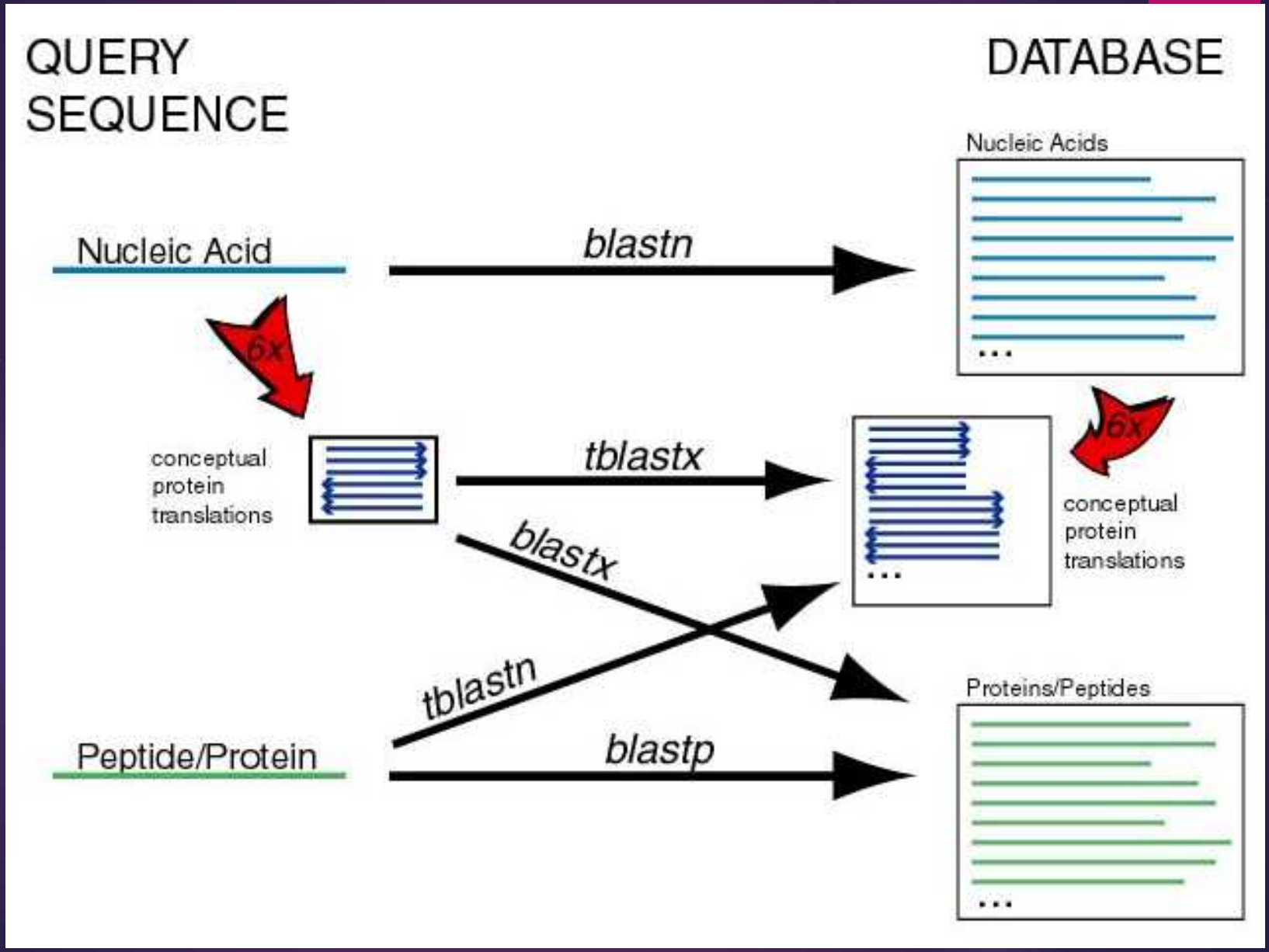
BLAST
(Basic Local Alignment Search Tool)

