

Basic Local Alignment Search Tool (BLAST)

Bioinformatics and Functional Genomics

Third Edition

Jonathan Pevsner

Introduction

- ❑ **BLAST** is the **main NCBI tool** for **comparing** a **protein** or **DNA** sequence to **other sequences** in various databases.
- ❑ BLAST searching is one of the **fundamental ways of learning** about a **protein** or **gene**: the search reveals **what related sequences** are **present** in the **same organism** and **other organisms**.
- ❑ The **NCBI website** includes **several excellent resources** for **learning** about **BLAST**.

Introduction

- ❑ BLAST searching allows the **user** to **select** one sequence (termed the *query*) and **perform pairwise sequence alignments** between the *query* and an **entire database** (termed the *target*).
- ❑ The **Needleman–Wunsch global alignment algorithm** is **not used** for **database searches** because we are usually more interested in **identifying locally matching regions** such as **protein domains**.
- ❑ The **Smith–Waterman local alignment algorithm** finds **optimal pairwise alignments**, but we **cannot use** it for **database searches** generally because it is **too computationally intensive**.
- ❑ BLAST offers a local alignment strategy having both **speed** and **sensitivity**. It also offers **convenient accessibility** on the **World Wide Web** or as a **command-line tool**.

Introduction

BLAST searching has a **wide variety** of uses:

- ❑ **Determining** what **orthologs** and **paralogs** are known for a **particular protein** or **nucleic acid sequence**.
- ❑ **Determining** what **proteins** or **genes** are **present** in a **particular organism**.
- ❑ **Determining** the **identity** of a **DNA** or **protein** sequence. (For example, in an **RNAseq experiment**: a **particular RNA sequence** is **dramatically regulated** under the **experimental conditions** that you are using. This **sequence** may be **searched against** a **protein database** to learn **what proteins** are **most related** to the **protein encoded** by your **nucleotide sequence**.
- ❑ **Discovering new genes**. For example, a BLAST search of **genomic DNA** may **reveal** that the **DNA** encodes a **protein** that **has not been described before**.
- ❑ **Determining** what **variants** have been described for a **particular gene** or **protein**. For example, many **viruses** are **extremely mutable**; what **HIV-1 Pol variants** are known?
- ❑ **Investigating** expressed sequence tags (**ESTs**) that may exhibit **alternative splicing**. There is an **EST database** that can be **explored** by **BLAST searching**.
- ❑ **Exploring amino acid residues** that are **important** in the **function** and/or **structure** of a **protein**.

Main page for a BLASTP search at NCBI.

The sequence can be input as an **accession number**, **GI identifier**, or **FASTA-formatted sequence** (arrow 1).

The database must be selected (arrow 2) if the default setting is **not selected** (as here, in which the database is set to **RefSeq proteins**)

The **search** can be **restricted** to a **particular organism** or **taxonomic group**, and **Entrez queries** can be used to **further focus** the search (arrow 3)

The screenshot shows the NCBI BLASTP search page. The interface is titled "Standard Protein BLAST" and includes navigation links for "blastn", "blastp", "blastx", "tblastn", and "tblastx". The "blastp" tab is selected. The "Enter Query Sequence" section contains a text box with a FASTA-formatted sequence:

```
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAMGNPKVKAH
GKKVLGAFTSDGLAHLDNLKGTFAITSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQ
AAYQKVVAAGVANALAHKYH
```

. An arrow labeled "1" points to this text box. Below the text box is a "Browse..." button. The "Job Title" field contains the text "gi|4504349|ref|NP_000509.1| hemoglobin subunit...". The "Choose Search Set" section has a "Database" dropdown menu set to "Reference proteins (refseq_protein)". An arrow labeled "2" points to this dropdown. Below the database selection are fields for "Organism" and "Exclude" options. The "Entrez Query" field contains the text "perutz mf[Author]". An arrow labeled "3" points to this field. The "Program Selection" section has a "blastp" radio button selected. An arrow labeled "4" points to this section. At the bottom, there is a "BLAST" button and a "Show results in a new window" checkbox. An arrow labeled "5" points to the "Algorithm parameters" link at the bottom left. The page also includes a "Reset page" and "Bookmark" link in the top right corner.

BLAST Search Steps

Step 1: Specifying Sequence of Interest

There are **two main forms** of data **input**:

