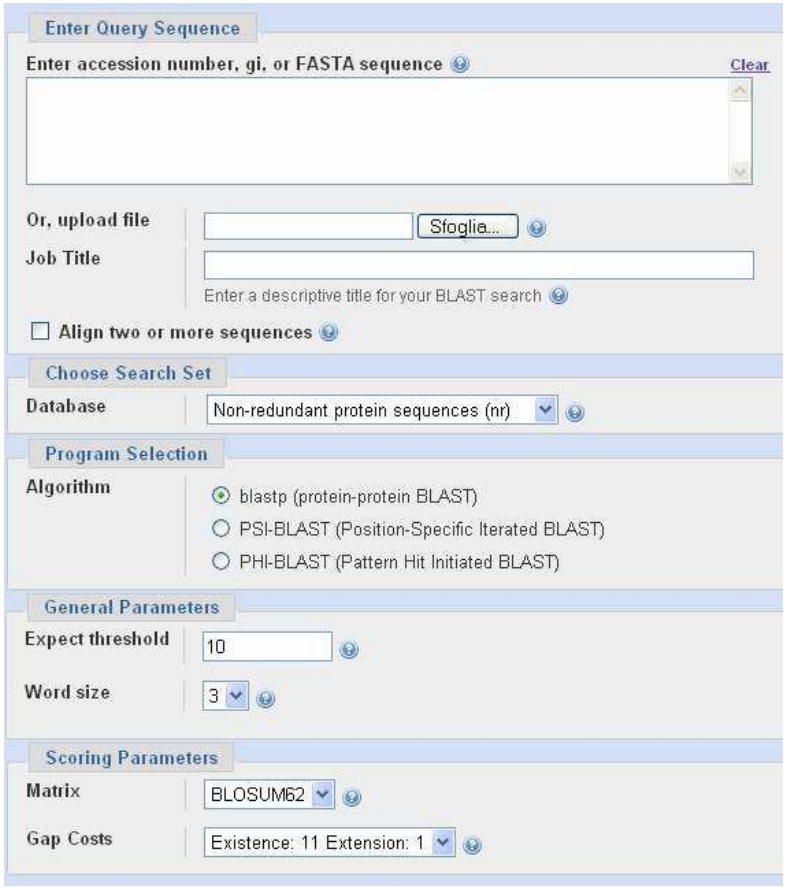
BLAST

26.04.2018

Basic Local Alignment Search Tool (BLAST)

BLAST (Altshul-1990) is an heuristic Pairwise Alignment composed by six-steps that search for local similarities.



The most used access point to BLAST is NCBI.

As for the FASTA mask allows set the search parameters and the minimum significance required.

It is also possible to choose the type of BLAST algorithm to use.

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

Basic Local Alignment Search Tool (BLAST)

Phases 1 to 3:

Also BLAST bases its heuristic on indexing, but introduces a sophistication level of the k-tuples, which are also called "word".

It starts by creating a list of words of length W and creation of "w-mers", i.e.. words of length W. W are words that, according a substitution matrix, have a score >T if aligned with the input sequence.

```
TFDER LSH GVQQTFWECIKGD
      VSH = 16
      ISH = 14
      LAH = 13
      LTH = 13
      LSH = 13
```

Word size (W) = 3
Threshold (T) = 13

For each word they are also generated all possible w-mers, so there are many more words than in FASTA.

