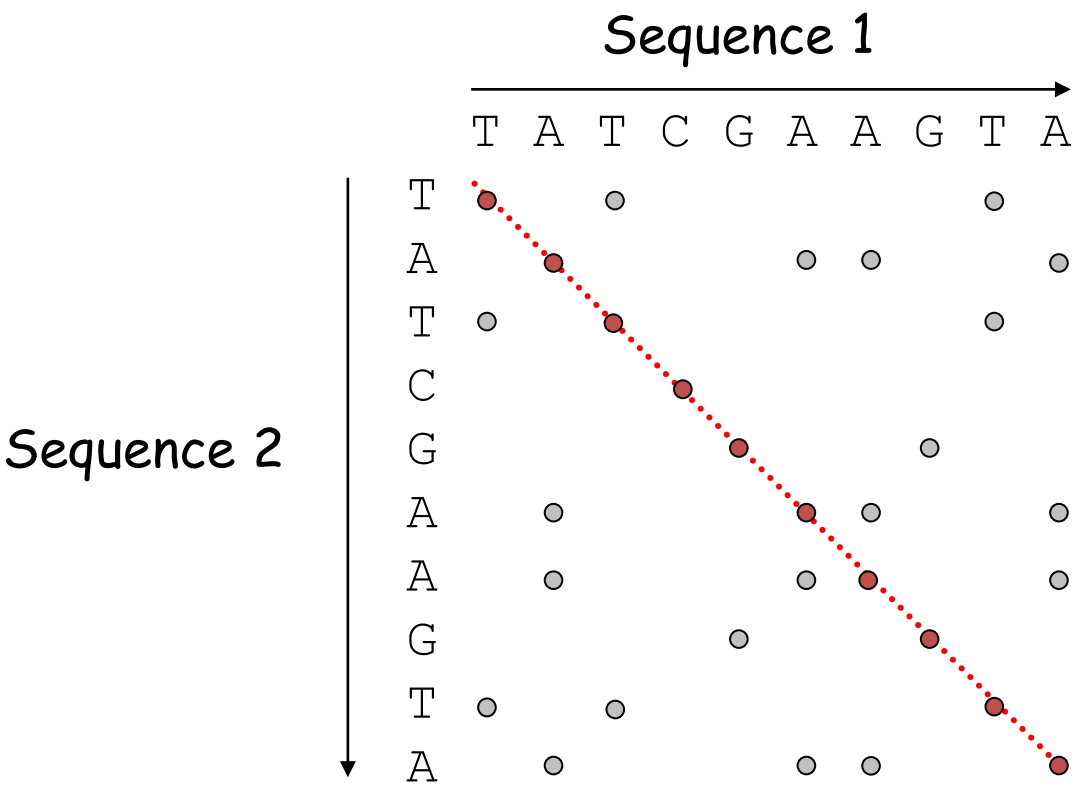
which are either meaningful, or noise

- **Popular freeware package is Dotter**
  http://sonnhammer.sbc.su.se/Dotter.html

