

Making the most of a sequence search – sequence and keyword searching in USGENE

With the number of sequence patent documents and published sequence listings constantly on the rise it becomes more and more important not only to rely on mere sequence searching but to also include keyword search strategies comprising reasonable data content (depending on the used data set). Simple sequence searching more frequently results in huge answer sets, especially if the focus lies on sequences of high commercial interest and value and may evtl. be surrounded by already heavy patenting activity. Focusing the primary sequence search results according to contextual specifications may in many cases enhance sequence search strategies and decisively improve precision of results.

STN offers multiple sequence databases that not only include the sequence itself but also additional bibliographic, patent and sequence related, numeric as well as text data. Such a database design where the sequence as well as related bibliographic and text data reside in one database record offers the possibility for more organized and specialized sequence searching. USGENE, the USPTO Genetic Sequence Database on STN, is a sequence database that comprises the following relevant sequence and patent publication related information:

- Bibliographic information related to the patent assignee and inventors of the patent publication (PA, IN)
- Patent publication, application and priority application numbers and dates (PN, PD, PY, AP, AY etc.)
- The organism name field indicating the source organism (ORGN)
- Original publication title (TI), author abstract (AB) and first/exemplary claim (ECLM) of the patent publication allowing text based searches
- A description field (DESC) with a concise description of the given sequence

- The sequence source field (SSO) indicating if a published sequence is from a published application or a granted/issued patent publication. This field also includes information on the data source (e.g. USPTO PSIPS server, USPTO full-text, NCBI Genbank etc.)
- The sequence count (SEQC) field. This field gives the overall number of sequences for any retrieved publication and thus, for instance, allows identifying sequences from mega patents.

In the USGENE database the basic index (BI), which is used for keyword term searches if no database field is explicitly specified, comprises the following fields: The original publication title (TI), the author abstract (AB), organism species (ORGN), molecule type (MTY) and the description field (DESC). In the USGENE basic index keyword terms may be searched using simultaneous left and right truncation (SLART).

The following search strategy begins with a BLAST online sequence search for homo sapiens tumor necrosis factor receptor superfamily, member 11b (TNFRSF11b) mRNA CDS region (NCBI: NM_002546). The resulting primary data set is focused on sequences from publications filed (applied for) after 1998. Further refinement uses the organism species field (ORGN) to concentrate on human sequences. Text based keyword term search is applied in an additional step in the basic index (BI) and the exemplary claims (ECLM) fields in order to specify a relation to tumor necrosis factor receptor and further topical terms. Using the sequence count field (SEQC) sequence documents from mega patents may be identified (and/or excluded). With the sequence source (SSO) field the resulting answer set may be limited to sequences from granted/issued patent publications. Final sorting according to

similarity (SCORE) and according to patent family (FSORT) groups together all sequences belonging to one publication in quality order. You may organize the display of multiple patent family members with the DISPLAY PFAM (display patent family) command which allows flexible display options for any family member of each separate family.

The different steps in the strategy below lead to an ever finer and more focused answer set:

- Online BLAST search for homo sapiens tumor necrosis factor receptor superfamily, member 11b with de-selected low complexity filter (/SQN -F f): 1606
- Refining the primary search result to sequences from publications filed after 1998: 1385

- Focus on different expressions for specific organism types (including unspecific or "not provided" types): 1348
- Text based keyword term search in the basic index and the exemplary claims for tumor necrosis factor receptor specifics: 619
- Check for publications from mega patents: 0
- Refinement to sequences from granted/issued patent publications: 226
- Sorting according to SCORE and AN: 226
- Sorting according to patent family (FSORT): 11 inventions comprising 10 multi-member families with 225 sequence records plus 1 single member family

Search Example: USGENE BLAST® search for homo sapiens tumor necrosis factor receptor superfamily, member 11b (TNFRSF11b) mRNA CDS region (NCBI: NM_002546) (Commands are given in blue, hit term highlighting is in red)

=> **FIL USGENE**

=> **UPL R BLAST**

Uploading S:\Eigene Dateien\Biosequences\USGENE\Doku\NM_002546_CDS.txt

UPLOAD SUCCESSFULLY COMPLETED

L1 GENERATED

Use the „Upload Query Wizard“ from the Discover! button in STN Express (version 8.2 or higher) to upload your query sequence. These upload commands will then run automatically.