- Developed by Steven Altschul and Samuel Karlin in 1990.
- Compares nucleotide/aminoacid sequences
- Is a heuristic method.
- Is a fast but approximate method of alignment.
- Locates local alignments/short matches called **words**

Uses of BLAST:

- **Search a database for sequences similar to an input sequence.**
- **Identify previously characterized sequences.**
- **Find phylogenetically related sequences.**
- **Identify possible functions based on similarities to known sequences.**

Types of BLAST:



Types of BLAST:

- **blastp**: compares a protein sequence against a protein sequence database.
- **blastn**: compares a nucleotide sequence against a nucleotide sequence database.
- **blastx**: compares a six frame translation of a nucleotide sequence against a protein database
- **tblastn**: compares a protein sequence against a six frame translation of a nucleotide database
- **tblastx**: compares a six frame translation of a nucleotide sequence against a six frame translation of a nucleotide database.

How BLAST works

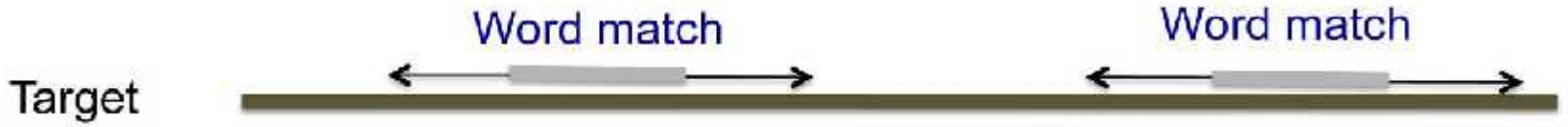
- Blast searches begin with a query sequence that will be matched against sequence databases specified by the user.
- Begins by breaking down the query sequence into a series of short overlapping "words"
- Default word size for BLAST N is 28 nucleotides
- Default word size for BLAST P is 3 amino acids
- Results obtained depend on the scoring matrix used.
- BLOSUM 62 matrix is the default scoring matrix for BLASTP

The basic strategy used by the BLAST algorithms

1. The **query** sequence is broken into "**words**" that will act as seeds in alignments

2. BLAST searches for matches (or synonyms) in **target** entries in the database

3. If a **target** entry has two or more matches to "**words**" from the query, the alignment is extended in both directions looking for additional similarity



The BLASTP algorithm

- Query sequence is broken into all possible 3-letter words using a moving window
- Numerical score is calculated for each word by adding up the values for the amino acids from the BLOSUM62 matrix
- Words with a score of 12 or more are collected into the initial BLASTP search set.
- The search set is broadened by adding synonyms that differ from the words at one position.
- Only synonyms with scores above a threshold value are added to the search set. NCBI BLASTP uses a default threshold of 10 for synonyms

