

STN[®]

To PSIPS and Beyond

Exploring the Content and Utility of
USGENE[®]

Robert Austin – FIZ Karlsruhe

Agenda

- STN[®] sequence databases
- USGENE[®] database content
- The 7 basic steps of USGENE BLAST[®]
- Comparisons and conclusions

BLAST is a registered trademark of the U.S. National Library of Medicine (NLM)

STN

2

STN sequence searchable databases

- CAS REGISTRYSM
 - Chemical Abstracts Service (CAS) Registry File
- DGENE
 - Thomson Scientific GENESEQTM
- PCTGEN
 - WIPO/PCT Patent Application Biosequences
- USGENE
 - The USPTO Genetic Sequence Database

See *Effective patent sequence searching on STN*:

http://www.stn-international.com/training_center/bioseq/epss.pdf

STN

3

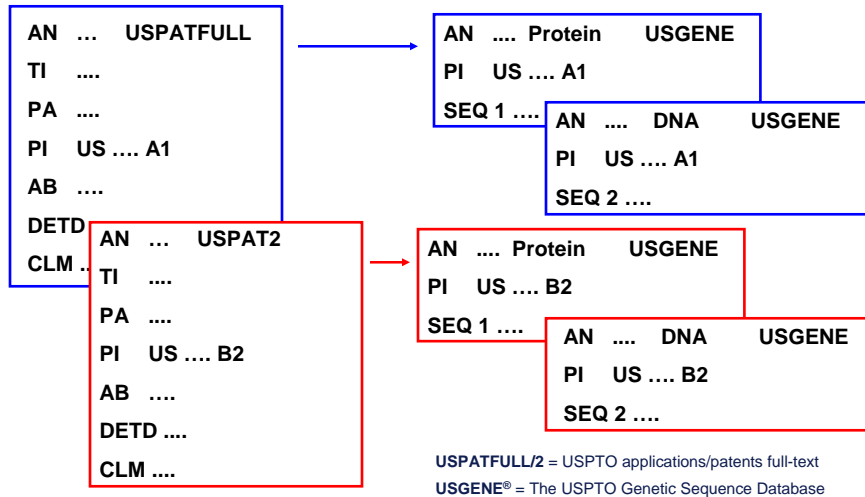
USGENE is the USPTO Genetic Sequence Database

- Sequences from all relevant USPTO published patent applications and issued (granted) patents
- Assignee and full inventor names; publication, application, and parent case PCT numbers and dates; original publication **title**, **abstract**, and **claims**
- Organism name, sequence length, Molecule Type, SEQ ID, and feature tables for features/annotations
- Produced by the SequenceBase Corporation
- Updated weekly – within **7 days** of publication
- 1982 – present

STN

4

Relationship between USPATFULL/2 and USGENE databases



STN

5

USGENE consolidates unique USPTO sequence data from different sources

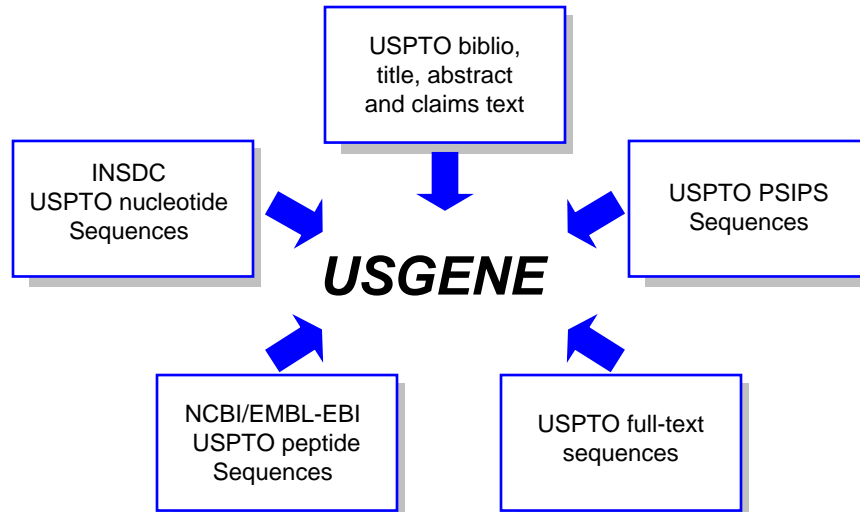
- USPTO Publication Site for Issued and Published Sequences (PSIPS)
- International Nucleotide Sequence Database Collaboration (INSDC) (NCBI/EMBL/DDBJ)
- USPTO Protein Database (NCBI/EMBL)
- USPTO Patents and Applications Full-Text

The USGENE Sequence Source (/SSO) field indicates which source any given USGENE sequence record was derived from.