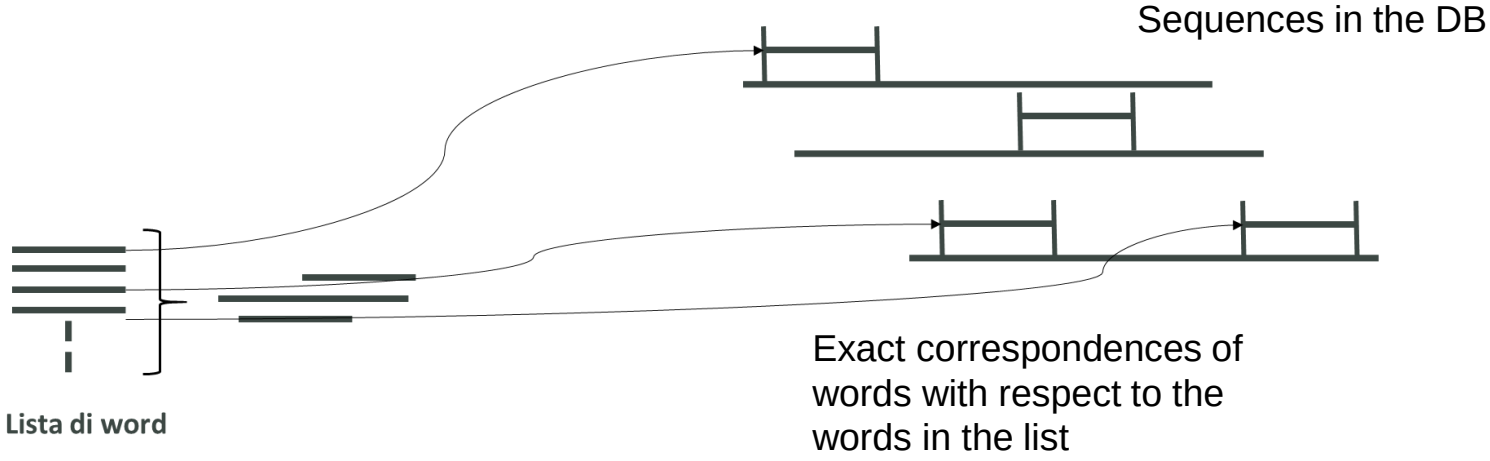
BLAST

Phase 1:

It is extracted from the query sequence a list of words, each having the same length w. We obtain L-W+1 words.

Phase 2:

Each word in the list is compared with sequences of the same lenght belonging to the all sequences of the considered database.



BLAST

Phase 3:

For each alignment it is assigned a score using a substitution matrix (for example, BLOSUM 62) and is compared with a pre-established value T.

The sequences with values equal to or greater than T (with values between 13 and 15) are further analyzed, while those with values less than T are discarded.

The substitution matrices assigned a positive score for each identity or for a substitution with amino acids of the same type (hydrophobic with hydrophobic, positively charged with positively charged etc ...) and negative for a substitution with amino acids of different type (eg. Basic amino acid with amino acid etc ..).

BLAST

Example:

word

KRHRKVLRDNIQGITKPAIRRLARRGGVKRISGLIYEETRGVLKIFLENVIRD

GVK	18
GAK	16
GIK	16
GGK	14
GLK	13
GNK	12
GRK	11
GEK	12
GDK	11

Similar words

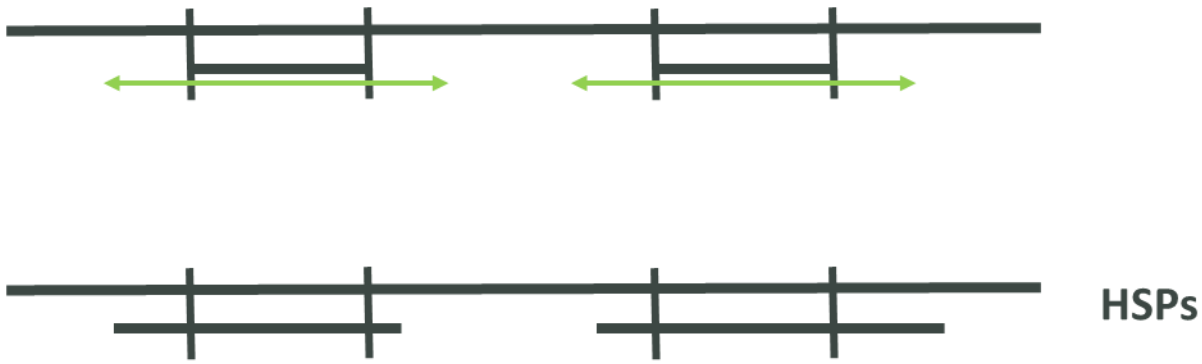
Threshold T=13

Allineamento singolo euristico: BLAST

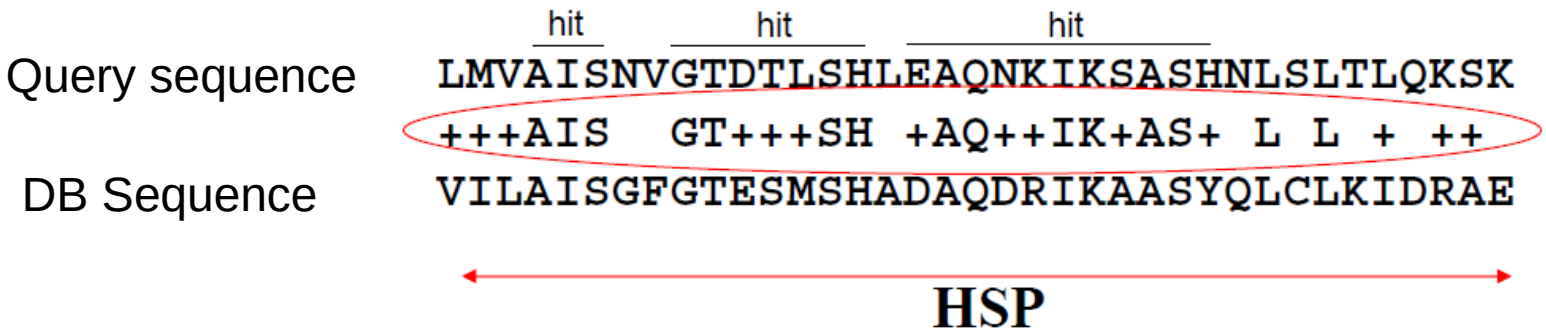
Phase 4:

Extending the aligned zone. In this phase are taken into account, one at a time, the characters immediately adjacent (both right and left) to the **word**.

For each character added to the word you recalculate the score with the sequence in the database, and as long as the similarity score increases will be added to the adjacent characters the word (the pairing areas so found are called **HSP - high scoring segment pair**) .



BLAST



Phase 5:

The HSP score (S) below a certain threshold (chosen empirically) are discarded. The output of BLAST phase 5 will contain the set of sequences whose HSP were not discarded with their scores.

A threshold X is used to determined the maximum tolerated score decrease.

BLAST

Phase 6:

BLAST generally operates on databases that contain millions of sequences

So not all alignments have a biological relevance, or not all of the output sequences are homologous to the query sequence / input.

Therefore it is necessary a statistical evaluation of alignment obtained by BLAST

BLAST adds to each sequence present in the output the **E-Value**, a value that, properly interpreted by the researcher, will indicate how likely it is that the S score indicates a biological correlation between the two sequences.

The more E-value is close to zero the more attest to a biological correlation between the sequences.

BLAST

```
 Score =  302 bits (163),  Expect = 5e-79
 Identities = 238/273 (87%), Gaps = 9/273 (3%)
 Strand=Plus/Plus

Query  1    ATGATGAAGACTGCAGTAATGTTCATATTAGTAGTCGTGATCAGTTTGACGTATTCAAGC  60
            |||||||||||| |||||||| |||| |||||||||||||||||| ||| || |||| ||
Sbjct  1    ATGATGAAGACTTCAGTAATGCTCATGTTAGTAGTCGTGATCAGTCTGATGTGTTCAGGC  60

Query  61   GAGGAGCAGGAGGTAGCGCGCACGTACTGCGGCCGACACTTGGCGAATATACTCGCTTAC  120
            |||| ||||||||||||||||||||||||||||||||||||||| ||| ||||||  ||
Sbjct  61   GAGGCGCAGGAGGTAGCGCGCACGTACTGCGGCCGACACTTGGCGGATACACTCGCCGAC  120

Query  121  GTGTGCTTCGGAGTGGAGAAACGCGGCGGTGCGCAGTACGCGCCTTACT---GGCAGGAG  177
             |||||||||| ||||||||||||||||||||||||||||||||| ||||   || | |||
Sbjct  121  TTGTGCTTCGGGGTGGAGAAACGCGGCGGTGCGCAGTACGCGCCCTACTTCTGG-ACGAG  179

Query  178  ACA-TACTTGCGATCTCGCAAGGGAC-C-CGGCGTAGTGGACGAATGTTGTTTCCGCCCT  234
            ||| |||||| ||||||||  ||||  | |||||||||||||||||| |||||||| |||
Sbjct  180  ACAATACTTGGGATCTCGC--GGGAAACGCGGCGTAGTGGACGAATGCTGTTTCCGGCCT  237

Query  235  TGCAAACTAGAGGTTCTCAAATCGTACTGCGGT  267
            |||| ||| || ||||||  ||| |||||||||
Sbjct  238  TGCACACTTGACGTTCTCTTATCTTACTGCGGT  270
```