=> **D L1 LQUE**

```
L1 ANSWER 1 USGENE COPYRIGHT 2011 SEQUENCEBASE CORP
LQUE ATGAACAACCTTGCTGTGCTGCGCGCTCGTGTTCCTGGACATCTCCATTAAGT
CTCCAAAGTACCTTCATTATGACGAAGAAACCTCTCATCAGCTGTTGTGTGTA
CCTAAACAACACTGTACAGCAAAGTGGAAAGCCGTGTGCGCCCCCTGCCCCGACCACTACACAGACAGC
TGCCACACCACTGACGAGTGTCTATACTGCAGCCCCGTGTGCAAGGAGCTGCAGTACGTCAGCAGGAGTGCA
ATCGCACCCACAACCCGCTGTGCGAATGCAAGGAAGGGCGCTACCTTGAGATAGAGTTCTGCTGAAACATAG
GAGCTGCCCTCCTGGATTTGGAGTGGTGCAAGCTGGAACCCAGAGCGAAATACAGTTTGCAAAAAGATGTCCA
GATGGGTTCTTCTCAAATGAGACGTCATCTAAAGCACCCCTGTAGAAAACACACAAAATGCAGTGTCTTTGGTC
TCCTGCTAACTCAGAAAGGAAATGCAACACACGACACACATATGTTCCGGAAACAGTGAATCAACTCAAAAATG
TGGAATAGATGTTACCCTGTGTGAGGAGGCATCTTCAGGTTGTGCTTCCACAAAGTTTACGCCTAACTGG
CTTAGTGTCTTGGTAGACAATTTGCCTGGCACCAGTAAACGCAGAGAGTGTAGAGAGGATAAAACGGCAAC
ACAGCTCACAAGAACAGACTTCCAGCTGCTGAAGTTATGGAAACATCAAAACAAAAGACCAAGATATAGTCAA
GAAGATCATCCAAGATATTGACCTCTGTGAAAACAGCGTGCAGCGGCACATGGACATGCTAACCTCACCTTC
GAGCAGCTTCGTAGCTTGATGGAAAGCTTACCGGAAAGAAAGTGGGAGCAGAAGACATTGAAAAACAATAA
AGGCATGCAAACCCAGTGACCAGATCCTGAAGCTGCTCAGTTGTGGCGAATAAAAAATGGCGACCAAGACAC
CTTGAAGGGCTAATGCACGCACTAAAGCACTCAAAGCAGTACCACCTTCCAAAACCTGCTCACTCAGAGCTA
AAGAAGACCATCAGTTCTTACAGCTTCAAAATGTATCAGAAATGTATCAGAAAGTTATTTTAGAAATGATAG
GTAACCAGGTTCAAATCAGTAAAAATAAGCTGCTTATAA
```

=> **RUN BLAST L1/SQN -F f**

BLAST Version 2.2

The BLAST software is used herein with permission of the National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM). See also, Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res. 25:3389-3402

This human tumor necrosis factor receptor superfamily, member 11b is searched with the BLAST® algorithm in a nucleotide search (/SQN) with de-selected low complexity filter (-F f).

```

Database USGENE US1_NA
Posted date: Sep 16, 2011 6:57 PM
Lambda K H
1.37 0.711 1.31
Gapped
Lambda K H
1.37 0.711 1.31
Matrix: blastn matrix:1 -3
Gap Penalties: Existence: 5, Extension: 2
Number of Hits to DB: 11,298,502
Number of Sequences: 22143792
Number of extensions: 11298502
Number of successful extensions: 807922
Number of sequences better than 10.0: 1608
Number of HSP's better than 10.0 without gapping: 1608
Number of HSP's successfully gapped in prelim test: 0
Number of HSP's that attempted gapping in prelim test: 805080
Number of HSP's gapped (non-prelim): 2668
length of query: 2414
length of database: 17,458,453,314
effective HSP length: 23
effective length of query: 1183
effective length of database: 16,949,146,098
effective search space: 20050839833934
effective search space used: 20050839833934
T: 0
A: 0
X1: 11 (21.8 bits)
X2: 15 (29.7 bits)
S1: 12 (24.3 bits)
S2: 21 (42.1 bits)
Total Execution Time: 20.9967

