

Frequently Asked Questions (FAQ) concerning sequence similarity searching using
NCBI BLAST[®] and GETSIM in USGENE[®] on STN[®]

NCBI BLAST[®] SIMILARITY SEARCHING IN USGENE[®]	3
1. Q: What are the search limits for BLAST ?	3
2. Q: How do I upload sequences longer than the 256 character line length limit?	3
3. Q: How can I view an uploaded sequence before I submit it to BLAST ?	3
4. Q: Do spaces, carriage returns and header characters, from formats such as FASTA or WIPO ST.25, in uploaded queries harm the sequence search ?	4
5. Q: How can I alter the advanced options for a BLAST search ? How do they effect the answer set ? ...	4
6. Q: What does "EXCEEDS MAXIMUM FIELD LENGTH, WILL BE SEARCHED AS 'XXXXX'" mean when I use the Query command and what can I do about it ?	4
7. Q: What does the graphic summary appearing after a similarity search represent ?	5
8. Q: What is the message " Maximum answer limit of 10,000 reached" trying to convey, and do I need to be concerned about it ?	5
9. Q: How do I display BLAST alignments and what information do they provide ?	5
10. Q: Is it possible to limit BLAST searches in USGENE with patent text, date terms and to preferred document types – especially to issued U.S. patents ?	6
11. Q: Unlike GETSIM, the BLAST search method does not report a Query Self Score value. Why is that and is there anyway to calculate this myself?	6
12. Q: How can I calculate the similarity percentage of the aligned sequence ?	6
13. Q: How can I see only one result per patent family but sorted by similarity ?	7
14. Q: How many BATCH searches can be queued up at any one time?	7
15. Q: Will I be billed for a result if I delete it from the batch queue without retrieving it ?	7
16. Q: What status is possible for a BATCH and how long are the results available ?	7
17. Q: I would like to monitor patenting activities for biosequences. Can I run a BLAST search within STN's standard SDI command procedure ?	8
18. Q: Does the ALERT search work any differently from the standard BLAST search ?	8
19. Q: Are the search options BATCH and ALERT also available for BLAST ?	8
20. Q: Which similarity matrices are used by the BLAST algorithm?	8

Frequently Asked Questions (FAQ) concerning sequence similarity searching using
NCBI BLAST[®] and GETSIM in USGENE[®] on STN[®]

GETSIM SIMILARITY SEARCHING IN USGENE[®]:	12
21. Q: What are the search limits for GETSIM ?	12
22. Q: What shall I do with sequences longer than 500/750 residues?	12
23. Q: Which algorithm should I use with similarity searching in USGENE ?	12
24. Q: Is there any difference in the command syntax when conducting a similarity search with BLAST compared to GETSIM ?	12
25. Q: What are the main differences between the BLAST and GETSIM ?	12
26. Q: After a similarity search with BLAST I get a totally different number of answers than with GETSIM for the same query sequence. What is the cause of that ?	13
27. Q: What does the graph appearing after a GETSIM search represent ?	13
28. Q: Is the graph after a GETSIM search different to the one after a BLAST search ?	13
29. Q: What is the Smith-Waterman score and how can I calculate it myself ?	13
30. Q: How is the score threshold for the candidate answer set determined ?	13
31. Q: Could you explain what the Query Self Score value means?	13
32. Q: What is the message "Incomplete Search" trying to convey ?	13
33. Q: What does the ALIGN display format following a GETSIM search show ?	14
34. Q: How can I calculate the similarity percentage of the aligned sequence ?	14
35. Q: Which similarity scoring matrices are used in a GETSIM search ?	15
36. Q: Which similarity scoring matrices are used in a translated (/TSQN) search ?	17

Frequently Asked Questions (FAQ) concerning sequence similarity searching using
NCBI BLAST® and GETSIM in USGENE® on STN®

NCBI BLAST® SIMILARITY SEARCHING IN USGENE®

1. **Q: What are the search limits for BLAST ?**

A: Minimum sequence query length : 5
Maximum sequence query length - command line input : 256
Maximum sequence query length - uploaded input for /SQP : 10000
Maximum sequence query length - uploaded input for /SQN : 10000
Maximum sequence query length - uploaded input for /TSQN : 10000
Maximum sequence query length - uploaded input for BATCH processing : 10000
Maximum sequence query length - uploaded input for ALERT processing : 10000

2. **Q: How do I upload sequences longer than the 256 character line length limit?**

A: When using STN Express you should use the Upload Query Wizard, which is accessed via the Discover! Button bottom left of the STN Online and Results window when online. Follow the step-by-step prompts.

See: [STN Express Upload Query Wizard](#)

A: When using STN on the Web you must use the Upload Sequence Query from the Search Assistants Folder, or use the menu-driven STN on the Web Sequence Search Assistant.

See: [STN on the Web Sequence Search Assistant](#)

3. **Q: How can I view an uploaded sequence before I submit it to BLAST ?**

A: Uploaded queries can be conveniently viewed using the D L# LQUE command.

=> UPL R BLAST

UPLOAD SUCCESSFULLY COMPLETED
L1 GENERATED

=> D L1 LQUE

```
L1 ANSWER 1 USGENE COPYRIGHT 2007 SEQUENCEBASE CORP
LQUE gtgttcaaaaaataccaatacctcgctttggcagcactgtgtgcccgcctcgctggcaggctgcgacaaagccg
gcagctttttcggtgcggaacaaaaagaagcatcctttgtagaacgcatcaaacacacaccaaagacgacggcag
cgtcagtatgctgctgcccgaactttgtccaactgggtcaaagcgaaggcccgagtcgtcaatattcaggca
gccccgccccgcgacccccaaaaacggcagcagcaatgccgaaaccgattccgacccgcttgccgacagcgacc
cgttctacgaatttttcaaacgcctcgtcccgaacatgccgaaaatcccccaagaagaagcagatgacggngg
attgaacttcggttcgggcttcatcatcagcaaaagacggctatattctgaccaatacgcacgctcgttaccggc
atgggcagtatcaaagtcctgctcaacgacaagcgcgaatataaccgccaaactcatcggttcggatgtccaat
gcaactccggcgcccgcgtgttcaacttaaaaggacaggtcgtcggcatcaactcgaaaatatacagccgcag
cggcggattcatgggcatttcccttcgccatcccgattgacgcttgccatgaatgtcgccgaacagctgaaaaac
accggcaaaagccaacgcggacaactgggcgtgattattcaagaagtatcctacggtttggcacaatcgttcg
gtttggacaaaagccggcggcgcactgattgccaaaatcctgcccggcagccccgcagaacgctgcccgcctgcg
ggcggcgacatcgtcctcagcctcgacggcggagaaatacgttcttccggcgaccttcccgttatggtcggc
gccattacgcccgggaaaaagaagtacgacctcggcgtatggcgcgaaaggcgaagaaatcacaatcaaagtcaagc
tgggcaacgcccggagcagatcggcgcacatcctcaaaaacagatgaagccccctacaccgaacagcaatccgg
tacgttctcggtcgaatccgcaggcattacccttcagacacataccgacagcagcggcggacacctcgtcgtc
gtacgggtttccgacgcggcagaacgcgcaggcttgaggcgcggcgacgaaaattcttgccgtcgggcaagtcc
ccgtcaatgacgaagccgggtttccgaaaagctatggacaaggcaggcaaaaacgctccccctgctgatcatgcg
ccgtggcaacacgctgttatcgcattaaacctgcaataa
```