Dotplot

Dotplot gives an overview of all possible alignments

The ideal case: two identical sequences



Every word in one sequence is aligned with each word in the second sequence

The dotplot generates a diagonal

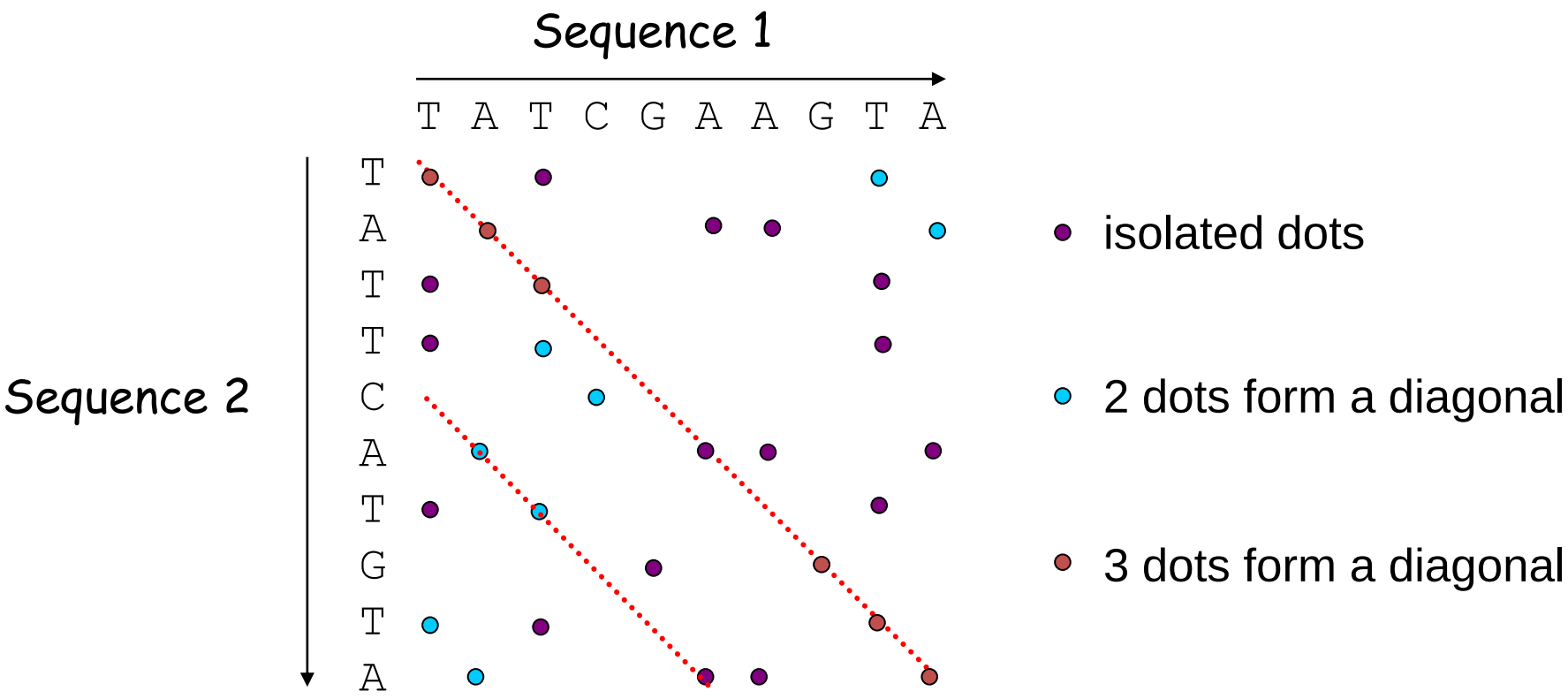
But there are more matches

which are either meaningful, or noise

Dotplot

Dotplot gives an overview of all possible alignments

The normal case: two somewhat similar sequences