STN

6

USGENE combines these sequences with bibliographic data and claims text



STN

7

USGENE records include full patent bibliography, title, and abstract

```

L1 ANSWER 1 OF 1 USGENE COPYRIGHT 2007 SEQUENCEBASE CORP on STN
AN 6881821.58 (1) Protein (2) USGENE
TI Hepatitis-C virus type 4, 5, and 6 (Patent) (3) ALL display format.
IN Simmonds Peter (Edinburgh, GB) (4)
  Yap Peng Lee (Edinburgh, GB)
  Pike Ian Hugo (Bromley, GB)
PA Common Services Agency (Edinburgh GB) (5)
  Murex Diagnostics International Inc (Bridgetown BB)
PI US 6881821      B2  20050419
  US 2005032047   A1  20050210
  WO 9425602      A   19941110
AI  US 1995-537802  19951221
RLI WO 1994-GB957  19940505
ED  20070328
DT  Patent
AB  Newly elucidated sequences of hepatitis C virus type 4 and type 5 are
    (7) described, together with those of a newly discovered type 6. Unique
    type-specific sequences in the NS4, NS5 and core regions enable HCV
    detection and genotyping into types 1 to 6. Antigenic peptides and
    immunoassays are described.
  
```

STN

8

USGENE records also include patent or published application claims text

CLM US6881821 B2: What is claimed is: ALL display format (cont.).

(8) 1. An isolated peptide having an antigenic sequence selected from the following: a) QPAVIPDREVLYQQFDEN (SEQ ID NO:32); and, b) ECSKHLPLVEHGLQLAEQF (SEQ ID NO:46).

2. A peptide according to claim 1 which is bound to a multiple antigen peptide core.

3. A peptide according to claim 2 having a sequence selected from the following: a) [H.sub.2 N-QPAVIPDREVLYQQFDEN].sub.8 K.sub.4 K.sub.2 K-COOH (SEQ ID NO:32); and, b) [H.sub.2 N-ECSKHLPLVEHGLQLAEQF].sub.8 K.sub.4 K.sub.2 K-COOH (SEQ ID NO:46); where K.sub.4 K.sub.2 K is the multiple antigen peptide core.

4. A peptide according to claim 1 which is fused to another peptide to form a fusion peptide.

5. A peptide according to claim 4 fused to another peptide

STN

9

All USGENE sequences are provided in STN standardized format

SSO PROTEIN; USPTO; GRANTED (9) ALL display format (cont.).

ORGN Hepatitis C Virus (10)

SQL 19 (11)

SEQ

1 ecaskaalie egqrmaeml (12)

FEATURE TABLE: (13)

Key	Location
VARIANT	(0)...(0) HCV TYPE 2 NS4 REGION

See (8) - (13) on slide 12.

STN

10

USGENE sample record annotations

- 1) USGENE Accession Number (AN), including the sequence identity number (SEQ ID NO)
- 2) Molecule Type (MTY)
- 3) Original publication title – a “PublishedApplication” or “Patent” indication is given in parentheses
- 4) Full inventor names, city and state/country
- 5) Patent assignee name, city and state/country
- 6) Publication, application and related PCT parent case application details and dates
- 7) Original patent or published application abstract

STN

11

USGENE sample record annotations

- 8) Published application or granted patent claims
- 9) The Sequence Source (SSO) – nucleic or protein; PSIPS/USPTO, NCBI, etc; granted or application
- 10) Organism (where given) – providing the name of the organism from which the sequence is derived
- 11) Searchable and sortable Sequence Length (SQL)
- 12) Standardized patent sequence (SEQ) – each USGENE record is based upon a sequence
- 13) Feature table including sequence modifications, features and/or annotations, as provided by the patent applicant or assignee

STN

12

USGENE Original Sequence (SEQO)

=> D SEQO

L1 ANSWER 1 OF 1 USGENE COPYRIGHT
SEQO

```
cgctcgcagt ctgtgggccc tccgggagcc ggcggagcc acccgggga gggggcggg 60  
cgcagc atg gca gcc tcc tta cgg ctc ctc gga gct gcc tcc ggt ctc 108  
Met Ala Ala Ser Leu Arg Leu Leu Gly Ala Ala Ser Gly Leu  
1 5 10  
cgg tac tgg agc cgg cgg ctg cgg ccg gca gcc ggc agc ttt gca gcg 156  
Arg Tyr Trp Ser Arg Arg Leu Arg Pro Ala Ala Gly Ser Phe Ala Ala  
15 20 25 30  
gtg tgt tct agg tca gtg gct tca aag act cca gtt gga ttc att gga 204  
Val Cys Ser Arg Ser Val Ala Ser Lys Thr Pro Val Gly Phe Ile Gly  
35 40 45  
ctg gcc aac atg ggg aat cca atg gc  
Leu Gly Asn Met Gly Asn Pro Met Al  
50 55
```

The original input format of a USGENE sequence is available for display using the **SEQO** display field.