Frequently Asked Questions (FAQ) concerning sequence similarity searching using NCBI BLAST® and GETSIM in USGENE® on STN®

4. **Q: Do spaces, carriage returns and header characters, from formats such as FASTA or WIPO ST.25, in uploaded queries harm the sequence search ?**

A: When using the STN Express Upload Query Wizard (see 2.) any spaces, carriage returns and header information from a number of standard formats are stripped out before the plain text query is uploaded to STN. In addition, any three letter amino acid codes are converted to the corresponding one letter codes. However, it is necessary to ensure that there is at least one carriage return (line break) every 300 characters. If this is not done, the following error message will occur in [USGENE](#).

```
=> UPL R BLAST
```

```
UPLOAD SUCCESSFULLY COMPLETED  
L1 GENERATED
```

```
=> RUN BLAST L1 /SQP -F F
```

```
WARNING: Your query may have been truncated. A single line in a file for UPLOAD  
cannot have more than 300 characters
```

```
DO YOU WISH TO CONTINUE ? (NO): NO
```

A: When using the STN on the Web (see 2.) any unwanted characters are NOT stripped out before the plain text query is uploaded to STN. So it is essential to do this manually in advance, before uploading the query. FASTA format is therefore recommended. Like STN Express, it is also necessary to ensure that there is at least one carriage return (line break) every 300 characters (see above).

5. **Q: How can I alter the advanced options for a BLAST search ? How do they effect the answer set ?**

A: For the experienced user of BLAST, a variety of advanced options are available via the STN command line.

The BLAST advanced options are specified with a single letter code preceded by a hyphen, followed by a space and the required setting or value. You may alter the expectation value (-e), the word size (-w), the filter (-f), the searched strand (-s), the gap cost (-g) and the gap extension (-x). For a polypeptide search the matrix (-m), e.g. -m BLOSUM80, may be defined, and for a nucleotide search the penalty for a nucleotide mismatch (-q) and the reward for a nucleotide match (-r) may be specified. For example:

```
=> RUN BLAST GPFQAFXCDDPDYAKTLRTPKSYKFSPPLLGKLDG/SQP -E 1000 -W 2 -F F -M PAM30
```

With several BLAST advanced options only a certain range of predefined values are acceptable (e.g. the gap cost and gap extension settings with the different matrices). Online you may use `HELP BLAST` and `HELP OPTIONS` to view the default settings and accepted value. This HELP is also available as a [PDF file](#).

Altering default BLAST parameters will have a profound effect on the outcome of the search. It is therefore highly recommended that users are completely familiar with [NCBI documentation](#) regarding these parameters, before embarking on customizing any of the BLAST advanced options.

6. **Q: What does "EXCEEDS MAXIMUM FIELD LENGTH, WILL BE SEARCHED AS 'XXXXX'" mean when I use the Query command and what can I do about it ?**

A: If you use the query command to store a biosequence query, but do not give a field code, the system does not know that it has to deal with a biosequence. By default it assumes a text search in the basic index. Since the basic index is restricted to a certain term length, above message is issued. If you give a field code like /SQP the system knows that it has to handle a biosequence and hence does not issue above message, e.g. :

```
=> QUE MIQPVFRKVDLSLSEDISLTQSIYDKKLVLMQKNLQGLDPKALNNCSFCHEAGQ/SQSP  
L1 QUE MIQPVFRKVDLSLSEDISLTQSIYDKKLVLMQKNLQGLDPKALNNCSFCHEAGQ/SQSP
```


Frequently Asked Questions (FAQ) concerning sequence similarity searching using NCBI BLAST® and GETSIM in USGENE® on STN®

Example: peptide sequence

```
BLASTALIGN
Query   = 39 letters
Length  = 8946
Score   = 24.3 bits (51), Expect = 0.007
Identities = 13/36 (36%), Positives = 20/36 (55%), Gaps = 4/36 (11%)
Query: 2   GPFQAFXCDPDYAKTLRTDPKSQYKFSPLLGKLDG 37
          G +A   D D+A+TLR   + + +F   LG +DG
Sbjct: 1358 GDLRAGVIDADHARTLRQHVVQVESRF----LGAMDG 1389
```

The alignment may also be readily combined with other [USGENE](#) display formats, e.g. D BRIEF ALIGN.

10. **Q: Is it possible to limit BLAST searches in USGENE with patent text, date terms and to preferred document types – especially to issued U.S. patents ?**

A: Yes. To limit a BLAST search in [USGENE](#) with date or text terms, simply AND them with the BLAST answer set L-number. To limit your search to issued (granted) patents, as opposed to published applications sequences, search GRANTED in the Sequence Source (/SSO) field. After using text or date limitations, it is important to remember to SORT SCORE D on the newly created L-numbered answer set.

For example: limit a BLAST search by application year range, text terms and to granted patents.

```
=> RUN BLAST L1 /SQP -F F
```

```
. . . . .
```

```
HOW MANY ANSWERS WOULD YOU LIKE TO KEEP ? (ALL) OR ?: ALL
```

```
L2   RUN STATEMENT CREATED
```

```
L2       1350 VQTVPLSRLFDHAMLEAHRAHELALDITYQEFEEETYIPKDQKYSFLHDSQT
          SFCFSDSIPTSPNMEETQQKSNLELLRISLLLLIESWLEPVRFLRSMFANN
          LVYDTSDDSDDYHLLKDLLEEGIQTLMGRLEDGSRRTGQILKQTYSKFDTNS
          HNHDAALLKNYGLLYCFRKDMDKVETFLRMVQCRSVEGSCGF/SQP.-F F
```

Answer set arranged by accession number; to sort by descending similarity score, enter at an arrow prompt (=>) "sor score d".

```
=> S L2 AND SOMATOMAMMOTROPIN AND AY<1996 AND GRANTED/SSO
```

```
L3       7 L2 AND SOMATOMAMMOTROPIN AND AY<1996 AND GRANTED/SSO
```

```
=> SOR SCORE D
```

```
PROCESSING COMPLETED FOR L3
```

```
L4       7 SOR L3 SCORE D
```

11. **Q: Unlike GETSIM, the BLAST search method does not report a Query Self Score value. Why is that and is there anyway to calculate this myself?**

A: BLAST has been implemented as provided by the NCBI, and as such it does not feature a Query Self Score to assist in assessing search results. If you wish to calculate your own Query Self Score use the [NCBI BLAST 2 SEQUENCES](#) tool. There you can compare the query against itself to manually calculate a Query Self Score.