```

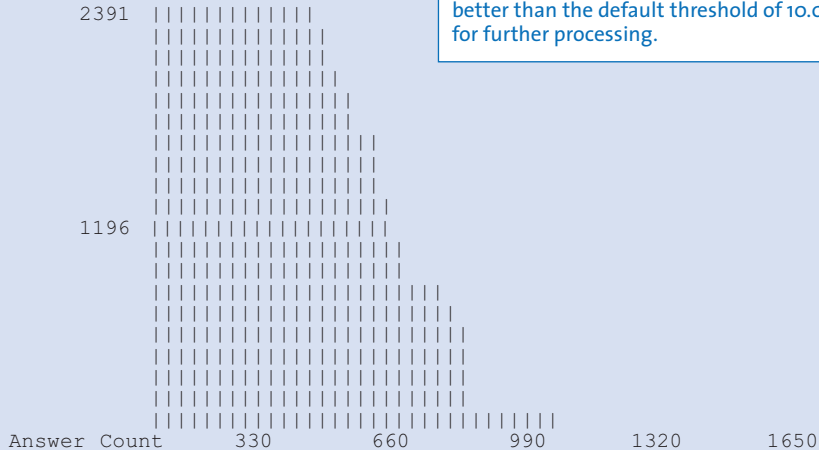
1606 ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0

```

QUERY SELF SCORE VALUE IS 2391
BEST ANSWER SCORE VALUE IS 2391

```

Similarity
Score



1606 answers are found in the nucleotide BLAST® search to be better than the default threshold of 10.0. ALL answers are kept for further processing.

```

ENTER EITHER THE NUMBER OF ANSWERS YOU WISH TO KEEP
OR ENTER MINIMUM PERCENT OF SELF SCORE FOLLOWED BY %
(BEST ANSWER PERCENTAGE OF SELF SCORE IS 100%)

```

ENTER (ALL) OR ? :**ALL**

L2 RUN STATEMENT CREATED

```

L2 1606 ATGAACAACCTTGCTGTGCTGCGCGCTCGTGTTCCTGGACATCTCCATTA
GTGGACCACCCAGGAAACGTTTCTCCAAAGTACCTTCATTATGACGAAG
AAACCTCTCATCAGCTGTTGTGTGACAAATGTCTCCTGGTACCTACCTA
AAACAACACTGTACAGCAAAGTGAAGACCGTGTGCGCCCTTGCCCTGA
CCACTACTACACAGACAGCTGGCACACCAGTGACGAGTGTCTATACTGCA
GCCCCGTGTGCAAGGAGCTGCAGTACGTCAAGCAGGAGTGCAATCGCACC
CACAACCGCGTGTGCGAATGCAAGGAAGGGCGCTACCTTGAGATAGAGTT
CTGCTTGAAACATAGGAGCTGCCCTCCTGGATTGGAGTGGTCAAGCTG
GAACCCAGAGCGAAATACAGTTTGCAAAAGATGTCCAGATGGGTTCTTC
TCAAATGAGACGTCATCTAAAGCACCTGTAGAAAAACACAAAATTGCAG
TGTCTTGGTCTCCTGCTAACTCAGAAAGGAAATGCAACACACGACAACA
TATGTTCCGGAAACAGTGAATCAACTCAAAAATGTGGAATAGATGTTACC
CTGTGTGAGGAGGCATTCTTCAGGTTTGTGTTCCTACAAAGTTTACGCC
TAACTGGCTTAGTGTCTTGGTAGACAATTTGCCTGGCACCAGTAAACG
CAGAGAGTGTAGAGAGGATAAAACGGCAACACAGCTCACAAAGCAGACT