- (1) cutting and pasting DNA or protein sequence (e.g., in the **FASTA format**)
- (2) using an **accession number** (e.g., a **RefSeq** or **GenBank Identification (GI)** number)

```
>6F34_1|Chain A|Amino acid transporter|Geobacillus kaustophilus (235909)
MNLFRKKPIQLLMKESGAKGASLRKELGAFDLTMLGIGAIIGTGIFVLTGVAAAEHAGPALVLSFILSGLACVFAALCY
AEFASTVPVSGSAYTYSYATFGELIAWILGWDLILEYGVASSAVAVGWSGYFQGLLSGFGIELPKALTSAYDPAKGTFI
DLPAAIIVLFITFLLNLGAKKSARFNAVIVAIAKVAVVLLFLAVGVWYVKPENWTPFMPYGFSGVATGAATVFFAYIGFD
AVSTAAEEVRNPQRDMPIGIIVSLLVCTLLYIAVSLVLTGIVPYEQLNVKNPVAFALNYIHQDWWAGFISLGAIAAGITT
VLLVSMYGQTRLFYAISRDGLLPKVFARISPTRQVPYVNTWLTGAAVAVFAGIIPLNKLAELTNIGTLFAFITV SIGVL
VLRKTQPDLKRAFRVPFVPVVPILAVLFCGYLV LQLPAMTWIGFVSWLLIGLVIYFIYGRKHSELN
```

The **BLAST search** also **allows** you to **select** a **subset** of an **entire query sequence**, such as a region or domain of interest.

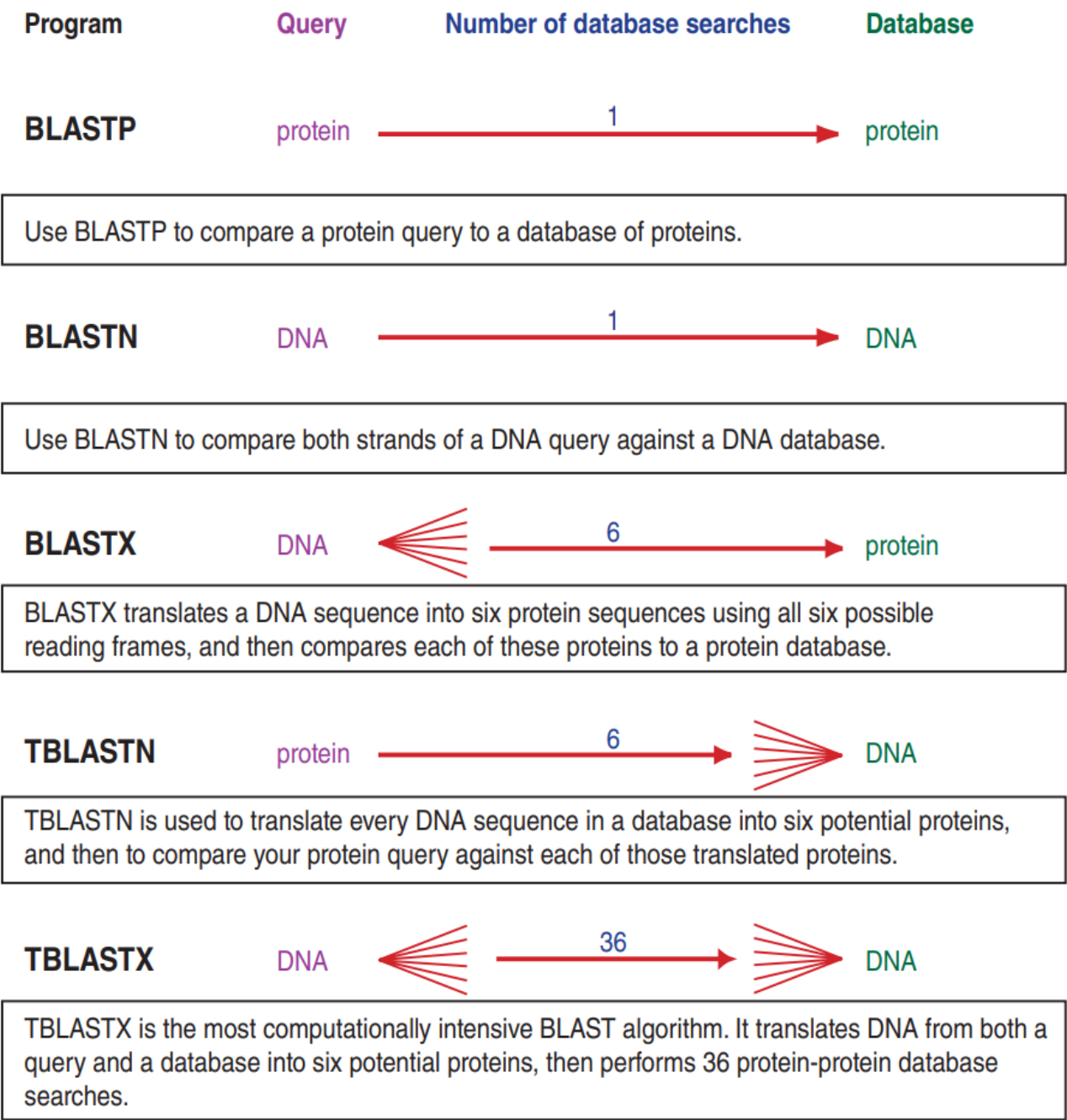
Step 2: Selecting BLAST Program

The NCBI BLAST family of programs includes **five main programs**

Note that

- suffix **P** refers to **protein** (as in BLASTP)
- **N** refers to **nucleotide**
- **X** refers to a **DNA query** that is dynamically **translated** into **six protein sequences**
- The prefix **T** refers to “**translating**,” in which a **DNA database** is dynamically **translated** into **six proteins**

In all three cases, these **algorithms** perform **protein–protein alignments**.