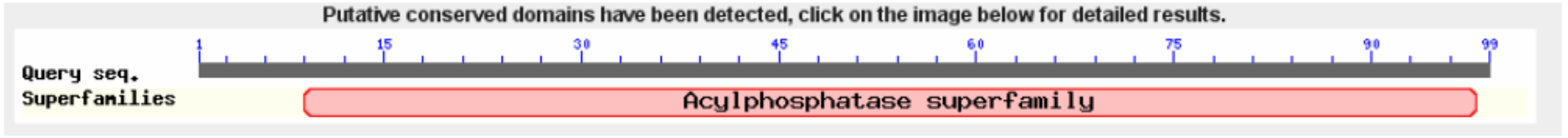
Figura 4

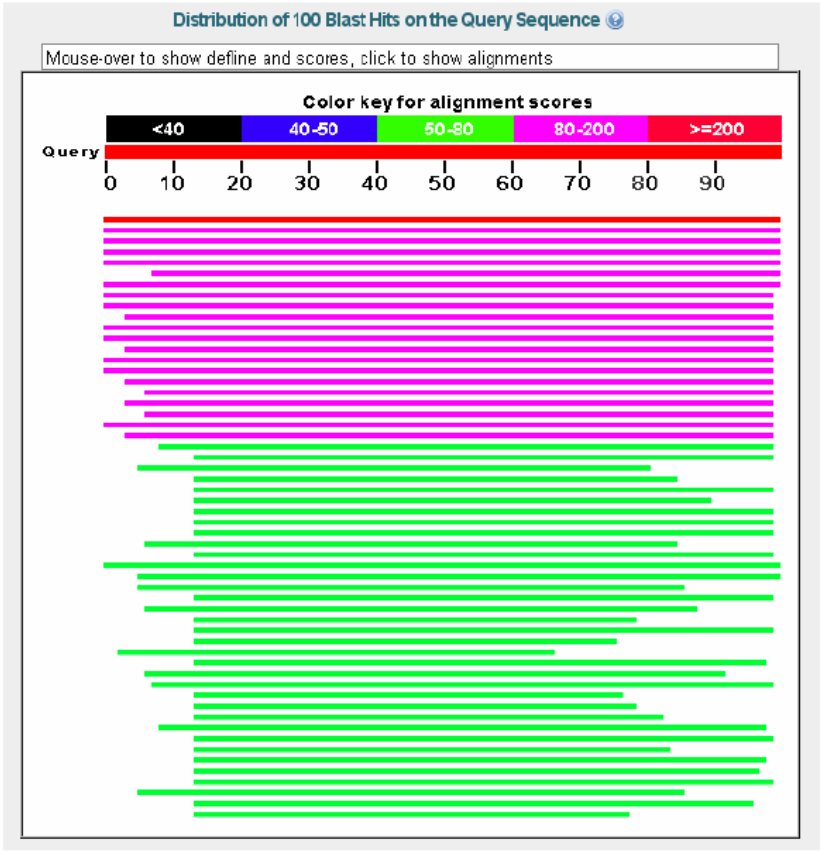
- **Score**: is the alignment score;
- **Expect:** is the E-value of the alignment;
- **Identities:** is the number of identical characters / the alignment length; in brackets, it is indicated the resulting percentage of identity over that length of alignment;
- **Gaps:** indicates the number of gaps present in the array / alignment length; in brackets it indicates the percentage of resulting gap on the length of alignment;
- **Strand:** is the sequence orientation "query" compared to the sequence of the database with which aligns. Two possibilities: plus / plus and plus / minus