12. **Q: How can I calculate the similarity percentage of the aligned sequence ?**

A: BLAST is a local alignment tool and does not provide a global similarity percentage between query and answer. A local identity percentage provided within the ALIGN display (see 9.). It is possible for you to calculate a Query Self Score (see 11.) and from that an ideal to actual answer score percentage.

Frequently Asked Questions (FAQ) concerning sequence similarity searching using NCBI BLAST[®] and GETSIM in USGENE[®] on STN[®]

13. **Q: How can I see only one result per patent family but sorted by similarity ?**

A: USGENE records are based on sequences. Sorting according to the similarity score is accomplished using SOR SCORE D. Sequences from one patent application can be grouped together with the [patent family sort \(FSORT\) command](#). If a similarity sorted answer set is subsequently re-grouped using FSORT, the similarity sort order is retained separately within the multi-record and individual record subgroups provided by FSORT. Selective display from each family (application) group can be accomplished by using the [D PFAM function](#).

14. **Q: How many BATCH searches can be queued up at any one time?**

A: Up to 16 BATCH queries maybe queued and/or stored per STN Login ID.

15. **Q: Will I be billed for a result if I delete it from the batch queue without retrieving it ?**

A: You will be billed the small fee for the initiation of the BATCH, but not the larger fee for collection of results.

16. **Q: What status is possible for a BATCH and how long are the results available ?**

A: If a batch search is conducted the status of the batch is indicated with "queued", "running", and "complete". If the search has completed the results can be obtained by retrieving the batch. A "retrieved" batch request is deleted automatically one week after the first retrieval. During this time it is possible to retrieve the same request several times and process the answer set. Only the first retrieval of a batch request will be charged. During the week period subsequent repeat retrievals of the batch result will be free of charge. Completed batches which have not yet been retrieved will remain available for three months from initiation.

Example: Get a BATCH list

```
=> RUN GETBATCH
Please enter your batch identifier
    or enter # for batch id list
    or enter * for batch id at top of list
    or enter - before batch id to delete
    or enter . for (end)
BATCH REQUEST:#
Batch result files remaining:
TEST1      Completed (blast)
TEST2      Completed (blast)
SPH1       Retrieved (blast)
IPST1      Retrieved (getsim)
IPST2      Completed (getsim)
IPST3      Running      (getsim)
IPST3      Queued       (getsim)
-----
Please enter your batch identifier
    or enter # for batch id list
    or enter * for batch id at top of list
    or enter - before batch id to delete
    or enter . for (end)
BATCH REQUEST:
```

Frequently Asked Questions (FAQ) concerning sequence similarity searching using NCBI BLAST® and GETSIM in USGENE® on STN®

17. **Q: I would like to monitor patenting activities for biosequences. Can I run a BLAST search within STN's standard SDI command procedure ?**

A: We provide the ALERT function for current awareness searching with the BLAST similarity search tool. Add the word ALERT to the end of the RUN BLAST command and give the ALERT a name when prompted.

```
=> RUN BLAST L1 /SQP -F F ALERT
```

For USGENE text and bibliography SDI queries you should use the standard [SDI command](#) procedure.

18. **Q: Does the ALERT search work any differently from the standard BLAST search ?**

A: Each BLAST ALERT works on a smaller search space than the standard BLAST search, i.e. just the latest database update. Other than that BLAST ALERT is the same as the standard BLAST run online.

19. **Q: Are the search options BATCH and ALERT also available for BLAST ?**

A: Yes, both options are available for similarity search with GETSIM as well as BLAST. Although the BATCH search mode is not needed with BLAST as normally the search is completed within about 30 seconds. The BATCH feature is more useful with GETSIM when the online search may take considerably longer. When retrieving the BATCH or ALERT list you are pointed at the search mode conducted.

20. **Q: Which similarity matrices are used by the BLAST algorithm?**

A: The similarity of nucleotides is represented by the settings of the penalty of a nucleotide mismatch (default: -3) and the reward for a nucleotide match (default: 1). The calculation of similarity between nucleotides may be changed by altering these settings on the STN command line (see [5](#)).

A: For the similarity searching of peptide sequences several matrices are available. BLOSUM62 is used as default, and can be changed to either the BLOSUM45, BLOSUM80, PAM30 or PAM70 matrices. The matrix of choice is set on the STN command line using -M option. See [5](#) for more information on command syntax.

The general advice on usage of BLAST matrices is given from the NCBI as follows:

<u>Query Length</u>	<u>Substitution Matrix</u>	<u>Gap costs</u>
<35	PAM-30	(9,1)
35 – 50	PAM-70	(10,1)
50 – 85	BLOSUM-80	(10,1)
>85	BLOSUM-62	(11,1)

The PAM family:

PAM matrices are based on global alignments of closely related proteins. The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence. Other PAM matrices are extrapolated from PAM1.

The BLOSUM family:

BLOSUM matrices are based on local alignments. BLOSUM 62 is a matrix calculated from comparisons of sequences with no less than 62% divergence. All BLOSUM matrices are based on observed alignments; they are not extrapolated from comparisons of closely related proteins. BLOSUM 62 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.

BLOSUM matrices with higher numbers and PAM matrices with low numbers are both designed for comparisons of closely related sequences. BLOSUM matrices with low numbers and PAM matrices with high numbers are designed for comparisons of distantly related proteins.