Often the original format includes the patent applicant's alignment of the nucleotide sequence coding region with the corresponding protein sequence.

STN

13

USGENE represents a new tool for tackling business critical searches

- DGENE and REGISTRY sequences are indexed by Thomson from the DWPISM basic and by CAS from the CPlusSM basic respectively:
 - 65% of basic patents are PCT published applications
- Sequence listing variation often occurs between published application and granted patent stage:
 - Especially important, e.g. for freedom-to-operate
- USGENE provides sequences from both USPTO **published applications** and **granted patents**

STN

14

Example: sequence listing variation between patent family members

```

LL ANSWER 1 OF 1 WPINDEX COPYRIGHT 2007 THE THOMSON CORP on STN
AN 1994-358278 [44] WPINDEX
TI New polynucleotide(s) specific for hepatitis C virus types 4, 5 and 6 -
and related antigenic peptide(s) and antibodies, useful in vaccines,
diagnosis, HCV typing and treatment
DC B04; D16; S03
IN PIKE I H; SIMMONDS P; YAP P L
PA (COMM-N) COMMON SERVICES AGENCY; (MURE-N) MUREX DIAGNOSTICS INT INC; . . .
PI WO 9425602 A1 19941110 (199444)* EN 70[5]
AU 9465797 A 19941121 (199508) EN
FI 9505224 A 19951221
EP 698101 A1 19960224
JP 09500009 W 19970107
AU 695259 B 19980811
EP 698101 B1 20041110
DE 69434116 E 20041204
US 20050032047 A1 20050210 (200512) EN
US 6881821 B2 20050419 (200527) EN
. . . . .
ADT WO 9425602 A1 WO 1994-GB957 19940505 . . . .
PRAI GB 1994-263 19940107
GB 1993-9237 19930505

```

In this example the patent family has:

- 9 sequences from WO 9425602 in DGENE
- 58 sequences from US 6881821 in USGENE

STN

15

Agenda

- STN sequence databases
- USGENE database content
- The 7 basic steps of USGENE BLAST®
- Comparisons and conclusions

STN

16

USGENE offers the same sequence search options as DGENE

- NCBI BLAST similarity
 - RUN BLAST
- FASTA similarity
 - RUN GETSIM
- Sequence Code Match (SCM)
 - RUN GETSEQ
- Offline BATCH and ALERT options

The *DGENE Workshop Manual* is the complete guide:
http://www.stn-international.com/training_center/bioseq/dgene_wm.pdf



17

The 7 steps of USGENE BLAST

- 1) SAVE, UPLOAD and VERIFY a query text file (L1)
- 2) RUN the BLAST search (/SQP or /SQN)
- 3) Decide how many answers to keep (L2)
- 4) SORT SCORE in Descending order (L3)
- 5) Review answers in a free-of-charge format
e.g. D L3 TRI ORGN ALIGN 1-
- 6) Display selected answers in bibliographic format, e.g. D L3 BIB AB CLM ALIGN 1,3,10
- 7) Ensure session transcript was captured and Logoff



18

1) SAVE, UPLOAD and VERIFY

The screenshot shows three overlapping windows from the STN software:

- Select Discover! Wizard (1):** The 'Upload Query' option is highlighted in the 'Choose a search wizard' section.
- STN Upload Query Wizard (2):** The 'File Name' field is set to 'C:\Documents and Settings\Robert Austin\...', and a 'Browse...' button is visible.
- STN Upload Query Wizard (3):** A list of databases is shown, with 'PCTGEN World Patent Application Biosequences' selected.

Three callout boxes provide instructions:

- (1) Click **Upload Query**.
- (2) Choose file of interest.
- (3) Select database.

A larger callout box states: "The sequence becomes a **Query L-number** in the database of choice for use with RUN BLAST."

STN logo is in the bottom left, and the number 19 is in the bottom right.

1) SAVE, UPLOAD and VERIFY (cont.)

```

=> FILE PCTGEN
=> UPL R BLAST
  
```

These commands are automatically run by the STN Express® Sequence Query Upload wizard.

```

UPLOAD SUCCESSFULLY COMPLETED
L1 GENERATED

=> D L1 LQUE

L1 ANSWER 1 PCTGEN COPYRIGHT 2007 WIPO on STN
LQUE vqtvplsrldhamleahrahelaidtyqefeetyipkdqkysflhdsqtsfcfsdsi
      ptpsnmeetqkksnlellrislllleswlepvrflrsmfannlvdytdsddyhllkd
      leeqigtlmgrledgsrrtgqilkqtyskfdtnshhdallknyglycfrkdmkve
      tflrmvqcrsvvegscgf

=>
  
```

The sequence query is now ready for searching directly in USGENE using the L-number (L1).