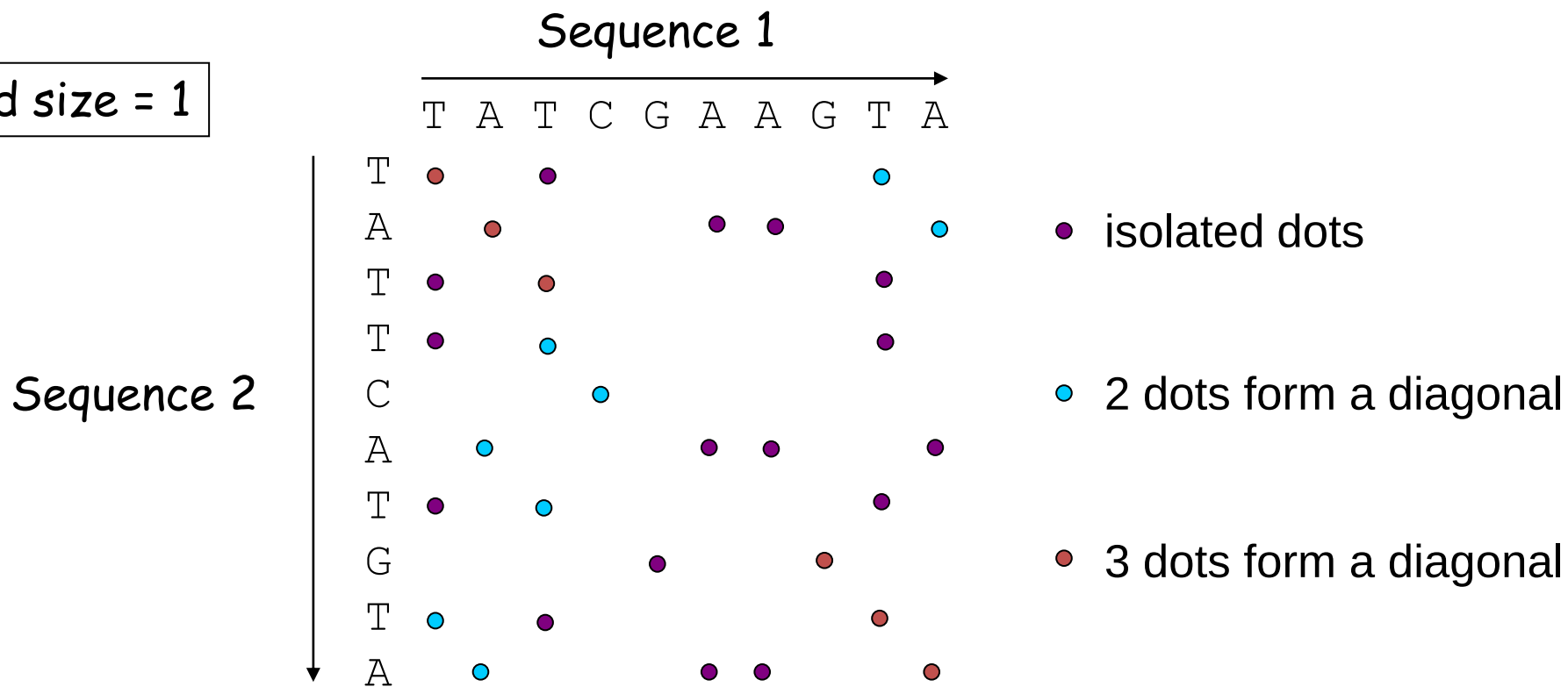


Dotplot

Dotplot gives an overview of all possible alignments

Filters (word size) can be introduced to get rid of noise

Word size = 1

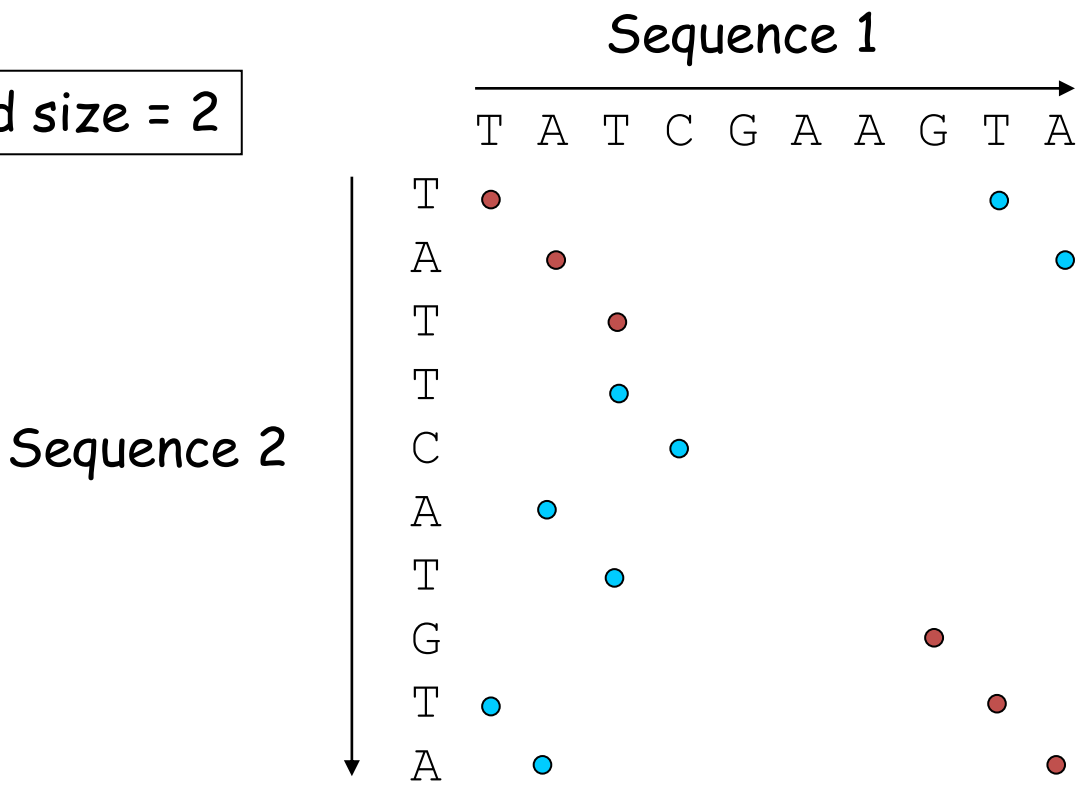


Dotplot

Dotplot gives an overview of all possible alignments

Filters (word size) can be introduced to get rid of noise

Word size = 2



2 dots form a diagonal

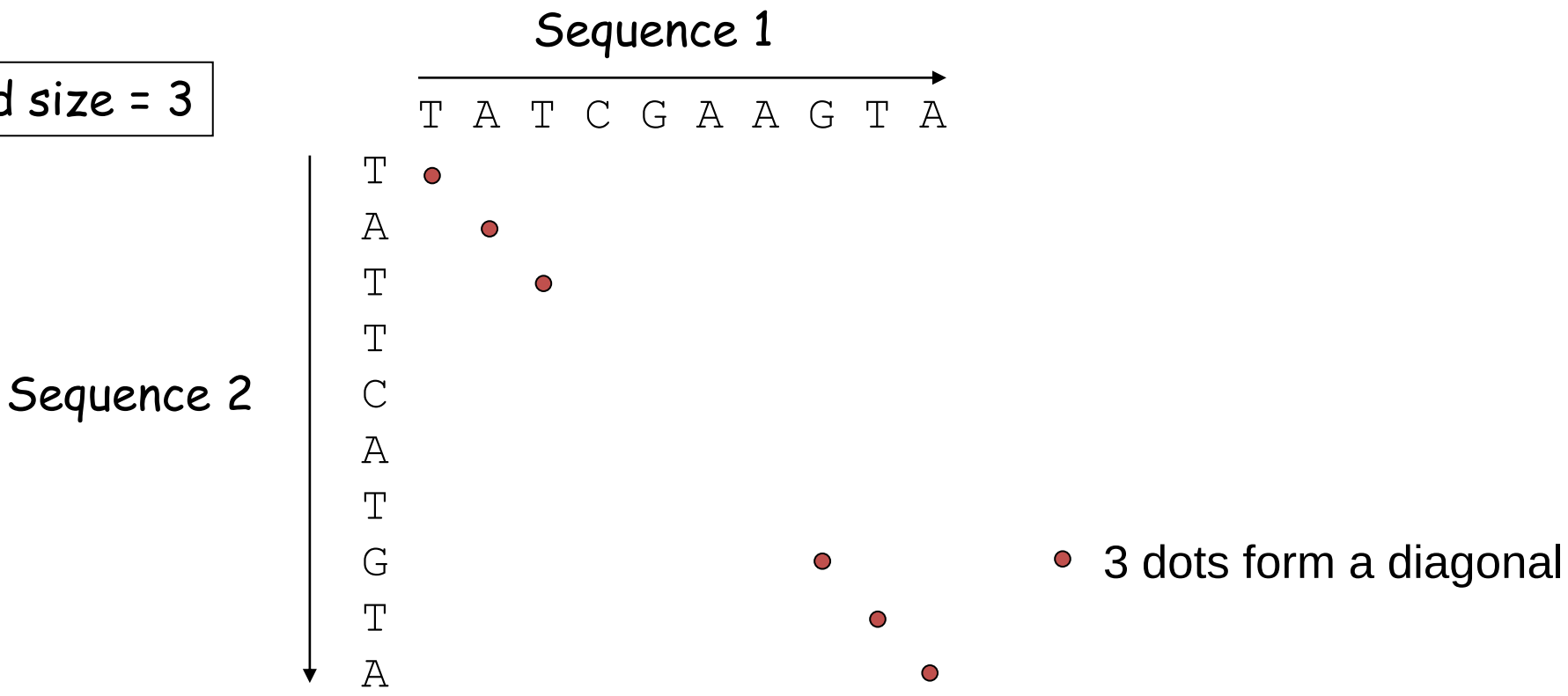