```

```

TTCCAGCTGCTGAAGTTATGGAACATCAAACAAAGACCAAGATATAGT
CAAGAAGATCATCCAAGATATTGACCTCTGTGAAAACAGCGTGCAGCGGC
ACATTGGACATGCTAACCTCACCTTCGAGCAGCTTCGTAGCTTGATGGAA
AGCTTACCGGAAAGAAAGTGGGAGCAGAAGACATTGAAAAACAATAAA
GGCATGCAAACCCAGTGACCAGATCCTGAAGCTGCTCAGTTTGTGGCGAA
TAAAAATGGCGACCAAGACACCTTGAAGGGCCTAATGCACGCTAAAG
CACTCAAAGACGTACCACTTTCCAAAACCTGTCACTCAGAGTCTAAAGAA
GACCATCAGGTTCCCTTACAGCTTCACAATGTACAAAATGTATCAGAAGT
TATTTTAGAAAATGATAGGTAACCAGGTCCAATCAGTAAAAATAAGCTGC
TTATAA/SQN.-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 AY>1998

```

29327536 AY>1998
              (AY>1998)
L3          1385 L2 AND AY>1998

```

Refine the search results to sequences from publications filed (applied for) after 1998 and to human or unknown organism sequences.

=> S L3 AND (HOMO OR HUMAN OR PROBE OR PRIMER OR UNKNOWN OR UNIDENTIFIED OR ARTIFICIAL OR "NOT PROVIDED")/ORGN

```

L4          1348 L3 AND (HOMO OR HUMAN OR PROBE OR PRIMER OR UNKNOWN OR UNIDENTIF
              IED OR ARTIFICIAL OR "NOT PROVIDED")/ORGN

```

=> D L4 TRI ORGN 100 300

```

L4          ANSWER 100 OF 1348 USGENE COPYRI
TI          OSTEOPROTEGERIN (PublishedApplicat
DESC       Artificial DNA; Synthetic oligonud
MTY        DNA
SQL        54
ORGN       Artificial sequence

```

Note: Although the majority of human sequences are indexed as "Homo sapiens", it is important to take into account the minority indexed as "human", and those which do not indicate a source organism at all.

```

L4          ANSWER 300 OF 1348 USGENE COPYRIGHT 2011 SEQUENCEBASE CORP on STN
TI          Osteoprotegerin in milk (Patent)
DESC       Homo sapiens DNA; sequence 7 of 8
MTY        DNA
SQL        1182
ORGN       Homo sapiens

```

Use SET SFIELDS BI ECLM to add ECLM to the default USGENE Basic Index for text searching.

=> SET SFIELDS BI ECLM

SET COMMAND COMPLETED

Automatically add plurals and English-language spelling variations using two SET commands.

=> SET PLURALS ON

SET COMMAND COMPLETED

If you want to use these settings permanently add P&R to the commands to set them for future sessions.

=> SET SPELLINGS ON

SET COMMAND COMPLETED

=> S L4 AND ((TNF## OR TUMOR NECROSIS FACTOR?) (2A) (RECEPTOR? OR BINDING PROTEIN?) OR (OSTEOCLASTOGENESIS INHIBITORY FACTOR OR OCIF OR OPG OR OSTEOPROTEGERIN))

```

L5          619 L4 AND ((TNF##/BI,ECLM OR TUMOR NECROSIS FACTOR?/BI,ECLM) (2A) (RE
              CEPTOR?/BI,ECLM OR BINDING PROTEIN?/BI,ECLM) OR (OSTEOCLASTOGENE
              SIS INHIBITORY FACTOR/BI
              OSTEOPROTEGERIN/BI,ECLM))

```

Determine if any sequences in the results are from "mega" publications with the sequence count field (/SEQC).

=> S L5 AND SEQC>10000

```

16427827 SEQC>10000
L6          0 L5 AND SEQC>10000

```

In this search example, no documents with more than 10,000 sequences were retrieved.

=> S L5 AND GRANTED/SSO

```

7555947 GRANTED/SSO
L7          226 L5 AND GRANTED/SSO

```

The sequence source (/SSO) field indexing provides a simple way to narrow the USGENE search to sequences from issued (granted) patents.

=> SOR SCORE D AN D

```

PROCESSING COMPLETED FOR L7
L8          226 SOR L7 SCORE D AN D

```

Tip: Using the secondary sort AN D puts answers with the same BLAST score into descending accession number (i.e. publication) order, with recent patents before older patents. This can be useful in any subsequent FSORT.

=> FSORT L8

SET SMARTSELECT ON
SET COMMAND COMPLETED

SET HIGHLIGHTING OFF
SET COMMAND COMPLETED

SET AUDIT OFF
SET COMMAND COMPLETED

FSORT can be used to group together sequence hits from the same publication, application and/or patent family for a more efficient review of the results retrieved.