STN logo is in the bottom left, and the number 20 is in the bottom right.

The 7 basic steps of USGENE BLAST

2) RUN the BLAST search:

- Protein search: RUN BLAST L1 /SQP
- Nucleotide search: RUN BLAST L1 /SQN
- Translated search: RUN BLAST L1 /TSQN

STN

21

2) RUN the USGENE BLAST search

```
=> FILE USGENE
```

```
FILE 'USGENE' ENTERED AT 07:52:24 ON 04 MAY 2007  
COPYRIGHT (C) 2007 SEQUENCEBASE CORP
```

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

Turn the Low Complexity Filter off
with the syntax... /SQP -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
```

```
BLAST SEARCHING . . .
```

Disclaimer: this search was conducted in a pre-release
USGENE test-file and the results may not be complete.

STN

22

RUN BLAST command syntax

Similarity Searching with BLAST (protein/polypeptides)

=> **RUN BLAST L1** (sequence or L-number)

/SQP (protein) (default)

-e (Expect-value)
-f (Filter) (on by default)
-w (Word size)
-m (Matrix)
-g (Gap penalty)
-x (Gap extension)
 BATCH (offline)
 ALERT (Alert/SDI)

STN

23

RUN BLAST command syntax

Similarity Searching with BLAST (Nucleic acids)

=> **RUN BLAST L1** (sequence or L-number)

/SQN (nucleotide)

SIN (single strand)

COM (complementary strand)

BOTH (both strands) (default)

-e (Expect-value)
-f (Filter)
-w (Word size)
-g (Gap penalty)
-x (Gap extension)
-q (penalty for mismatch)
-r (reward for match)
 BATCH (offline)
 ALERT (Alert/SDI)

STN

24

RUN BLAST advanced options

Expectation Value (-E)

Expectation value (E-Value) is the statistical significance threshold for reporting matches against a sequence database. The E-value can be any positive number, and the default value is 10. This means that 10 matches may be expected to be found merely by chance. In general E-value is lowered to make the search more precise and raised to retrieve more answers.

Word Size (-W)

Word Size is the length of the character string fragments of a sequence query which are used as the basis for a BLAST search. For SQN the default is 11 and the range 7-23. For all other BLAST searches the default is 3 and the range 2-3. For short search queries, reducing the default word size can give improved search results.



25

RUN BLAST advanced options (cont.)

Low Complexity Filtering (on by default) (-F)

The low complexity filter can eliminate biologically uninteresting segments that have low compositional complexity and are statistically significant, as determined by specific programs for peptide or nucleotide sequences in nature. Filtering is applied to the query sequence and is indicated by a series of Xs for peptide sequences and Ns for nucleotide sequences. Low complexity filtering can be turned off (i.e. set to F - false).

Peptide similarity matrices (-M)

For peptide based searches SQP and TSQN the advanced options provide additional scoring matrices to the default BLOSUM62 (next slide)



26

Guidelines from NCBI on the use of Advanced Settings for peptide sequence searching are as follows:

<u>Query Length</u>	<u>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) (BLAST default)

STN

27

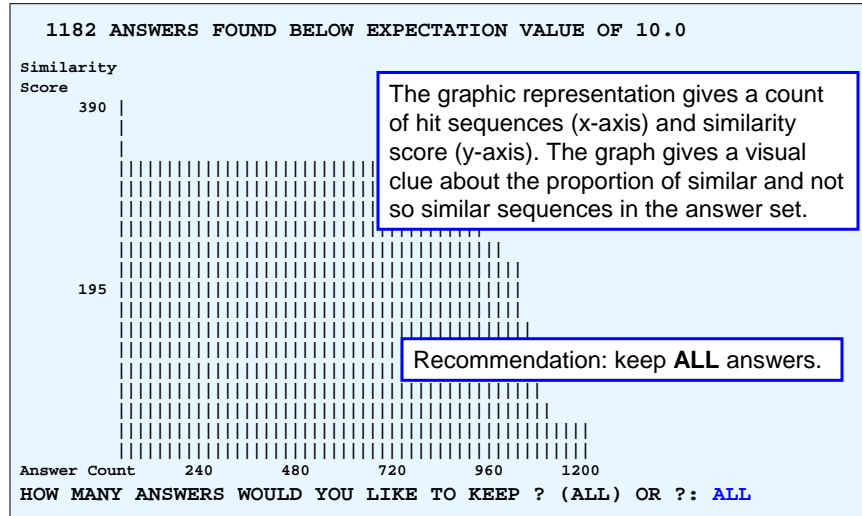
The 7 basic steps of USGENE BLAST

- 3) Decide how many answers to keep (L2):
- How many answers would you like to keep? (ALL) or ?:
 - Recommendation: Keep **ALL** answers

STN

28

3) Decide how many answers to keep



STN

29

7 basic steps of USGENE BLAST

4) SORT by SCORE descending (L3):

- SOR L2 SCORE D
- Option: limit using text terms and/or dates (L4)
- Remember to SORT L4 SCORE D !! (L5)

STN

30

4) SORT by SCORE descending