Frequently Asked Questions (FAQ) concerning sequence similarity searching using NCBI BLAST® and GETSIM in USGENE® on STN®

BLOSUM 62 Matrix (the BLAST default)

* column uses minimum score

BLOSUM Clustered Scoring Matrix in 1/2 Bit Units

Cluster Percentage: = 62

Entropy = 0.6979, Expected = -0.5209

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

The characters in the database of peptide sequences represent the following amino acids:

1-Letter Code	3-Letter Code	Name
A	Ala	Alanine
B	Asx	Aspartic acid or Asparagine
C	Cys	Cysteine
D	Asp	Aspartic acid
E	Glu	Glutamic acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
X	Xxx	Uncommon
Y	Tyr	Tyrosine
Z	Glx	Glutamic acid or Glutamine

**Frequently Asked Questions (FAQ) concerning sequence similarity searching using
NCBI BLAST® and GETSIM in USGENE® on STN®**

BLOSUM 45 Matrix

* column uses minimum score

BLOSUM Clustered Scoring Matrix in 1/3 Bit Units

Cluster Percentage: = 45

Entropy = 0.3795, Expected = -0.2789

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-2	-2	0	-1	-1	0	-5
R	-2	7	0	-1	-3	1	0	-2	0	-3	-2	3	-1	-2	-2	-1	-1	-2	-1	-2	-1	0	-1	-5
N	-1	0	6	2	-2	0	0	0	1	-2	-3	0	-2	-2	-2	1	0	-4	-2	-3	4	0	-1	-5
D	-2	-1	2	7	-3	0	2	-1	0	-4	-3	0	-3	-4	-1	0	-1	-4	-2	-3	5	1	-1	-5
C	-1	-3	-2	-3	12	-3	-3	-3	-3	-3	-2	-3	-2	-4	-1	-1	-5	-3	-1	-2	-3	-2	-5	
Q	-1	1	0	0	-3	6	2	-2	1	-2	-2	1	0	-4	-1	0	-1	-2	-1	-3	0	4	-1	-5
E	-1	0	0	2	-3	2	6	-2	0	-3	-2	1	-2	-3	0	0	-1	-3	-2	-3	1	4	-1	-5
G	0	-2	0	-1	-3	-2	-2	7	-2	-4	-3	-2	-2	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-5
H	-2	0	1	0	-3	1	0	-2	10	-3	-2	-1	0	-2	-2	-1	-2	-3	2	-3	0	0	-1	-5
I	-1	-3	-2	-4	-3	-2	-3	-4	-3	5	2	-3	2	0	-2	-2	-1	-2	0	3	-3	-3	-1	-5
L	-1	-2	-3	-3	-2	-2	-2	-3	-2	2	5	-3	2	1	-3	-3	-1	-2	0	1	-3	-2	-1	-5
K	-1	3	0	0	-3	1	1	-2	-1	-3	-3	5	-1	-3	-1	-1	-1	-2	-1	-2	0	1	-1	-5
M	-1	-1	-2	-3	-2	0	-2	-2	0	2	2	-1	6	0	-2	-2	-1	-2	0	1	-2	-1	-1	-5
F	-2	-2	-2	-4	-2	-4	-3	-3	-2	0	1	-3	0	8	-3	-2	-1	1	3	0	-3	-3	-1	-5
P	-1	-2	-2	-1	-4	-1	0	-2	-2	-2	-3	-1	-2	-3	9	-1	-1	-3	-3	-3	-2	-1	-1	-5
S	1	-1	1	0	-1	0	0	0	-1	-2	-3	-1	-2	-2	-1	4	2	-4	-2	-1	0	0	0	-5
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-1	2	5	-3	-1	0	0	-1	0	-5	
W	-2	-2	-4	-4	-5	-2	-3	-2	-3	-2	-2	-2	1	-3	-4	-3	15	3	-3	-4	-2	-2	-5	
Y	-2	-1	-2	-2	-3	-1	-2	-3	2	0	0	-1	0	3	-3	-2	-1	3	8	-1	-2	-2	-1	-5
V	0	-2	-3	-3	-1	-3	-3	-3	-3	3	1	-2	1	0	-3	-1	0	-3	-1	5	-3	-3	-1	-5
B	-1	-1	4	5	-2	0	1	-1	0	-3	-3	0	-2	-3	-2	0	0	-4	-2	-3	4	2	-1	-5
Z	-1	0	0	1	-3	4	4	-2	0	-3	-2	1	-1	-3	-1	0	-1	-2	-2	-3	2	4	-1	-5
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	-2	-1	-1	-1	-1	-1	-5
*	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1

BLOSUM 80 Matrix

* column uses minimum score

BLOSUM Clustered Scoring Matrix in 1/2 Bit Units

Cluster Percentage: = 80

Entropy = 0.9868, Expected = -0.7442

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

Frequently Asked Questions (FAQ) concerning sequence similarity searching using NCBI BLAST® and GETSIM in USGENE® on STN®

PAM 30 Matrix

PAM 30 substitution matrix, scale = $\ln(2)/2 = 0.346574$
 Expected score = -5.06, Entropy = 2.57 bits
 Lowest score = -17, Highest score = 13

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	6	-7	-4	-3	-6	-4	-2	-2	-7	-5	-6	-7	-5	-8	-2	0	-1	-13	-8	-2	-3	-3	-3	-17
R	-7	8	-6	-10	-8	-2	-9	-9	-2	-5	-8	0	-4	-9	-4	-3	-6	-2	-10	-8	-7	-4	-6	-17
N	-4	-6	8	2	-11	-3	-2	-3	0	-5	-7	-1	-9	-9	-6	0	-2	-8	-4	-8	6	-3	-3	-17
D	-3	-10	2	8	-14	-2	2	-3	-4	-7	-12	-4	-11	-15	-8	-4	-5	-15	-11	-8	6	1	-5	-17
C	-6	-8	-11	-14	10	-14	-14	-9	-7	-6	-15	-14	-13	-13	-8	-3	-8	-15	-4	-6	-12	-14	-9	-17
Q	-4	-2	-3	-2	-14	8	1	-7	1	-8	-5	-3	-4	-13	-3	-5	-5	-13	-12	-7	-3	6	-5	-17
E	-2	-9	-2	2	-14	1	8	-4	-5	-5	-9	-4	-7	-14	-5	-4	-6	-17	-8	-6	1	6	-5	-17
G	-2	-9	-3	-3	-9	-7	-4	6	-9	-11	-10	-7	-8	-9	-6	-2	-6	-15	-14	-5	-3	-5	-5	-17
H	-7	-2	0	-4	-7	1	-5	-9	9	-9	-6	-6	-10	-6	-4	-6	-7	-7	-3	-6	-1	-1	-5	-17
I	-5	-5	-5	-7	-6	-8	-5	-11	-9	8	-1	-6	-1	-2	-8	-7	-2	-14	-6	2	-6	-6	-5	-17
L	-6	-8	-7	-12	-15	-5	-9	-10	-6	-1	7	-8	1	-3	-7	-8	-7	-6	-7	-2	-9	-7	-6	-17
K	-7	0	-1	-4	-14	-3	-4	-7	-6	-6	-8	7	-2	-14	-6	-4	-3	-12	-9	-9	-2	-4	-5	-17
M	-5	-4	-9	-11	-13	-4	-7	-8	-10	-1	1	-2	11	-4	-8	-5	-4	-13	-11	-1	-10	-5	-5	-17
F	-8	-9	-9	-15	-13	-13	-14	-9	-6	-2	-3	-14	-4	9	-10	-6	-9	-4	2	-8	-10	-13	-8	-17
P	-2	-4	-6	-8	-8	-3	-5	-6	-4	-8	-7	-6	-8	-10	8	-2	-4	-14	-13	-6	-7	-4	-5	-17
S	0	-3	0	-4	-3	-5	-4	-2	-6	-7	-8	-4	-5	-6	-2	6	0	-5	-7	-6	-1	-5	-3	-17
T	-1	-6	-2	-5	-8	-5	-6	-6	-7	-2	-7	-3	-4	-9	-4	0	7	-13	-6	-3	-6	-4	-4	-17
W	-13	-2	-8	-15	-15	-13	-17	-15	-7	-14	-6	-12	-13	-4	-14	-5	-13	13	-5	-15	-10	-14	-11	-17
Y	-8	-10	-4	-11	-4	-12	-8	-14	-3	-6	-7	-9	-11	2	-13	-7	-6	-5	10	-7	-6	-9	-7	-17
V	-2	-8	-8	-8	-6	-7	-6	-5	-6	2	-2	-9	-1	-8	-6	-6	-3	-15	-7	7	-8	-6	-5	-17
B	-3	-7	6	6	-12	-3	1	-3	-1	-6	-9	-2	-10	-10	-7	-1	-3	-10	-6	-8	6	0	-5	-17
Z	-3	-4	-3	1	-14	6	6	-5	-1	-6	-7	-4	-5	-13	-4	-5	-6	-14	-9	-6	0	6	-5	-17
X	-3	-6	-3	-5	-9	-5	-5	-5	-5	-5	-6	-5	-5	-8	-5	-3	-4	-11	-7	-5	-5	-5	-5	-17
*	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	1