The numbers indicate the position of the amino acids within the respective sequences: for example, the number **237** is circled in red to indicate that the base "t", indicated by the red arrow, is the base number 237 in the database sequence. If in a given alignment position of the base sequence in the "query" and the corresponding database sequence coincide, the line between the two sequences shows the character "|".

The result page of BLAST

If BLAST recognizes a wide similarity with one or more known protein families, it indicates it in a graphical way.

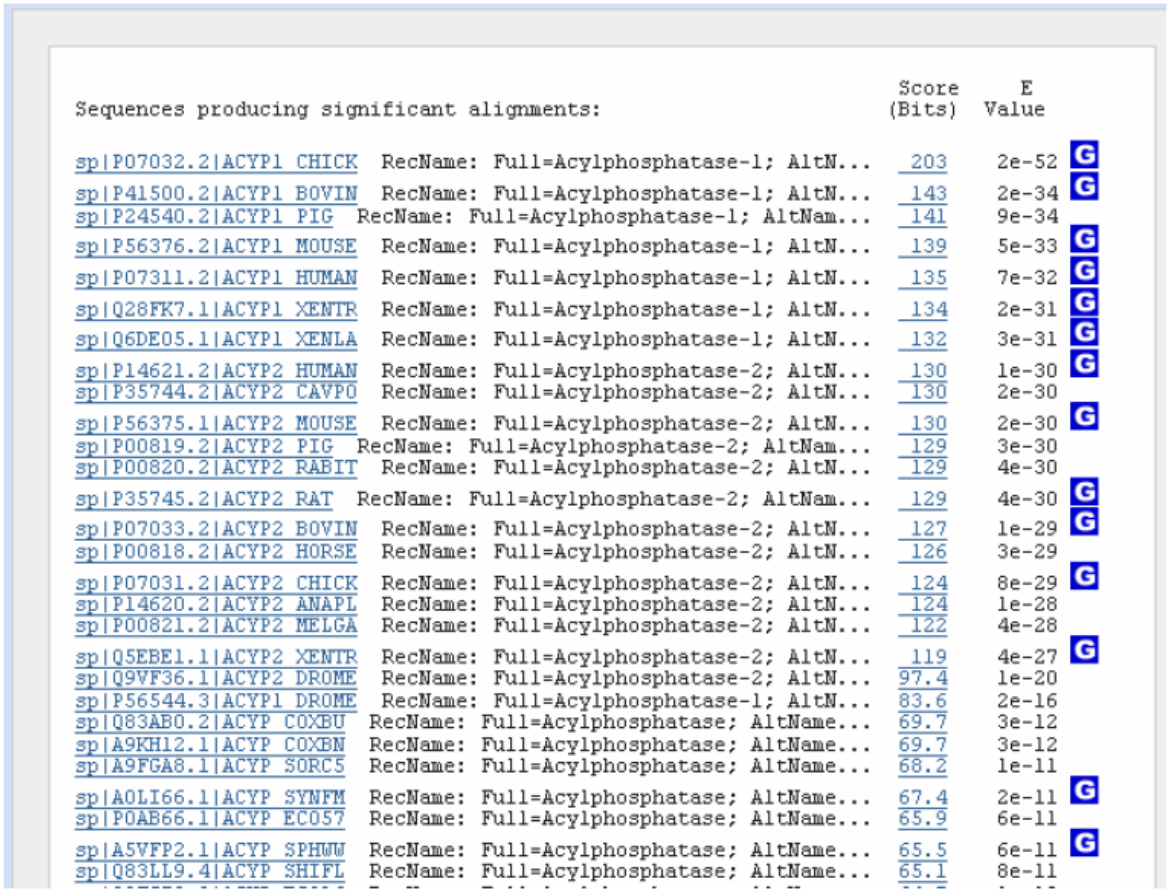




It is also present a summary graph that allows to quickly explore the first 100 hits, showing them aligned with the query sequence.