```
HOW MANY ANSWERS WOULD YOU LIKE TO KEEP ? (ALL) OR ? : ALL
L2      RUN STATEMENT CREATED
L2      1182 VQTVPLSRFLFDHAMLEAHRRAHELAIPTYQFEETYIPKDQKYSFLHDSQT
          SFCFSDSIPTPSNMEETQQKSNLELLRISLLLLIESWLEPVRFLRSMFANN
          LVYDTSDDYHLLKDLLEGIQTLMGRLEDGSRRTGQILKQTYSKFDTNS
          HNHDAKLLKNYGLLYCFRKDMDKVETFLRMVQCRSVEGSCGF/SQP.-F F
```

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

```
=> SOR SCORE D
PROCESSING COMPLETED FOR L2
L3      1182 SOR L2 SCORE D
```

Use SORT SCORE D to sort by descending BLAST score.

STN

31

The 7 basic steps of USGENE BLAST

5) Review answers using a *free-of-charge* format including alignment (ALIGN), while "parked" in the STNGUIDESM file:

- D L5 TRI ORGN ALIGN 1-
- FILE STNGUIDE

STN

32

5) Review answers with a free-of-charge format including alignment

=> D L3 TRI ORGN ALIGN 1-30; FILE STNGUIDE

```
L3 ANSWER 1 OF 1182 USGENE COPYRIGHT 2007 SEQUENCEBASE CORP on STN
TI Recombinant DNA transfer vectors (Patent)
MTY Protein
SQL 191
ORGN Unknown
BLASTALIGN
  Query = 191 letters
  Length = 191
  Score = 387 bits (995), Expect = e-113
  Identities = 189/191 (98%), Positives = 191/191 (99%)
  Query: 1 VQTVPLSRLFDHAMLQAHRAHQLAIDTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
           VQTVPLSRLFDHAML+AHRAH+LAIDTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
  Sbjct: 1 VQTVPLSRLFDHAMLQAHRAHQLAIDTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
  Query: 61 PSNMEETQQKSNLELLRISLLLIESWLEPVVFLRSMFANNLVYDTSDDYHLLKDLEEG
           PSNMEETQQKSNLELLRISLLLIESWLEPVVFLRSMFANNLVYDTSDDYHLLKDLEEG
  Sbjct: 61 PSNMEETQQKSNLELLRISLLLIESWLEPVVFLRSMFANNLVYDTSDDYHLLKDLEEG
  . . . .
```

This top hit comes from a U.S. issued patent.

STN

33

5) Review answers with a free-of-charge format including alignment

```
L3 ANSWER 3 OF 1182 USGENE COPYRIGHT 2007 SEQUENCEBASE CORP on STN
TI Genetic polymorphisms associated with myocardial infarction, methods
  of detection and uses thereof (PublishedApplication)
MTY Protein
SQL 217
ORGN Homo Sapiens
BLASTALIGN
  Query = 191 letters
  Length = 217
  Score = 387 bits (995), Expect = e-113
  Identities = 189/191 (98%), Positives = 191/191 (99%)
  Query: 1 VQTVPLSRLFDHAMLQAHRAHQLAIDTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
           VQTVPLSRLFDHAML+AHRAH+LAIDTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
  Sbjct: 1 VQTVPLSRLFDHAMLQAHRAHQLAIDTYQEFEEETYIPKDQKYSFLHDSQTSFCFSDSIPT
  Query: 61 PSNMEETQQKSNLELLRISLLLIESWLEPVVFLRSMFANNLVYDTSDDYHLLKDLEEG
           PSNMEETQQKSNLELLRISLLLIESWLEPVVFLRSMFANNLVYDTSDDYHLLKDLEEG
  Sbjct: 61 PSNMEETQQKSNLELLRISLLLIESWLEPVVFLRSMFANNLVYDTSDDYHLLKDLEEG
  Query: 121 IQTLMGRLEDGSRRTGQILKQTYSKFDTNSHNHDALLKNYGLLYCFRKMDKRVETFLRMV
           IQTLMGRLEDGSRRTGQILKQTYSKFDTNSHNHDALLKNYGLLYCFRKMDKRVETFLRMV
  Sbjct: 147 IQTLMGRLEDGSRRTGQILKQTYSKFDTNSHNHDALLKNYGLLYCFRKMDKRVETFLRMV
  . . . .
```

The third from top hit comes from a U.S. published application.

BLAST alignment details are explained on the next slide. . . .

STN

34

Understanding BLAST alignments

Query	the length of the query sequence
Length	the length of the answer sequence
Score	a relative score assigned by BLAST
Expect	Expectation Value – a value representing the chance that an answer is a random hit. The closer to zero, the less likely the hit is random
Identities	the number of exact letter matches between query and answer within the displayed local alignment. The amino acid letter is repeated* in the display
Positives	a combination of identities and amino acid family matches shown with + (plus) in the alignment
Gaps	shown as dashes - where BLAST must break the query or answer to maintain an alignment

(* For nucleic acid searches a vertical bar is used to indicate nucleotide identities in the alignment display.)



35

Option: refine USGENE BLAST results with text and/or date search terms