3 dots form a diagonal

Dotplot

Dotplot gives an overview of all possible alignments

Filters (word size) can be introduced to get rid of noise

Word size = 3



Dotplot

Dotplot gives an overview of all possible alignments

Filters (word size) can be introduced to get rid of noise

Word size = 4

Sequence 1

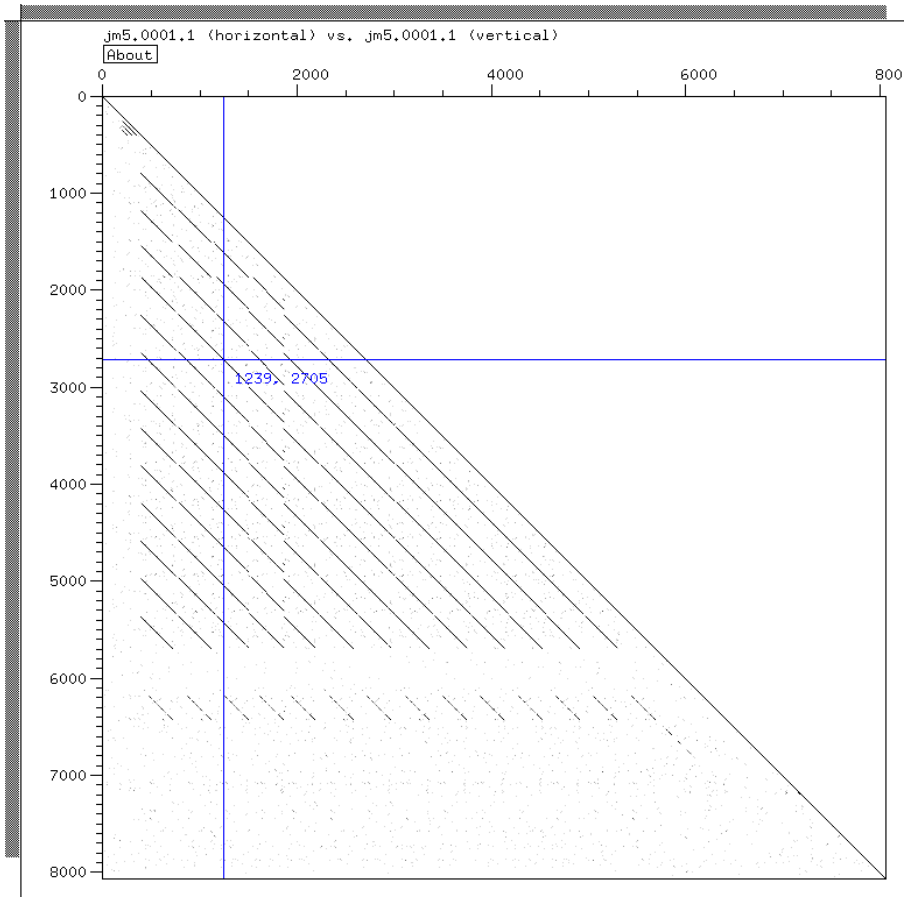
T A T C G A A G T A

Sequence 2

T
A
T
T
C
A
T
G
T
A

conditions too stringent !!