The result page of BLAST

BLAST includes in its output a textual summary of the sequences that were found: for each it shows a link (in this case to Entrez Protein), the name, the score (in bits) and The Expect value, according to which the results are sorted .



Sequences producing significant alignments:

	Score (Bits)	E Value	
sp|P07032.2|ACYP1_CHICK RecName: Full=Acylphosphatase-1; AltN...	203	2e-52	G
sp|P41500.2|ACYP1_BOVIN RecName: Full=Acylphosphatase-1; AltN...	143	2e-34	G
sp|P24540.2|ACYP1_PIG RecName: Full=Acylphosphatase-1; AltNam...	141	9e-34	
sp|P56376.2|ACYP1_MOUSE RecName: Full=Acylphosphatase-1; AltN...	139	5e-33	G
sp|P07311.2|ACYP1_HUMAN RecName: Full=Acylphosphatase-1; AltN...	135	7e-32	G
sp|Q28FK7.1|ACYP1_XENTR RecName: Full=Acylphosphatase-1; AltN...	134	2e-31	G
sp|Q6DE05.1|ACYP1_XENLA RecName: Full=Acylphosphatase-1; AltN...	132	3e-31	G
sp|P14621.2|ACYP2_HUMAN RecName: Full=Acylphosphatase-2; AltN...	130	1e-30	G
sp|P35744.2|ACYP2_CAVPO RecName: Full=Acylphosphatase-2; AltN...	130	2e-30	
sp|P56375.1|ACYP2_MOUSE RecName: Full=Acylphosphatase-2; AltN...	130	2e-30	G
sp|P00819.2|ACYP2_PIG RecName: Full=Acylphosphatase-2; AltNam...	129	3e-30	
sp|P00820.2|ACYP2_RABIT RecName: Full=Acylphosphatase-2; AltN...	129	4e-30	
sp|P35745.2|ACYP2_RAT RecName: Full=Acylphosphatase-2; AltNam...	129	4e-30	G
sp|P07033.2|ACYP2_BOVIN RecName: Full=Acylphosphatase-2; AltN...	127	1e-29	G
sp|P00818.2|ACYP2_HORSE RecName: Full=Acylphosphatase-2; AltN...	126	3e-29	
sp|P07031.2|ACYP2_CHICK RecName: Full=Acylphosphatase-2; AltN...	124	8e-29	G
sp|P14620.2|ACYP2_ANAPL RecName: Full=Acylphosphatase-2; AltN...	124	1e-28	
sp|P00821.2|ACYP2_MELGA RecName: Full=Acylphosphatase-2; AltN...	122	4e-28	
sp|Q5EBE1.1|ACYP2_XENTR RecName: Full=Acylphosphatase-2; AltN...	119	4e-27	G
sp|Q9VF36.1|ACYP2_DROME RecName: Full=Acylphosphatase-2; AltN...	97.4	1e-20	
sp|P56544.3|ACYP1_DROME RecName: Full=Acylphosphatase-1; AltN...	83.6	2e-16	
sp|Q83AB0.2|ACYP_COXBU RecName: Full=Acylphosphatase; AltName...	69.7	3e-12	
sp|A9KH12.1|ACYP_COXBN RecName: Full=Acylphosphatase; AltName...	69.7	3e-12	
sp|A9FGA8.1|ACYP_SORC5 RecName: Full=Acylphosphatase; AltName...	68.2	1e-11	
sp|A0LI66.1|ACYP_SYNFM RecName: Full=Acylphosphatase; AltName...	67.4	2e-11	G
sp|P0AB66.1|ACYP_ECO57 RecName: Full=Acylphosphatase; AltName...	65.9	6e-11	
sp|A5VFP2.1|ACYP_SPHWW RecName: Full=Acylphosphatase; AltName...	65.5	6e-11	G
sp|Q83LL9.4|ACYP_SHIFL RecName: Full=Acylphosphatase; AltName...	65.1	8e-11	