```
HOW MANY ANSWERS WOULD YOU LIKE TO KEEP ? (ALL) OR ? : ALL
L2  RUN STATEMENT CREATED
L2  1182 VQTVPLSRFLFDHAMLEAHRAHELAIPTYQEFEEITYIPKDQKYSFLHDSQT
    SFCFSDSIPTPSNMEETQQKSNLELLRISLLLIESWLEPVRFRLRSMFANN
    LVYDTSDDDYHLLKDLEEGIQTLMGRLEDGSRRTGQILKQYTSKFDTNS
    HNH DALLKNYGLLYCFRKDMR VETFLRMVQCRSVEGSCGF/SQP.-F F
```

Answer set arranged by accession number; to sort by descending similarity score, enter at an arrow

```
=> SOR SCORE D
PROCESSING COMPLETED FOR L2
L3  1182 SOR L2 SCORE D
```

The BLAST search (L2) is further refined to sequences from granted patents, with application year prior to 1996, and to a specific text search term (L4).

```
=> S L2 AND SOMATOMAMMOTROPIN AND AY<1996 AND GRANTED/SSO
L4  7 L2 AND SOMATOMAMMOTROPIN AND AY<1996 AND GRANTED/SSO
```

```
=> SOR SCORE D
PROCESSING COMPLETED FOR L4
L5  7 SOR L4 SCORE D
```

If you limit using text and/or date terms remember to SORT SCORE D again!



36

The 7 basic steps of USGENE BLAST

- 6) Display selected relevant answers in a bibliographic format including alignment:
 - D L5 BIB AB CLM ALIGN 1 5 6
- 7) Ensure your STN Express session transcript was captured and then logoff

STN

37

Display selected USGENE answers in a preferred bibliographic format

=> D BIB AB CLM ORGN SSO ALIGN 1 3 5

```
L5 ANSWER 1 OF 7 USGENE COPYRIGHT 2007 SEQUENCEBASE CORP on STN
AN 4363877.1 Protein USGENE
TI Recombinant DNA transfer vectors (Patent
IN Goodman Howard M. (San Francisco, CA)
Shine John (San Francisco, CA)
Seeburg Peter H. (San Francisco, CA)
PA The Regents of the University of California
PI US 4363877 A 19821214
AI US 1978-897710 19780419
AB Recombinant DNA transfer vectors containing codons for human
somatomammotropin and for human growth hormone.
CLM US4363877 A: What is claimed is:
1. A recombinant DNA transfer vector comprising codons for human
chorionic somatomammotropin comprising
ORGN Unknown
SSO PROTEIN; EMBL; GRANTED
BLASTALIGN . . . .
```

This sequence hit comes from a U.S. granted patent, with an application date prior to 1996, and a key concept in the abstract and claims.

Note: this USGENE sequence record, sourced from EMBL, is an example of one which is not indexed in DGENE or REGISTRY.

STN

38

Review: 7 steps of USGENE BLAST

- 1) SAVE, UPLOAD and VERIFY a query text file (L1)
- 2) RUN the BLAST search (/SQP or /SQN)
- 3) Decide how many answers to keep (L2)
- 4) SORT SCORE in Descending order (L3)
- 5) Review answers in a free-of-charge format
e.g. D L3 TRI ALIGN 1-
- 6) Display selected answers in bibliographic format, e.g. D L3 BIB AB ECLM ALIGN 1,3,10
- 7) Ensure session transcript was captured and Logoff

STN

39

The importance of using the correct BLAST advanced options

```
=> RUN BLAST GSSFLSPEHQR/SQP
```

```
BLAST Version 2.2 . . . .
```

```
NO ANSWERS FOUND BELOW THRESHOLD OF 10
```

Changing BLAST options is especially important for short sequence queries!

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

```
BLAST Version 2.2 . . . .
```

```
690 ANSWERS FOUND BELOW EXPECTATION VALUE OF 1000.0
```

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

```
L1 RUN STATEMENT CREATED
```

```
L1 690 GSSFLSPEHQR/SQP.-M PAM30 -W 2 -E 1000 -F F
```

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

STN

40

The importance of using the correct BLAST advanced options (cont.)