PAM 70 Matrix

PAM 70 substitution matrix, scale = $\ln(2)/2 = 0.346574$
 Expected score = -2.77, Entropy = 1.60 bits
 Lowest score = -11, Highest score = 13

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	5	-4	-2	-1	-4	-2	-1	0	-4	-2	-4	-4	-3	-6	0	1	1	-9	-5	-1	-1	-1	-2	-11
R	-4	8	-3	-6	-5	0	-5	-6	0	-3	-6	2	-2	-7	-2	-1	-4	0	-7	-5	-4	-2	-3	-11
N	-2	-3	6	3	-7	-1	0	-1	1	-3	-5	0	-5	-6	-3	1	0	-6	-3	-5	5	-1	-2	-11
D	-1	-6	3	6	-9	0	3	-1	-1	-5	-8	-2	-7	-10	-4	-1	-2	-10	-7	-5	5	2	-3	-11
C	-4	-5	-7	-9	9	-9	-9	-6	-5	-4	-10	-9	-9	-8	-5	-1	-5	-11	-2	-4	-8	-9	-6	-11
Q	-2	0	-1	0	-9	7	2	-4	2	-5	-3	-1	-2	-9	-1	-3	-3	-8	-8	-4	-1	5	-2	-11
E	-1	-5	0	3	-9	2	6	-2	-2	-4	-6	-2	-4	-9	-3	-2	-3	-11	-6	-4	2	5	-3	-11
G	0	-6	-1	-1	-6	-4	-2	6	-6	-6	-7	-5	-6	-7	-3	0	-3	-10	-9	-3	-1	-3	-3	-11
H	-4	0	1	-1	-5	2	-2	-6	8	-6	-4	-3	-6	-4	-2	-3	-4	-5	-1	-4	0	1	-3	-11
I	-2	-3	-3	-5	-4	-5	-4	-6	-6	7	1	-4	1	0	-5	-4	-1	-9	-4	3	-4	-4	-3	-11
L	-4	-6	-5	-8	-10	-3	-6	-7	-4	1	6	-5	2	-1	-5	-6	-4	-4	-4	0	-6	-4	-4	-11
K	-4	2	0	-2	-9	-1	-2	-5	-3	-4	-5	6	0	-9	-4	-2	-1	-7	-7	0	-1	-2	-3	-11
M	-3	-2	-5	-7	-9	-2	-4	-6	-6	1	2	0	10	-2	-5	-3	-2	-8	-7	0	-6	-3	-3	-11
F	-6	-7	-6	-10	-8	-9	-9	-7	-4	0	-1	-9	-2	8	-7	-4	-6	-2	4	-5	-7	-9	-5	-11
P	0	-2	-3	-4	-5	-1	-3	-3	-2	-5	-5	-4	-5	-7	7	0	-2	-9	-9	-3	-4	-2	-3	-11
S	1	-1	1	-1	-1	-3	-2	0	-3	-4	-6	-2	-3	-4	0	5	2	-3	-5	-3	0	-2	-1	-11
T	1	-4	0	-2	-5	-3	-3	-3	-4	-1	-4	-1	-2	-6	-2	2	6	-8	-4	-1	-1	-3	-2	-11
W	-9	0	-6	-10	-11	-8	-11	-10	-5	-9	-4	-7	-8	-2	-9	-3	-8	13	-3	-10	-7	-10	-7	-11
Y	-5	-7	-3	-7	-2	-8	-6	-9	-1	-4	-4	-7	-7	4	-9	-5	-4	-3	9	-5	-4	-7	-5	-11
V	-1	-5	-5	-5	-4	-4	-4	-3	-4	3	0	-6	0	-5	-3	-3	-1	-10	-5	6	-5	-4	-2	-11
B	-1	-4	5	5	-8	-1	2	-1	0	-4	-6	-1	-6	-7	-4	0	-1	-7	-4	-5	5	1	-2	-11
Z	-1	-2	-1	2	-9	5	5	-3	1	-4	-4	-2	-3	-9	-2	-2	-3	-10	-7	-4	1	5	-3	-11
X	-2	-3	-2	-3	-6	-2	-3	-3	-3	-3	-4	-3	-3	-5	-3	-1	-2	-7	-5	-2	-2	-3	-3	-11
*	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	-11	1

**Frequently Asked Questions (FAQ) concerning sequence similarity searching using
NCBI BLAST[®] and GETSIM in USGENE[®] on STN[®]**

GETSIM SIMILARITY SEARCHING IN USGENE[®]:

21. Q: What are the search limits for GETSIM ?

A: Minimum sequence query length : 5.
 Maximum sequence query length - command line input : 256
 Maximum sequence query length - uploaded input for /SQP : 750
 Maximum sequence query length - uploaded input for /SQN : 500
 Maximum sequence query length - uploaded input for /TSQN : 500
 Maximum sequence query length - uploaded input for BATCH processing : 2000
 Maximum sequence query length - uploaded input for ALERT processing : 2000

22. Q: What shall I do with sequences longer than 500/750 residues?

A: Use the BATCH mode for your search, which allows up to 2,000 characters in the query sequence. If your query is longer than 2,000 characters it might be advisable to restrict the search to the most significant parts.