Step 2: Selecting BLAST Program

- ❑ **BLASTX:** If you have a **DNA sequence** and you **want to know** what **protein** (if any) it **encodes**, you can perform a BLASTX search.
 - The **BLASTX program** then **compares** each of the **six translated protein sequences** to **all** the **members** of a **protein database** (look up [NM_000518 in Nucleotide NCBI](#)).

- ❑ **TBLASTN:** One might use this program to **ask** whether **a DNA database** **encodes** a **protein** that **matches your protein query** of **interest**. Does a query with **beta globin** yield any **matches** in **a database** of **genomic DNA** from the **genome-sequencing project** of a **particular organism**?

- ❑ **TBLASTX:** Consider a situation in which you have **a DNA sequence** with **no obvious database matches** and you **want to know** if it **encodes** a **protein** with **distant, statistically significant database matches** in a **database of expressed sequence tags**.

Note:

A **BLASTX** search would be **more sensitive** than **BLASTN**, and **therefore useful** to **reveal genes** that **encode proteins** **homologous** to **your query**.

Step 3: Selecting a Database

❑ For **protein database searches** (**BLASTP** and **BLASTX**), the **default** option is the **nonredundant (nr) database**.

Database	Title	# sequences
nr	All nonredundant GenBank CDS translations + PDB + SwissProt + PIR + PRF excluding environmental samples from WGS projects	65 million
Reference proteins	NCBI protein reference sequences	50 million
UniProtKB/SwissProt	Nonredundant UniProtKB/SwissProt sequences	450,000
Patented protein sequences	Protein sequences derived from the Patent division of GenBank	1.3 million
Protein Data Bank	PDB protein database	77,000
Metagenomic proteins	Proteins from WGS metagenomic projects (env_nr)	6.5 million
Transcriptome	Transcriptome Shotgun Assembly (TSA) sequences	770,000

- **Another option** is to search only **Refseq proteins**.

Step 3: Selecting a Database

- ❑ For **DNA database searches** (**BLASTN, TBLASTN, TBLASTX**) the **default** option is to search the nucleotide **nr/nt database**.
- The **nr databases** are often the **preferred sites** for **searching** the **majority of available sequences**.

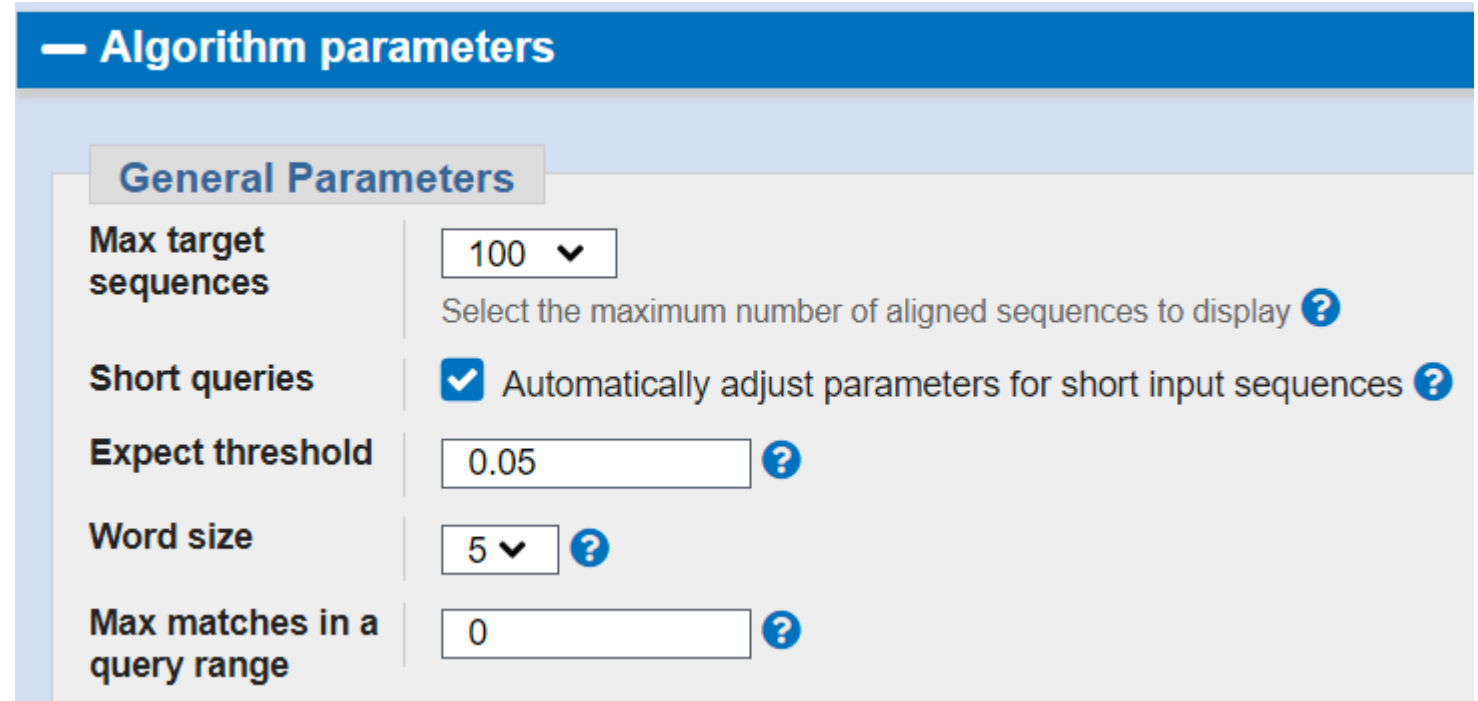
Database	Title	# sequences
Human Genomic + Transcript	Homo sapiens NCBI Annotation Release 104 RNAs; Homo sapiens all assemblies	55,000
Mouse Genomic + Transcript	Mus musculus NCBI Annotation RNAs; Mus musculus all assemblies	N/A
nr/nt	All GenBank+EMBL+DDBJ+PDB+RefSeq sequences, but excludes EST, STS, GSS, WGS, TSA, patent sequences as well as phase 0, 1, and 2 HTGS sequences	25 million
refseq_rna	NCBI transcript reference sequences	3.5 million
refseq_genomic	NCBI genomic reference sequences	2.7 million
NCBI Genomes	NCBI chromosome sequences	28,000
Expressed sequence tags (EST)	Database of GenBank+EMBL+DDBJ sequences from EST Divisions	75 million
Genomic survey sequences (gss)	Genome survey sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences	36 million
High-throughput genomic sequences (HTGS)	Unfinished high-throughput genomic sequences; sequences: phases 0,1 and 2	153,000
Patent sequences	Nucleotide sequences derived from the Patent division of GenBank	21 million
Protein Data Bank	PDB nucleotide database	8000
alu	Human <i>Alu</i> repeat elements	325
Sequence tagged sites (STS)	Database of GenBank+EMBL+DDBJ sequences from STS Divisions	1.3 million
Whole-genome shotgun (wgs)	Whole-genome-shotgun contigs	116 million
Transcriptome Shotgun Assembly (TSA)	Transcriptome shotgun assembly (TSA) sequences	15 million
16S ribosomal RNA sequences (Bacteria and Archaea)	16S ribosomal RNA sequences (bacteria and archaea)	7500