The results page of BLAST

Under the descriptions begin the actual results, i.e., HSP: in addition to the description, the result shows the identity, similarity (Positives) and the inserted gaps, if any.

```
>sp|P14620.2|ACYP2_ANAPL  RecName: Full=Acylphosphatase-2; AltName: Full=Acylphosphate
phosphohydrolase 2; AltName: Full=Acylphosphatase, muscle type
isozyme
Length=103

 Score =  124 bits (311),  Expect = 1e-28, Method: Compositional matrix adjust.
 Identities = 56/92 (60%), Positives = 69/92 (75%), Gaps = 0/92 (0%)

Query  7    LMSVDYEVSGRVQGVFFRKYTQSEAKRLGLVGWVRNTSHGTVQGQAQGPAARVRELQEWL  66
            L SVDYEV GRVQGV FR YT+ EA++LG+VGWV+NTS GTV GQ QGP  +V  ++ WL
Sbjct  11   LKSVDYEVFGRVQGVCFRMYTEEEARKLGVVGWVKNTSQGTVTGQVQGPEDKVNAMKSWL  70

Query  67   RKIGSPQSRISRAEFTNEKEIAALEHTDFQIR  98
             K+GSP SRI R  F+NEKEI+ L+ + F  R
Sbjct  71   TKVGSPSSRIDRTNFSNEKEISKLDFSGFSTR  102


>sp|P00821.2|ACYP2_MELGA  RecName: Full=Acylphosphatase-2; AltName: Full=Acylphosphate
phosphohydrolase 2; AltName: Full=Acylphosphatase, muscle type
isozyme; AltName: Full=Acylphosphatase isozyme TU1
Length=103

 Score =  122 bits (306),  Expect = 4e-28, Method: Compositional matrix adjust.
 Identities = 56/95 (58%), Positives = 69/95 (72%), Gaps = 0/95 (0%)

Query  4    SEGLMSVDYEVSGRVQGVFFRKYTQSEAKRLGLVGWVRNTSHGTVQGQAQGPAARVRELQ  63
            S  L SVDYEV GRVQGV FR YT+ EA++LG+VGWV+NT  GTV GQ QGP  +V  ++
Sbjct  8    SGALKSVDYEVFGRVQGVCFRMYTEEEARKLGVVGWVKNTRQGTVTGQVQGPEDKVNAMK  67

Query  64   EWLRKIGSPQSRISRAEFTNEKEIAALEHTDFQIR  98
             WL K+GSP SRI R  F+NEKEI+ L+ + F  R
Sbjct  68   SWLSKVGSPSSRIDRTNFSNEKEISKLDFSGFSTR  102
```

BLAST releases

BLAST has recently been implemented with a two-hit method, which provides for the extension of the HSP regions can only happen if two independent hits occur within a number of residues (represented by threshold A), with no gaps in between. There were also implemented several specialized versions:

- **blastp**: look for similarities in protein databases from a query of amino acids.
- **blastn**: look for similarities in nucleotide databases from a query of nucleotides.
- **blastx**: look for similarities in protein databases from a query of nucleotides that is translated into all possible frame.
- **tblastn**: look for similarities in nucleotide databases from a query of amino acids, translating the entries of the database into nucleotides.
- **tblastx**: look for similarities in nucleotide databases from a query of nucleotides, translating all the entries of the database into amino acids

BLAST

- BLAST has 4 fundamental parameters:
  - **W**: word size, the greater the number, the smaller the number of words generated, the lower the execution time. But the sensitivity decreases considerably.
  - **T**: threshold, the lower the number, the greater the number of w-mers included in the list, the higher the running time. But it has an enhanced sensitivity.
  - **S**: score, the smaller the number, the greater the HSP length.
  - **X**: the lower the number, the greater the number of w-mers included in the list, the higher the running time. But it has an enhanced sensitivity.