```
=> SOR L1 SCORE D
PROCESSING COMPLETED FOR L1
L2          690 SOR L1 SCORE D
```

Correct use of BLAST options
finds relevant sequence hits.

```
=> D TRI ORGN ALIGN
```

```
L2  ANSWER 1 OF 690  USGENE COPYRIGHT 2007 SEQUENCEBASE CORP on STN
TI  Genetic polymorphisms associated with myocardial infarction, methods
    of detection and uses thereof (PublishedApplication)
MTY  Protein
SQL  117
ORGN Homo Sapiens
BLASTALIGN
      Query  = 11 letters
      Length = 117
      Score  = 26.9 bits (58), Expect = 2e-05
      Identities = 11/11 (100%), Positives = 11/11 (100%)
Query: 1  GSSFLSPEHQK 11
          GSSFLSPEHQK
Sbjct: 24 GSSFLSPEHQK 34
```

STN

41

Exploring USGENE search fields

- USGENE is similar in design to DGENE, but has a number of unique additional search fields:
 - /CLM Full claims text
 - /ECLM Exemplary (1st) claim text
 - /SEQC Sequence count (total number of sequences)
 - /SSO Sequence source (NCBI, USPTO, etc)
 - /SEQN Sequence Identity Number (SEQ ID NO)
- The USGENE Basic Index (/BI) comprises:
 - Title (/TI), abstract (/AB), organism name (/ORGN) and molecule type (/MTY) fields
 - Add claims (/CLM), e.g. => S VIRUS/BI,CLM

STN

42

Useful USGENE display fields/formats

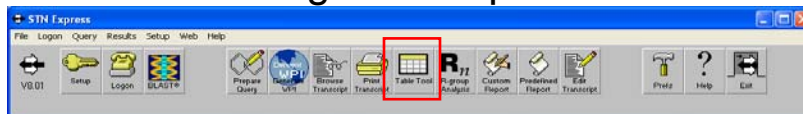
TRIAL*	Title, Molecule Type, Sequence Length
SCAN*	Random Title
ALIGN*	BLAST/GETSIM Sequence Alignment
SCORE*	Similarity Score (for post-processing)
BIB	Inventors, Assignees, numbers, dates
AB	Original abstract
ECLM	Exemplary (1 st) claim text
CLM	All claims text
ALL	BIB + AB + CLM, sequence, sequence source (SSO), feature table (FEAT)

(* Free of charge display formats in USGENE.)

STN

43

USGENE results can be post-processed into tables using STN Express 8.01+



Accession Number	Title	Assignee	Abstract	Exemplary Claim	BLAST Alignment	BLAST Score
7141547.2216 Protein USGENE	Albumin fusion proteins comprising GLP-1 polypeptides (Patent)	Human Genome Sciences Inc (Rockville MD)	The present invention encompasses albumin fusion proteins. Nucleic acid molecules encoding the albumin fusion proteins of the invention are also encompassed by the invention, as are vectors containing these nucleic acids, host cells transformed with these nucleic acids, vectors, and/or host cells. Additionally, the present invention encompasses pharmaceutical compositions comprising albumin fusion proteins and methods of treating, preventing, or ameliorating disease, disorders or conditions using albumin fusion proteins of the invention.	US7141547 B2: What is claimed is: 1. An albumin fusion protein comprising two or more tandemly oriented GLP-1 polypeptides, wherein said GLP-1 polypeptides are selected from wild-type GLP-1, GLP-1 fragments, and GLP-1 variants, fused to albumin comprising the amino acid sequence of SEQ ID NO 1039, an albumin fragment, or albumin variant thereof, wherein said albumin fragment or albumin variant increases the serum plasma half-life of the GLP-1 polypeptides, and wherein said fusion protein has GLP-1 activity.	Query = 11, EntrezSeq Search = 22, Score = 26.9 bits (58), Expect = 5e-06 Identifiers = 11/11 (100%), Positives = 11/11 (100%) Query: 1 GGGFLPFRSD 11 GGGFLPFRSD 11 Subject: 1 GGGFLPFRSD 11	38
7074910.442 Protein USGENE	PRO4340 nucleic acids (Patent)	Genentech Inc (South San Francisco CA)	The present invention is directed to novel polypeptides and to nucleic acid molecules encoding those polypeptides. Also provided herein are vectors and host cells comprising those nucleic acid sequences, chimeric polypeptide molecules comprising the polypeptides of the present invention fused to heterologous polypeptide sequences, antibodies which bind to the polypeptides of the present invention and to methods for producing the polypeptides of the present invention.	US7074910 B2: What is claimed is: 1. An isolated nucleic acid comprising: (a) the nucleic acid sequence of SEQ ID NO. 120 or the complement thereof; (b) the full-length coding sequence of the nucleic acid of SEQ ID NO. 120 complement full-length coding sequence of the cDNA of ATCC accession 203067 or thereof.	Query = 11, EntrezSeq Search = 13, Score = 22.9 bits (58), Expect = 2e-05 Identifiers = 11/11 (100%), Positives = 11/11 (100%) Query: 1 GGGFLPFRSD 11 GGGFLPFRSD 11 Subject: 24 GGGFLPFRSD 04	30