```
SEL L8 1- PN,APPS
L9      SEL L8 1- PN APPS :      64 TERMS
```

```
'L9' DELETED
L9      226 FSO L8
```

```
      10 Multi-record Families   Answers 1-225
          Family 1               Answers 1-96
          Family 2               Answers 97-144
          Family 3               Answers 145-147
          Family 4               Answers 148-152
          Family 5               Answers 153-158
          Family 6               Answers 159-183
          Family 7               Answers 184-210
          Family 8               Answers 211-212
          Family 9               Answers 213-223
          Family 10              Answers 224-225
          1 Individual Record    Answer 226
          0 Non-patent Records
```

10 multi-record families comprise 225 of the hit sequences and 1 individual sequence record forms an additional single member patent family. The overall 226 hit sequences are thus pooled in 11 patent inventions.

```
SET SMARTSELECT OFF
SET COMMAND COMPLETED
```

```
SET HIGHLIGHTING ON
SET COMMAND COMPLETED
```

```
SET AUDIT ON
SET COMMAND COMPLETED
```

Review answers in a free-of-charge format using D PFAM and check for hit terms and sequence hit quality (e.g. with similarity score and BLAST percent identity).

```
=> D PFAM=1-4 1 TRI ORGN IDENT SCORE
```

```
L9      ANSWER 1 OF 226  USGENE COPYRIGHT 2011 SEQUENCEBASE CORP on STN FAMILY1
TI      Nucleic acid molecules encoding osteoclastogenesis inhibitory
        factor proteins (Patent)
DESC   Homo sapiens DNA; sequence 6 of 108
MTY    DNA
SQL    1206
ORGN   Homo sapiens
IDENT  100%
SCORE  2391          100% of query self score 2391
```

```
L9      ANSWER 97 OF 226  USGENE COPYRIGHT 2011 SEQUENCEBASE CORP on STN FAMILY2
TI      Monoclonal antibodies that bind OCIF (Patent)
MTY    DNA
SQL    1206
ORGN   Unknown
IDENT  100%
SCORE  2391          100% of query self score 2391
```

```
L9      ANSWER 145 OF 226  USGENE COPYRIGHT 2011 SEQUENCEBASE CORP on STN FAMILY3
TI      Osteoprotegerin variant proteins (Patent)
DESC   Homo sapiens DNA; sequence 1 of 49
MTY    DNA
SQL    2291
ORGN   Homo sapiens
IDENT  99%
SCORE  2383          99% of query self score 2391
```

```
L9      ANSWER 148 OF 226  USGENE COPYRIGHT 2011 SEQUENCEBASE CORP on STN FAMILY4
TI      Antibodies to human tumor necrosis factor receptor-like genes (Patent)
DESC   Homo sapiens DNA; sequence 1 of 11
MTY    DNA
SQL    1527
ORGN   Homo sapiens
IDENT  99%
SCORE  2383          99% of query self score 2391
```

For in-depth review display selected answers in a preferred bibliographic format. Hit terms show the relevancy of answers in view of the textual specifications whereas the alignment provides information concerning the similarity to the query sequence and thus the hit sequence quality.