Step 4a: Selecting Optional Search Parameters

BLAST

2. Short queries.

If you **select** this option, the **expect value** and **word** size are **automatically adjusted**.



The screenshot shows the 'Algorithm parameters' section of the BLAST interface. Under the 'General Parameters' tab, the following settings are visible:

Parameter	Value	Help
Max target sequences	100	?
Short queries	<input checked="" type="checkbox"/> Automatically adjust parameters for short input sequences	?
Expect threshold	0.05	?
Word size	5	?
Max matches in a query range	0	?

3. The **expect value E** is the **number of different alignments** with scores **equal** to or **greater** than **some score S** that are **expected** to **occur** in a **database** search by **chance** (look up (NP_000198.1 as a query).

- A reasonable general guideline is that database matches having **E values** of ≤ 0.05 are **statistically significant**.

By **changing** the **expect option** to a **lower number**, **fewer database hits** are **returned**; **fewer chance matches** are **reported**. **Increasing E** returns **more hits**.

Step 4a: Selecting Optional Search Parameters

4. Word size.

When a query is used to search a database, the **BLAST algorithm** first **divides** the **query** into a **series** of smaller sequences (**words**) of a **particular length** (**word size**). For **BLASTP**, a **larger word size** yields a **more accurate search**.

In practice, the **word size** can **remain** at **3** and should be reduced to **2** only when **your query** is a **very short peptide** (i.e., a short string of amino acids).

Algorithm parameters

General Parameters

Max target sequences	100 ▼	Select the maximum number of aligned sequences to display ?
Short queries	<input checked="" type="checkbox"/> Automatically adjust parameters for short input sequences ?	
Expect threshold	0.05 ?	
Word size	5 ▼ ?	
Max matches in a query range	0 ?	

4. Word size

BASIC LOCAL ALIGNMENT SEARCH TOOL (BLAST)

The **BLAST program** was developed by **Stephen Altschul** of **NCBI** in **1990** and has since become one of the **most popular programs for sequence analysis**.

The **objective** is to find **high-scoring ungapped segments** among **related sequences**.