Pairwise Alignment Considerations: FASTA vs BLAST

- FASTA differs from BLAST since it searches the best alignment between the entire sequence under investigation and the sequences of the database used as a reference. Instead BLAST search only local sequence homology.
- FASTA also uses a scoring matrix only during the confrontation extension while BLAST uses a scoring matrix during several phases of the alignment (that is, scanning and extension).
- FASTA examines the amino acids in pairs (ktup = 2) or individually taken (ktup = 1), BLAST uses for comparison of 3-4 amino groups (words).
- Both the algorithm that governs the early stages of FASTA and the one of BLAST do not permit the presence of gaps within each sequence segment taken into consideration.
- Unlike BLAST, however, the FASTA algorithm contemplates the possibility of insertions and deletions in the last phase alignment.

Multiple alignment vs Pairwaise alignment

A

```
1: EAGFPPGVVNVIPGFGPTAGAAHASHEDVDKVAFTGSTEVGHLIQVA
2: EAGFPPGVVNIVPGFGPTAGAAIASHEDVDKVAFTGSTEIGRVIQVA
3: QYMDQNLYLVVKGG-VPETTELL--KERFDHIMYTGSTAVGKIVMAA
4: NVFSPAWA-TVVEGDETISQQLL--QEKFDHIFFTGSPRVGRLIMAA
5: EAGVPVGLVNVVQG-GAETGSLLCHHPNVAKVSFTGSVPTGKKVMEM
6: DI-FPAGVINILFGRGKTVGDPLTGHPKVRMVSLTGSIATGEHIISH
```

B

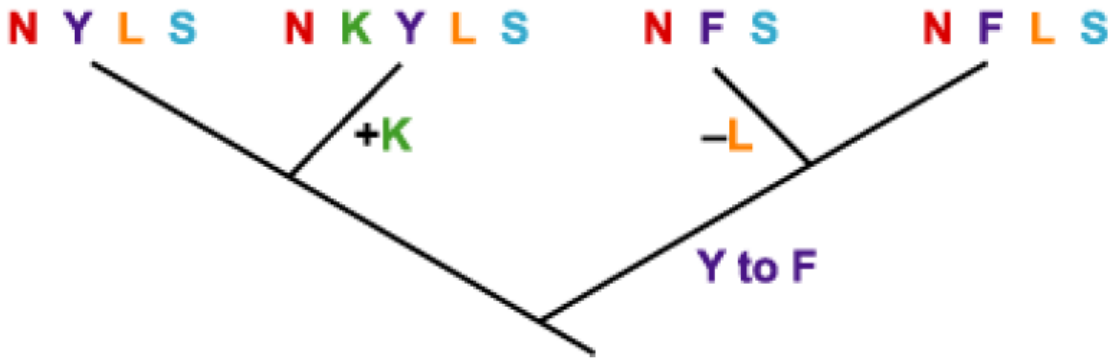
```
1: EAGFPPGVVNVIPGFGPTAGAAIASHEDVDKVAFTGSTEVGHLIQVA
2: EAGFPPGVVNIVPGFGPTAGAAIASHEDVDKVAFTGSTEIGRVIQVA
3: QYMDQNLYLVVKGG-VPETTELL--KERFDHIMYTGSTAVGKIVMAA
4: NVFSPAWA-TVVEGDETISQQLL--QEKFDHIFFTGSPRVGRLIMAA
5: EAGVPVGLVNVVQG-GAETGSLLCHHPNVAKVSFTGSVPTGKKVMEM
6: DI-FPAGVINILFGRGKTVGDPLTGHPKVRMVSLTGSIATGEHIISH
```

Multiple Alignment

Conservation of the catalytic amino acids in some members of the trypsin family

Multiple Alignment

seqA	N	•	F	L	S
seqB	N	•	F	–	S
seqC	N	K	Y	L	S
seqD	N	•	Y	L	S



The evolutionary history is deducted from the alignment

Multiple Alignment

- Difficulty
  - objective scoring Function: weight to be assigned to the various sequences
  - Complexity of the problem: computation time
- Clustal, ClustalW and ClustalX