```
=> D PFAM=4,6 1 BIB AB ECLM SCORE ALIGN
```

```
L9      ANSWER 148 OF 226  USGENE COPYRIGHT 2011 SEQUENCEBASE CORP on STN FAMILY4
AN      7078493.1  DNA      USGENE Full-text
TI      Antibodies to human tumor necrosis factor receptor-like genes (Patent)
IN      Greene John M. (Gaithersburg, MD); Fleischmann Robert D. (Gaithersburg,
        MD); Ni Jian (Rockville, MD)
PA      Human Genome Sciences Inc (Rockville MD)
PI      US 7078493          B1 20060718
AI      US 2000-526437      20000315
PRAI   US 1999-136248P    19990526
        US 1999-124489P    19990315
```

```

US 1996-718737          19960918
US 1995-469637          19950606
WO 1995-US3216          19950315
XPD 20150315 (calculated)
PSL SEQ ID NO 1
DESC Homo sapiens DNA; sequence 1 of 11
DT Patent
AB The present inventors have discovered novel receptors in the Tumor Necrosis
Factor (TNF) receptor family. In particular, receptors having homology to the
type 2 TNF receptor (TNF-RII) are provided. Isolated nucleic acid molecules
are also provided encoding the novel receptors of the present invention.
Receptor polypeptides are further provided as are vectors, host cells and
recombinant methods for producing the same.
ECLM US7078493 B1: 1. An isolated antibody, which specifically binds a protein
selected from the group consisting of: (a) a protein whose sequence
consists of the amino acid sequence of SEQ ID NO:2; (b) a protein whose
sequence consists of amino acids 1 to 380 of SEQ ID NO:2.
SCORE 2383          99% of query self score 2391
BLASTALIGN
Query = 1206 letters
Length = 1527
Score = 2383 bits (1202), Expect = 0.0
Identities = 1205/1206 (99%)
Strand = Plus / Plus

Query: 1      atgaacaacttgctgtgctgctgcgctcgctggttcttgacatctccattaagtggaccacc
            |||
Sbjct: 46     atgaacaagtgtgctgtgctgctgcgctcgctggttcttgacatctccattaagtggaccacc

Query: 61     caggaaacgtttcctccaaagtaccttcattatgacgaagaacacctcatcagctgttg
            |||
Sbjct: 106    caggaaacgtttcctccaaagtaccttcattatgacgaagaacacctcatcagctgttg

.
.
.
Query: 1081   gtcactcagagtctaagaagaccatcaggttccttcacagcttcacaatgtacaaattg
            |||
Sbjct: 1126   gtcactcagagtctaagaagaccatcaggttccttcacagcttcacaatgtacaaattg

Query: 1141   tatcagaagtatttttagaaatgataggtaccaggtccaatcagtaaaaaaagctgc
            |||
Sbjct: 1186   tatcagaagtatttttagaaatgataggtaccaggtccaatcagtaaaaaaagctgc

Query: 1201   ttataa 1206
            |||
Sbjct: 1246   ttataa 1251

L9 ANSWER 159 OF 226 USGENE COPYRIGHT 2011 SEQUENCEBASE CORP on STN FAMILY6
AN 7632922.124 DNA USGENE Full-text
TI Osteoprotegerin (Patent)
IN Boyle William J. (Moorpark, CA); Lacey David L. (Thousand Oaks, CA);
Calzone Frank J. (Westlake Village, CA); Chang Ming-Shi (Newbury Park,
CA)
PA Amgen Inc (Thousand Oaks CA)
PI US 7632922 B1 20091215
AI US 2000-718725 20001122
PRAI US 2000-718725 20001122
US 1998-132985 19980812
US 1996-771777 19961220
US 1996-706945 19960903
US 1995-577788 19951222
XPD 20151222 (calculated)
NTE Subject to any Disclaimer, the term of this patent is extended or
adjusted under 35 USC 154(b) by 1275 days.
PSL SEQ ID NO 124
DESC Homo sapiens DNA; sequence 124 of 179
DT Patent
AB The present invention discloses a novel secreted polypeptide, termed
osteoprotegerin, which is a member of the tumor necrosis factor receptor
superfamily and is involved in the regulation of bone metabolism. Also
disclosed are nucleic acids encoding osteoprotegerin, polypeptides,
recombinant vectors and host cells for expression, antibodies which bind OPG,
and pharmaceutical compositions. The polypeptides are used to treat bone
diseases characterized by increased resorption such as osteoporosis.
ECLM US7632922 B1: 1. An isolated antibody or fragment thereof which
specifically binds to an OPG polypeptide consisting of residues 22-401
inclusive of SEQ ID NO:121.
SCORE 2375          99% of query self score 2391

```

Note: The ALIGN displays in this example have been truncated for brevity.