Dot matrix: example of a repetitive DNA sequence



- In addition to the main diagonal, there are several other diagonal.
- Only one half of the matrix is shown because of the symmetry.

perfect tool to visualize repeats

Problems with Dot matrices

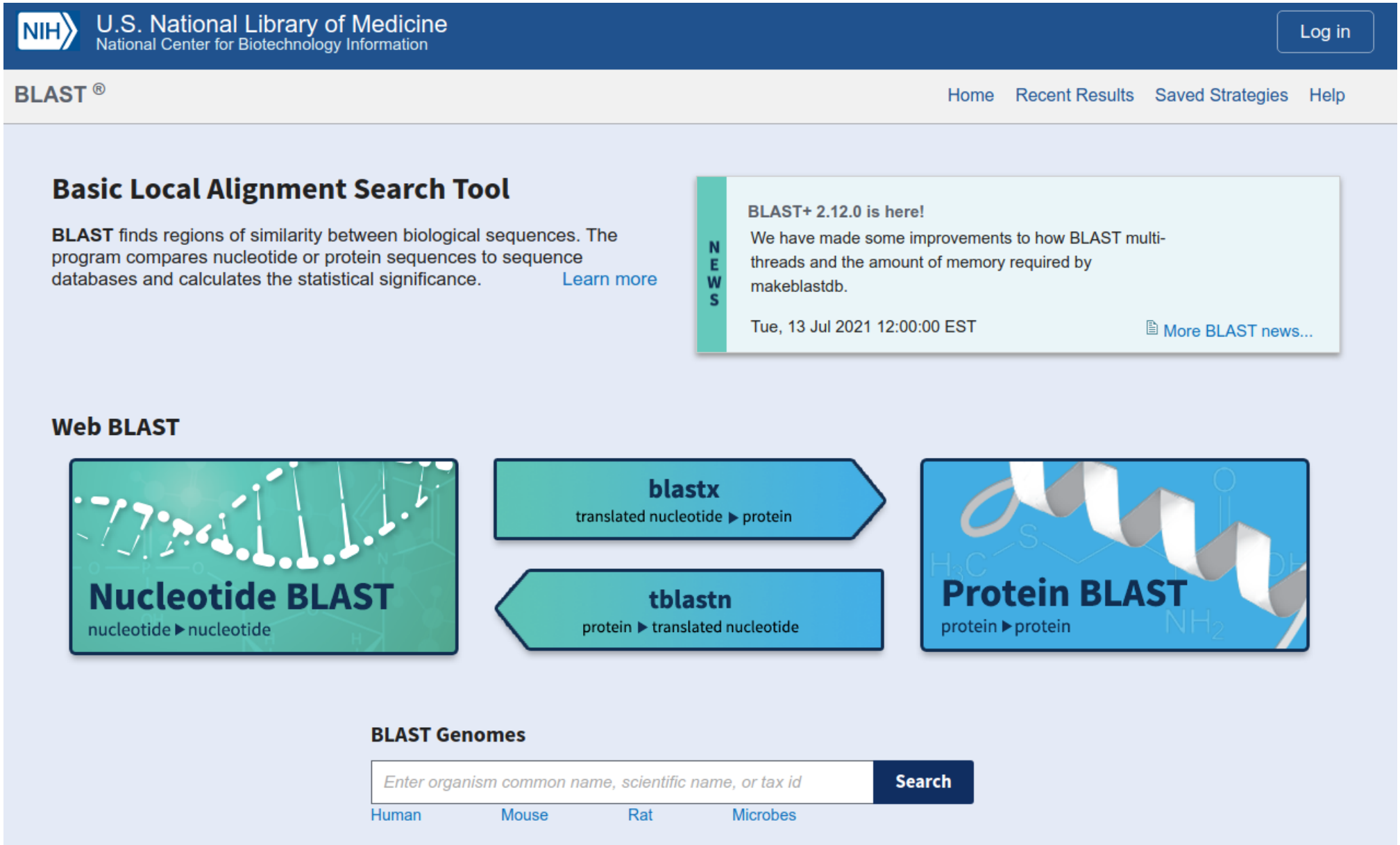
- Rely on visual analysis
  (necessarily merely a screen dump due to number of operations)
  Improvement: Dotter (Sonnhammer et al.)
- Difficult to find optimal alignments
- Difficult to estimate significance of alignments
- Insensitive to conserved substitutions (e.g. L ↔ I or S ↔T) if no substitution matrix can be applied
- Compares only two sequences (*vs*. multiple alignment)
- Time consuming (1,000 bp *vs*. 1,000 bp = 10^6 operations,
  1,000,000 *vs*. 1,000,000 bp = 10^{12} operations)