Sequence alignment procedure in BLAST:

- **First step: Seeding**; creating a list of words from the **query sequence**
(Each word is typically **three residues** for **protein sequences** and **eleven residues** for **DNA sequences**.)
- **Second step**; searching a **sequence database** for the **occurrence** of **these words**
(This step is to **identify database sequences** containing the matching words.)
- **Third step**; scoring of the **matching** of the **words** by a **given substitution matrix**
(A **word** is considered a **match** if it is above a **threshold**.)
- **Fourth step**; pairwise alignment by **extending** from the words in **both directions** while counting the **alignment score** using the **same substitution matrix**
(The **extension** continues **until** the score of the alignment **drops** below a **threshold** due to **mismatches** (the drop threshold is **twenty-two** for **proteins** and **twenty** for **DNA**).)

BASIC LOCAL ALIGNMENT SEARCH TOOL (BLAST)

1. Query: **MRD**PYN**KLIS**
2. Scan every three residues to be used in searching BLAST word database.
3. Assuming one of the words finds matches in the database.

Query	PYN	PYN	PYN	PYN	...
Database	PYN	PFN	PFQ	PFE	...

4. Calculate sums of match scores based on BLOSUM62 matrix.

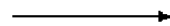
Query	PYN	PYN	PYN	PYN	...
Database	PYN	PFN	PFQ	PFE	...
Sum of score	20	16	10	10	...

5. Find the database sequence corresponding to the best word match and extend alignment in both directions.

Query	M R D	PYN	K L I S
Database	M H E	PYN	D V P W



extension to left



extension to right

6. Determine high scored segment above threshold (22).

Query	M R D	PYN	K L I S
Database	M H E	PYN	D V P W
	5 0 2	20	-1 1 -3 -3

HSP, total score 24

The **resulting contiguous aligned segment pair without gaps** is called **high-scoring segment pair (HSP)**

5. Max matches in a query range (Max hits per query).

Limits the number of target sequences that BLAST will return for a given portion of your query sequence.

What it means:

- It restricts the number of alignments BLAST reports for *any local region* of the query sequence.
- It's particularly useful when you have **long or multiple-domain queries**, and you don't want an overwhelming number of similar hits from redundant or closely related sequences.
- Helps **reduce computational load** and improve the **relevance** of results.

Example:

Suppose you have a query sequence with two domains, and you set "**Max matches in a query range**" to 1:

- For each domain region, BLAST will **only return the top-scoring match**, not all possible matches.
- You might get only **2 total hits**: 1 per domain (range), even if there are dozens of high-scoring ones.

If set to a **higher number** (e.g. 100):

- For each local region of the query, BLAST could return up to 100 matches, depending on what's available in the database.

Algorithm parameters

General Parameters

Max target sequences	100 ▼ <small>Select the maximum number of aligned sequences to display ?</small>
Short queries	<input checked="" type="checkbox"/> Automatically adjust parameters for short input sequences ?
Expect threshold	0.05 ?
Word size	5 ▼ ?
Max matches in a query range	0 ?

Use cases:

- **Set low** (e.g., 1–5) if you're looking for the best matches only, like in species identification or fast annotation.
- **Set high** (or leave at default) for more comprehensive homology searches.

Step 4a: Selecting Optional Search Parameters

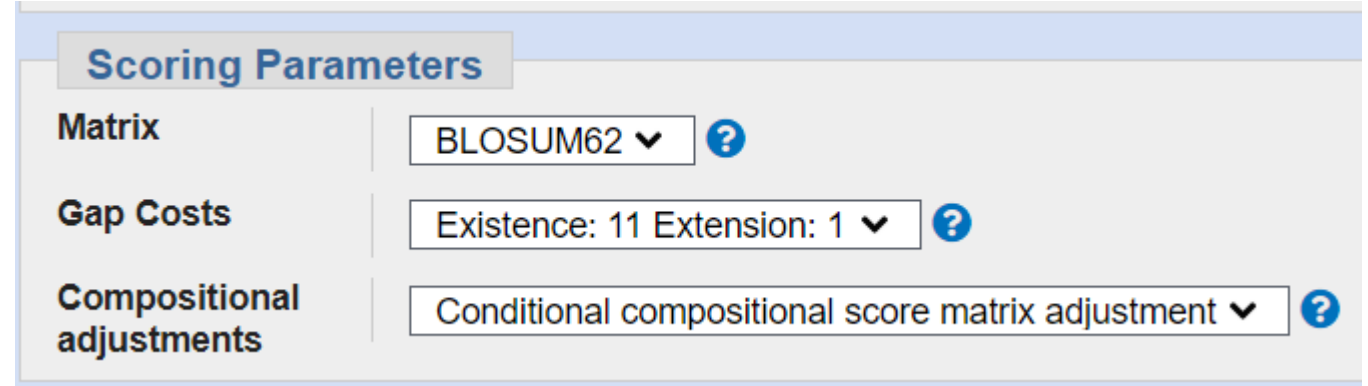
6. Matrix.

For **very short queries** (e.g., **15 or fewer amino acid residues**), a **PAM30** matrix is **recommended** (and is **automatically invoked** at the NCBI BLASTP site).