1. BLASTP begins with a query sequence

E A G L E S

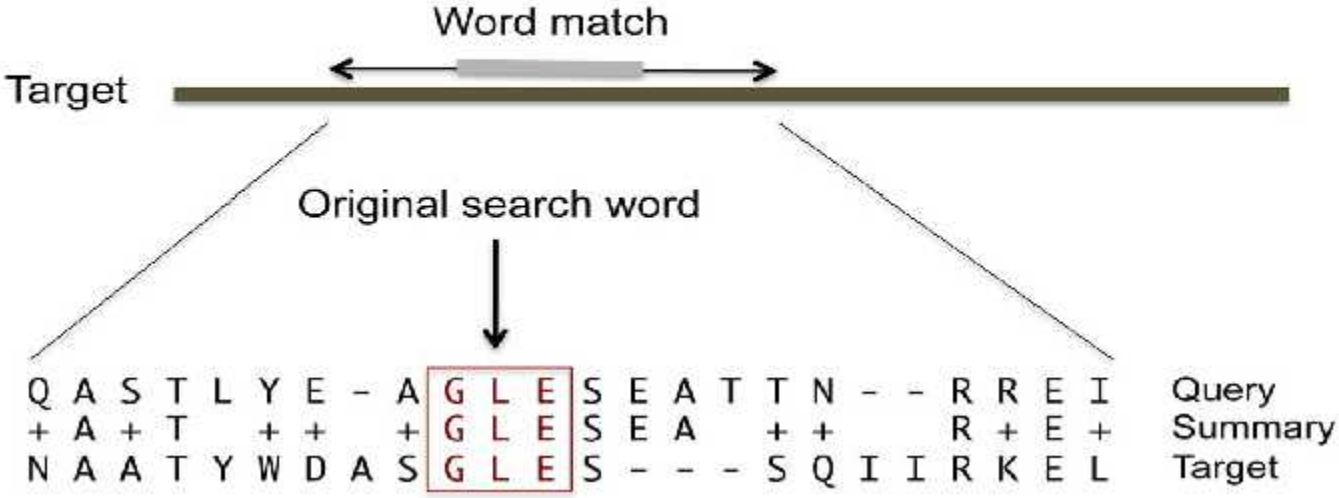
2. Query is divided into words, which are assigned a score.

E	A	G		5 + 4 + 6 = 15
	A	G	L	4 + 6 + 4 = 14
		G	L E	6 + 4 + 5 = 15
			L E S	6 + 4 + 5 = 15

3. Synonyms with scores above 10 are added to the search set.

E A G	A G L	G L E	L E S
K A G (11)	S G L (11)	G I E (13)	I E S (13)
E S G (12)	A G I (12)	G L D (12)	
E C G (11)		G L Q (12)	
E T G (11)			
E V G (11)			

4. Word matches are extended until running scores drop too low.



Contd….

- Using this search set, BLAST scans a database and identifies word hits/matches that score above the threshold.
- These short matches serve as seeds. The BLAST algorithm attempts to extend the match in the immediate sequence neighborhood
- BLAST keeps a running raw score, using scoring matrices, as it extends the matches. Each new amino acid either increases or decreases the raw score
- Penalties are assigned for mismatches and for gaps between the two alignments.

Contd….

- In the NCBI default settings, a gap brings an initial penalty of 11, which increases by 1 for each missing amino acid.
- Once the score falls below a set level, the alignment ceases and blast stops trying to extend the alignment.
- An extended sequence alignment that was initially seeded by a word hit is produced -called an hsp, or high-scoring segment pair

Contd….

- **All HSPs that have a cumulative score above the threshold score are reported in BLAST results.**
- **Raw scores are then converted into bit scores by correcting for the scoring matrix used**