23. Q: Which algorithm should I use with similarity searching in USGENE ?

A: As the two algorithms are fundamentally different in their calculation of the similarity between two sequences, and thus may result in a substantially different number and content of answers in the answer set, it is highly recommended to use both search modes when conducting a similarity search. Thereby the most comprehensive set(s) of answers can be retrieved.

24. Q: Is there any difference in the command syntax when conducting a similarity search with BLAST compared to GETSIM ?

A: The basic command syntax is the same. If you want to conduct a similarity search in **USGENE** use either the command **RUN BLAST** or **RUN GETSIM**. The sequence qualifiers **/SQP** for a polypeptide sequence (default), **/SQN** for a nucleotide sequence, and **/TSQN** for translated search are identical for both. When displaying the alignment between the query sequence and a hit sequence use **D ALIGN** for both. Beyond this, the only difference is that **BLAST** has a series of user-definable advanced options for expert searchers (see **5.**).

25. Q: What are the main differences between the BLAST and GETSIM ?

A: The table below shows a comprehensive summary of the main differences between FASTA-based GETSIM and NCBI BLAST.

BLAST	FASTA
Faster than FASTA	Slower than BLAST
Equivalent for highly similar sequences	
Misses some less similar sequences	Better for less similar sequences
Comparison of shorter sequence parts	Comparison over entire sequence length
Less sensitive when using default settings	More sensitive; misses less homologs
Less separation between true homologs and random hits	More separation between true homologs and random hits
Calculates probabilities	Calculates significance "on the fly" from the given dataset

Frequently Asked Questions (FAQ) concerning sequence similarity searching using NCBI BLAST[®] and GETSIM in USGENE[®] on STN[®]

26. **Q: After a similarity search with BLAST I get a totally different number of answers than with GETSIM for the same query sequence. What is the cause of that ?**

A: As the two algorithms FASTA and BLAST use slightly different default settings and a basically different calculation of similarity they produce a different answer set when conducted. Depending on the length of a query sequence or global or local sequence alignments this can result in a considerable difference in the number of retrieved similar sequences and the content of the answer set. In the BLAST search the number of retrieved answers can be changed by altering the expectation value (see 5.).

27. **Q: What does the graph appearing after a GETSIM search represent ?**

A: After a similarity search has been completed a graphical presentation of the results appears. The vertical axis shows the similarity score calculated for the hit sequences compared to the query sequence and the horizontal axis represents the number of retrieved hit sequences. To the left there is(are) the sequence(s) with the highest score value(s) with descending similarity score (similarity to the query sequence) to the right thus expressing how many answers were found with high, middle, or low similarity to the query sequence.

28. **Q: Is the graph after a GETSIM search different to the one after a BLAST search ?**

A: The graphic account represents basically the same, i.e. the similarity score on the vertical axis and the number of retrieved hit sequences on the horizontal axis. The difference in the representation is based on the different calculation of similarity and score value by the two algorithms.

29. **Q: What is the Smith-Waterman score and how can I calculate it myself ?**

A: GETSIM is based upon the FASTA algorithm which employs the Smith-Waterman scoring system. The Smith-Waterman score is a value calculated for the optimized alignment taking the weighted match values and gap penalties of the sequence and query symbols into account. This value cannot be easily calculated by hand, and the value of the query against itself (query self score) is included in the GETSIM output for convenience.

30. **Q: How is the score threshold for the candidate answer set determined ?**

A: The threshold value is a function of the query sequence length and it is adjusted to yield a reasonable number of candidate answers. Values are adjusted to character set, database size and search method, and they have to be individually set for the various search types, e.g. SQP, TSQN or SQN.BOTH. Also the increasing size of the database has to be taken into account and the threshold value parameters occasionally adjusted accordingly. Hence the absolute threshold values vary and on their own are not likely to provide an indication of the quality of a match unless put in relation to others.

31. **Q: Could you explain what the Query Self Score value means?**

A: The query self-score is gives you the value obtained when the query is compared to itself. Or in other words, if there is an exact match in your answer set – same sequence of characters and length – that answer should have the same Smith-Waterman score as the query self score. If the sequence in the database is shorter than the query, the self score might be higher than the Smith-Waterman score if the answer retrieved.

32. **Q: What is the message "Incomplete Search" trying to convey ?**

A: The above message appears when the search results in more than 10,000 candidate answers with the same score and the search has been discontinued prematurely. Usually this occurs when fairly unspecific query sequences are being searched which results in a too large amount of hit sequences. Please try to make the query sequence more specific until the candidate answer set is smaller than 10000. The GETSIM fee is not charged for incomplete answer sets created via an online search. If you were running a BATCH search the smaller initialization fee will be charged, but not the larger results collection fee. Also see 8.

Frequently Asked Questions (FAQ) concerning sequence similarity searching using NCBI BLAST® and GETSIM in USGENE® on STN®

33. **Q: What does the ALIGN display format following a GETSIM search show ?**

A: It is possible to see the alignment between the retrieved sequence and the query sequence with the display format ALIGN. The top line is the query sequence and the bottom line the hit sequence. In this display format a line between the two sequences gives the information about the degree of similarity: two dots represent identical nucleotides/amino acids, and a blank occurs if there is no match. One dot indicates a chemical “family” match. Gaps inserted in the query or answer sequence for alignment purposes are shown with an underscore.

Example: Nucleotide sequence

```
ALIGN Smith-Waterman score: 136
      50 na overlap starting at 45
      acatccttgtg__gcagct_gtcgaagccat_gagaggtcc__aagtcag
      :::::::::::  :::::  :::::::::::  :::::::::::  :::::::::::
      acatccttgtagcasctggtcgaagccatrgagaggtccctgaagtcag
```

Example: Peptide sequence

```
ALIGN Smith-Waterman score: 61
      22 aa overlap starting at 559
      qlqetlxqylcasteddpskcp
      .:  ::  .:::::  ...:  :
      kllatlqhvltcsltnspqmp
```

34. **Q: How can I calculate the similarity percentage of the aligned sequence ?**

A: The query sequence self score and a calculation of a similarity percentage are provided. A local identity percentage is not provided. Both can be displayed using the display command D SCORE. For example:

=> **D SCORE**

```
SCORE 239 15% of query self score 1496
```

**Frequently Asked Questions (FAQ) concerning sequence similarity searching using
NCBI BLAST® and GETSIM in USGENE® on STN®**

35. Q: Which similarity scoring matrices are used in a GETSIM search ?

A: For peptide and nucleotide searches two different similarity matrices are being employed.