The BLAST algorithm

- The BLAST programs (**B**asic **L**ocal **A**lignment **S**earch **T**ools) are a set of sequence comparison algorithms introduced in 1990 that are used to search sequence databases for optimal local alignments to a query.
  - Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) “Basic local alignment search tool.” J. Mol. Biol. 215:403-410.
  - Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.” NAR 25:3389-3402.

Several different BLAST programs

Program	Description
blastp	Compares an amino acid query sequence against a protein sequence database.
blastn	Compares a nucleotide query sequence against a nucleotide sequence database.
blastx	Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.
tblastn	Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
tblastx	Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. Please note that the tblastx program cannot be used with the nr database on the BLAST Web page because it is too computationally intensive.



https://blast.ncbi.nlm.nih.gov/Blast.cgi

Standalone and API BLAST

Download BLAST
Get BLAST databases and executables

Use BLAST API
Call BLAST from your application

Use BLAST in the cloud
Start an instance at a cloud provider

Specialized searches

SmartBLAST
Find proteins highly similar to your query

Primer-BLAST
Design primers specific to your PCR template

Global Align
Compare two sequences across their entire span (Needleman-Wunsch)

CD-search
Find conserved domains in your sequence

IgBLAST
Search immunoglobulins and T cell receptor sequences

VecScreen
Search sequences for vector contamination

CDART
Find sequences with similar conserved domain architecture

Multiple Alignment
Align sequences using domain and protein constraints

MOLE-BLAST
Establish taxonomy for uncultured or environmental sequences

MegaBLAST

- megaBLAST
  - For aligning very similar sequences
  - Nucleotide only
  - Very efficient for long query sequences
  - Uses big word (k-tuple) sizes to start search
    - Very fast

http://www.ncbi.nlm.nih.gov/BLAST/

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with Primer-BLAST
- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)
- Search sequences that have gene expression profiles (GEO)
- Search immunoglobulins (IgBLAST)
- Search using SNP flanks
- Screen sequence for vector contamination (vecscreen)
- Align two (or more) sequences using BLAST (bl2seq)
- Search protein or nucleotide targets in PubChem BioAssay
- Search SRA transcript and genomic libraries
- Constraint Based Protein Multiple Alignment Tool
- Needleman-Wunsch Global Sequence Alignment Tool
- Search RefSeqGene
- Search WGS sequences grouped by organism

http://www.ncbi.nlm.nih.gov/BLAST/

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with Primer-BLAST
- Search trace archives
- Find conserved domains in your sequence (cds)
- Find sequences with similar conserved domain architecture (cdart)
- Search sequences that have gene expression profiles (GEO)
- Search immunoglobulins (IgBLAST)
- Search using SNP flanks
- Screen sequence for vector contamination (vecscreen)
- Align two (or more) sequences using BLAST (bl2seq)
- Search protein or nucleotide targets in PubChem BioAssay
- Search SRA transcript and genomic libraries
- Constraint Based Protein Multiple Alignment Tool
- Needleman-Wunsch Global Sequence Alignment Tool
- Search RefSeqGene
- Search WGS sequences grouped by organism

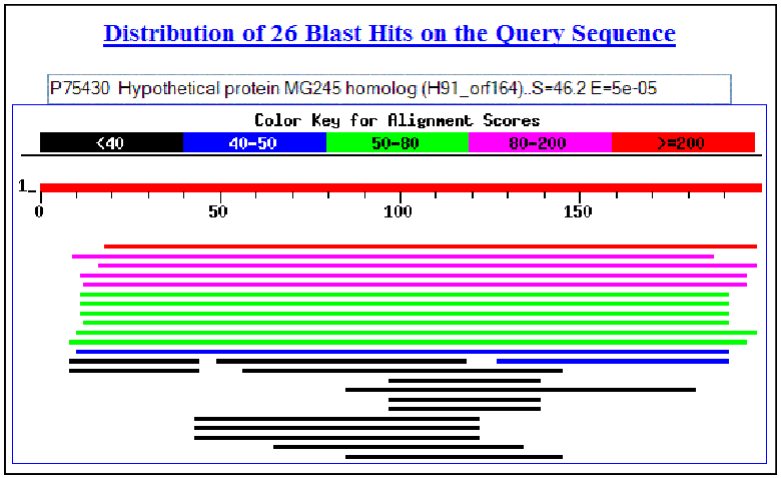
QUERY sequence(s)

```
>gi|15237380|ref|NP_197163.1| myb family transcription factor (MYB43) [Arabidopsis thaliana]
MGRQPCCDKVGLKKGPWTIEEDKKLINFILTNGHCCWRALPKLSGLLRCGKSCRLRWINYLRPDLKRGLL
SEYEEQKVINLHAQLGNRWSKIASHLPGRTDNEIKNHWNTHIKKKLRKMGIDPLTHKPLSEQEASQQAQG
RKKSLVPHDDKNPKQDQQTKDEQEQHQLEQALEKNNTSVSGDGFCIDEVPLLNPHEILIDISSSHHHHSN
DDNVNINTSKFTSPSSSSSSTSSCISSVVPGDEFSKFFDEMEILDLKWLSSDDSLGDDISKDGKFNNSTV
DTMNLWDINDLSSLDMFMNEHDDGFIGNGNGCSRMVLDQDSWTFDLL
```