BLOSUM 62 scoring matrix

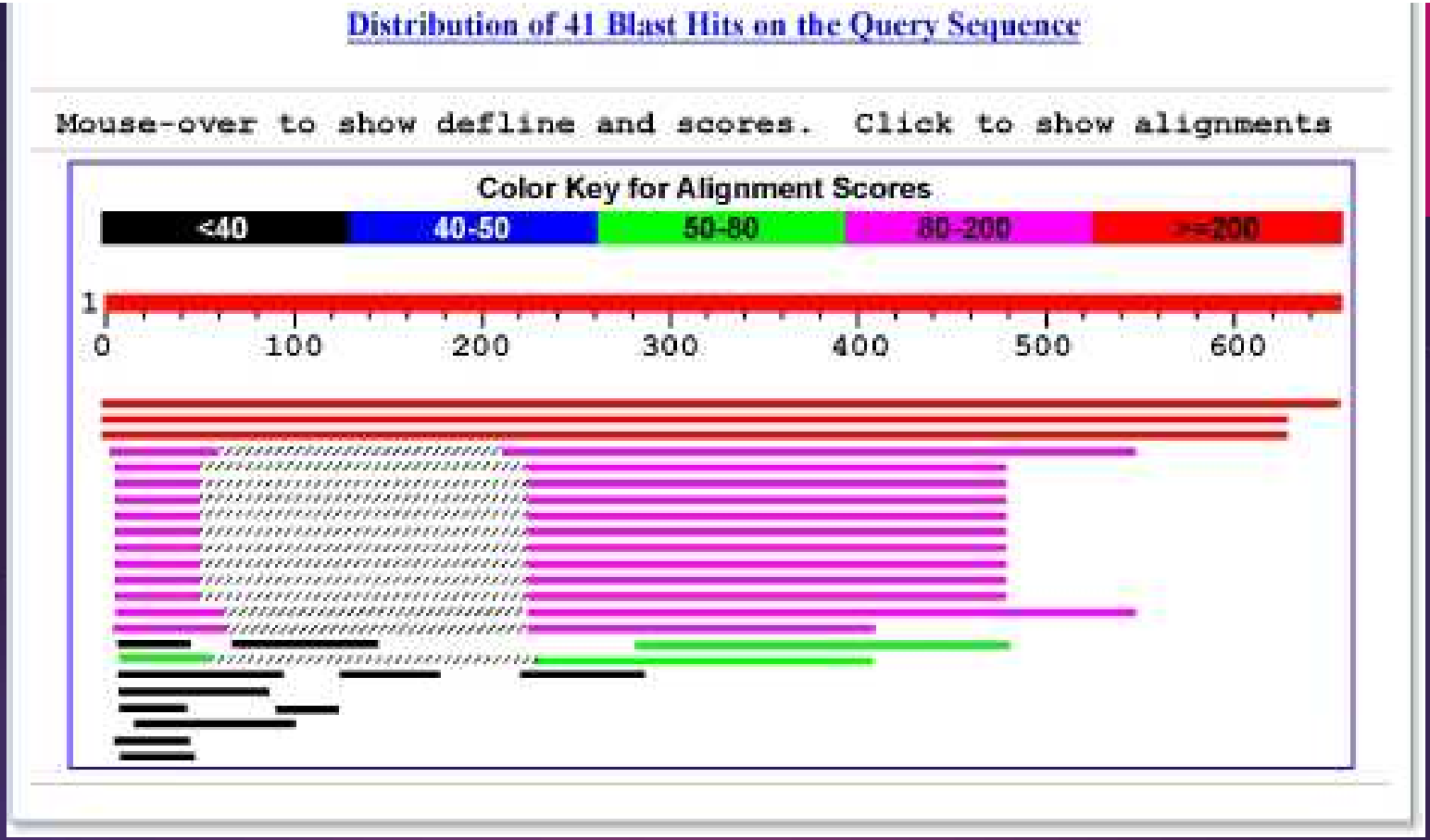
(positive values are shaded)