7160993.442 Protein USGENE	Nucleic acids encoding PRO4400	Genentech Inc (South San Francisco CA)	The present invention is directed to novel polypeptides and to nucleic acid molecules encoding those polypeptides of the present invention.	US7160993 B2: What is claimed is: 1. A nucleic acid molecule encoding a polypeptide of the present invention.		

STN Express 8.01+ tables can be saved, e.g., as MS Excel® files for forwarding to other colleagues.

STN

44











Agenda

- STN sequence databases
- USGENE database content
- The 7 basic steps of USGENE BLAST®
- Comparisons and conclusions

STN

45



How does USGENE compare to other USPTO sequence data sources?

	USPTO PGP	USPTO Patents	USPTO claims text	Value added
USGENE				
DGENE (DWPI basics)				
REGISTRY (CAplus basics)				
EMBL-EBI				

STN

46

How does USGENE compare to other USPTO sequence data sources? (cont.)

	Update Frequency	Typical Timeliness	Value added
USGENE	Weekly	7 days	
REGISTRY	Daily	27 days	
DGENE	Biweekly	65 days	
EMBL-EBI	Daily	1-3 months	

STN

47

Comparing STN databases...

- **DGENE**
 - The most comprehensive patent sequence database
 - Implemented in-house at major patent offices
- **REGISTRY**
 - More timely than DGENE; complementary indexing
 - Unique non-patent literature coverage
- **USGENE**
 - More timely than DGENE and REGISTRY (7 days)
 - Sequences from equivalent USPTO applications and patents
- **PCTGEN**
 - The most timely database (24 hours)
 - Sequences from equivalent WIPO/PCT publications

STN

48

Functionality / Options	CAS REGISTRY Access	DGENE, PCTGEN, USGENE Access
<i>Sequence Code Match (SCM)</i>	SEARCH command	RUN GETSEQ
<i>FASTA Homology</i>	Not available	RUN GETSIM
<i>Blast Homology</i>	CAS REGISTRY BLAST software	RUN BLAST
<i>Command line search</i>	SCM only	All 3 options (GETSEQ, GETSIM, BLAST) with RUN
<i>STN Express</i>	SCM and CAS REGISTRY BLAST	All 3 options (GETSEQ, GETSIM, BLAST) with RUN
<i>STN[®] on the WebSM</i>	SCM and CAS REGISTRY BLAST (slightly different implementation)	All 3 options (GETSEQ, GETSIM, BLAST) with RUN or with Sequence Search Assistant

STN 49

Several factors contribute to the concept of “comprehensiveness”

- Backfile and diversity of authority coverage
- Timeliness from publication to online update
- Indexed patent family member (basic patent, published application, granted patent, etc.)
- Value-added indexing versus applicant data
- Editorial indexing rules (e.g. claimed, example, disclosure or derived sequences, etc.)

See *Effective patent sequence searching on STN*:

http://www.stn-international.com/training_center/bioseq/epss.pdf

STN

50

Conclusions

- USGENE is a vital new tool for business critical patent searches, providing a complete collection of U.S. Issued Patent sequences with searchable claims text
- USGENE also provides a collection of published application sequence data, not covered by EMBL-EBI
- DGENE remains an “industry-standard” database and must be used in every patent sequence search
- REGISTRY also offers complementary value-added indexing and is typically more timely than DGENE
- USGENE, REGISTRY, and DGENE should all be used for a comprehensive search of USPTO sequence data

STN

51

Visit www.stn-international.com for the latest USGENE reference materials

The screenshot shows the STN International website homepage. The browser window title is "Databases in Science and Technology - STN International - Mozilla Firefox". The address bar shows "http://www.stn-international.com/". The page features the STN logo and navigation links. A central banner reads "Your Connection to Science and Technology" with a "Get Connected!" section. A sidebar on the right contains a "What's new" section with a red circle around the "Training Center" link.

STN

52

The logo consists of the letters 'STN' in a bold, blue, sans-serif font. The letters are three-dimensional, with a slight shadow underneath them, and they are set against a white background with a blue gradient at the top. A registered trademark symbol (®) is located to the upper right of the 'N'.

To PSIPS and Beyond

**Exploring the Content and Utility of
USGENE®**

Robert Austin – FIZ Karlsruhe