BLAST program

BLAST database

BLAST results



Considerations for choosing a BLAST database

- First consider your research question:
  - Are you looking for an particular gene in a particular species?
    - BLAST against the genome of that species.
  - Are you looking for additional members of a protein family across all species?
    - BLAST against the non-redudant database (nr), if you can’t find hits check wgs, htgs, and the trace archives.
  - Are you looking to annotate genes in your species of interest?
    - BLAST against known genes (RefSeq) and/or ESTs from a closely related species.

When choosing a database for BLAST…

- Changing your choice of database is changing your search space
- Database size affects the BLAST statistics
- Databases change rapidly and are updated frequently

Where does the score (S) come from?

- The quality of each pair-wise alignment is represented as a score and the scores are ranked.
- **Scoring matrices** are used to calculate the score of the alignment base by base (DNA) or amino acid by amino acid (protein).
- **The alignment score will be the sum of the scores for each position**.

What's a scoring matrix?

- Substitution matrices are used for amino acid alignments.
  - each possible residue substitution is given a score
- A simpler unitary matrix is used for DNA pairs
  - each position can be given a score of +1 if it matches and a score of -1 if it does not.

	A	C	D	E	F	G	H
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	[illegible]
G	0	-3	-1	-2	-3	[illegible]	
H	-2	-3	-1	0			

BLOSUM 62

BLOSUM vs. PAM

BLOSUM 45	**BLOSUM 62**	**BLOSUM 90**
PAM 250	**PAM 160**	**PAM 100**
More Divergent		***Less Divergent***

- BLOSUM62 is the default matrix in BLAST. Though it is tailored for comparisons of moderately distant proteins, it performs well in detecting closer relationships. A search for distant relatives may be more sensitive with a different matrix.

Sequence Similarity Searching – The statistics are important

- Discriminating between real and artifactual matches is done using an estimate of probability that the match might occur by chance.

What do the Score and the e-value really mean?

- The quality of the alignment is represented by the Score.
  - **Score (S)**
    - The score of an alignment is calculated as the sum of substitution and gap scores. Substitution scores are given by a look-up table (PAM, BLOSUM) whereas gap scores are assigned empirically .
- The significance of each alignment is computed as an E value.
  - **E value (E)**
    - Expectation value. The number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score.

I'm confused! What does the E-value mean again?

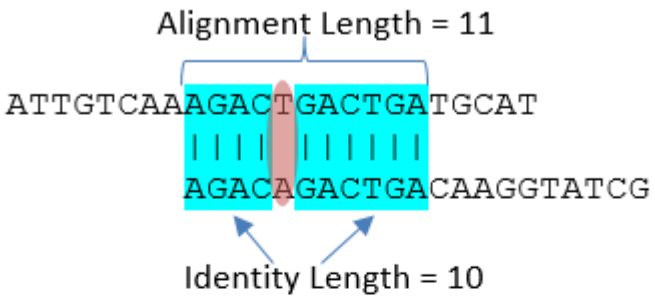
- **E value (E)**
  - Expectation value. The number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score.
- When $E < 0.01$, *P*-values and *E*-value are nearly identical.
  - So, the E-value is the number of times you expect to see your hit occur in the database (with as good as or better score) due to random chance alone.

Notes on E-values

- Low E-values suggest that sequences are homologous
  - Can’t show non-homology
- Statistical significance depends on both the size of the alignments and the size of the sequence database
  - Important consideration for comparing results across different searches
  - E-value increases as database gets bigger
  - E-value decreases as alignments get longer

Coverage

- Coverage: The proportion of the aligned length with respect to the length of the query or subject.
- Example
  - Your gene is 1000bp, and you have a Blast alignment from 250-500. What is the query coverage?