7. Gap costs.

Since a **single mutational event** may **cause the insertion or deletion of more than one residue**, the **presence of a gap** is **frequently ascribed** more **significance** than the **length** of the **gap**. Thus:

- **gap introduction is penalized heavily**
- **gap extension is lesser penalty**



Scoring Parameters	
Matrix	BLOSUM62 ?
Gap Costs	Existence: 11 Extension: 1 ?
Compositional adjustments	Conditional compositional score matrix adjustment ?

□ gap cost calculation

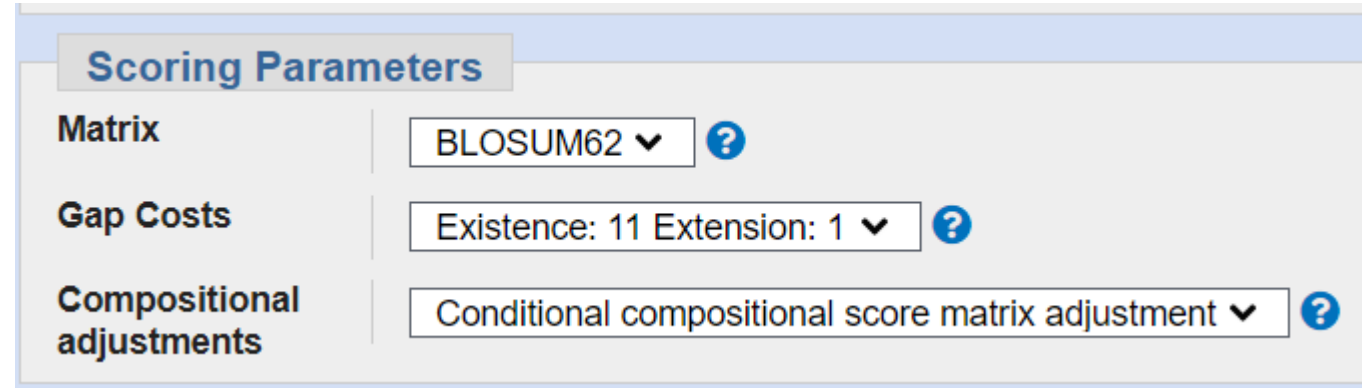
$$G + Ln$$

- G, the gap-opening penalty
 - L, the gap extension penalty
 - n, gap length
- The choice of **gap costs** is typically **10–15** for **G** and **1–2** for **L**. These are called **affine gap penalties**.

Step 4a: Selecting Optional Search Parameters

8. Compositional adjustments

- ☐ A **standard matrix** such as **BLOSUM62** is **not appropriate** for the **comparison of two proteins** with **nonstandard composition**.
- **Compositional adjustments** generally **increase** the **accuracy** of **BLAST searches** considerably.
- ☐ For **queries of very different lengths**:
 - **Compositional adjustments** can **reduce false positive search results** in **specialized circumstances** such as subjects matching.
 - In that case the **longer sequence** may **have a substantially different composition** than the **shorter**.
- ☐ In BLAST, "**Compositional adjustments**" refers to corrections made to the scoring system based on the **amino acid (or nucleotide) composition** of the sequences being compared. This is especially relevant for **protein BLAST** searches (e.g., **blastp**, **tblastn**), where the **amino acid composition** can vary widely across sequences.



The screenshot shows the 'Scoring Parameters' section of a BLAST search interface. It contains three rows of settings, each with a label on the left and a dropdown menu on the right, followed by a blue question mark icon. The first row is 'Matrix' with 'BLOSUM62' selected. The second row is 'Gap Costs' with 'Existence: 11 Extension: 1' selected. The third row is 'Compositional adjustments' with 'Conditional compositional score matrix adjustment' selected.

Scoring Parameters	
Matrix	BLOSUM62 ?
Gap Costs	Existence: 11 Extension: 1 ?
Compositional adjustments	Conditional compositional score matrix adjustment ?

What It Does:

BLAST assumes, by default, a certain **background frequency** of amino acids. However, suppose your query or database sequences have unusual composition (e.g., very **rich** in certain residues like **glycine** or **lysine**). In that case, this can **inflate similarity scores** just by chance, not true homology.

Compositional adjustments help correct for this bias to **reduce false positives** and **improve the biological relevance** of alignments.

Types of Compositional Adjustments in BLAST:

Depending on the BLAST version and search type, options may include:

1. No adjustment:

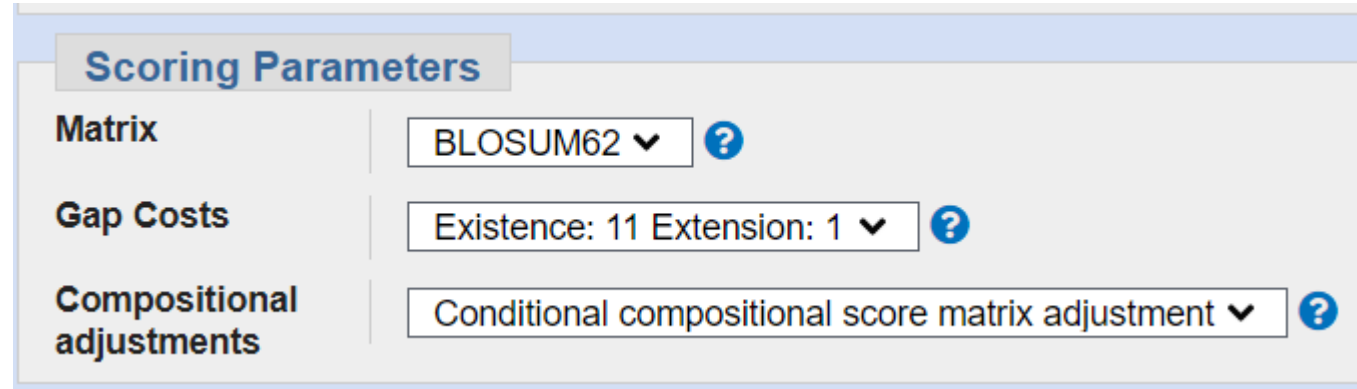
1. The original (older) behavior.
2. Scores are based on fixed background amino acid frequencies.
3. May overestimate similarity for biased sequences.

2. Composition-based statistics (conditional compositional score matrix adjustment):

1. Adjusts substitution scores based on the observed composition of the query and/or subject.
2. This is the **default for blastp** and considered more accurate.

3. Compositional score matrix adjustment:

1. A less conservative version compared to conditional adjustment.
2. May be used in some contexts for better sensitivity.



The image shows a screenshot of the 'Scoring Parameters' section in a BLAST web interface. It contains three rows of settings:

- Matrix:** A dropdown menu set to 'BLOSUM62' with a blue question mark icon to its right.
- Gap Costs:** A dropdown menu set to 'Existence: 11 Extension: 1' with a blue question mark icon to its right.
- Compositional adjustments:** A dropdown menu set to 'Conditional compositional score matrix adjustment' with a blue question mark icon to its right.

When to Use:

- **Default behavior is generally recommended** unless you have a specific reason to disable it.
- If you're working with **low-complexity or compositionally biased sequences**, this adjustment helps suppress **spurious hits**.
- Turning it off might be useful if you're trying to recover **remote homologs**, but it increases the risk of false positives.

(a) Default: **conditional compositional score matrix adjustment**

❑ Pairwise alignments from BLASTP

searches:

- The effects of changing compositional matrices and filtering options.
- Query: Human insulin (NP_000198.1)
- Database: RefSeq restricted to proteins in *Drosophila*

Insulin-like peptide 3 [*Drosophila melanogaster*]

Sequence ID: [ref|NP_648360.2|](#) Length: 120 Number of Matches: 1

Range 1: 32 to 114 [GenPept](#) [Graphics](#)

Score	Expect	Method	Identities	Positives	Gaps
31.6 bits(70)	0.050	Compositional matrix adjust.	21/88(24%)	40/88(45%)	12/88(13%)
Query 29	HLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQ-- 87				
	LCG L E L +C + + T+R + + Q++ G L+ L + S+Q				
Sbjct 32	KLCGRKLPETLSKLCV---YGFNAMTKRTLDPVNFNQID--GFEDRSLLERLLSDSSVQM 86				
Query 88	-----KRGIVEQCCTSICSLYQLENYC 109				
	+ G+ ++CC C++ ++ YC				
Sbjct 87	LKTRRLRDGVFDECCLKSCTMDEVLYYC 114				

(b) No adjustment (by default, filter low complexity regions)

Insulin-like peptide 3 [*Drosophila melanogaster*]

Sequence ID: [ref|NP_648360.2|](#) Length: 120 Number of Matches: 1

Range 1: 33 to 114 [GenPept](#) [Graphics](#)

Score	Expect	Identities	Positives	Gaps
33.5 bits(75)	0.009	21/87(24%)	40/87(45%)	12/87(13%)
Query 30	LCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQ-- 87			
	LCG L E L +C + + T+R + + Q++ G L+ L + S+Q			
Sbjct 33	LCGRKLPETLSKLCV---YGFNAMTKRTLDPVNFNQID--GFEDRSLLERLLSDSSVQML 87			
Query 88	-----KRGIVEQCCTSICSLYQLENYC 109			
	+ G+ ++CC C++ ++ YC			
Sbjct 88	KTRRLRDGVFDECCLKSCTMDEVLYYC 114			

- ❑ **Removing** compositional adjustments **lowers** the *E* value from 0.05 to 0.009

(c) Composition-based statistics

- ❑ Invoking a **composition-based** statistics option **improves** the ***E* value** by **500-fold** to 1×10^{-4}

Insulin-like peptide 3 [Drosophila melanogaster]

Sequence ID: [ref|NP_648360.2|](#) Length: 120 Number of Matches: 1

Range 1: 33 to 114 [GenPept](#) [Graphics](#)

Score	Expect	Method	Identities	Positives	Gaps
30.4 bits(67)	1e-04	Composition-based stats.	21/87(24%)	40/87(45%)	12/87(13%)
Query 30	LCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQ--	87			
	LCG L E L +C + + T+R + + Q++ G L+ L + S+Q				
Sbjct 33	LCGRKLPETLSKLCV---YGFNAMTKRTLDPVNFNQID--GFEDRSLLERLLSDSSVQML	87			
Query 88	-----KRGIVEQCCTSICSLYQLENYC	109			
	+ G+ ++CC C++ ++ YC				
Sbjct 88	KTRRLRDGVFDECCLKSCTMDEVLRVC	114			

- ❑ The **magnitude** of these effects *depends on* the **composition** of the particular **query**.