	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**

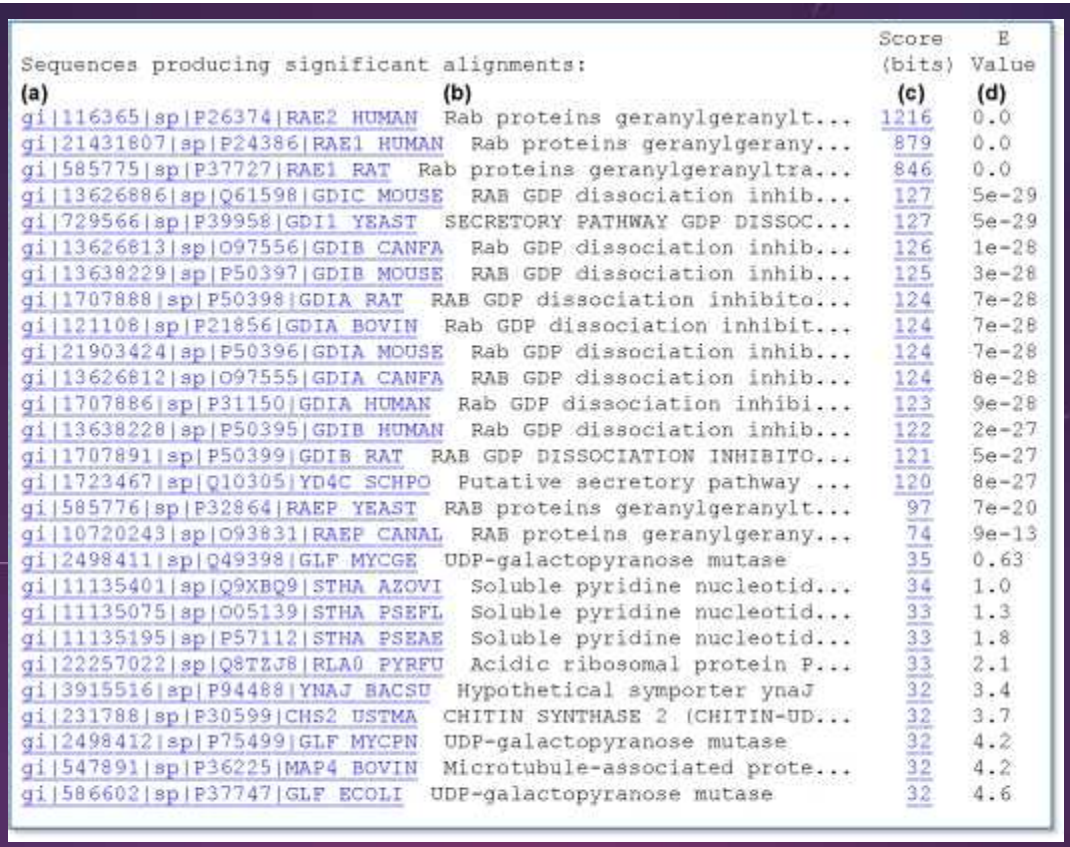
The Blast output

- Includes a table with the bit scores (S) for each alignment and its E-value, or "expect score"
- the score (S) is a measure of the quality of an alignment (calculated as the sum of substitution and gap scores for each aligned residue)
- E-value (E), or expectation value is a measure of the significance of the alignment. The E-value is 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 alignment result.
- Alignments with the highest bit scores and lowest E-values are listed at the top of the table.

How a BLAST result looks



The query sequence - numbered red bar at the top of the figure. Database hits are shown aligned to the query, below the red bar. Of the aligned sequences, the most similar are shown closest to the query. In this case, there are three high scoring database matches that align to most of the query sequence. The next twelve bars represent lower-scoring matches that align to two regions of the query, from about residues 3–60 and residues 220–500. The cross-hatched parts of the these bars indicate that the two regions of similarity are on the same protein, but that this intervening region does not match. The remaining bars show lower-scoring alignments. Mousing over the bars displays the definition line for that sequence to be shown in the window above the graphic.