$$Alignment\ Identity\ \% = \frac{Identity\ Length}{Alignment\ Length} = \frac{10}{11}$$

$$Query\ \ Identity\ \% = \frac{Identity\ Length}{Query\ Length} = \frac{10}{25}$$

$$Query\ \ Coverage\ \% = \frac{Alignment\ Length}{Query\ Length} = \frac{11}{25}$$

$$Subject\ \ Identity\ \% = \frac{Identity\ Length}{Subject\ Length} = \frac{10}{21}$$

$$Subject\ \ Coverage\ \% = \frac{Alignment\ Length}{Subject\ Length} = \frac{11}{21}$$

https://www.stn.org/help/stnext/Managing_Your_STNext_Files/Biosequences/Biosequence_Results/BLAST/Filter_Results_-_BLAST.htm

FASTA File Format

- Plain text file (e.g. don't open with Word!) Each
- sequence has 2 parts.
  - One header line starts with “>”
    - e.g. “>This is a fasta header. Any text goes here.”
  - One or more sequence lines:
    - e.g. “ATTCTCGCTCGAATCGATCGCATAGTAGCA”
- Each file can contain multiple sequences
- Sequences can be DNA or protein (not a mixture)

```
>AT1G09780|1|training
GTGGAGTAGAAGAATTGAGAGCCTTATCAG
TTTTTGAAGAGAGGGCTGAAACTCTCTAGT
TATCTTTTGTTGCTTTTCTAATAATAAGAG
TTTACACACAG
>AT1G31812|0|testing
TCCTCATCTGCAGTAACTTTATCTTAAGCA
TCAAAATAACATTGCATAAGACTTGTTCTT
GCTCTTGTGTTTCTATCATATTTAAGCTAT
CTACTTTGTGA
```

Part 1

Part 2

Part 3

Alignments

```
>ref|YP_496553.1| recombinase A [Novosphingobium aromaticivorans DSM 12444]
Length=356

 GENE ID: 3917906 recA | recombinase A
[Novosphingobium aromaticivorans DSM 12444]

 Score =  483 bits (1244),  Expect = 2e-173, Method: Compositional matrix adjust.
 Identities = 236/332 (71%), Positives = 282/332 (85%), Gaps = 6/332 (2%)

Query  1    ALAAALAQIEKQFGKGSIMRMGDGEATENIQVVSTGSLGLDIALGVGGLPRGRVVEIYGP  60
            AL AALAQI++ FGKGS MR+G  EA + ++ VSTGSLGLDIALG+GGLPRGR++EIYGP
Sbjct  21   ALDAALAQIDRAFGKGSAMRLGSKEAMQ-VEAVSTGSLGLDIALGIGGLPRGRIIEIYGP  79

Query  61   ESSGKTTLTLQVIAELQKIGGTAAFIDAEHALDVQYAAKLGVNVPELLISQPDTGEQALE  120
            ESSGKTTL L  IAE QK GGTAAFIDAEHALD  YA KLGV++  L++SQPDTGEQALE
Sbjct  80   ESSGKTTLALHAIAEAQKGGGTAAFIDAEHALDPVYARKLGVDIDNLIVSQPDTGEQALE  139

Query  121  ITDALVRSGSIDMIVIDSVAALVPKAEIEGEMGDSLPGLQARLMSQALRKLTGTIKRTNC  180
            ITD LVRS +ID++V+DSVAALVP+AEIEGEMGDS  GLQARLMSQALRKLTG+I R+ C
Sbjct  140  ITDTLVRSNAIDVLVVDSVAALVPRAEIEGEMGDSHVGLQARLMSQALRKLTGSISRSRC  199

Query  181  LVIFINQIRMKIGVMFGNPETTTGGNALKFYSSVRLDIRRIGSIKKNDEVIGNETRVKVV  240
            +VIFINQ+RMKIGVM+GNPETTTGGNALKFY+SVRLDIRR G IK  DE++GN TRVKVV
Sbjct  200  MVIFINQVRMKIGVMYGNPETTTGGNALKFYASVRLDIRRTGQIKDRDEIVGNATRVKVV  259

Query  241  KNKVSPPFREAIFDILYGEGISRQGEIIDLGVQAKIVDKAGAWYSYNGEKIGQGKDNARE  300
            KNKV+PPF++  FDI+YGEGIS+ GEI+DLGV+A +V+K+GAW+SY+  +IGQG++NA+
Sbjct  260  KNKVAPPFKQVEFDIMYGEGISKIGEILDLGVKAGLVEKSGAWFSYDSIRIGQGRENAKN  319

Query  301  FLRENPEIAREIENRIRESL-----GVVAMPD  327
            FLRENPE+   +E  IR        G++A PD
Sbjct  320  FLRENPEVCSRLEAAIRGRTDQVAEGLMAGPD  351
```

Databases

- NR “non-redundant” database
  - Sequences from various experiments (not just completed genomes)
  - May not be that “non-redundant”
- RefSeq
  - Curated sequences by NCBI
  - Does not contain duplicates
- Swissprot
  - A manually curated sequence of proteins
- Protein Data Bank
  - Contains protein sequences that have 3D structures available