MATRIX for the search of nucleotides

	A	B	C	D	G	H	K	M	N	R	S	T	U	V	W	X	Y
A	5	-2	-4	1	-4	1	-1	2	-1	2	-1	-4	-4	1	2	-1	-1
B	-2	1	1	-1	1	-1	1	-1	-1	-1	1	1	1	-1	-1	-1	1
C	-4	1	5	-2	-4	1	-1	2	-1	-1	2	-4	-4	1	-1	-1	2
D	1	-1	-2	1	1	-1	1	-1	-1	1	-1	1	1	-1	1	-1	-1
G	-4	1	-4	1	5	-2	2	-1	-1	2	2	-4	-4	1	-1	-1	-1
H	1	-1	1	-1	-2	1	-1	1	-1	-1	-1	1	1	-1	1	-1	1
K	-1	1	-1	1	2	-1	2	-1	-1	1	1	2	2	-1	1	-1	1
M	2	-1	2	-1	-1	1	-1	2	-1	-1	1	-1	-1	1	1	-1	-1
N	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
R	2	-1	-1	1	2	-1	1	-1	-1	2	1	-1	-1	1	1	-1	-2
S	-1	1	2	-1	2	-1	1	1	-1	1	2	-1	-1	1	-1	-1	1
T	-4	1	-4	1	-4	1	2	-1	-1	-1	-1	5	5	-2	2	-1	2
U	-4	1	-4	1	-4	1	2	-1	-1	-1	-1	5	5	-1	2	-1	2
V	1	-1	1	-1	1	-1	-1	1	-1	1	1	-2	-1	1	-1	-1	-1
W	2	-1	-1	1	-1	1	1	1	-1	1	-1	2	2	-1	2	-1	1
X	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
Y	-1	1	2	-1	-1	1	1	-1	-1	-2	1	2	2	-1	1	-1	2

The characters in the database of nucleotide sequences represent the following nucleotides :

Codes	Name or Definition
A	Adenine
G	Guanine
U	Uracil
R	A or G
S	C or G
K	G or T/U
H	A, C or T/U; not G
B	C, G or T/U; not A
C	Cytosine
T	Thymine
M	A or C
W	A or T/U
Y	C or T/U
V	A, C or G; not T/U
D	A, G or T/U; not C
N	Unknown or Other

Frequently Asked Questions (FAQ) concerning sequence similarity searching using NCBI BLAST® and GETSIM in USGENE® on STN®

MATRIX for the search of peptides and proteins

	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	X	Y	Z
A	5	-2	-1	-2	-1	-3	0	-2	-1	-1	-2	-1	-1	-1	-1	-2	1	0	0	-3	-1	-2	-1
B	-2	5	-3	5	1	-4	-1	0	-4	0	-4	-3	4	-2	0	-1	0	0	-4	-5	-1	-3	2
C	-1	-3	13	-4	-3	-2	-3	-3	-2	-3	-2	-2	-2	-4	-3	-4	-1	-1	-1	-5	-2	-3	-3
D	-2	5	-4	8	2	-5	-1	-1	-4	-1	-4	-4	2	-1	0	-2	0	-1	-4	-5	-1	-3	1
E	-1	1	-3	2	6	-3	-3	0	-4	1	-3	-2	0	-1	2	0	-1	-1	-3	-3	-1	-2	5
F	-3	-4	-2	-5	-3	8	-4	-1	0	-4	1	0	-4	-4	-4	-3	-3	-2	-1	1	-2	4	-4
G	0	-1	-3	-1	-3	-4	8	-2	-4	-2	-4	-3	0	-2	-2	-3	0	-2	-4	-3	-2	-3	-2
H	-2	0	-3	-1	0	-1	-2	10	-4	0	-3	-1	1	-2	1	0	-1	-2	-4	-3	-1	2	0
I	-1	-4	-2	-4	-4	0	-4	-4	5	-3	2	2	-3	-3	-3	-4	-3	-1	4	-3	-1	-1	-3
K	-1	0	-3	-1	1	-4	-2	0	-3	6	-3	-2	0	-1	2	3	0	-1	-3	-3	-1	-2	1
L	-2	-4	-2	-4	-3	1	-4	-3	2	-3	5	3	-4	-4	-2	-3	-3	-1	-1	-2	-1	-1	-3
M	-1	-3	-2	-4	-2	0	-3	-1	2	-2	3	7	-2	-3	0	-2	-2	-1	1	-1	-1	0	-1
N	-1	4	-2	2	0	-4	0	1	-3	0	-4	-2	7	-2	0	-1	1	0	-3	-4	-1	-2	0
P	-1	-2	-4	-1	-1	-4	-2	-2	-3	-1	-4	-3	-2	10	-1	-3	-1	-1	-3	-4	-2	-3	-1
Q	-1	0	-3	0	2	-4	-2	1	-3	2	-2	0	0	-1	7	1	0	-1	-3	-1	-1	-1	4
R	-2	-1	-4	-2	0	-3	-3	0	-4	3	-3	-2	-1	-3	1	7	-1	-1	-3	-3	-1	-1	0
S	1	0	-1	0	-1	-3	0	-1	-3	0	-3	-2	1	-1	0	-1	5	2	-2	-4	-1	-2	0
T	0	0	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	2	5	0	-3	0	-2	-1
V	0	-4	-1	-4	-3	-1	-4	-4	4	-3	1	1	-3	-3	-3	-3	-2	0	5	-3	-1	-1	-3
W	-3	-5	-5	-5	-3	1	-3	-3	-3	-3	-2	-1	-4	-4	-1	-3	-4	-3	-3	15	-3	2	-2
X	-1	-1	-2	-1	-1	-2	-2	-1	-1	-1	-1	-1	-1	-2	-1	-1	-1	0	-1	-3	-1	-1	-1
Y	-2	-3	-3	-3	-2	4	-3	2	-1	-2	-1	0	-2	-3	-1	-1	-2	-2	-1	2	-1	8	-2
Z	-1	2	-3	1	5	-4	-2	0	-3	1	-3	-1	0	-1	4	0	0	-1	-3	-2	-1	-2	5

The characters in the database of peptide sequences represent the following amino acids:

1-Letter Code	3-Letter Code	Name
A	Ala	Alanine
B	Asx	Aspartic acid or Asparagine
C	Cys	Cysteine
D	Asp	Aspartic acid
E	Glu	Glutamic acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
X	Xxx	Uncommon
Y	Tyr	Tyrosine
Z	Glx	Glutamic acid or Glutamine

**Frequently Asked Questions (FAQ) concerning sequence similarity searching using
NCBI BLAST® and GETSIM in USGENE® on STN®**

36. Q: Which similarity scoring matrices are used in a translated (/TSQN) search ?

A: For the translated search of a peptide query in the database of nucleotides, whose nucleic acids are translated into peptides, a translation table based on the Universal Genetic Code is used. For the similarity search itself a procedure is used which is based on the peptide search. Generic codes indexed in **USGENE** using the IUB symbols are taken into consideration.