Sequences producing significant alignments:		Score (bits)	E Value
(a)	(b)	(c)	(d)
gi|116365|sp|P26374|RAE2_HUMAN	Rab proteins geranylgeranylt...	1216	0.0
gi|21431807|sp|P24386|RAE1_HUMAN	Rab proteins geranylgerany...	879	0.0
gi|585775|sp|P37727|RAE1_RAT	Rab proteins geranylgeranyltra...	846	0.0
gi|13626886|sp|Q61598|GDIC_MOUSE	RAB GDP dissociation inhib...	127	5e-29
gi|729566|sp|P39958|GDI1_YEAST	SECRETORY PATHWAY GDP DISSOC...	127	5e-29
gi|13626813|sp|O97556|GDIB_CANFA	Rab GDP dissociation inhib...	126	1e-28
gi|13638229|sp|P50397|GDIB_MOUSE	RAB GDP dissociation inhib...	125	3e-28
gi|1707888|sp|P50398|GDIA_RAT	RAB GDP dissociation inhibito...	124	7e-28
gi|121108|sp|P21856|GDIA_BOVIN	Rab GDP dissociation inhibit...	124	7e-28
gi|21903424|sp|P50396|GDIA_MOUSE	Rab GDP dissociation inhib...	124	7e-28
gi|13626812|sp|O97555|GDIA_CANFA	RAB GDP dissociation inhib...	124	8e-28
gi|1707886|sp|P31150|GDIA_HUMAN	Rab GDP dissociation inhibi...	123	9e-28
gi|13638228|sp|P50395|GDIB_HUMAN	Rab GDP dissociation inhib...	122	2e-27
gi|1707891|sp|P50399|GDIB_RAT	RAB GDP DISSOCIATION INHIBITO...	121	5e-27
gi|1723467|sp|Q10305|YD4C_SCHPO	Putative secretory pathway ...	120	8e-27
gi|585776|sp|P32864|RAEP_YEAST	RAB proteins geranylgeranylt...	97	7e-20
gi|10720243|sp|O93831|RAEP_CANAL	RAB proteins geranylgerany...	74	9e-13
gi|2498411|sp|Q49398|GLF_MYCGE	UDP-galactopyranose mutase	35	0.63
gi|11135401|sp|Q9XBQ9|STHA_AZOVI	Soluble pyridine nucleotid...	34	1.0
gi|11135075|sp|O05139|STHA_PSEFL	Soluble pyridine nucleotid...	33	1.3
gi|11135195|sp|P57112|STHA_PSEAE	Soluble pyridine nucleotid...	33	1.8
gi|22257022|sp|Q8TZJ8|RLA0_PYRFU	Acidic ribosomal protein P...	33	2.1
gi|3915516|sp|P94488|YNAJ_BACSU	Hypothetical symporter ynaJ	32	3.4
gi|231788|sp|P30599|CHS2_USTMA	CHITIN SYNTHASE 2 (CHITIN-UD...	32	3.7
gi|2498412|sp|P75499|GLF_MYCPN	UDP-galactopyranose mutase	32	4.2
gi|547891|sp|P36225|MAP4_BOVIN	Microtubule-associated prote...	32	4.2
gi|586602|sp|P37747|GLF_ECOLI	UDP-galactopyranose mutase	32	4.6

One-line descriptions in the BLAST report

Each line is composed of four fields: (*a*) *the gi number, database designation, accession number, and locus name for the matched* sequence, separated by vertical bars (appendix 1); (*b*) *a brief textual description of the sequence, the definition. This usually* includes information on the organism from which the sequence was derived, the type of sequence (e.G., mRNA or DNA), and some information about function or phenotype. The definition line is often truncated in the one-line descriptions to keep the display compact; (*c*) *the alignment score in bits. Higher scoring hits are found at the top of the list; and (d) the e-value, which provides* an estimate of statistical significance. For the first hit in the list, the gi number is *116365, the database designation is sp (for* SWISS-PROT), the accession number is *P26374, the locus name is RAE2_HUMAN, the definition line is rab proteins, the score* is *1216, and the e-value is 0.0. Note that the first 17 hits have very low e-values (much less than 1) and are either RAB proteins* or GDP dissociation inhibitors. The other database matches have much higher e-values, 0.5 and above, which means that these sequences may have been matched by chance alone.

A pairwise sequence alignment from a BLAST report

The alignment is preceded by the sequence identifier, the full definition line, and the length of the matched sequence, in amino acids. Next comes the bit score (the raw score is in *parentheses) and then the E-value. The following line contains information* on the number of identical residues in this alignment (*Identities), the number of conservative substitutions (Positives), and if* applicable, the number of gaps in the alignment. Finally, the actual alignment is shown, with the query on *top, and the database* match is labeled as *Sbjct, below. The numbers at left and right refer to the position in the amino acid sequence. One or more dashes* (–) within a sequence indicate insertions or deletions. Amino acid residues in the query sequence that have been masked because of low complexity are replaced by Xs (see, for example, the *fourth and last blocks). The line between the two sequences indicates* the similarities between the sequences. If the query and the subject have the same amino acid at a given location, the residue itself is shown. Conservative substitutions, as judged by the substitution matrix, are indicated with +.