Note that **both** of the **altered settings** (Fig. b, c) had **minimal effects** on the **lengths** of the **aligned regions** or the **gaps**.

Step 4a: Selecting Optional Search Parameters

9. Filters.

- ☐ Filtering **masks** portions of the **query sequence** with **low complexity** (or *highly biased compositions*).
 - **Low-complexity sequences** are **defined** as **having commonly found stretches of amino acids** (or **nucleotides**) with **limited information content**.
- ☐ **Examples:**
 - **Dinucleotide repeats** (e.g., the repeating nucleotides **CACACACA...**)
 - ***Alu* sequences**
 - **Protein regions** that are **extremely rich** in **one** or **two amino acids**

Note:

Stretches of **hydrophobic amino acid residues** that form a **transmembrane domain** are **very common**, and a **database search** with such sequences **results in many database matches** that are **statistically significant** **but** **biologically irrelevant**.

- Other **motifs** that are **masked** by **filtering** include **acidic-**, **basic-**, and **proline-rich regions**.

- ☐ The **BLASTP** and **BLASTN** programs offer several main options:
 - For **protein sequence queries**, the **SEG program** is used
 - For **nucleic acid sequences**, the **DUST program** is employed.
 - Another approach is to **filter repeats** (for **BLASTN only**). This is **useful** to **avoid matching a query** with ***Alu* repeats** or other **repetitive DNA** to spurious database entries.

Note:

The **filtering** is **applied** to the **query sequence**, and **not** to the **entire database**.

Step 4a: Selecting Optional Search Parameters

10. Masking.

In BLAST, the "**Mask for lookup table only**" option refers to how **low-complexity regions** in your **query sequence** are treated during the **initial word lookup phase** of the alignment.

What It Means:

BLAST has a **two-phase process**:

- 1. Word lookup:** BLAST first creates a **lookup table** of "words" (short sequence fragments) from the query to quickly find potential matches in the database.
- 2. Extension and scoring:** BLAST then extends and scores these matches to find high-scoring segment pairs (HSPs).

The "**Mask for lookup table only**" setting controls whether **low-complexity regions** are:

- **Masked only during the lookup phase** (this option **enabled**)
- **Masked for both lookup and alignment** (option **disabled**)

How It Works:

- When enabled, **low-complexity regions** (e.g., stretches of a **single amino acid, repeats**) are **ignored during the initial word lookup**, so they don't generate misleading hits.
- But these regions **are still included in the alignment and scoring** once a hit is found.

The image shows a screenshot of the BLAST search interface. The top section is titled "Scoring Parameters" and contains three rows: "Matrix" with a dropdown set to "BLOSUM62", "Gap Costs" with a dropdown set to "Existence: 11 Extension: 1", and "Compositional adjustments" with a dropdown set to "Conditional compositional score matrix adjustment". The bottom section is titled "Filters and Masking" and contains two rows: "Filter" with a checkbox for "Low complexity regions" and "Mask" with two checkboxes, "Mask for lookup table only" and "Mask lower case letters". All checkboxes are currently unchecked.

Why Use It:

- **Reduces spurious alignments** caused by simple repetitive sequences.
- **Preserves biologically meaningful alignments** in low-complexity regions.
- Useful when you **don't want to exclude low-complexity regions entirely**, but still want to **avoid them biasing the search**.

Example Use Case:

You're aligning a protein with long stretches of **glutamine (Q)** or **proline (P)**, common in **transcription factors** or **structural proteins**. Using this setting helps **avoid** flooding your results with **low-quality matches** while still considering those regions in final alignments.

Step 4a: Selecting Optional Search Parameters

10. Masking.

☐ Mask lower-case letters:

- **Allows** you to **enter** a **query** in the **FASTA format** using *upper case characters for the search*, **but** **filtering** those **residues you choose** to filter by entering them in **lower case**.
- These **particular options** have **dramatic effects** for some queries (including those having **transmembrane spans** that can **potentially match thousands** of database entries).

Scoring Parameters	
Matrix	BLOSUM62 ▼ ?
Gap Costs	Existence: 11 Extension: 1 ▼ ?
Compositional adjustments	Conditional compositional score matrix adjustment ▼ ?

Filters and Masking	
Filter	<input type="checkbox"/> Low complexity regions ?
Mask	<input type="checkbox"/> Mask for lookup table only ?
	<input type="checkbox"/> Mask lower case letters ?