**GENETIC CODE
Standard Translation Table**

Symbol	3-letter	Codons	! IUPAC	..
A	Ala	GCT GCC GCA GCG	! GCX	
B	Asx		! RAY	
C	Cys	TGT TGC	! TGY	
D	Asp	GAT GAC	! GAY	
E	Glu	GAA GAG	! GAR	
F	Phe	TTT TTC	! TTY	
G	Gly	GGT GGC GGA GGG	! GGX	
H	His	CAT CAC	! CAY	
I	Ile	ATT ATC ATA	! ATH	
J	???	...	! ...	
K	Lys	AAA AAG	! AAR	
L	Leu	TTG TTA CTT CTC CTA CTG	! TTR CTX YTR	; YTX
M	Met	atg	! ATG	
N	Asn	AAT AAC	! AAY	
O	???	...	! ...	
P	Pro	CCT CCC CCA CCG	! CCX	
Q	Gln	CAA CAG	! CAR	
R	Arg	CGT CGC CGA CGG AGA AGG	! CGX AGR MGR	; MGX
S	Ser	TCT TCC TCA TCG AGT AGC	! TCX AGY	; WSX
T	Thr	ACT ACC ACA ACG	! ACX	
U	???	...	! ...	
V	Val	GTT GTC GTA GTG	! GTX	
W	Trp	TGG	! TGG	
X	Xxx		! XXX	
Y	Tyr	TAT TAC	! TAY	
Z	Glx		! SAR	
.	! ...	
O	End	TAA TAG TGA	! TAR TRA	; TRR

Frequently Asked Questions (FAQ) concerning sequence similarity searching using NCBI BLAST® and GETSIM in USGENE® on STN®

MATRIX for TSQN searches

The matrix used for GETSIM TSQN searches is based on the peptide matrix (above) with an additional entry to take into account stop codons (indicated in the table below by an asterisk).

	A	B	C	D	E	F	G	H	I	K	L	M	N	*	P	Q	R	S	T	V	W	X	Y	Z
A	5	-2	-1	-2	-1	-3	0	-2	-1	-1	-2	-1	-1	-5	-1	-1	-2	1	0	0	-3	-1	-2	-1
B	-2	5	-3	5	1	-4	-1	0	-4	0	-4	-3	4	-5	-2	0	-1	0	0	-4	-5	-1	-3	2
C	-1	-3	13	-4	-3	-2	-3	-3	-2	-3	-2	-2	-2	-5	-4	-3	-4	-1	-1	-1	-5	-2	-3	-3
D	-2	5	-4	8	2	-5	-1	-1	-4	-1	-4	-4	2	-5	-1	0	-2	0	-1	-4	-5	-1	-3	1
E	-1	1	-3	2	6	-3	-3	0	-4	1	-3	-2	0	-5	-1	2	0	-1	-1	-3	-3	-1	-2	5
F	-3	-4	-2	-5	-3	8	-4	-1	0	-4	1	0	-4	-5	-4	-4	-3	-3	-2	-1	1	-2	4	-4
G	0	-1	-3	-1	-3	-4	8	-2	-4	-2	-4	-3	0	-5	-2	-2	-3	0	-2	-4	-3	-2	-3	-2
H	-2	0	-3	-1	0	-1	-2	10	-4	0	-3	-1	1	-5	-2	1	0	-1	-2	-4	-3	-1	2	0
I	-1	-4	-2	-4	-4	0	-4	-4	5	-3	2	2	-3	-5	-3	-3	-4	-3	-1	4	-3	-1	-1	-3
K	-1	0	-3	-1	1	-4	-2	0	-3	6	-3	-2	0	-5	-1	2	3	0	-1	-3	-3	-1	-2	1
L	-2	-4	-2	-4	-3	1	-4	-3	2	-3	5	3	-4	-5	-4	-2	-3	-3	-1	1	-2	-1	-1	-3
M	-1	-3	-2	-4	-2	0	-3	-1	2	-2	3	7	-2	-5	-3	0	-2	-2	-1	1	-1	-1	0	-1
N	-1	4	-2	2	0	-4	0	1	-3	0	-4	-2	7	-5	-2	0	-1	1	0	-3	-4	-1	-2	0
*	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5
P	-1	-2	-4	-1	-1	-4	-2	-2	-3	-1	-4	-3	-2	-5	10	-1	-3	-1	-1	-3	-4	-2	-3	-1
Q	-1	0	-3	0	2	-4	-2	1	-3	2	-2	0	0	-5	-1	7	1	0	-1	-3	-1	-1	-1	4
R	-2	-1	-4	-2	0	-3	-3	0	-4	3	-3	-2	-1	-5	-3	1	7	-1	-1	-3	-3	-1	-1	0
S	1	0	-1	0	-1	-3	0	-1	-3	0	-3	-2	1	-5	-1	0	-1	5	2	-2	-4	-1	-2	0
T	0	0	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-5	-1	-1	-1	2	5	0	-3	0	-2	-1
V	0	-4	-1	-4	-3	-1	-4	-4	4	-3	1	1	-3	-5	-3	-3	-3	-2	0	5	-3	-1	-1	-3
W	-3	-5	-5	-5	-3	1	-3	-3	-3	-2	-1	-4	-5	-4	-1	-3	-4	-3	-3	15	-3	2	-2	-2
X	-1	-1	-2	-1	-1	-2	-2	-1	-1	-1	-1	-1	-1	-5	-2	-1	-1	-1	0	-1	-3	-1	-1	-1
Y	-2	-3	-3	-3	-2	4	-3	2	-1	-2	-1	0	-2	-5	-3	-1	-1	-2	-2	-1	2	-1	8	-2
Z	-1	2	-3	1	5	-4	-2	0	-3	1	-3	-1	0	-5	-1	4	0	0	-1	-3	-2	-1	-2	5